sociological science

# Sharing Compromising Information as a Cooperative Strategy

Diego Gambetta,[a] Wojtek Przepiorka[b]

a) Collegio Carlo Alberto, Torino; b) Utrecht University

**Abstract:** Well-enforced norms create an opportunity for norm breakers to cooperate in ventures requiring trust. This is realized when norm breakers, by sharing evidence of their breaches, make themselves vulnerable to denunciation and therefore trustworthy. The sharing of compromising information (SCI) is a strategy employed by criminals, politicians, and other actors wary of their partners' trustworthiness in which the cost of ensuring compliance is offloaded on clueless norm enforcers. Here we introduce SCI as a sui generis cooperative strategy and test its functioning experimentally. In our experiment, subjects first acquire the label "dove" or "hawk" depending on how cooperative or uncooperative they are, respectively. Hawks acquire compromising information embodied in their label and can reveal it before an interaction with trust at stake. Unlike doves, hawks who reveal their label make themselves vulnerable to their partners, who can inflict a penalty on them after interaction. We find that even students in as artificial a setting as a computerized decision laboratory grasp the advantage of SCI and use it to cooperate. Our results corroborate the idea that compromising information can be conceived as a "hostage" that, when mutually exchanged, makes each party to the interaction vulnerable and therefore trustworthy in joint endeavours.

**Keywords:** information sharing; trust; cooperation; social norms; norm enforcement; venture game

THE idea of sharing compromising information (SCI) originates from an intuition of Thomas Schelling (1960), who conceived of it in the extreme case of a hostage desperate to be trusted: "Both the kidnapper who would like to release his prisoner, and the prisoner, may search desperately for a way to commit the latter against informing on his captor once released, without finding one." But Schelling came up with a solution: "If the victim has committed an act whose disclosure could lead to blackmail, he may confess it [to the kidnapper]; if not, he might commit one in the presence of his captor, to create the bond that will ensure his silence" (43–44).

All types of agents who benefit from cooperation, but have reasons to distrust one another and cannot or do not want to rely on the law, face dilemmas similar to that faced by the characters in Schelling's illustration. How can criminals trust one another to share the robbery's loot fairly? How can politicians trust their comrades' promises of future rewards to be delivered in return for their present support? How can adulterers trust that their lovers will never blackmail them? SCI, we argue, is one powerful answer (see also Baccara and Bar-Isaac 2008; Cook, Hardin, and Levi 2007).

SCI induces trust by binding agents to behave trustworthily because of the punishment they would face otherwise. Hence, the primary condition for SCI to induce trust is the belief that once informed of a breach, the norm enforcers have a high probability of meting out the punishment. In other words, all parties to

**Citation:** Gambetta, Diego, and Wojtek Przepiorka. 2019. "Sharing Compromising Information as a Cooperative Strategy." Sociological Science 6: 352-379.

**Received:** February 20, 2019

**Accepted:** March 20, 2019

**Published:** May 8, 2019

**Editor(s):** Jesper Sørensen, Delia Baldassarri

**DOI:** 10.15195/v6.a14

the interaction must expect that the enforcers, unaware of their part in the ploy, will do the "dirty work" for the party informing on the norm breaker. Another condition that strengthens the reliability of SCI is the expectation that the breach is unlikely to be independently discovered by the norm enforcers, whether directly or via the media, nosy neighbors, or enterprising hackers. An independent discovery of the breach would deactivate the bonding force of the compromising information, which rests on it remaining a secret shared only by those concerned. The ideal norm enforcers for SCI, rather than proactively pursuing norm breakers, should spring into action only when informed of their breaches.

Anecdotal evidence of the relevance of SCI has been found in a variety of extralegal domains. It refers both to *asymmetric* and *symmetric* variants of SCI. Examples of the former are found in organizations (criminal or political) in which eager novices, in order to gain admission, are asked to confess or commit a crime, not just as proof of their skills and mettle but also as incriminating evidence that could be used against them. Skull and Bones, a Yale University students' fraternity, demand that novices, as part of their initiation, disclose a highly embarrassing secret (Rosenbaum 2000:155ff).[1] This variant of SCI is asymmetric because recruiters gather incriminating information on recruits but not vice versa. Asymmetric SCI thus carries high risks for those who reveal their breaches and therefore requires high payoffs.

Examples of the symmetric variant of SCI have been observed in corrupt networks and pedophile Internet rings (Gambetta 2018, 2009:60–66) and, one imagines, must be rife among adulterers. Noel Biderman, CEO of Ashley Madison, made it into a business proposition: "If you want to keep an affair discreet, you need to work on the basis of mutually assured destruction—you need someone who's taking the same level of risk as you" (Walden 2010).[2] Symmetric SCI also seems a driving force in the choice of friends among deviant teenagers (Flashman and Gambetta 2014). When studying different strategies to deter cartel formation, Bigoni et al. (2015) also found that "the possibility to report to law enforcers [is] used [in the experiment] as a costly punishment to discipline cartel deviation" (665–6). Varese, Wang, and Wong (2018) report evidence that Chinese businessmen bond by jointly using prostitutes' services, a compromising act not only because spouses could be informed but also because prostitution is against the law in China. Schelling's fictional example, too, describes a case in which trust is achieved when the victim reveals compromising information and shares it with the kidnapper, thereby making the potential for blackmail symmetrical. Unlike the asymmetric variant, symmetric SCI creates balanced bonds and trust through mutually assured destruction.

SCI shares some of its features with other enforcement strategies, in particular those that rely on punishment and on vulnerability (Kopanyi-Peuker, Offerman, and Sloof 2017), such as hostage exchange (Nakayachi and Watabe 2005; Raub and Keren 1993; Williamson 1983). Still, its properties make it a *sui generis* alternative with attractive features.[3] Unlike peer punishment, the cost of punishment is borne by the clueless norm enforcer rather than by the agents, and unlike centralized punishment, unleashing castigation is in the gift of the agents themselves and does not require surrendering control to potentially dangerous entities, such as mafias.[4] Furthermore, exchanging compromising information in lieu of hostages has the

advantages that information can be more easily moved, hidden, and stored; its supply is much larger than that of prized family members; and it is more credible as the agents themselves, rather than their relatives, bear the burden of vulnerability. Lastly, unlike two well-studied mechanisms that support cooperation, repeated and reputation-based interactions,[5] SCI can support cooperation in one-off encounters and in the absence of a network of efficient information diffusion.

The anecdotal evidence showing SCI's pervasiveness as well as its attractive features provides a powerful encouragement to further explore SCI's viability. We chose to do this by means of a computerized laboratory experiment. The controlled setting of a lab experiment is well suited to test hypotheses derived from formal models, it allows for systematic variation of experimental conditions with the aim of isolating the causal mechanism in question, and lab experiments facilitate cumulative research and the replicability of results (Bader et al. 2019; Falk and Heckman 2009; Jackson and Cox 2013). Lab experiments have been criticized for producing results that lack external validity because of their artificial setup and use of convenience samples of university students as participants (Henrich, Heine, and Norenzayan 2010; Levitt and List 2007). However, our primary aim is to test the internal validity of SCI as a general cooperative strategy, and to our knowledge, we are the first to do so. Once we have established the scope conditions of SCI in the lab, we will be better equipped to scrutinize the working of SCI in the wild.

## Theory and Hypotheses

To test the working of SCI in the lab, we devised a variant of the trust game (TG; Dasgupta 1988; Kreps 1990), in which if trust is placed by both players, chance decides who plays the trustee. We call this version the venture game (VG; Figure 1). The VG is superior to the other two possible candidates for testing SCI, the standard TG and the prisoners' dilemma game (PD): Unlike the TG, the VG elicits trust from both players, and unlike the PD, it keeps players' decisions to trust separate from their decisions to cooperate (i.e., be trustworthy).[6]

A stylized account of the VG goes as follows: Two players (e.g., criminals) consider embarking on a joint endeavour (e.g., a robbery). Each player chooses independently and simultaneously whether to trust the other player or stay out. If either chooses to stay out, the robbery cannot take place, and both save the opportunity costs of committing the robbery ($P$). If both decide to trust, the robbery takes place, but after their decision, a coin toss determines who guards the loot. The chosen player can cooperate and share the loot with the other player (in which case, both earn $R$), or he or she can defect and run off with it (in which case, the runner earns $T$ and the other earns $S$). The VG constitutes a social dilemma because rational, self-regarding, and risk-neutral (RSN) players should choose to stay out (because $P > 0.5T + 0.5S$), whereas mutual trust and cooperation result in a Pareto efficient outcome (because $R > P$; Kollock 1998).

We conceptualise the VG as a one-shot game (rather than a repeated game) to rule out the shadow of the future as a possible mechanism enforcing cooperation (Axelrod 1984; Baccara and Bar-Isaac 2008; Kreps and Wilson 1982). How then can SCI promote cooperation in the one-shot VG? SCI implies that two stages are added

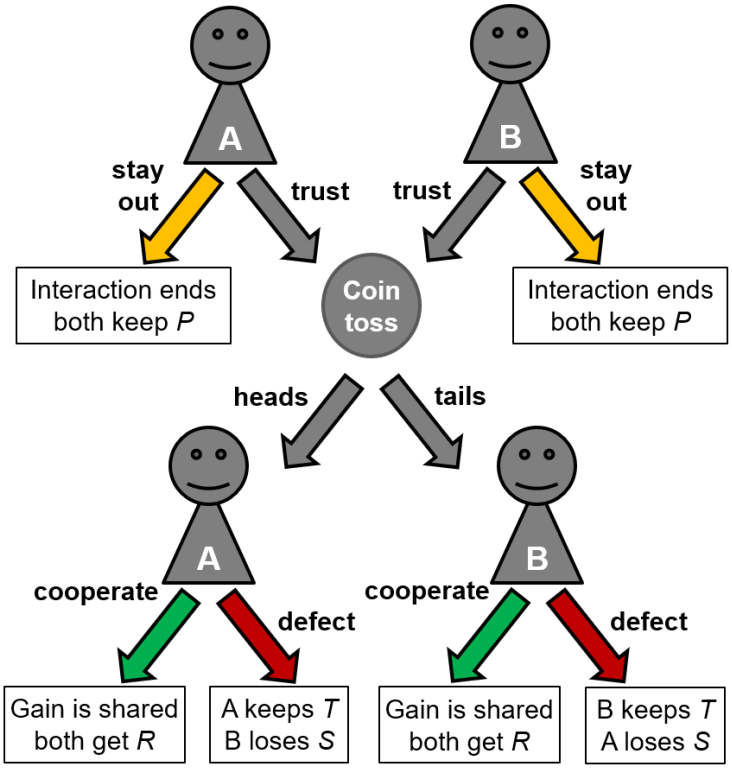


**Figure 1:** The VG ($T > R > P > S$ and $P > 0.5T + 0.5S$).

to the VG: one that enables the two agents to share compromising information before the VG (*information-sharing stage*) and one that enables agents who obtained compromising information from their interaction partners to instigate punishment after the VG (*norm-enforcement stage*). Let us start with adding the information-sharing stage only and consider the theoretical implications of this step first.

