The Ambivalence of Creating Peace Signals: How a Self-Binding Device Influences Aggression

Johanna Jauernig Matthias Uhl Christoph Luetge*

Technical University of Munich

*Chair of Business Ethics, Arcisstrasse 21, D-80333 Munich; Jauernig (corresponding author): Phone: +49.89.289.25132, E-Mail: johanna.jauernig@tum.de

Abstract

We investigated to which extent aggression between opponents after competition is triggered by preemptive retaliation or spite. To disentangle motives, we introduced a credible self-binding signal. In constrast to previous studies, we found that aggression against a defenceless subject plays an important role. Preemptive retaliation proofed to be another dominant motive. The role of self-binding signals to mitigate preemptive retaliation is ambivalent. They do lead to an overall de-escalation in our setting. Still, a considerable number of subjects left the self-binding device on the table, which constitutes a danger: If two unbound subjects meet, the knowingly unused signal leads to an escalation of conflict compared to a situation in which no self-binding device exists.

Keywords: Competition, money burning, spite, preemptive retaliation, self-binding, self-commitment.

JEL-Classification: C72, C90, C91

1 Introduction

In a society with deficient rule of law where people may carry weapons openly and use them in an act of self-administered justice at any time, the following dilemma arises: Two people can become involved in an exchange of fire without any intention of doing so. If one deems it probable that the other will draw a weapon, the first has no choice other than firing first to prevent being harmed himself. This rational Hobbesian reasoning may lead to a disastrous outcome irrespective of the peaceful intentions of the involved. In allusion to wild west mythology this situation is known as the gunfighter's dilemma (Cushing 1995).

This phenomenon of a preemptive strike or defensive aggression (Simunovic et al. 2013) is a prevalent reason for acts of aggression, be it military actions to prevent terrorist attacks or the assault of an alpha animal to deter challengers. To find out whether this is the actual motive for an attack, one pivotal property must be in place: fear of the opponent. In the introductory example this fear could be ruled out by a self-binding device which both gunfighters use as a commitment to peace (see Schelling 1980). A self-bound subject is no longer able to strike, therefore sending a strong and credible signal to the possible attacker. If an obviously bound subject is still harmed, spite is an important driver of aggression. Preemptive retaliation plays

a role if the punishment inflicted on an unbound subject is greater than the punishment inflicted on a bound subject.

If preemptive retaliation drives aggression, the presence of a self-binding device gives subjects the chance to credibly signal their peaceful intentions and may thus help to de-escalate conflict. But what if the device is not used? Saliently available self-binding signals which are not sent may not only be ineffective, but may even have a detrimental effect on aggression. In our introductory example this could be a deputy inviting the gunfighters to simultaneously throw away their weapons. If both hesitate, the situation may escalate as compared to a situation in which no deputy had ever entered the scene.

In experimental literature we find evidence for preemptive strikes. In the first-strike game, paired subjects have the option to strike against the other for a small fee, destroying the victim's previously accumulated earnings and severely reducing his future payoffs (Abbink and de Haan 2014). Furthermore, the victim loses the ability to deactivate the aggressor in later rounds. Thus subjects have strong incentives to strike first in order to prevent themselves from being harmed. Data shows that subjects seem to be so fearful of an aggressive opponent that they are willing to harm a fellow subject considerably. However, if there is no opponent to be feared, and spite were the only reason for a strike, the authors do not find a single instance. In a similar game which was played one shot costly money burning was analyzed (Simunovic et al. 2013). In one treament, participants could simultaneously burn each others' money, while in the other treatment one participant was passive. The findings are in line with Abbink and de Haan (2014). If the fear of the other is ruled out, almost no money burning remains.

As in Sinumovic et al. (2013), launching a preemptive strike does not hinder the opponent from attacking in our study, since the opponents strike simultaneously. It only prevents that the striking party is ultimately worse off than the opponent. The strike is therefore an act of preemptive retaliation caused by the fear of being the sucker (see, for instance, Yamagishi and Sato 1986 in the context of public good games). In psychological literature this phenomenon is discussed as "sugrophobia" (for an overview see Vohs and Baumeister 2007).

