ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Roider, Andreas; Muehlheusser, Gerd

Working Paper

Black Sheep and Walls of Silence

IZA Discussion Papers, No. 1171

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Roider, Andreas; Muehlheusser, Gerd (2004) : Black Sheep and Walls of Silence, IZA Discussion Papers, No. 1171, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
http://hdl.handle.net/10419/20410

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

DISCUSSION PAPER SERIES

IZA DP No. 1171

Black Sheep and Walls of Silence

Gerd Muehlheusser
Andreas Roider

June 2004

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Black Sheep and Walls of Silence

Gerd Muehlheusser
University of Bern
and IZA Bonn

Andreas Roider
University of Bonn

Discussion Paper No. 1171
June 2004

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of the institute. Research disseminated by IZA may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit company supported by Deutsche Post World Net. The center is associated with the University of Bonn and offers a stimulating research environment through its research networks, research support, and visitors and doctoral programs. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available on the IZA website (www.iza.org) or directly from the author.

IZA Discussion Paper No. 1171
June 2004

ABSTRACT

Black Sheep and Walls of Silence*

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.

JEL Classification: D82, C73

Keywords: teams, misbehavior, wall of silence, asymmetric information

Corresponding author:

Andreas Roider
Department of Economics
University of Bonn
Adenauerallee 24-42
53113 Bonn
Germany
Email: roider@uni-bonn.de

August 2004-July 2004:
Graduate School of Business
Stanford University
Stanford, CA 94305
USA

* We are grateful to Mathias Drehmann, Eberhard Feess, Lilo Locher, Georg Nöldeke, Jörg Oechssler, Wendelin Schnedler, Urs Schweizer, and Uwe Sunde for helpful comments and suggestions. Both authors gratefully acknowledge financial support from the Graduiertenkolleg ”Quantitative Economics” at the University of Bonn.

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.) and massively tailgated 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.[1]

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 affects 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 examples being police

[1] See Süddeutsche Zeitung, February 17, 2004, p. 17.

departments and hospitals. In police departments, it is frequently observed that police officers are extremely reluctant to testify against their colleagues and, consequently, such behavior is often referred to as the "blue wall of silence". For example, Chevigny (1995, p. 92) reports that according to civilian members of the Civilian Complaint Review Board (CCRB) of New York, "it had never had a case in which a police witness testified against another". Further evidence is provided by Kleinig (2001, p. 1) who quotes an anonymous source according to which "it's unwritten law in police departments that police officers must never testify against their brother officers". Finally, according to the report of the Mollen Commission that investigated police violence in New York "the vast majority of honest police officers still protect the minority of corrupt officers".[2] In a medical context, there is the phenomenon of the so-called "white wall of silence" referring to the fact that doctors seem to be reluctant to testify against colleagues in cases of malpractice. For example, Benoit and Dubra (2004, p. 784f.) report that, while the number of deaths per year in US hospitals due to malpractice is estimated to lie between 44,000 and 98,000, "two thirds of the nation's hospitals haven't reported a single adverse incident involving a physician in the last eight years". In an economic context, walls of silence allegedly sometimes emerge in auditing relationships.[3]

The explanation we propose for this phenomenon is based on benefits that agents

[2]See *Commision to Investigate Allegations of Police Corruption and the Anti-Corruption Procedures of the Police Department Report*, New York, 1994, p. 51.

[3]Additionally, in a social context, there is abound evidence that community members are often extremely reluctant to cooperate with the police during the investigation of a crime. For example, Freeman (1999) and Donohue and Levitt (2001) report such reluctance for members of minority communities, where the police force is frequently perceived as being racially biased (on the related issue of racial profiling in motorist searches, see Knowles, Persico, and Todd 2001 and Persico 2002). As a consequence, there is a debate whether more own race policing should be introduced in some neighborhoods in order to increase cooperation with the police and to reduce crime rates. In their empirical analysis, Glaeser and Sacerdote (2000) indeed find that part of the geographic variation in crime may be attributed to lower arrest probabilities in cities resulting from lower reporting rates.

potentially reap from cooperating with other agents within a team. Such benefits may be substantial. In the police example, a police officer needs to be backed up by his colleagues in dangerous situations, i.e., it is important for him to have an attentive colleague around. In the medical context, a surgeon's probability of conducting a surgery successfully might strongly depend on the quality of support he receives from other members of his team. In the test driver example, the above-mentioned statement by the judge indicates that there might exist substantial benefits from being an accepted member of the team. While in all these case there is some benefit from cooperation within the team, in the absence of cooperation each team member is on its own and foregoes these benefits. At the same time, such benefits from cooperation may well be asymmetric. For example, a new member may consider being accepted more important than a more established one. In addition, with the emergence of more attractive outside options for some team members, such asymmetries may also arise over time

Given that walls of silence may emerge it is interesting how, in turn, they influence 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, in the medical context, doctors may save on effort costs when not taking appropriate care thereby causing harm to patients. In the police context, some cops may handle suspects in a manner that, while acceptable to themselves, may be considered unduly harsh or even brutal by others.