Would the mere presence of an information-sharing stage induce our two criminals to trust each other and embark on a robbery together? The answer is no because without a norm-enforcement stage, the sharing of evidence on their hitherto undiscovered crimes has no consequences and therefore would not be compromising. Hence, it can be expected that the two criminals, even if they learned about each other's misdeeds by accident, would not trust one another and would abstain from embarking on the robbery together. In fact, learning about each other's malicious abilities would make them even more distrustful.

This is quite unlike the sharing of reliable information about good deeds. Although we conceived the VG as a model for the interaction between criminals, its structural properties equally apply to interactions in legal contexts. Even to a larger extent than with criminals, law-abiding citizens embark on joint endeavours requiring trust. Short of criminal records, these agents can share evidence of their good deeds and reveal something about their unobserved properties that make them trustworthy (Gambetta and Przepiorka 2014). Hence, in a world populated by both criminals and law-abiding citizens (let's call them *hawks* and *doves*, respectively),

the existence of an information-sharing stage only has favorable consequences for the latter, not the former. We can thus state our first two hypotheses:

> Hypothesis 1 (H1): In the VG with an information-sharing stage only, a dove will trust and cooperate with another dove more than a dove with a hawk, a hawk with a dove, or a hawk with a hawk.
>
> Hypothesis 2 (H2): If given the option, doves will more often engage in information sharing about their good deeds than hawks will about their bad deeds.

Let us now add the norm-enforcement stage and turn to our two criminals again. With a norm-enforcement stage, evidence about one's interaction partner's unpunished crimes can be used to inform the enforcers and trigger the pending punishment on them. Hence, by sharing such evidence, agents make themselves vulnerable; because punishment is costly, agents would regret having shared the evidence if their partners gave it away to enforcers. However, if the sharing of compromising information is symmetric (i.e., reciprocal), the norm-enforcement stage establishes a state of mutually assured destruction. And this is how SCI becomes a cooperative strategy.

> Hypothesis 3 (H3): In the VG with an information-sharing stage and a norm-enforcement stage, a dove will trust and cooperate with another dove as much as a hawk with a hawk and more than a dove with a hawk or a hawk with a dove.
>
> Hypothesis 4 (H4): If given the option, doves will engage in information sharing about their good deeds as much as hawks will engage in information sharing about their bad deeds.

H3 and H4 are strict in the sense that they imply that with a norm-enforcement stage, hawks will start behaving like doves. This expectation is at the limit and can be relaxed by comparing hawks with hawks across conditions rather than comparing hawks with doves within conditions. We therefore also propose lenient versions of hypotheses 3 and 4:

> Hypothesis 3b (H3b): In the VG with an information-sharing stage and a norm-enforcement stage, a hawk will trust and cooperate with another hawk more than a hawk with a hawk in the VG with an information-sharing stage only.
>
> Hypothesis 4b (H4b): If given the option, hawks will engage in information sharing about their bad deeds more than hawks in the VG with an information-sharing stage only.

The properties of the one-shot VG with an information-sharing stage and a norm-enforcement stage make it a suitable paradigm to test the working of SCI experimentally. An extended game-theoretic analysis of the VG and a formal derivation of H1 through H4b are provided in Appendix A.

## Experimental Design, Procedures, and Data

We use the so-called strategy method to let subjects decide in the VG whether to stay out, trust and cooperate, or trust and defect (Brandts and Charness 2011; Selten 1967). That is, subjects make their decisions knowing that nature will randomly select who among them has the power to decide how to share the profit in case they and their partners chose to trust each other. However, the decision situation presented to subjects is not hypothetical, as subjects' earnings are determined based on their and their interaction partners' actual decisions. In our experiment, the VG payoffs are set to $T = €8.5$, $R = €6.5$, $P = €4.5$, and $S = (–)€1$. The experiment was programmed and conducted using z-Tree (Fischbacher 2007).

Our experiment is divided in three parts, and in each, pairs of subjects interact in the VG over several rounds (see Figure 2). Subjects are told from the start that the experiment comprises three parts, but they receive the part-specific instructions only at the start of each part. We do this to rule out strategic considerations in the first part and also to avoid overwhelming subjects with too much information from the start. We do not vary the sequence of the three parts because the three parts build on each other.

In all part-specific instructions, subjects are told that they will be randomly paired with another partner in every round (so-called stranger matching); subjects are not told the exact number of rounds they will play but only that it will be between seven and 10. They are told, moreover, that the total amount they earn will be the sum of what they earn in each part (plus the show-up fee of €3) and that their earnings in each part will be determined by the outcome of the VG in which they participated in one randomly chosen round. The English translation of the instructions (translated from Italian by the authors) is reproduced in figures S1 through S7 in the online supplement. Subjects received printed instructions and completed a quiz about the instructions at the start of each part. If at least one subject answered a control question incorrectly, the question was read out loud, and the correct answer was provided and explained to all subjects.

In the first part (T0), all subjects play eight rounds of the VG. Thereafter, they are informed that they are assigned a label that they will keep until the end of the experiment. Subjects are assigned the label based on how they played in these eight rounds. Subjects who chose to defect three times or more obtain the label "hawk," whereas subjects who chose to defect fewer than three times obtain the label "dove." In other words, in T0, subjects unbeknownst to them produce their type "naturally," as it were (Gambetta and Przepiorka 2014); furthermore, this allows those who behave as hawks to acquire "compromising information," embodied by their multiple defections. We employ the VG to create the two types of players because the VG elicits subjects' cooperative intent but also because, in this way, subjects become acquainted with the strategic properties of the VG before having to consider the strategic implications of SCI (i.e., VG with the information-sharing and norm-enforcement stages added to it).

After the first part, subjects are assigned to be in a condition without or with a norm-enforcement stage (see Figure 2). Subjects stay in their respective conditions in the second part (T1 or T3, respectively) and in the third part (T2 or T4, respectively).

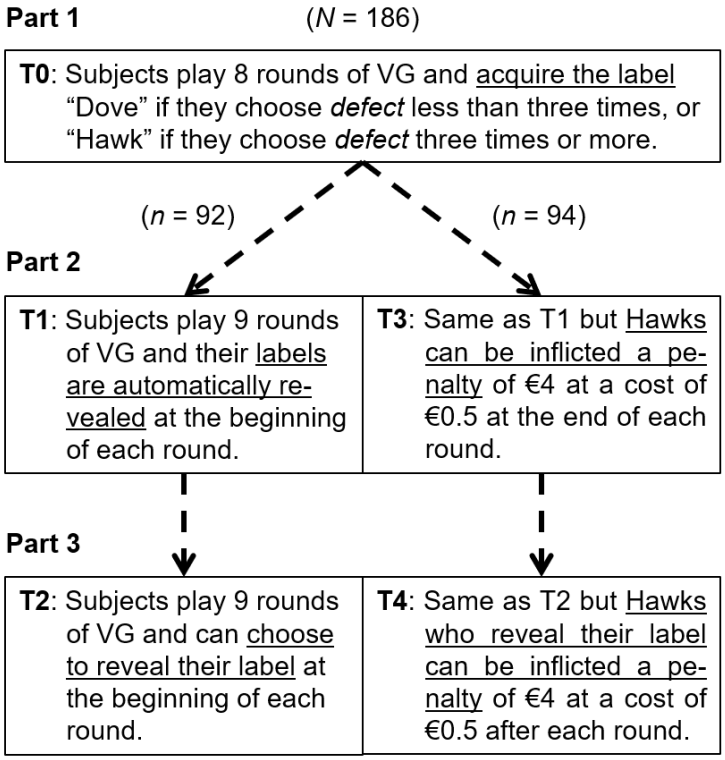


**Figure 2:** Experimental design.

In the second part, subjects play nine rounds of the VG, and their labels are automatically revealed at the information-sharing stage at the beginning of each round. Unlike in T1, in T3, hawks (and only hawks) can be inflicted a penalty by their interaction partners after being told their partners' VG decision and outcome in that round. If inflicted, the punished and the punisher incur a cost of €4 and €0.50, respectively. Conditions T1 and T3 thus re-create the situation in which subjects unwittingly share information about their past deeds or misdeeds.

In the third part, subjects play another nine rounds of the VG, either without (T2) or with (T4) a norm-enforcement stage, but unlike in the second part, they can now decide whether to reveal their label in the information-sharing stage at the beginning of each round. That is, unlike in T2, in T4, hawks who reveal their label (and only hawks who reveal their label) can be inflicted a penalty by their interaction partners. Thus, although in T3 hawks can exploit the information automatically transmitted to support cooperation with other hawks, only in T4 can they deliberately invoke SCI to promote cooperation among each other.

The norm-enforcement stage available in T3 and T4 is a key feature of our design. If a hawk meets another hawk, and both learn that they are hawks and choose to trust each other in the VG, they have three options at the end of the round: They can not only choose to say yes or no to inflicting the penalty on their interaction partner but also "no unless my partner chooses yes." If both choose the conditional option, then no penalty can be inflicted. But if only one chooses the conditional

option and the other chooses yes, then both are penalized. This conditional option allows hawks to create the state of mutually assured destruction: the quintessence of symmetric SCI. In interactions in which only one is revealed to be a hawk and both choose to trust each other, only the other party has the option to inflict the penalty on the unveiled hawk.

The experiment was conducted at the Bologna Laboratory for Experiments in Social Science with 186 student subjects from University of Bologna. Subjects were 54 percent female and 24.7 years old on average ($sd$ = 4.31) and gave informed consent to participate in our experiment. The experiment comprised eight sessions with 22 to 26 subjects per session. A session lasted 1.5 hours on average, and subjects earned €22 on average. Subjects were recruited via e-mail using the Online Recruitment System for Economic Experiments (Greiner 2015).

All test statistics and figures reported in the results section are based on regression model estimations provided in tables S1 through S3 in the online supplement. Statistical significance is set at the 5 percent level (i.e., $\alpha = 0.05$) for two-sided tests, and we account for the repeated measures obtained on the same subjects by estimating cluster-robust standard errors. Except for hawks in T0, subjects' behavior is fairly stable over the course of each part (see figures S9 and S10 in the online supplement). We therefore refrain from an in-depth analysis of the behavioral dynamics. We use Stata's margins command to calculate proportions from (multinomial) logistic regressions and test the statistical significance of the differences between proportions using linear combinations of parameters. The figures are created using the coefplot command (Jann 2014), and the regression tables are created using the esttab command in Stata (Jann 2007). The data and replication files are available from the authors on request.