Empirical data indicate that there are considerable amounts of aggression after competition (Adachi and Willoughby 2011, Muller et al. 2012). In a study by Jauernig et al. (unpublished) 40 % of subjects burned parts of their opponent's endowment after competition in a one shot interaction. After competition means that this money burning did not give them any strategic advantage in the process of competition itself as in experiments allowing for sabotage (e.g.,

Harbring and Irlenbusch 2008, Harbring and Irlenbusch 2011). It is particularly noteworthy that destructive behavior after competition occurred although subjects could not hide their intentions behind a random destruction mechanism as was the case in the original joy-of-destruction game (Abbink and Sadrieh 2009). This finding suggests that spite might be a motivating driver for burning money. As cited above preemptive strikes are another mediator of aggression. We consider post-competitive aggression as a particularly fruitful field to explore the effectiveness of a self-binding device for reducing the level of aggression. Substantially, competition is a widespread feature of modern societies which emphasizes how relevant it is to adequately address its dark side. Methodologically, since the focus on aggression *after* competition abstracts from any strategic intentions, it allows for a clear disentanglement of the motives of preemptive retaliation and spite.

The creation of a costless self-binding device serves two purposes: First, it creates a setting where unbound and bound subjects meet within and across groups, which allows us to disentangle the motives of preemptive retaliation and spite. Second, it enables us to test whether self-binding is a mechanism fit to mitigate or even resolve aggressive behavior after competition.

2 Hypotheses

We find money burning in various studies. In this form of punishment, any extrinsic incentive to harm, such as a monetary gain for the attacker, is ruled out. Assume that subjects engage in a competition and know that they will soon have the possiblity to burn their opponent's money. Would they use a self-binding signal to abstain from money burning, if offered after competition? It is plausible that the rationale to self-bind or to abstain from doing so is driven by the following motives.

Other-regarding preferences exercise an influence in two different ways. Subjects may have an urge to punish their former opponent. Their reasoning is less driven by preventing their own endowment from destruction than by inflicting harm on others. Consequently they would not bind themselves and engage in punishment, regardless of whether or not the target is bound. We refer to this behavior as spite.

A second other-regarding preference is the fear of being the sucker. In this case subjects do not have an intention to harm the other subject, but want to avoid being worse off than the other one under any circumstances. These subjects would consequently not bind them-

selves. However, they would only have a reason to punish an unbound opponent, since they only exercise punishment as an act of preemptive retaliation.
If subjects are mainly motivated by self-regarding preferences, they will primarily care about preventing their own endowment from destruction. They are thus likely to use the self-binding device to signal their own peaceful intentions to the opponent. By that they reduce the risk of being attacked. Although using the device does not prevent being attacked by spiteful opponents, it rules out preemptive retaliatory motives in the opponent.

In order to determine whether the presented motives actually play a role, we put forward the following hypotheses.

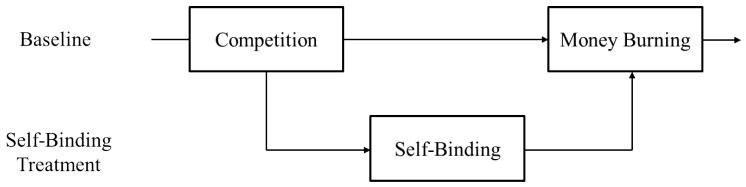
Hypothesis 1: Spite. Unbound subjects will punish bound subjects.

Hypothesis 2: Preemptive Retaliation. Unbound subjects will punish other unbound subjects more severely than they will punish bound subjects.