Importantly, note that 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. There seem to be many disciplining devices in the real world such as the threat of physical retaliation that might explain such behavior for the case of criminal

teams. In our model, we focus on the question why *honest* team members might set up a wall of silence, and the "threat of retaliation" will be much more indirect via a reduced amount of cooperation.

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 the model, honest team members differ with respect to their (privately known) willingness to report misbehavior. As a consequence, their reporting decision may convey information about their 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). Anticipating that reporting may not occur leads black sheep to misbehave in the first place. From a game theoretical point of view, we thus analyze a signaling game in which the receiver of a signal (the black sheep) chooses an action (a level of misbehavior) *before* the sender of the signal (the honest team member) chooses his first action (his reporting decision).

We provide conditions under which black sheep indulge in misbehavior and honest team members set up a wall of silence. Our results on the existence of walls of silence are most pronounced in the case of asymmetric teams where cooperation is more important for honest team members.

Relation to the Literature Our analysis is closely related to the work by Benoit and Dubra (2004) who ask "Why Do Good Cops Defend Bad Cops?". Similar to this

paper, we aim at explaining why honest team fellows protect dishonest ones. Benoit and Dubra (2004) model this by enquiring under which circumstances a single (potentially honest) agent would favor the representation of all agents (called the union) to *indiscriminately* defend misbehaving colleagues (i.e., to set up a wall of silence) over employing a *candid* strategy in which the union honestly reports all (stochastic) 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. This might reduce the probability for an individual agent of being subject to a type II error. Our analysis is different in at least four aspects: i) we explicitly model the incentives of black sheep to indulge in misbehavior and the incentive of honest team member to file a report, ii) we stress the role of cooperation benefits in determining equilibrium outcomes, iii) we consider private information on part of honest team members with respect to their willingness to file reports, and iv) in our analysis walls of silence are not driven by potentially false decisions that courts may make when assessing whether misbehavior has indeed occurred. Thus while Benoit and Dubra (2004) show that walls of silence may emerge due to enforcement errors, we identify an additional channel that also gives rise to this phenomenon.

Furthermore, Donohue and Levitt (2001) analyze a model in which the effect of race on the behavior of police, criminals and the community is explored. The sequence of events is similar to ours in that in a first stage, a criminal opportunity arises and a potential criminal decides whether or not to commit a crime. Then, the victim chooses whether to report the crime to the police. If reporting takes place, it may lead to a false arrest. In their work the impact of enforcement errors is again at center stage.

The remainder of the paper is organized as follows: In Section 2 the model is set up,

which is then analyzed in Section 3. Section 4 concludes. All proofs are relegated to an appendix.

2 The Model

We consider two risk-neutral individuals, B and G who may derive benefits $b > 0$ and $g > 0$, respectively, from cooperating with each other in a team. As spelled out in the introduction, b and g might, for example, be thought of as the benefits for surgeons from working together with attentive and careful colleagues during the course of a surgery. In the police context, b and g might represent the benefits for cops due to additional backup provided by team mates in dangerous situations.

We consider the case that B, the potential "black sheep", might engage in activities disliked by the "good guy" G. As explained above, in the police context, this might represent that B tends to treat suspects unduly harsh; in the medical context, B might be prone to lacking an adequate care level when treating patients. In an auditing context, B, a client, might engage in irregular accounting practices. Throughout the paper, we will refer to such activities as "misbehavior".

Given that G dislikes misbehavior by B, he may choose to work on his own thereby foregoing the benefit from cooperation. The value of this outside option is denoted by $\underline{g} > 0$. Similarly, B might refuse to cooperate and work on his own in which case he derives an outside option $\underline{b} > 0$.

Stage game In order 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, which are illustrated in Figure 1 and laid out in more detail below.

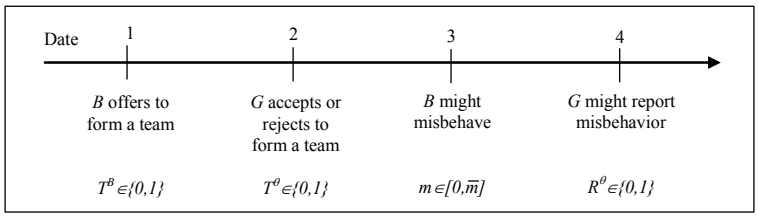

Figure 1: Stage game

Dates 1 and 2 (team formation). To model the issue of team formation and cooperation as simply as possible, we assume that at a date 1 B decides whether or not to offer G to form a team (i.e., $T^B \in \{1,0\}$). In case an offer has been made, G decides at date 2 whether or not to accept.[4] As will be laid out below, G may be one of two possible types $\theta \in \{H,D\}$, and hence G's decision is denoted by $T^\theta \in \{1,0\}$. When a team is formed (i.e., if $T^B \cdot T^\theta = 1$), B and G receive cooperation benefits b and g respectively, while if $T^B \cdot T^\theta = 0$, they receive their reservation payoffs $\underline{b}$ and $\underline{g}$ respectively.