## Results

At the end of the first part (T0), 92 subjects (49.5 percent) had defected fewer than three times and received the label "dove," and 94 subjects (50.5 percent) had defected three times or more and received the label "hawk." Figure 3 shows the proportion of decisions to stay out, cooperate, and defect made by doves, by hawks, and overall in the first part of our experiment. By definition, the defection rate of doves is much lower than that of hawks. Hence, the label "hawk" is a strong sign of a subject's inclination to defect, and the label "dove" is a strong sign of the subject's inclination to cooperate. However, subjects who mostly stayed out in T0 obtained the label "dove" also. Doves therefore are a less "select" group than hawks.

Figure 4 shows the proportions of VG decisions across experimental conditions and for the four types of pairings in which both subjects' labels are revealed, either by design (T1 and T3) or by choice (T2 and T4). Note first that in T1, doves cooperate with doves and hawks defect with hawks (Figures 4a and 4b) as much as they did when they interacted with unlabeled others in T0; comparing the same subjects across the two parts shows that the differences in cooperation rates are statistically insignificant (doves: z = 0.81 [$p$ = 0.418]; hawks: z = 1.09 [$p$ = 0.277]).

Based on the (corroborated) assumption that doves are significantly more other-regarding than hawks, our first hypothesis (H1) is that in the absence of a norm-

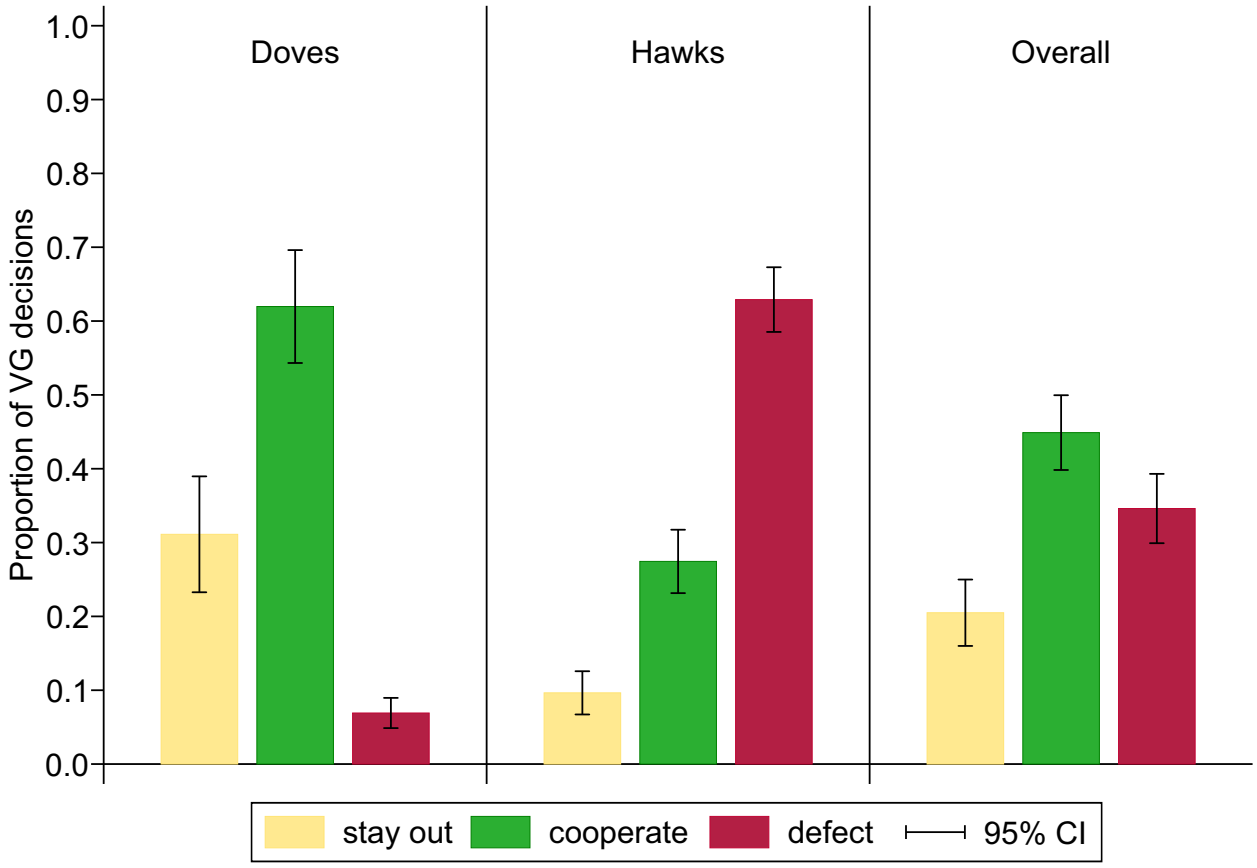


**Figure 3:** Proportions of VG decisions in part 1 of the experiment.

enforcement stage, doves are more cooperative with doves (Figure 4a) than hawks are with hawks (Figure 4b), doves are with hawks (Figure 4c), and hawks are with doves (Figure 4d). We find clear support for this hypothesis both in T1 (Figure 4b: z = 5.39 [$p < 0.001$]; Figure 4c: z = 6.18 [$p < 0.001$]; Figure 4d: z = 3.18 [$p = 0.001$]) and in T2 (Figure 4b: z = 5.44 [$p < 0.001$]; Figure 4c: z = 6.19 [$p < 0.001$]; Figure 4d: z = 3.71 [$p < 0.001$]). This is reassuring. If one is cooperative, kindred spirits happily embark in joint ventures together. Only doves gain in the absence of the norm-enforcement stage, whereas hawks are avoided (by doves) or exploited (by other hawks). But do hawks reach similar cooperation levels as doves once a norm-enforcement stage makes SCI work?

Our third and most important hypothesis (H3) is that once hawks can use SCI, they cooperate to the same extent as doves. This hypothesis is supported for hawks who choose to reveal their labels (T4); they cooperate with each other to a similar extent as doves (the difference in cooperation rates is insignificant: z = 0.53 [$p = 0.596$]). When subjects' labels are revealed automatically (T3), H3 is not supported. Hawks cooperate with hawks significantly less than doves do with doves (z = 5.22; $p < 0.001$). However, the lenient version of H3 (H3b)—comparing hawks across conditions rather than comparing hawks and doves within conditions—is

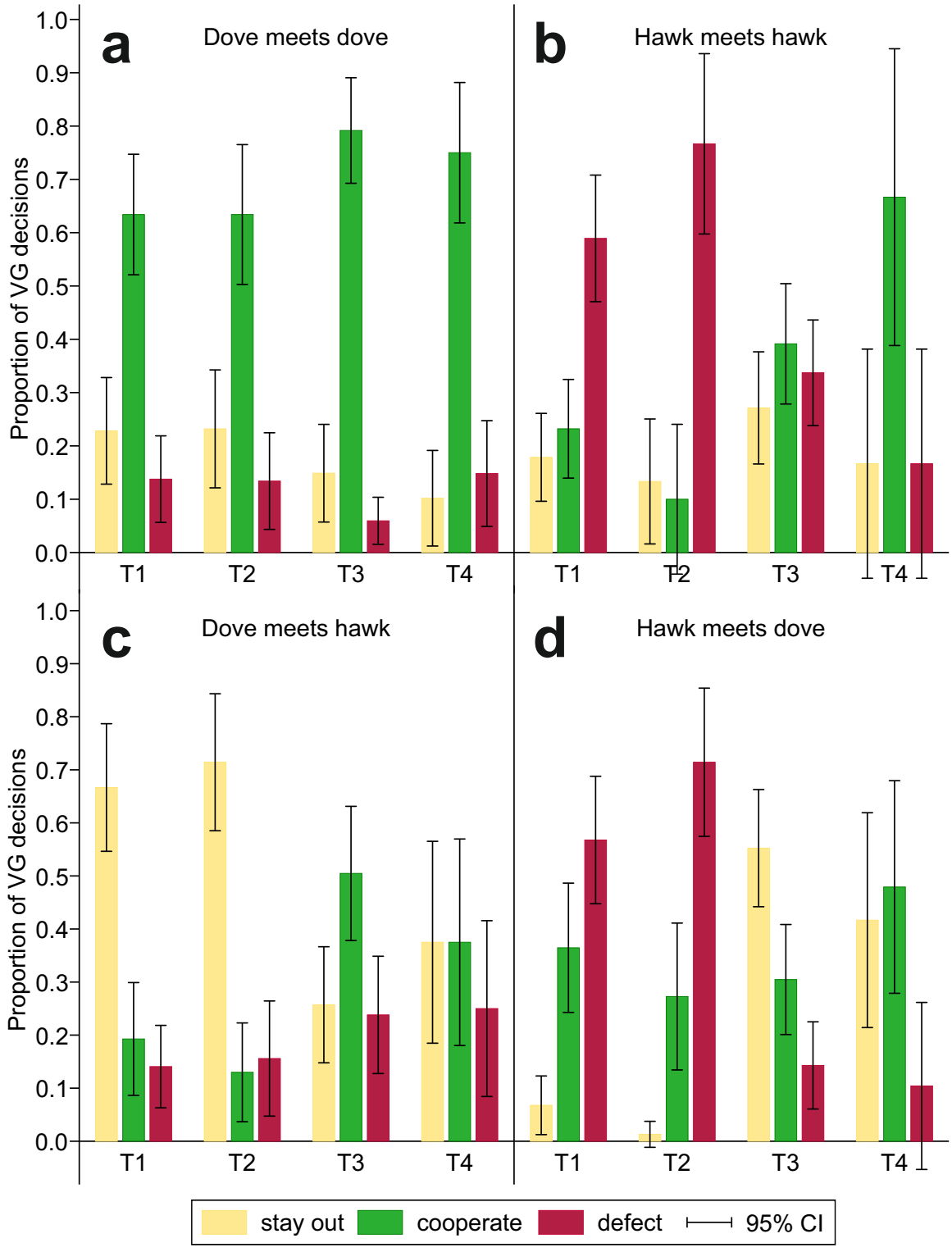


**Figure 4:** Proportion of subjects' VG decisions across experimental conditions.

clearly supported (Figure 4b). As soon as hawks can be penalized (T3 and T4), they cooperate significantly more (T1 vs. T3: z = 2.14 [$p = 0.032$]) and even more so when they can choose to reveal their label and jointly do (T2 vs. T4: z = 3.56 [$p < 0.001$]). Overall, these results corroborate that SCI allows norm breakers to cooperate.

Unsurprisingly, doves are afraid of hawks and prefer to stay out, but with a "stick" in their hands, they even try to cooperate with hawks (Figure 4c). When hawks meet doves, they prefer to stay out if doves can inflict a penalty on them, but they mostly defect in the absence of a norm-enforcement stage (Figure 4d).

SCI thus provides hawks with the necessary means to overcome their cooperation problem, but how well do hawks recognize the strategic advantage of revealing their label? Our prediction is that hawks reveal their label less than doves do in the absence of a norm-enforcement stage in T2 (H2) but do so as much as doves when inflicting a penalty on them becomes an option in T4 (H4), or at least more than they do in the absence of a norm-enforcement stage (H4b).

As expected (Figure 5), doves reveal their label at a very high rate irrespective of condition, and hawks reveal their label significantly less than doves do when they have a choice in the absence of a norm-enforcement stage (T2; z = 5.73 [$p < 0.001$]). But contrary to our expectation, rather than increasing, the rate at which hawks reveal their label (when both revelation and penalty are options) decreases in T4 compared with in T2 (z = 2.91; $p = 0.004$). Hence, it is not the case that hawks in general realize the advantage they have by making themselves vulnerable to other hawks so as to reach cooperation in the VG. Several hawks, once their compromising information is automatically revealed, do make a virtue of it and cooperate. But only a small, select flock of hawks grasp the value of revelation, whereas the majority are afraid to show their true nature.

When applying common sense rather than theory, the "shyness" shown by many hawks is unsurprising: Naïve subjects do not normally run around advertising their misdeeds. They rather act on their first instinct and hide them, even though hawks who do not reveal their label are in any case assumed to be hawks by their interaction partners (see Figure S8 in the online supplement). In the artificial setting of a lab, to grasp that one's norm violations offer an opportunity that can be strategically exploited takes either game theory training or some uncommon, harsh real-life experience in which trust was at once very hard to come by and badly needed, as it was for the fictional kidnapper's victim in Schelling's original example. Moreover, for hawks, keeping mum does not seem a sign of guilt over how they played in T0 because they persist in their hawkishness and continue to defect. Rather than astute criminals or shrewd politicians, most of our hawks seem to be unsophisticated opportunists.

## Conclusion

The striking result of our experiment is that even among naïve subjects, some do grasp the advantage of SCI. In the condition in which compromising information is automatically shared, cooperation among hawks grows significantly: Although shy to share it, once the compromising information is out, some hawks know how to use it to cooperate provided that the penalty is available (T3; Figure 4b). More

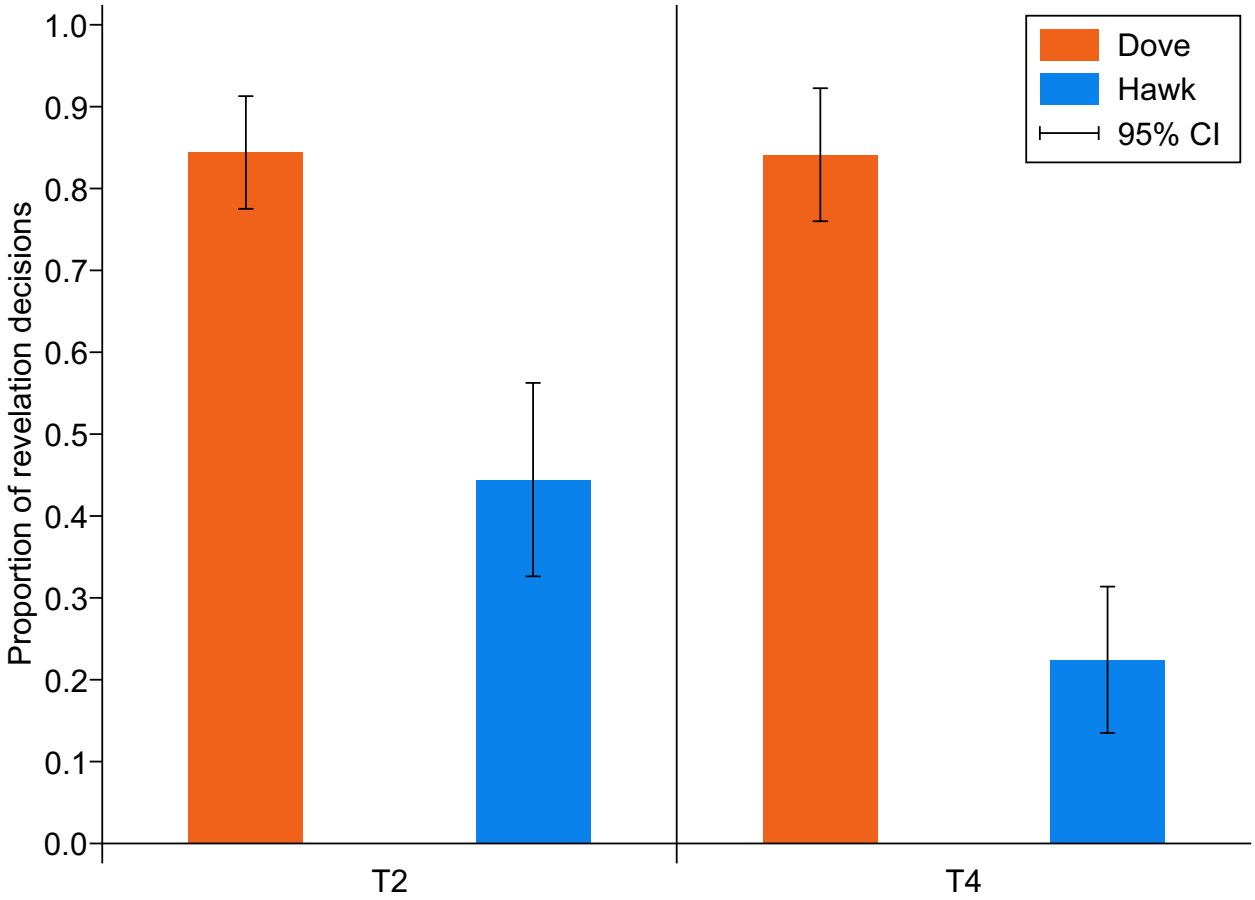


**Figure 5:** Proportion of revelation decisions.

striking still is the handful of hawks who do share the compromising information willingly (T4, in which they have the option to both reveal and inflict a penalty on each other), and their cooperation shoots up to the same level as that between doves. A transparent system with the option of "mutually assured destruction" makes (savvy) hawks as good at cooperating as doves are.

One might object and invoke alternative explanations for these findings stemming from other theories that have been used to explain cooperation among deviants. For example, social identity theory could offer an alternative explanation to the increase in cooperation between hawks (e.g., Aksoy 2015; Sherif 1966; Simpson 2006). If, after being assigned their labels, hawks cooperate more with hawks and doves cooperate more with doves, such evidence could be indicative of the working of the minimal group paradigm—the idea that even an arbitrary assignment of individuals to groups would induce these individuals to be positively biased toward members of their own group (Tajfel et al. 1971). However, this explanation cannot account for the fact that cooperation among hawks is only high in the conditions in which revealed hawks can be inflicted a penalty (T3 and T4); in conditions without the norm-enforcement stage (T1 and T2), hawks mostly defect with each other.

Appendix B provides a formal test of the alternative hypothesis stemming from social identity theory and reports results from other robustness checks.

Although remarkable, our findings are less reassuring. One does not need to be prosocial by character to embark on successful joint ventures (see Simpson and Willer 2015). The age-old question of what makes criminals cooperate finds a novel answer here: They can thrive parasitically in the shadow of norms put in place for a different purpose, exploiting them in a subtle way to support their joint ventures. Whenever there is a norm that is enforced, its breaches, once shared, have the potential to create a bond with partners.

This perverse effect of norms might explain why antisocial behaviors are far from extinct. Norms, whether legal or social, are part of the fabric of all societies, making the opportunities for SCI widespread—and not just for grand antisocial endeavours. Nothing in this strategy confines it to career criminals or to individuals, such as politicians, whose professions often require extralegal agreements. SCI is a strategy that can be adapted to a variety of ventures. In some cases, the norm breach that, once revealed, cements the bond and the cooperation that is achieved is in different activities: "I tell you about my crime X in order to cooperate with you on activity Y." Sometimes, the bonding breach is the same regarding which cooperation is pursued: "I tell you about my crime X to continue to do X in the knowledge that you will not betray me." This is one reason why SCI can be of value to a host of petty deviants, who count on the mutual complicity that they achieve after they deviate together at least once: adulterers, shirkers trading card-punching favors, or dope-taking teenagers. One of the costs that parties resorting to SCI are spared is having to deal with violence or getting involved with violent enforcers. This is what makes it a favored strategy of enforcement for "bourgeois" wrongdoers, such as tax evaders, embezzlers, patrons of prostitutes, or consumers of indecent material.

SCI belongs to the unintended consequences of normative systems. It surreptitiously exploits breaches of norms, whether legal or social, to strengthen cooperative bonds. Even norms that are designed rather than emerge spontaneously, however, do not take the possibility of SCI into consideration as a potential downside.

Can we draw any implication on how to design norms that minimize the opportunities of SCI? In order to thrive, SCI requires well-enforced normative systems (Baccara and Bar-Isaac 2008), such as making the mutual threat of denunciation credible. Weakening enforcement to counter SCI would of course defy the purpose of norms and have costs greater than benefits; it would be like hiding one's wallet in other people's pockets to prevent theft. It is the other condition that makes SCI reliable, namely the expectation that a breach is unlikely to be independently discovered by the norm enforcers, that can be targeted: SCI thrives when norms activate their enforcement reactively, as a result of denunciation rather than proactively pursuing norm breakers. If the risk exists that norm breakers are caught independently of denunciation, the bonds resting on the mutual threat become looser. By contrast, a normative system that is inefficient in pursuing norm breakers but efficient enough in punishing them when they are brought to attention is optimal from the point of view of SCI partners. Systems that produce many more norms than they can afford to enforce generate breaches that can be exploited by SCI. To reduce SCI opportunities, a close balance between a normative system and enforcement ought

to be maintained, whether by the norm enforcers, an investigating free media, or law-abiding inquisitive communities that cooperate with norm enforcers (Gambetta 2018).

SCI is also useless when the desired cooperation is among agents who, for various reasons, are immune to enforcement or are under competing normative systems. These agents must resort to alternative enforcement mechanisms. Peer punishment has been observed to promote cooperative ventures in groups as diverse as gangsters and medieval merchants (Gambetta 1993; Greif, Milgrom, and Weingast 1994). Hostage posting, for instance, was widely used until early-modern Europe to secure a truce, secure a peace treaty, or guarantee safe passage (Kosto 2012). Mafias use hostages even now, for instance, by making publicly known the whereabouts of the family of an imprisoned boss suspected of "ratting," thus making his family an easy target for retaliation; in the cocaine trade, "The Colombian 'family' used to send to Europe its members, to guarantee with their lives the family's promises to their European partners. Likewise, the European partners [the Calabrian Mafia] sent their hostages to Colombia" (Direzione Investigativa Antimafia, Reggio Calabria, testimony of state witness Giacomo Lauro on December 8, 1992, quoted in Barbazza 2011:67).