Date 3 (misbehavior). Given that a team has been formed, B might choose to "misbehave". That is, he might take an action $m \in [0,\overline{m}]$ that generates a private gain $\widehat{b}(m)$, where $\widehat{b}(\cdot)$ is an increasing, concave function satisfying $\widehat{b}(0) = 0$. 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 potential misbehavior

[4] 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 always prefer not to form a team, and hence exercise their outside option. As a consequence, we do not consider the team formation and cooperation decisions separately, and use these terms interchangably. Note that our results below are not sensitive to the order in which B and G decide over team formation, and hence they would continue to hold if it is G who moves first.

by B to some authority that may then investigate the case. However, 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 hawkish type H has low reporting costs (which we set equal to zero), a dovelike type D faces some fixed reporting costs $r > 0$. Thus while type H is readily willing to report any kind of behavior he dislikes, type D is more reluctant to do so: for example he might simply dislike to cooperate with the authorities, or, alternatively, he might face opportunity costs from filing a report. As a result type D will only cooperate with the authorities if the level of misbehavior is sufficiently large. We denote the reporting decision of type θ of G by $R^{\theta} \in \{1, 0\}$. The benefits that G derives from reporting are made explicit below. We assume that θ is G's private information. G learns his type in the first period after a team has been formed.[5] With prior $h \equiv Prob(\theta = H) > 0$, G is a hawkish type $\theta = H$ and with probability $(1 - h)$ he is a dovelike type $\theta = D$. The probability h is assumed to be common knowledge.

Stage game payoffs In case a team has been formed, in addition to their cooperation benefits, the payoffs of the parties depend on whether B decides to misbehave and on whether G decides to report such misbehavior. As explained above, B derives a positive benefit $\widehat{b}(m)$ from choosing some positive m. However, G may choose to report such misbehavior to the authorities. If this indeed happens, some authority (who is not a player in our model) inspects the case. We take the authority's enforcement technology as exogenously given, i.e., we assume that it is independent of the players' actions and

[5]This assumption allows us to focus on the potential signaling effect of the reporting decisions.

can be represented by a mapping from the level of misbehavior to expected penalties.[6] 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 $\widehat{p}(m)$ that B faces is determined by a function $p : m \to \Re_0^+$, satisfying $\widehat{p}'(m) > 0$ and $\widehat{p}''(m) \geq 0$.[7] Absent any misbehavior, B does not face a sanction so that $\widehat{p}(0) = 0$. Moreover, penalties are sufficiently high to deter misbehavior if reporting occurs with certainty, i.e., $\widehat{p}'(m) > \widehat{b}'(m)\ \forall m$ is assumed to hold.

The honest G dislikes any kind of misbehavior (recall that it causes him a utility loss of $-m$), and he derives a private benefit if B is convicted. For example, he might simply be satisfied to see B penalized.[8] This expected (gross) benefit from reporting is given by $\widehat{r}\,(m) \geq 0$ which is increasing in m satisfying $\widehat{r}(0) = 0$. Thus, the higher the level of misbehavior by B, the more satisfied is G when B is penalized. It follows that the total expected net benefit from reporting is $\widehat{r}(m)$ and $(\widehat{r}(m) - r)$ for types H and D, respectively. Finally, although there is a benefit from conviction, G prefers lower levels of misbehavior, which formally amounts to $\widehat{r}'(m) < 1\ \forall m$.[9]

To summarize, the total period t payoff of type H is given by

$$T_t^B \cdot T_t^H \cdot \left[g - m_t + R_t^H \cdot \widehat{r}(m_t)\right] + (1 - T_t^B \cdot T_t^H) \cdot \underline{g}, \tag{1}$$

[6]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.

[7]Of course, in reality the expected penalty $\widehat{p}(m)$ 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).

[8]Benoit and Dubra (2004) report that moral considerations constitute a major reason for many (out of the few) police officers who testifiy against their colleagues.