In addition to having to rely on a reactively functioning normative system, a condition that is not under the control of SCI partners, SCI has other limitations. First, in the symmetrical form, it can be difficult to initiate, as no one wants to be the first to disclose compromising information. Next, once a deal is completed, if hostages are rescued, vulnerability ceases, but with compromising evidence, it can be hard to ensure that no copies exist and would not be misused. Finally, SCI is exposed to shocks that can transform the payoffs, such as the arrest of one of the partners, who is offered leniency in return for incriminating the others. Still, we argue, SCI constitutes a mechanism that promotes cooperation in difficult circumstances in a manner that is genuinely different from repeated encounters, reputation formation, peer or centralized punishment, or hostage posting (see also Cook et al. 2007).

SCI is also flexible in the ways in which it can come about: Norm breakers may accidentally discover their misdeeds and realize ex-post the cooperative opportunity that the accident offers. Alternatively, with greater foresight and enticed by a remunerative joint venture, they may open their cupboard to reveal a skeleton in order to persuade their partners of their trustworthiness (Gambetta and Przepiorka 2014). Our experiment replicated both of these versions of SCI. There remains a third version we did not explore, the most strategically demanding one: the case in which the compromising information is deliberately produced in order to be shared. This corresponds with Schelling's case, in which the victim commits a crime in order to have something to share with and be trusted by the kidnapper.

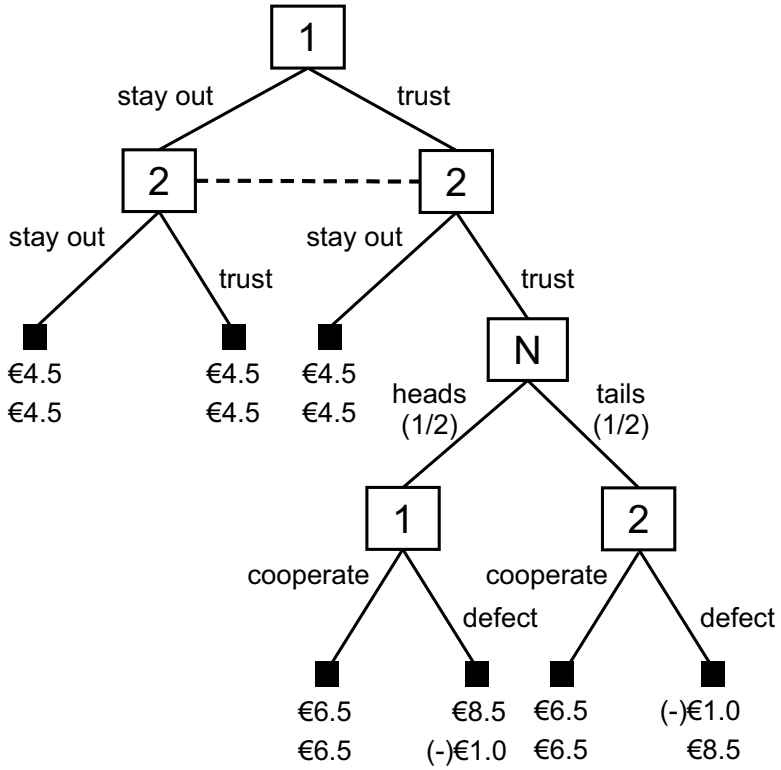


**Figure 6:** Extensive form of the VG without a norm-enforcement stage.

# Appendix A: Game Theoretic Analysis and Formal Derivation of Hypotheses

## *Self-Regarding Hawks and Other-Regarding Doves*

A rational, self-regarding, and risk-neutral (RSN) agent who believes that everybody is RSN will always stay out in the VG (Figure 6).

Proof: $4.5 > 0.5 \times 8.5 + 0.5 \times (-1) = 3.75$.

Under what conditions can we expect the VG to produce a balanced proportion of hawks and doves in T0? First, we assume the existence of agents who also have other-regarding preferences (e.g., Becker 1976).

A rational, (also) other-regarding and risk-neutral (RON) agent who believes that a proportion $p > 0$ of agents is RON to the extent that these agents would cooperate in the VG subgame[7] will trust and cooperate in the VG if $p > 0.467$. This is assuming that $1 - p$ is the proportion of RSN agents who trust and defect in the VG.

Proof:

$4.5 < 0.5 \times 6.5 + 0.5 \times (p \times 6.5 + [1 - p] \times [-1])$

$4.5 < 3.25 + p \times 3.25 - 0.5 + p \times 0.5$

$$1.75 < p \times 3.75 \Rightarrow p > 0.467.$$

Under these assumptions, does an RSN agent indeed trust and defect rather than stay out in the VG? An RSN agent who believes that a proportion $q$ of agents are RON will trust and defect rather than stay out in the VG if $q > 0.2$. This is assuming that $1 - q$ is the proportion of RSN agents who trust and defect in the VG.

Proof:

$$4.5 < 0.5 \times 8.5 + 0.5 \times [q \times 6.5 + (1 - q) \times (-1)]$$

$$4.5 < 4.25 + q \times 3.25 - 0.5 + q \times 0.5$$

$$0.75 < q \times 3.75 \Rightarrow q > 0.2.$$

Because $p > q$, we can ascertain that if the proportion of RON agents who trust and cooperate is greater than 47 percent, the remainder of the population will consist of RSN agents who trust and defect in the VG.

These calculations make it plausible that our parametrization of the VG will produce a balanced proportion of doves and hawks by the end of T0. According to our reasoning thus far, a dove corresponds to an RON agent, and a hawk corresponds to an RSN agent. In other words, the label "dove" is a sign of an agent's other-regarding preferences, and the label "hawk" is a sign of a lack thereof. The fact that in T0, agents can trust and defect up to two out of nine times and still be labeled "dove" merely accounts for the possibility of mistakes. Next, we derive our hypotheses for the different treatment conditions (T1 through T4).

## *Information Sharing*

Recall that in T1, at the start of every interaction, agents' labels are automatically revealed. That is, each agent knows whether they are a dove or a hawk, they know whether their interaction partner is a dove or a hawk, and they know that their interaction partner knows these facts as well. It is now relatively easy to see that a dove who meets a dove will trust and cooperate, a dove who meets a hawk will stay out, a hawk who meets a dove will trust and defect, and a hawk who meets a hawk will stay out. Note that a hawk who meets a dove will trust and defect (rather than stay out) because in the VG subgame, a dove will cooperate.

In T2, agents can choose whether or not to reveal their label. Revealing one's label is cost free. Those who do not reveal their label are labeled as such (i.e., "chose not to reveal label") but are not precluded from learning the label of their interaction partners who reveal their labels. Conditional on both parties to the interaction having revealed their labels, the predictions are the same as for T1. We can thus state our first hypothesis.

> H1: In conditions T1 and T2, a dove meeting a dove will trust and cooperate more than a dove meeting a hawk, a hawk meeting a dove, or a hawk meeting a hawk.

Because the cooperation rate among doves is high, doves gain from revealing their label in T2 and will do so at a high rate. Consequently, hawks are indifferent between

revealing and not revealing their label, as in either case they can be assumed to be hawks (see also Gambetta and Przepiorka 2014). We can thus state our second hypothesis.

H2: In condition T2, doves will more often reveal their label than hawks.

Note that the reasoning leading to H2 implies that H1 applies to all interactions in T2 and not only to those in which both parties reveal their labels. Because agents who do not reveal their label can be assumed to be hawks, they will be treated as such in the VG.

### *Norm Enforcement*

Condition T3 differs from T1 and condition T4 differs from T2 in that hawks who have their label revealed by design or reveal their label by choice, respectively, can be inflicted a penalty of 4.0 monetary units (MUs) after the VG if both parties to the interaction chose trust. Recall that inflicting a penalty on a hawk after the VG is associated with a cost of 0.5 MU. Hence, hawks, who are RSN, will never inflict a penalty on each other, whatever the outcome of the VG. However, recall that doves, who are RON, have a Fehr-Schmidt $\beta$ parameter greater than 0.21 and therefore will inflict a penalty on a hawk who defects.[7]

Proof:

$-1 - \beta(8.5 + 1) < -1.5 - \beta(4.5 + 1.5)$

$0.5 < 3.5\beta \Rightarrow \beta > 0.14.$

Hence, because doves cannot be inflicted a penalty and hawks do not inflict a penalty on each other, whatever the outcome of the VG, our expectations when doves and hawks meet their kind in T3 or T4 are the same as when they meet in T1 or T2. Before we examine the case in which a hawk meets a hawk in more detail, let us first ask what a hawk and a dove do when they meet each other in T3 or T4.

### *Resentful Doves*

Would a hawk meeting a dove that he or she expects to cooperate and punish a defection, trust and defect, or trust and cooperate? The hawk's expected payoff from trusting and defecting in this case is EP(trust and defect) = $0.5 \times (8.5 - 4) + 0.5 \times 6.5 = 5.5$, and the hawk's expected payoff from trusting and cooperating is EP(trust and cooperate) = 6.5. Hence, a hawk who meets a dove will trust and cooperate, and a dove will prefer to do the same rather than stay out. However, some of the doves may be resentful and inclined to inflict a penalty on hawks for the misdeeds that earned hawks their label. A hawk's expected payoffs from meeting a resentful dove are EP(trust and defect) = $0.5 \times (8.5 - 4) + 0.5 \times (6.5 - 4) = 3.5$ and EP(trust and cooperate) = $0.5 \times (6.5 - 4) + 0.5 \times (6.5 - 4) = 2.5$. In other words, in T3 and T4, a hawk who meets a resentful dove will stay out, whereas a hawk who meets a nonresentful dove will trust and cooperate. If the proportion $r$ of resentful doves is greater than 0.5, hawks prefer to stay out rather than trust and cooperate.

Proof:

$4.5 > r \times (2.5) + (1 - r) \times 6.5$

$4.5 > 6.5 - r \times 4 \Rightarrow r > 0.5.$

Thus, 50 percent of resentful doves will keep hawks' cooperation with doves low in T3 and T4. Under what conditions will two hawks cooperate in T3 and T4?

### *Vengeful Hawks*

The case in which two hawks interact lends itself to a classical, game theoretic analysis. Figure 7 shows the extensive form of the VG with the penalty option. It was produced with the software Gambit (McKelvey, McLennan, and Turocy 2016) and shows the decision situation faced by two hawks who both had their label revealed in T3 or who both chose to reveal their label in T4. The dashed lines denote players' information sets. Recall that by inflicting a penalty on one's interaction partner, the interaction partner incurs a cost of €4, and the agent inflicting the penalty incurs a cost of €0.50.

In the VG with the possibility of inflicting a penalty on one's interaction partner (which both hawks have), each hawk has 128 pure strategies and an infinite number of mixed strategies to choose from. Given the infinite number of strategy profiles, we used the software Gambit (McKelvey et al. 2016) to obtain the Nash equilibria of the game. Gambit calculated 269 Nash equilibria, all but four involve one or both hawks staying out. Of the four equilibria in which both hawks trust, both also cooperate and inflict a penalty on the other hawk if and only if the other hawk defects. Three of these four equilibria involve mixing at the penalty stage; one equilibrium is in pure strategies. In the pure strategy Nash equilibrium, both hawks trust, cooperate, and inflict a penalty on the other hawk only if the other hawk defects. This analysis shows that SCI is a cooperative equilibrium strategy in the VG with the penalty option (henceforth, SCI equilibrium). However, as already indicated, the four equilibria in which both hawks trust and cooperate are not subgame perfect; if a hawk deviates and defects instead of cooperating (e.g., by mistake), the other hawk has no incentive to penalize this defection.

Although the SCI equilibrium is not subgame perfect, it may still be selected by hawks. Among the many pure strategy equilibria of the game, it is the only equilibrium in which both hawks choose trust. It therefore establishes a focal point (Schelling 1960). Moreover, the choice of the SCI equilibrium may be reinforced by the existence of vengeful hawks. We define a vengeful hawk as one who incurs a cost to inflict a penalty on the hawk they interact with if and only if that hawk defects in the VG. If the proportion of vengeful hawks is $v > 0.5$, then both vengeful and nonvengeful hawks prefer to use SCI rather than stay out.

Proof:

If $v = 0$, the nonvengeful hawks will stay out. If $v = 1$, the nonvengeful hawks are indifferent between staying out and trusting and defecting. In other words, the presence of a proportion $v < 1$ of vengeful hawks does not change nonvengeful hawks' inclination to stay out. However, with

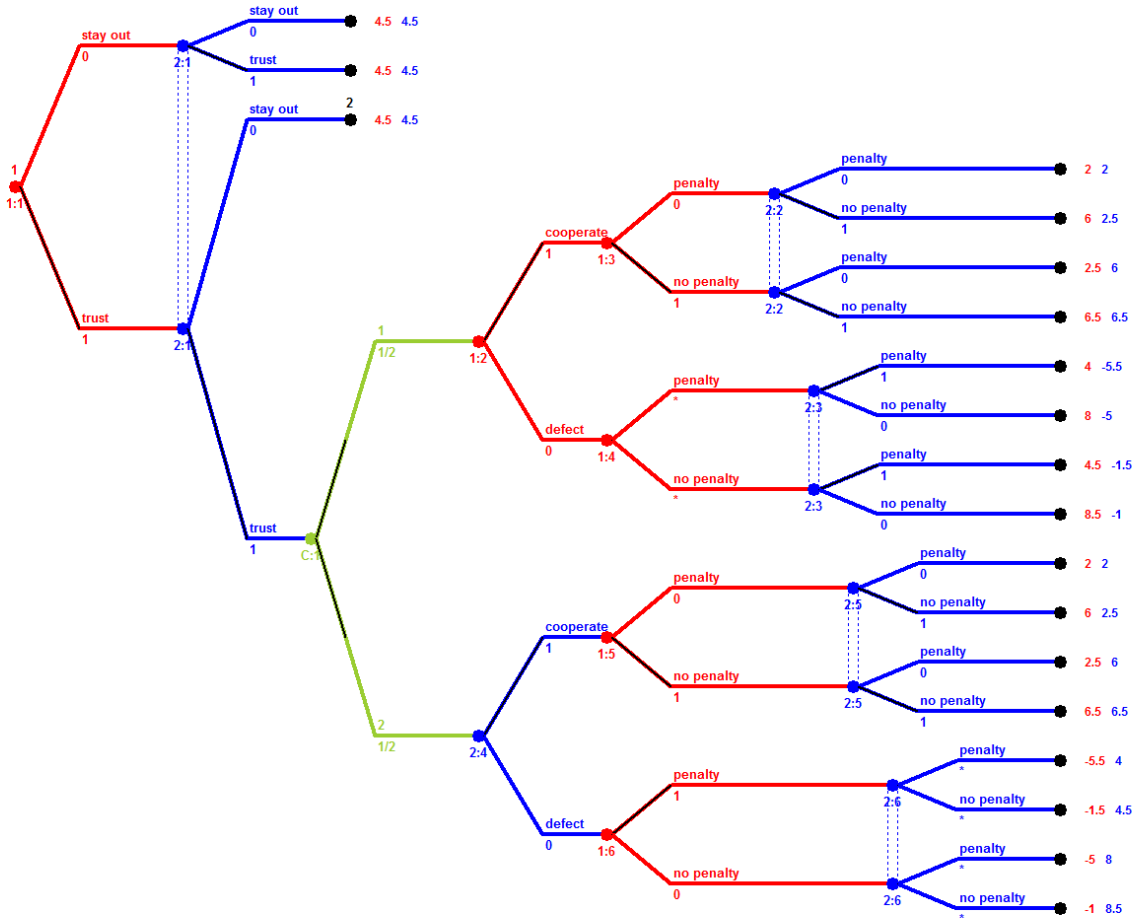


**Figure 7:** Extensive form of the VG with a norm-enforcement stage.

mostly vengeful hawks choosing to trust and cooperate, nonvengeful hawks are better off choosing to trust and cooperate than stay out. But now, as all hawks choose SCI, isn't it better for a hawk to trust and defect? Not as long as the proportion of vengeful hawks is $v > 0.5$.

$6.5 > v \times (8.5 - 4) + (1 - v) \times 8.5$

$6.5 > v \times (-4) + 8.5 \Rightarrow v > 0.5$.

Hence, assuming 50 percent or more resentful doves and 50 percent or more vengeful hawks, we can state our next hypothesis.

H3: In conditions T3 and T4, a dove meeting a dove will trust and cooperate more than a dove meeting a hawk or a hawk meeting a dove but not more than a hawk meeting a hawk.

Because now the cooperation rate among hawks is also high, both doves and hawks gain from revealing their label in T4 and will do so at a high rate. We can thus state our fourth hypothesis.

> H4: In condition T4, doves will not reveal their label more often than hawks.

Hypotheses H3 and H4 are strict in the sense that they imply that with a penalty option, hawks will start behaving like doves. Because this expectation is extreme, we relax both hypotheses by comparing hawks with hawks across conditions rather than comparing hawks with doves within conditions. The more lenient versions of hypotheses H3 and H4 are, respectively, as follows.

> H3b: In conditions T3 and T4, a hawk meeting a hawk will trust and cooperate more than a hawk meeting a hawk in conditions T1 and T2, respectively.
>
> H4b: In condition T4, hawks will reveal their label more often than hawks in condition T3.

## Appendix B: Testing Predictions from Labeling Theory and Social Identity Theory

Our experimental design also lends itself to test predictions made by two classic theories: labeling theory (Becker 1963) and social identity theory (Tajfel 1981). One of labeling theory's core predictions, that people's behavior is influenced by the terms used to describe them (the secondary deviance hypothesis), can be tested as a by-product of our experimental design. By contrast, a prediction derived from social identity theory could offer an alternative explanation if cooperation rates of hawks and doves with their own kind go against our hypotheses. The theory predicts a bias in favor of in-group members and has been shown to be present even when groups are arbitrarily defined (Tajfel et al. 1971). Although the aim of our experiment was not to test them, these theories have been very influential in explaining deviance and cooperation (Aksoy 2015; Bernburg, Krohn, and Rivera 2006; Hornsey 2016; Kroska, Lee, and Carr 2017; Simpson 2006).

### *Testing for Labeling Effects*

Labeling theory (Becker 1963) suggests that deviant individuals who are labeled as such will engage in deviant behavior not only because of their deviant personality but also by the mere fact that they have been labeled as deviant. According to the secondary deviance hypothesis, negative labels and stigmas change individuals' self-conceptions and behaviors to conform to the meaning of the labels. In other words, negative labels create self-fulfilling prophecies (Merton 1948). But how can the labeling effect be disentangled from individuals' natural inclination to engage in deviant behavior?

In our experiment, subjects who defected three times or more in T0 were subsequently labeled "hawks," and subjects who defected fewer than three times in T0

were labeled "doves." Our choice of three defections as the cutoff value in the labeling process is arbitrary, merely based on the previous finding that it will produce a similar number of hawks and doves in each session (which it does). Incidentally, our implementation of the labeling process constitutes a regression discontinuity design (Shadish, Cook, and Campbell 2002), which allows us to estimate the labeling effect net of subjects' natural inclination to defect.

If our labeling process induced a labeling effect, we should observe a jump (i.e., discontinuity) in the number of defections in T1 at the cutoff value. That is, hawks, who defected three times or more in T0, should defect disproportionately more in T1 than doves, who defected two times or fewer in T0. By "disproportionately" we mean that the function of the number of defections in T1 shows a displacement at the cutoff value. If present in T1, we can also expect to find evidence for the labeling effect in T3. However, the effect will be reduced in T3 compared with T1 because unlike in T1, in T3, hawks can be subjected to a penalty after the VG, whereas doves cannot.

Figure 8 shows the number of subjects' defections in T1 (panel a) and T3 (panel b) as a function of these subjects' defections in T0. Recall that different subjects participated in T1 than in T3. Figure 8a shows that there is a substantial and statistically significant positive relation between subjects' number of defections in T0 and T1 (b = 0.883; $p < 0.001$). This relation is much smaller between T0 and T3 (b = 0.199; $p = 0.048$), most likely due to the hampering effect of punishment threat directed at hawks in T3 (Figure 8b). However, neither in T1 nor in T3 is there evidence for a vertical discontinuity in this relation (d = 0.090 [$p = 0.938$] and d = –0.248 [$p = 0.778$], respectively). Hence, our results are not in line with predictions derived from labeling theory.

### *Testing for Social Identity Effects*

Social identity theory (Tajfel 1981) starts from the assumption that any social behavior can be conceived of as taking place on the continuum of being purely between individual and purely between group. Depending on the relative salience of individual characteristics and group categories, an individual's behavior will be motivated by their and their interaction partners' individual characteristics or group memberships, respectively (Turner et al. 1987). In a series of experiments, Tajfel et al. (1971) demonstrated that even subtle cues of group categorization, such as the sharing of a preference for an unknown painter, are sufficient to induce actors' behavior to be motivated by group memberships rather than by individual characteristics. What later became known as the minimal group paradigm had henceforth been used in experiments to induce social categorization, by which individuals would be more inclined to cooperate with members of their own group than with members of the other group (also see Aksoy 2015; Sherif 1966; Simpson 2006).