[9]Thus, we rule out the unrealistic case that G prefers high levels of misbehavior because the private benefit from conviction is so large.

while the total period t payoff of type D is given by

$$T_t^B \cdot T_t^D \cdot \left[g - m_t + R_t^D \cdot [\widehat{r}(m_t) - r]\right] + (1 - T_t^B \cdot T_t^D) \cdot \underline{g}. \tag{2}$$

Finally, given that a team has been formed, the expected period t payoff of B is given by

$$b + \widehat{b}(m_t) - ER_t \cdot \widehat{p}(m_t), \tag{3}$$

where $ER_t \equiv h_t \cdot R_t^H + (1 - h_t) \cdot R_t^D$ denotes the expected period t reporting decision given a belief h_t to face type H.

Information and equilibrium concept Throughout, we assume that, with the exception of θ, there is symmetric information between G and B. 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 h_1 equals h, based on the observed reporting behavior in the first period, B might hold a belief $h_2 = \beta \neq h$ 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).

3 Analysis of the Model

3.1 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 of both types of G only depend on m (and not on other parts of the history of the game).[10] Hence, while the hawkish type will report any positive level of m, the dovelike type will only report when his benefit from doing so is sufficiently large. To this end, it is useful to define a critical value m^D as the level of misbehavior where type D is indifferent between reporting and not reporting. Formally, m^D is implicitly given by $\widehat{r}(m^D) \equiv r$, where it follows from $r > 0$ and $\widehat{r}(0) = 0$ that $m^D > 0$. To avoid trivial results, we assume $m^D < \overline{m}$ (i.e., there exist sufficiently large levels of m for which type D reports).

[10] 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.

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 large, i.e.,* $R^{H*}(m) = 1 \ \forall m > 0$ *and* $R^{D*}(m) = 1 \Leftrightarrow m > m^D$.

When determining the optimal level of misbehavior, B takes G's subsequent reporting strategy into account.[11] This implies that in equilibrium, the period 2 level of misbehavior depends only on B's belief β at this point in time to face a hawkish type. It follows that the level of misbehavior optimally chosen is given by

$$m \equiv \underset{\widetilde{m}}{argmax} \ \{\widehat{b}(\widetilde{m}) - ER^*(\widetilde{m}) \cdot \widehat{p}(\widetilde{m})\}, \tag{4}$$

where, analogously to above, $ER^*(\widetilde{m}) \equiv \beta \cdot R^{H*}(m) + (1-\beta) \cdot R^{D*}(m)$. Denote the unique solution to (4) as a function of β 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* m^D. *In particular,* $m(0) = m^D$, $m(1) = 0$, *and* $m(\beta)$ *is weakly decreasing in* β.

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 our assumption that $\widehat{p}'(m) > \widehat{b}'(m) \ \forall m$). Finally, as the hawkish type always reports, a higher β induces B to choose a lower level of misbehavior.

Corollary 1 (partial wall of silence in the static case) *In the static case, type D does not report in equilibrium.*

[11] Note that if in cases of indifference G would report, equilibria might fail to exist.

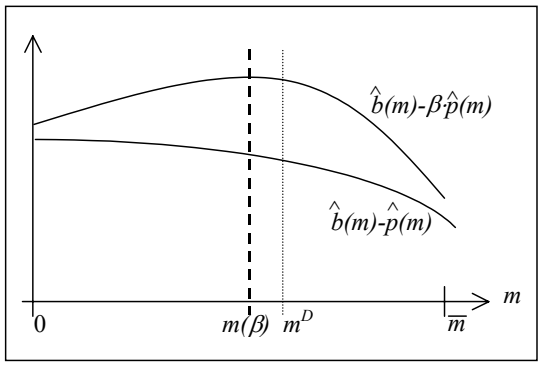

Figure 2: The optimal level of misbehavior: an example

Corollary 1 immediately follows from Lemmata 1 and 2. The fact 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 r, 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 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 join forces. 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., $g > \overline{m} + \underline{g}$.*

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. As B always has the option not to misbehave (in which case $\widehat{b}(0) - \beta \cdot \widehat{p}(0) = 0$), it follows that $\widehat{b}(m(\beta)) - \beta \cdot \widehat{p}(m(\beta)) \geq 0$ for all β. Hence, if $b \geq \underline{b}$ (the *symmetric case*), B will always want to form a team. On the other hand, if $b < \underline{b}$ (the *asymmetric case*), this does not necessarily hold true: party B will only propose to form a team if his belief β to face a hawkish type is sufficiently low such that his payoff inside the team, which is given by $b + \widehat{b}(m(\beta)) - \beta \cdot \widehat{p}(m(\beta))$, exceeds his outside option $\underline{b}$. For this asymmetric case it is useful to implicitly define a critical value $\overline{\beta}$ by $\underline{b} - b = \widehat{b}(m(\overline{\beta})) - \overline{\beta} \cdot \widehat{p}(m(\overline{\beta}))$, where $\overline{\beta} < 1$ holds.[12] 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, i.e., $T^{H*} = T^{D*} = 1$, and B's optimal team formation*

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

decision is given by

$$T^{B*}(\beta) = \begin{cases} 0 & \textit{if } b - \underline{b} < 0 \textit{ and } \beta > \overline{\beta}, \textit{ and} \\ 1 & \textit{otherwise.} \end{cases}$$

Lemma 3 implies that B might choose not to offer to cooperate with G when both, $\underline{b}$ and his belief β that he faces a hawkish type are sufficiently large: while a larger $\underline{b}$ makes exercising the outside option more attractive for B, a larger β reduces B's profit arising within a team. In order to avoid trivial outcomes (that might arise in the asymmetric team case), we assume that, B offers to cooperate given the prior belief h. That is, we assume that the critical value $\overline{\beta}$ is sufficiently large, i.e., $\overline{\beta} > h$ if $b < \underline{b}$.[13]

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

$$\left\{T^{B*} = T^{H*} = T^{D*} = 1,\ m^* = m\left(h\right),\ R^{D*} = 0,\ R^{H*} = 1\right\}. \tag{5}$$

To summarize, in the static version of the game, B would indeed become a black sheep and misbehave, and type D would tolerate B's behavior and set up a wall of silence by not reporting to the authorities. To the contrary, when facing a hawkish type, reporting would occur, and hence a wall of silence emerges with probability $(1-h)$. Moreover, in

[13] 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 static case m^D provides an upper bound for the equilibrium level of misbehavior, and this upper bound only depends on characteristics of type D. Hence, the equilibrium crime level is weakly increasing both in the belief to face type D and in this type's disincentive r to report.

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.

3.2 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.[14] In the following, we speak of a *separating equilibrium* if the parties cooperate in period 1 and $R_1^{H*}(m_1^*) \neq R_1^{D*}(m_1^*)$, 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 he receives no additional information through the period 1 reporting decision. Therefore, in a pooling equilibrium his belief β at the beginning of period 2 has to equal 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.

[14] 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 collegues) who observe the first period interaction between G and B and with whom G may want to interact in the second period.

Proposition 1 (no separating equilibria) *In any equilibrium it holds that* $R_1^{H*}(m_1) = R_1^{D*}(m_1)\ \forall m_1$*, 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 $R_1^\theta \in \{0, 1\}$ would be on the equilibrium path. Suppose, for example, that for a given level of first-period misbehavior m_1 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. Given this structure of the beliefs, it 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 the hawkish type is supposed to report on the equilibrium path, the dovelike type 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 a hawkish type which induces the latter to refrain from reporting.

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 (5). 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 r. In the second period (relative to type D) type H obtains a reporting benefit $[r(m(\beta)) - r(m(h))]$ that is smaller than r.[15] 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

[15] This follows from the fact that both $m(\beta)$ and $m(h)$ are not larger than m^D (see Lemma 2).

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).[16] In many cases, however, the Intuitive 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:[17]

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 3.2.2) would not change at all if Assumption 2 were not imposed. In the symmetric case (see Section 3.2.1), 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.[18] Although these equilibria survive the Intuitive Criterion, they involve

[16]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.

[17]Note that restricting off-equilibrium beliefs to $\beta^0(m_1) = 1$ and $\beta^1(m_1) = 0$ (as in the case when the Intuitive Criterion applies) would not alter our results in any way.

[18]It can be shown that a necessary and sufficient condition for the existence of such equilibria (that survive the Intuitive Criterion) is given by $m^D - r \geq r - \widehat{r}(m(h))$. This condition is, for example, satisfied if $r \leq \frac{1}{2}m^D$ holds, i.e., if type D's cost of reporting is not too high.

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.

3.2.1 The Symmetric Case

In the symmetric case the cooperation benefits of both G and B are sufficiently large ($b > \underline{b}$). In this case, $R_1^*(m_1) = 0$ can never be consistent with equilibrium because $b > \underline{b}$ 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 $\widehat{r}(m_1) - r$ from reporting in period 1, and face a level of misbehavior $m(h)$ in period 2. By deviating he would forego $\widehat{r}(m_1) - r$ and face $m(\beta^1(m_1))$ instead of $m(h)$. Hence, he has no incentive to deviate as long as

$$-m(h) + \widehat{r}(m_1) - r \geq -m(\beta^1(m_1)) \quad \Leftrightarrow \quad \widehat{r}(m_1) + m(\beta^1(m_1)) - m(h) - r \geq 0, \quad (6)$$

i.e., as long as the utility loss due to the higher level of misbehavior is larger than the reporting cost. If (6) 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 from cooperating with G, he nevertheless prefers this outcome to his (low) outside option $\underline{b}$. Off-equilibrium beliefs $\beta^1(m_1)$, such that (6) is satisfied, exist if the prior belief to face a hawkish type is sufficiently large, or equivalently, if dovelike types are sufficiently rare. Note that $\widehat{r}(m_1) + m(\beta^1(m_1)) - m(h) - r \geq m(\beta^1(m_1)) - m(h) - r$. Now, suppose that h is sufficiently large and that $\beta^1(m_1) = 0$ holds. In this case $m(\beta^1(m_1)) - m(h) - r \geq 0$ simplifies to $m^D \geq m(h) + r$ which is satisfied for sufficiently large h, because we have $m(1) = 0$ and $m^D \geq r$ (due to $\widehat{r}(m^D) = r$ and $\widehat{r}' < 1$).