In our experiment, subjects obtained the labels "dove" or "hawk" based on their behavior in T0. Moreover, subjects were made aware of what the labels stand for, namely that doves had defected fewer than three times and hawks had defected three times or more in T0. In other words, the labeling process we implemented in our experiment can be conceived of as an application of the minimal group

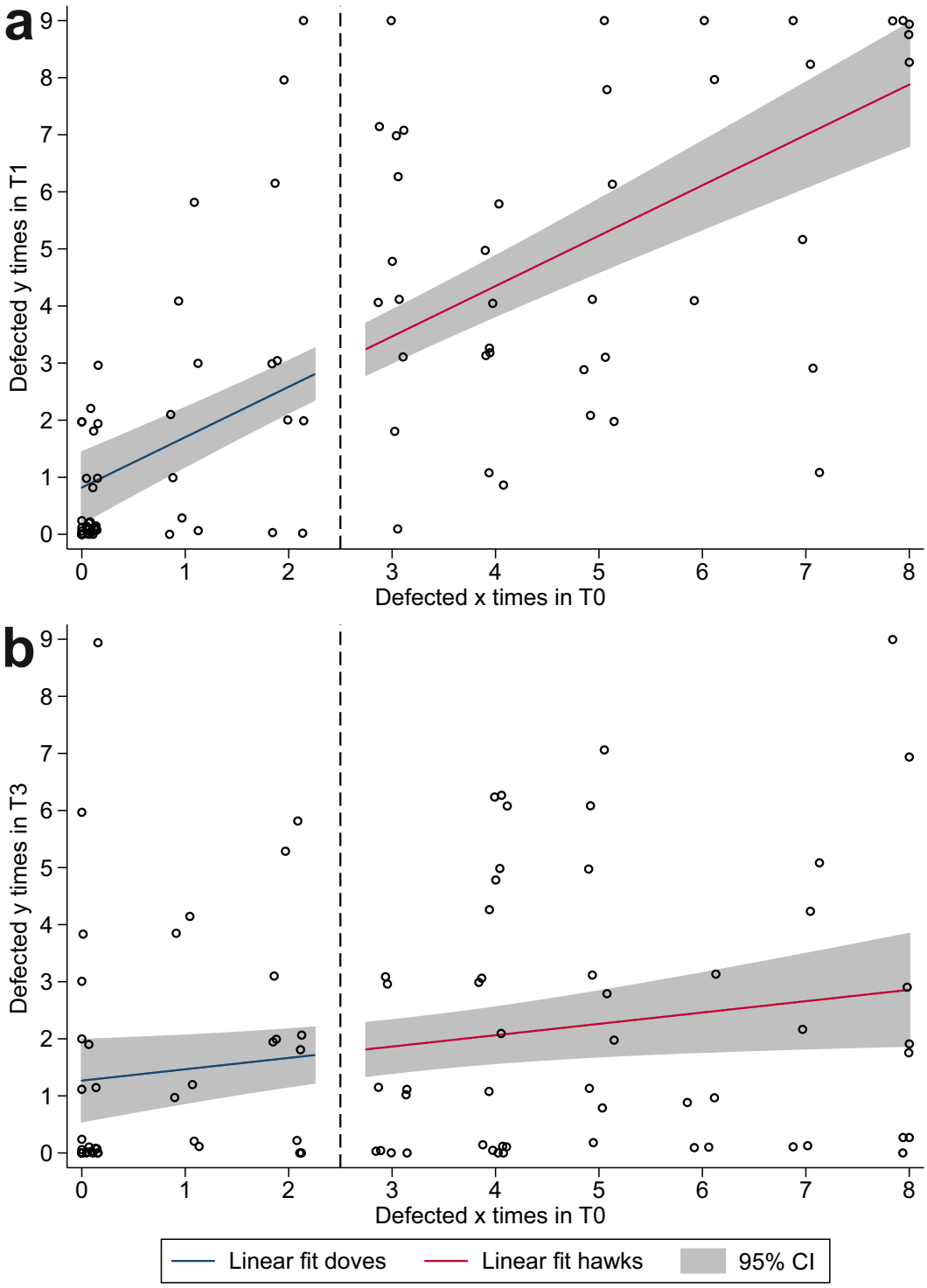


**Figure 8:** A, Regression discontinuity designs testing for labeling effects in T1. B, Regression discontinuity designs testing for labeling effects in T3.

paradigm. Hence, the fact that doves are more inclined to cooperate with doves than with hawks and that hawks are more inclined to cooperate with hawks than with doves in T3 and T4 could be due to a group categorization effect.

The best way to test this conjecture is by comparing doves' and hawks' behavior before and after they obtained and were informed about their labels. That is, by comparing doves' and hawks' cooperation rates in T0 with their cooperation rates in T1, we can directly test the group categorization effect. Note that T3 and T4 are not suitable for comparison with T0 because in these treatment conditions, the punishment threat faced by hawks also has a bearing on subjects' behavior. The only difference between T0 and T1 is that in T1, subjects are divided in two groups, whereas in T0, they are entirely anonymous. It could be argued, moreover, that in T2, the group categorization effect should be stronger, as actors also have the opportunity to signal their commitment to their group by revealing their label (Sosis 2005).

By comparing cooperation and defection rates of doves and hawks shown in Figure 3 with the corresponding rates in Figure 4a and 4b, respectively, we can see that group categorization does not make a difference—neither in T1 nor in T2. Doves continue to be as cooperative with each other in T1 as they were in T0 with an anonymous interaction partner without a label ($z = 0.81$; $p = 0.418$), and the same is true for hawks ($z = 1.09$; $p = 0.277$). The differences in cooperation rates remain insignificant if T0 and T2 are compared for doves ($z = 0.68$; $p = 0.499$) and even drops significantly for hawks ($z = 2.42$; $p = 0.016$). Based on this evidence, we can rule out that our results are driven by a group categorization effect.

## Notes

1 This is fictionalized in The Good Shepherd, a 2006 film directed by Robert De Niro.

2 Ashley Madison was eventually hacked and subscribers' personal details publicly disclosed with often catastrophic consequences for the would-be adulterers (Mansfield-Devine 2015).

3 For a comprehensive discussion of other mechanisms, see Cook et al. (2007) and Simpson and Willer (2015).

4 On peer punishment, see, for instance, Fehr and Gaechter (2002), Diekmann and Przepiorka (2015), and Horne (2009); on centralized punishment, see, for instance, Traulsen, Röhl, and Milinski (2012). There is also real-life evidence for centralized punishment, in which the interacting parties bear the cost by paying a third party, such as mafias or paramilitaries, who sell their protection (Gambetta 1993; Gambetta and Reuter 1995).

5 Both are mechanisms that link costly trustworthy behavior to higher payoffs: When interactions are repeated, cooperation can become a self-serving best response (Axelrod 1984; Dal Bó 2005; Fudenberg and Maskin 1986). A reputation for being trustworthy that is costly to acquire can pay off over time (Diekmann et al. 2014; Gambetta 2011; Hume [1740] 1969; Nelson 1974; Shapiro 1983). Even illegal drugs have been successfully transacted among anonymous buyers and sellers in "cryptomarkets," which are backed by an electronic reputation system (Przepiorka, Norbutas, and Corten 2017).

6 The standard PD is a two-person game in which both agents decide simultaneously whether to cooperate or defect. If both agents cooperate, both earn a payoff R ; if both

agents defect, both earn a payoff P ; and if one agent cooperates and the other agent defects, the cooperating agent earns a payoff S and the defecting agent earns a payoff T . The payoffs in the PD are ordered such that irrespective of what the other agent does, defection leads to a higher payoff ( $T > R > P > S$ ). However, if both agents follow this strategy, they end up earning less than what they would earn from cooperating with each other ( $P < R$ ). The standard TG has the same payoffs as the PD but differs from the PD in two important ways: (1) In the TG, one of the agents is the first mover (i.e., the truster), the other agent is the second mover (i.e., the trustee), and the second mover observes the first mover's choice before making theirs. (2) If the first mover defects, the second mover cannot make a decision, and both agents earn P . Both the PD and the TG are so-called social dilemmas because individually rational actions lead to collectively inferior outcomes (e.g., Kollock 1998).

7 According to the Fehr-Schmidt model (Fehr and Schmidt 1999), for example, agents with a relatively mild aversion to a lower payoff for their interaction partners will chose to cooperate rather than defect in the VG subgame. Proof: $6.5 > 8.5 - \beta(8.5 + 1)$ if $\beta > 0.21$.

## References


Aksoy, Ozan. 2015. "Effects of Heterogeneity and Homophily on Cooperation." *Social Psychology Quarterly* 78:324–44.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York, NY: Basic Books.

Baccara, Mariagiovanna, and Heski Bar-Isaac. 2008. "How To Organize Crime." *Review of Economic Studies* 75:1039–67. https://doi.org/10.1111/j.1467-937x.2008.00508.x.

Bader, Felix, Bastian Baumeister, Roger Berger, and Marc Keuschnigg. 2019. "On the Transportability of Laboratory Results." *Sociological Methods and Research*, first published on February 20, 2019, as doi.org/10.1177/0049124119826151.

Barbazza, Ruggero. 2011. *Le Alleanze Criminali - 'Ndrangheta e Clan Colombiani*. Tesi di Laurea, Facolta' di Scienze Politiche, Università di Milano.

Becker, Gary S. 1976. "Altruism, Egoism and Genetic Fitness: Economics and Sociobiology." *Journal of Economic Literature* 14:817–26.

Becker, Howard Saul. 1963. *Outsiders: Studies in the Sociology of Deviance*. Glencoe, IL: Free Press.

Bernburg, Jón Gunnar, Marvin D. Krohn, and Craig J. Rivera. 2006. "Official Labeling, Criminal Embeddedness, and Subsequent Delinquency: A Longitudinal Test of Labeling Theory." *Journal of Research in Crime and Delinquency* 43:67–88. https://doi.org/10.1177/0022427805280068.

Bigoni, Maria, Sven-Olof Fridolfsson, Chloe Le Coq, Giancarlo Spagnolo. 2015. "Trust, Leniency, and Deterrence." *Journal of Law, Economics, and Organization* 31:663–89. https://doi.org/10.1093/jleo/ewv006.

Brandts, Jordi, and Gary Charness. 2011. "The Strategy versus the Direct-Response Method: A First Survey of Experimental Comparisons." *Experimental Economics* 14:375–98. https://doi.org/10.1007/s10683-011-9272-x.

Cook, Karen S., Russell Hardin, and Margaret Levi. 2007. *Cooperation without Trust?* New York, NY: Russell Sage Foundation. https://doi.org/10.1093/sf/86.2.851.

Dal Bó, Pedro. 2005. "Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games." *American Economic Review* 95:1591–604. https://doi.org/10.1257/000282805775014434.

Dasgupta, Partha. 1988. "Trust as a Commodity." Pp. 49–72 in *Trust: Making and Breaking Cooperative Relations*, edited by D. Gambetta. Oxford, United Kingdom: Basil Blackwell.