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

(i) equilibria exist if the prior belief to face the hawkish 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., $T_1^{B*} = T_1^{H*} = T_1^{D*} = 1$), 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.[19]

3.2.2 The Asymmetric Case

We now turn to the case of asymmetric teams where B's outside option is assumed to be relatively attractive, such that $\underline{b} > b$ holds. An example for an asymmetric team might be the potential cooperation between a "young" G and an "old" B, where the young G's benefit from working with B is large (as he is eager to gain experience), while working

[19]Note, however, that this lack of misbehavior also comes at a cost. The equilibrium strategies require either type to confirm that no misbehavior has occurred even though this is costly for type D.

with the unexperienced G might impose some cost on the old B thereby making working with G less attractive.

Obviously, if $\underline{b}$ is too large relative to b, 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 $\underline{b}$ is not too large, i.e., where $\underline{b} < b + \widehat{b}(\overline{m})$ 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.

In order to prove these claims, in a first step, we will now characterize which first period 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 $\widehat{r}(m_1) - r + m^D - m(h) \geq 0$.

Intuitively, for a given m_1 where the equilibrium strategies prescribe $R_1^*(m_1) = 0$, the hawkish type can only be prevented from deviating if the future loss is sufficiently high. Only for off-equilibrium beliefs above the threshold $\overline{\beta}$ 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 $m_1^* = max\{m \mid R_1^*(m_1) = 0\}$ must be well defined, which implies $R_1^*(m_1^*) = 0$ for any $m_1^* > 0$. Moreover, given $\underline{b} > b$, 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 wall of silence in period 1. 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.

Summary of the results Let us now summarize and compare our results. In the static setting the equilibrium level of misbehavior is positive but only type D sets up a wall of silence and, as a consequence, B is still reported with positive probability. Contrary to that, in a dynamic setting a complete wall of silence may emerge as even the hawkish type might be willing to tolerate a strictly positive level of misbehavior and refrain from reporting it to the authorities. In the asymmetric case such wall of silence equilibria are payoff-dominant and survive a strong requirement on the off-equilibrium beliefs.

4 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.

5 Appendix

5.1 Proof of Lemma 2

We prove Lemma 2 by proving the following claim:

$$m(\beta) = \begin{cases} m^D & \text{if } \widehat{b}'(m^D) - \beta \cdot \widehat{p}'(m^D) \geq 0, \\ \widehat{m}(\beta) & \text{if } \widehat{b}'(m^D) - \beta \cdot \widehat{p}'(m^D) < 0 < \widehat{b}'(0) - \beta \cdot \widehat{p}'(0), \text{ and} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $\widehat{m}(\beta)$ is implicitly defined by $\widehat{b}'(\widehat{m}) - \beta \cdot \widehat{p}'(\widehat{m}) = 0$, and where $\widehat{m}(\beta) \in \left[0, m^D\right]$ holds for all β. *Proof of the claim:* Note that $\widehat{b}''(m) - \beta \cdot \widehat{p}''(m) < 0 \ \forall m, \beta$ by assumption. For a given β, B may choose some $m \leq m^D$ or some $m > m^D$ From Lemma 1 it follows that only type H reports for $m \leq m^D$, and that both types report for all $m > m^D$. Note that $m > m^D$ can never be optimal because $\widehat{b}\left(m^D\right) - \beta \cdot \widehat{p}\left(m^D\right) > \widehat{b}\left(m^D\right) - \widehat{p}\left(m^D\right)$ for all $\beta < 1$, and $\widehat{b}'\left(m\right) - \widehat{p}'\left(m\right) < 0$ by assumption. This observation also implies that $m = 0$ is optimal for $\beta = 1$. Therefore, we only need to consider $m \leq m^D$. Recall that the period 2 payoff of B is given by (3). Hence, if $\widehat{b}'(m^D) - \beta \cdot \widehat{p}'(m^D) \geq 0$, concavity implies that $\widehat{b}(m) - \beta \cdot \widehat{p}(m)$ is increasing for all $m \leq m^D$, and hence m^D is optimal. If $\widehat{b}'(0) - \beta \cdot \widehat{p}'(0) \leq 0$, concavity implies that $\widehat{b}(m) - \beta \cdot \widehat{p}(m)$ is decreasing for all $m \leq m^D$, and hence $m = 0$ is optimal. If $\widehat{b}'(0) - \beta \cdot \widehat{p}'(0) > 0 > \widehat{b}'(m^D) - \beta \cdot \widehat{p}'(m^D)$, concavity and the Intermediate Value Theorem imply that there exist some $\widehat{m}\left(\beta\right) \in (0, m^D)$ that solves $\widehat{b}(\widehat{m}) - \beta \cdot \widehat{p}(\widehat{m}) = 0$.

Finally, define a critical value β^D implicitly by $\widehat{b}'(m^D) - \beta^D \cdot \widehat{p}'(m^D) = 0$, and note that $\widehat{b}'(m^D) - \beta \cdot \widehat{p}'(m^D) < 0$ is equivalent to $\beta > \beta^D$. Hence, for all $\beta > \beta^D$ the optimal m is strictly below m^D and decreasing in β.

5.2 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^θ if and only if $T^B = 1$. It immediately follows from (1), (2) and Assumption 1 that both types of G strictly prefer $T^\theta = 1$ independent of the belief subsequently held by B. Given this equilibrium continuation $T^{B*}(\beta)$ immediately follows from the discussion above the Lemma.

5.3 Proof of Proposition 1

Suppose that in a candidate equilibrium $R_1^{H*}(m_1) \neq R_1^{D*}(m_1)$ for some $m_1 \in [0, \overline{m}]$. 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 ($b - \underline{b} \geq 0$). First, suppose that $R_1^{H*}(m_1) = 1$ and $R_1^{D*}(m_1) = 0$. In this case the incentive compatibility condition for type D is given by $-m_1 + g - m^D \geq -m_1 + \widehat{r}(m_1) - r + g \Leftrightarrow -m^D \geq \widehat{r}(m_1) - r$. If $m_1 > m^D$, then $\widehat{r}(m_1) - r > 0$, and D's incentive compatibility condition cannot be satisfied. If $m_1 \leq m^D$, then $\widehat{r}(m_1) - r \leq 0$. Note that $-m^D \geq \widehat{r}(m_1) - r \Leftrightarrow -\widehat{r}(m_1) \geq m^D - \widehat{r}(m^D)$. Moreover, $m^D - \widehat{r}(m^D) > 0$ because $\widehat{r}(0) = 0$ and $\widehat{r}' < 1$, which again yields a contradiction because $-\widehat{r}(m_1) \leq 0$ for all m_1. Second, suppose that $R_1^{H*}(m_1) = 0$ and $R_1^{D*}(m_1) = 1$. The incentive compatibility condition of type H is given by $-m_1 + g \geq -m_1 + \widehat{r}(m_1) + g - m^D + \widehat{r}(m^D) \Leftrightarrow m^D \geq \widehat{r}(m_1) + r$. The

incentive compatibility condition of type D is given by $-m_1 + \widehat{r}(m_1) - r + g - m^D \geq -m_1 + g \Leftrightarrow \widehat{r}(m_1) - r \geq m^D$. Hence, if both incentive compatibility conditions were satisfied simultaneously this would imply that $-r \geq r$ which is not possible.

Case 2 ($b - \underline{b} < 0$). First, suppose that $R_1^{H*}(m_1) = 1$ and $R_1^{D*}(m_1) = 0$. In this case the incentive compatibility condition of type H is given by $-m_1 + \widehat{r}(m_1) + \underline{b} \geq -m_1 + g - m^D + \widehat{r}(m^D) \Leftrightarrow 0 \geq [g - \underline{g}] + [\widehat{r}(m^D) - m^D] - \widehat{r}(m_1)$, which is violated for all levels of m_1 if it is violated for $m_1 = \overline{m}$. This is the case because $0 \geq [g - \underline{g}] + [\widehat{r}(m^D) - m^D] - \widehat{r}(\overline{m}) \Leftrightarrow 0 \geq [g - \overline{m} - \underline{g}] + [\widehat{r}(m^D) - m^D] - [\widehat{r}(\overline{m}) - \overline{m}]$ cannot be satisfied due to Assumption 1, $\widehat{r}' < 1$ and $m^D < \overline{m}$. Second, suppose that $R_1^{H*}(m_1) = 0$ and $R_1^{D*}(m_1) = 1$. In this case the incentive compatibility condition of type H is given by $-m_1 + \underline{g} \geq -m_1 + \widehat{r}(m_1) + g - m^D + \widehat{r}(m^D) \Leftrightarrow 0 \geq [g - m^D - \underline{g}] + \widehat{r}(m_1) + \widehat{r}(m^D)$ which cannot be satisfied due to Assumption 1.