Diekmann, Andreas, and Wojtek Przepiorka. 2015. "Punitive Preferences, Monetary Incentives and Tacit Coordination in the Punishment of Defectors Promote Cooperation in Humans." *Scientific Reports* 5:10321. https://doi.org/10.1038/srep10321.

Diekmann, Andreas, Ben Jann, Wojtek Przepiorka, and Stefan Wehrli. 2014. "Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets." *American Sociological Review* 79:65–85. https://doi.org/10.1177/0003122413512316.

Falk, Armin, and James J. Heckman. 2009. "Lab Experiments Are a Major Source of Knowledge in the Social Sciences." *Science* 326:535–8.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114:817–68.

Fehr, Ernst, and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415:137–40. https://doi.org/10.1038/415137a.

Fischbacher, Urs. 2007. "Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10:171–8. https://doi.org/10.1007/s10683-006-9159-4.

Flashman, Jennifer, and Diego Gambetta. 2014. "Thick as Thieves: Homophily and Trust among Deviants." *Rationality and Society* 26:3–45. https://doi.org/10.1177/1043463113512996.

Fudenberg, Drew, and Eric Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica* 54:533–54. https://doi.org/10.2307/1911307.

Gambetta, Diego. 1993. *The Sicilian Mafia*. Cambridge, MA: Harvard University Press.

Gambetta, Diego. 2009. *Codes of the Underworld. How Criminals Communicate*. Princeton, NJ: Princeton University Press. https://doi.org/10.1086/655451.

Gambetta, Diego. 2011. "'Response' to 'Roundtable on Codes of the Underworld."' *Global Crime* 12:150–9.

Gambetta, Diego. 2018. "Why Is Italy Disproportionally Corrupt? A Conjecture." Pp. 133–64 in *Institution, Governance and the Control of Corruption*, edited by K. Basu and T. Cordella. Basingstoke, United Kingdom: Palgrave Macmillan. https://doi.org/10.1007/978-3-319-65684-7_6.

Gambetta, Diego, and Peter Reuter 1995. "Conspiracy among the Many: The Mafia in Legitimate Industries." Pp. 116–36 in *The Economics of Organized Crime*, edited by G. Fiorentini and S. Peltzman. Cambridge, United Kingdom: Cambridge University Press. https://doi.org/10.1017/cbo9780511751882.010.

Gambetta, Diego, and Wojtek Przepiorka. 2014. "Natural and Strategic Generosity as Signals of Trustworthiness." *PLoS ONE* 9:e97533. https://doi.org/10.1371/journal.pone.0097533.

Greif, Avner, Paul Milgrom, and Barry R. Weingast. 1994. "Coordination, Commitment, and Enforcement: The Case of the Merchant Guild." *Journal of Political Economy* 102:745–76. https://doi.org/10.1086/261953.

Greiner, Ben. 2015. "Subject Pool Recruitment Procedures: Organizing Experiments with Orsee." *Journal of the Economic Science Association* 1:114–25. https://doi.org/10.1007/s40881-015-0004-4.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 3:61–83. https://doi.org/10.1017/s0140525x0999152x.

Horne, Christine. 2009. *The Rewards of Punishment: A Relational Theory of Norm Enforcement*. Stanford, CA: Stanford University Press. https://doi.org/10.1017/s0003975609990208.

Hornsey, Matthew J. 2016. "Dissent and Deviance in Intergroup Contexts." *Current Opinion in Psychology* 11:1–5. https://doi.org/10.1016/j.copsyc.2016.03.006.

Hume, David. [1740] 1969. *A Treatise of Human Nature*. Harmondsworth, United Kingdom: Penguin Books.

Jackson, Michelle, and David R. Cox. 2013. "The Principles of Experimental Design and Their Application in Sociology." *Annual Review of Sociology* 39:27–49.

Jann, Ben. 2007. "Making Regression Tables Simplified." *Stata Journal* 7:227–44.

Jann, Ben. 2014. "Plotting Regression Coefficients and Other Estimates." *Stata Journal* 14:708–37. https://doi.org/10.1177/1536867x1401400402.

Kollock, Peter. 1998. "Social Dilemmas: The Anatomy of Cooperation." *Annual Review of Sociology* 24:183–214. https://doi.org/10.1146/annurev.soc.24.1.183.

Kopanyi-Peuker, Anita, Theo Offerman, and Randolph Sloof. 2017. "Fostering Cooperation through the Enhancement of Own Vulnerability." *Games and Economic Behavior* 101:273–90. https://doi.org/10.1016/j.geb.2015.10.001.

Kosto, Adam J. 2012. *Hostages in the Middle Ages.* Oxford, United Kingdom: Oxford University Press. https://doi.org/10.1515/hzhz-2015-0224.

Kreps, David. 1990. "Corporate Culture and Economic Theory." Pp. 90–143 in *Perspectives on Positive Political Economy*, edited by J. E. Alt and K. A. Shepsle. Cambridge, MA: Cambridge University Press. https://doi.org/10.1017/cbo9780511571657.006.

Kreps, David M., and Robert Wilson. 1982. "Sequential Equilibria." *Econometrica* 50:863–94. https://doi.org/10.2307/1912767.

Kroska, Amy, James Daniel Lee, and Nicole T. Carr. 2017. "Juvenile Delinquency, Criminal Sentiments, and Self-Sentiments: Exploring a Modified Labeling Theory Proposition." Pp. 21–47 in *Advances in Group Processes*. Vol. 34, edited by S. R. Thye and E. J. Lawler. Bingley, United Kingdom: Emerald Publishing Limited. https://doi.org/10.1108/s0882-614520170000034002.

Levitt, Steven D., and John A. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" *Journal of Economic Perspectives* 21:153–74. https://doi.org/10.1257/jep.21.2.153.

Mansfield-Devine, Steve. 2015. "The Ashley Madison Affair." *Network Security* 2015:8–16. https://doi.org/10.1016/s1353-4858(15)30080-5.

McKelvey, Richard D., Andrew M. McLennan, and Theodore L. Turocy. 2016. "Software Tools for Game Theory, Version 16.0.1." Gambit. https://sourceforge.net/projects/gambit/files/gambit16/16.0.1/.

Merton, Robert K. 1948. "The Self-Fulfilling Prophecy." *Antioch Review* 8:193–210.

Nakayachi, Kazuya, and Motoki Watabe. 2005. "Restoring Trustworthiness after Adverse Events: The Signaling Effects of Voluntary 'Hostage Posting' on Trust." *Organizational Behavior and Human Decision Processes* 97:1–17. https://doi.org/10.1016/j.obhdp.2005.02.001.

Nelson, Phillip. 1974. "Advertising as Information." *Journal of Political Economy* 82:729–54.

Przepiorka, Wojtek, Lukas Norbutas, and Rense Corten. 2017. "Order without Law: Reputation Promotes Cooperation in a Cryptomarket for Illegal Drugs." *European Sociological Review* 33:752–64. https://doi.org/10.1093/esr/jcx072.

Raub, Werner, and Gideon Keren. 1993. "Hostages as a Commitment Device: A Game-Theoretic Model and an Empirical Test of Some Scenarios." *Journal of Economic Behavior and Organization* 21:43–67. https://doi.org/10.1016/0167-2681(93)90039-r.

Rosenbaum, Ron. 2000. *The Secret Parts of Fortune: Three Decades of Intense Investigations and Edgy Enthusiasms*. New York, NY: Random House.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Selten, Reinhard. 1967. "Die Strategiemethode zur Erforschung Des Eingeschränkt Rationalen Verhaltens Im Rahmen Eines Oligopolexperiments." Pp. 136–68 in *Beiträge zur Experimentellen Wirtschaftsforschung*, edited by H. Sauermann. Tübingen, Germany: Mohr Siebeck.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin. https://doi.org/10.1086/345281.

Shapiro, Carl. 1983. "Premiums for High Quality Products as Return to Reputation." *Quarterly Journal of Economics* 98:659–80.

Sherif, Muzafer, ed. 1966. *Group Conflict and Co-operation: Their Social Psychology*. London, United Kingdom: Routledge and Kegan Paul Ltd. https://doi.org/10.1177/030639686800900327.

Simpson, Brent. 2006. "Social Identity and Cooperation in Social Dilemmas." *Rationality and Society* 18:443–70. https://doi.org/10.1177/1043463106066381.

Simpson, Brent, and Robb Willer. 2015. "Beyond Altruism: Sociological Foundations of Cooperation and Prosocial Behavior." *Annual Review of Sociology* 41:43–63. https://doi.org/10.1146/annurev-soc-073014-112242.

Sosis, Richard. 2005. "Does Religion Promote Trust? The Role of Signaling, Reputation, and Punishment." *Interdisciplinary Journal of Research on Religion* 1:1–30.

Tajfel, Henri. 1981. *Human Groups and Social Categories: Studies in Social Psychology*. Cambridge, United Kingdom: Cambridge University Press. https://doi.org/10.1017/s0033291700041374.

Tajfel, Henri, M. G. Billig, R. P. Bundy, and Claude Flament. 1971. "Social Categorization and Intergroup Behavior." *European Journal of Social Psychology* 1:149–78. https://doi.org/10.1002/ejsp.2420010202.

Traulsen, Arne, Torsten Röhl, and Manfred Milinski. 2012. "An Economic Experiment Reveals That Humans Prefer Pool Punishment To Maintain the Commons." *Proceedings of the Royal Society B* 279:3716–21. https://doi.org/10.1098/rspb.2012.0937.

Turner, John C., Michael A. Hogg, Penelope J. Oaks, Stephan D. Reicher, and Margaret S. Wetherell. 1987. *Rediscovering the Social Group: A Self-Categorization Theory*. Oxford, United Kingdom: Basil Blackwell.

Varese, Federico, Peng Wang, and Rebecca W. Y. Wong. 2018. "'Why Should I Trust You with My Money?': Credible Commitments in the Informal Economy in China." *British Journal of Criminology* 59:594–613. https://doi.org/10.1093/bjc/azy061.

Walden, Celia. 2010. "Celia Walden and Ashley Madison, the Infidelity Website." *The Telegraph*, October 5. Retrieved February 7, 2017 (https://www.telegraph.co.uk/women/sex/8041989/Celia-Walden-and-Ashley-Madison-the-infidelity-website.html).

Williamson, Oliver E. 1983. "Credible Commitments: Using Hostages to Support Exchange." *American Economic Review* 73:519–40.

**Acknowledgments:** Both authors contributed equally to this work and thank Ozan Aksoy, Maria Bigoni, Manfred Milinski, and Werner Raub for their helpful comments on earlier versions. D. G. gratefully acknowledges financial support from the Research Council of the European University Institute.

**Diego Gambetta:** Collegio Carlo Alberto, Università di Torino.
E-mail: diego.gambetta@carloalberto.org.

**Wojtek Przepiorka:** Department of Sociology, Utrecht University.
E-mail: w.przepiorka@uu.nl.