5.4 Proof of Lemma 4

As discussed in Section 3.1, 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 $b - \underline{b} \geq 0$ 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 $[g-m(h)+\widehat{r}(m(h))]-[\widehat{r}(m_1)+g-m(\beta)+\widehat{r}(m(\beta))]$ is smaller than the difference between type D's candidate equilibrium payoff and his payoff following a deviation, which is given by $[g-m(h)]-[\widehat{r}(m_1)-r+g-m(\beta)]$, where $\beta \in [0,1]$ denotes the off-equilibrium belief. As $\widehat{r}(m(h))-r-\widehat{r}(m(\beta)) \leq \widehat{r}(m^D)-r-\widehat{r}(m(\beta)) = -\widehat{r}(m(\beta)) \leq 0 \ \forall \beta$ this is indeed the case.

5.5 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 $-\widehat{r}(m_1) + [\widehat{r}(m(h)) - m(h)] - [\widehat{r}(m(\beta^0(m_1))) - m(\beta^0(m_1))] \geq 0$, which, however, is violated due to Assumption 2 and $\widehat{r}' < 1$. Second, Lemma 4 implies that $R_1^*(m_1) = 1$ is consistent with equilibrium if and only if $\widehat{r}(m_1) - r + m(\beta^1(m_1)) - m(h) \geq 0$. 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 $b \geq \underline{b}$ both parties choose to cooperate.

5.6 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 5.5 that $R_1^*(m_1) = 0$ is not consistent with equilibrium if $\beta^0(m_1) \leq \overline{\beta}$. However, if $\beta^0(m_1) > \overline{\beta}$, the parties do not cooperate in period 2, and hence the incentive compatibility condition of type H is given by $g - m(h) + \widehat{r}(m(h)) \geq \widehat{r}(m_1) + \underline{g} \Leftrightarrow [g - \underline{g}] + [\widehat{r}(m(h)) - m(h)] - \widehat{r}(m_1) \geq 0$. The above inequality is satisfied if it is satisfied for $m_1 = \overline{m}$: $[g - \underline{g}] + [\widehat{r}(m(h)) - m(h)] - \widehat{r}(\overline{m}) \geq 0 \Leftrightarrow [g - \underline{g} - \overline{m}] + [\widehat{r}(m(h)) - m(h)] - [\widehat{r}(\overline{m}) - \overline{m}] \geq 0$, which holds due to Assumption 1 and $\widehat{r} < 1$.

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 5.5 implies that $R_1^*(m_1) = 1$ is consistent with equilibrium if $\widehat{r}(m_1) - r - m(h) + m(\beta^1(m_1)) \geq 0$. Off-equilibrium beliefs $\beta^1(m_1)$ such that this inequality is satisfied exist if and only if $\widehat{r}(m_1) - r - m(h) + m(0) \geq 0$.

5.7 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.

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 which is smaller than $\underline{b}$. 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 $\widehat{b}(m_1) > \widehat{b}(m_1) - \widehat{p}(m_1)$ for all $m_1 > 0$, B will choose the highest level of m_1 such that $R_1^*(m_1) = 0$.[20] 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 $b + \widehat{b}(\overline{m}) > \underline{b}$.

[20] If $R_1^*(m_1) = 1$ for all m_1, B will choose $m_1 = 0$.

References

Benoit, J.-P., and J. Dubra (2004): "Why Do Good Cops Defend Bad Cops?," *International Economic Review*, 45(3), 783–805.

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

Cho, I.-K., and D. Kreps (1987): "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, 102(2), 179–221.

Donohue, J. J., and S. D. Levitt (2001): "The Impact of Race on Policing, Arrest Patterns, and Crime," *Journal of Law and Economics*, 44(2), 367–394.

Freeman, R. B. (1999): "The Economics of Crime," in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, vol. 3, pp. 3529–3571. Elsevier, Amsterdam.

Fudenberg, D., and J. Tirole (1991): *Game Theory*. MIT Press, Cambridge, Mass.

Glaeser, E., and B. Sacerdote (2000): "Why is there more crime in cities?," *Journal of Political Economy*, 107(6), 225–258.

Kleinig, J. (2001): "The Blue Wall of Silence: An Ethical Analysis," *International Journal of Applied Philosophy*, 15(1), 1–23.

Knowles, J., N. Persico, and P. Todd (2001): "Racial Bias in Motor-Vehicle Searches: Theory and Evidence," *Journal of Political Economy*, 109(1), 203–229.

Mookherjee, D., and I. P. Png (1992): "Marginal Deterrence in Enforcement of Law," *Journal of Political Economy*, 102(5), 1039–1066.

Persico, N. (2002): "Racial Profiling, Fairness, and Effectiveness of Policing," *American Economic Review*, 92(5), 1472–1497.

Stigler, G. J. (1970): "The Optimum Enforcement of Laws," *Journal of Political Economy*, 78(3), 526–536.