What is the overall effect of creating self-binding devices in such a situation? The existence of a self-binding device reduces the uncertainty about an opponent's motives in comparison to a situation where no such device exists. Consider the reasoning of an unbound subject who is about to decide. If her opponent used the device, it is clear that his intentions are peaceful. If her opponent has not used the device, it is clear that his preferences must be other-regarding in one way or the other. If preemptive retaliation explains at least parts of the decider's behavior, this reasoning increases the probability that the decider will strike against her opponent. Why is this the case? Either the opponent did not self-bind, because he is spiteful or because he fears being the sucker. If spite is prevalent, the decider should punish in order to avoid being worse off than her opponent. If fear of being the sucker is prevalent, the decider knows that her opponent may strike against her because the oppenent fears to be worse off if he does not. Anticipating this also leads the decider to carry out a preemptive strike. Although neither have any spiteful intentions, their mutual fear puts them in a dilemma: the situation escalates.

Hypothesis 3: Escalation Through Unused Signals. If two intentionally unbound subjects encounter each other, punishment will be higher than in a situation in which no self-binding device exists.

Clip 1: Course of the Experiment



If this hypothesis is confirmed, the impact of creating a self-binding device is ambivalent depending on the prevalence of other-regarding preferences. If other-regarding preferences play a role in the population, the creation of a salient self-binding device may backfire. If the self-binding device is intentionally not used, this establishes a signal which reduces uncertainty in the sense that it has become more likely that the opponent constitutes a threat. The overall effect of establishing a self-binding device is therefore not obvious.

3 Experiment Design

Subjects are randomly matched to pairs and engage in a speed-based calculation task. In this task, they are confronted with a sequence of 19 matrices and have to identify two numbers adding up to 10. There are exactly two pairs of numbers adding up to 10 in each matrix. The first to click on a correct pair of numbers scores one point. The slower scores one point, if the faster has clicked on a wrong pair. After 19 rounds a winner and a loser are determined according to the total score. In a second stage, both are equally endowed with 10 ECU (10 ECU = €0.70) and have the chance to simultaneously reduce their opponent's money by any integer down to zero. This setting without a self-binding option taken from Jauernig et al. (unpublished) serves as a baseline.

To disentangle the motives of spite and preemptive retaliation, we introduce a self-binding stage between the competition part and money burning (see Clip 1). In this intermediary stage, subjects may choose to bind themselves by deactivating a future possibility to burn their opponent's money. This commitment is public and free of charge. Subjects simultaneously choose to self-bind or not by clicking the respective button. Thereafter, choices are reported and obligatory. Subjects may then reduce their opponent's endowment depending on their choice of whether to keep the option to burn money or not.

Subjects' choices are implemented with varying probabilities: A

loser is harmed by a winner's punishment with 60 % probability, whereas a winner is harmed by a loser's punishment with 40 % probability. Asymmetric probabilities which expose the winner to a lower risk of being harmed were implemented to safeguard against the possibility that a subject would purposely lose in order to avoid the opponent's wrath. Draws are independent, thus one, both or no destruction choice may be implemented per pair. Subjects learn about the outcomes of the random draws and the reduction of their endowment at the end of the experiment. They are then privately paid and leave the lab. Subjects are given all details about the exact procedure before the start of the experiment.

4 Results

The experiment took place in October and December 2014 as well as in July and August 2015 in the economics lab of a German university. The experiment was programmed with Python 2.7, subjects were invited via ORSEE (Greiner 2004) and received a show-up fee of €4.00.

The "Outgroup Stranger Treatment" in Jauernig et al. (unpublished), in which subjects could burn each other's endowment after competition, served as a baseline for our study. No self-binding device existed in this baseline. Seventy subjects participated in the treatment of which 28 (40.0 %) harmed their opponent. On average, 20.11 ECU (sd = 34.37 ECU) of the 100 ECU endowment were burned.

In the self-binding treatment, a total of 222 subjects participated. Of these 222 subjects, 148 bound themselves (66.7 %). Seventy-four subjects (33.3 %) thus kept the option to burn their opponent's money. Ultimately, 49 of all 222 subjects (22.1 %) burned money. Due to the high percentage of subjects who bound themselves, the overall punishment level is reduced to 8.16 ECU (sd = 22.99 ECU). This is significantly less than the punishment in a setting where no self-binding device exists ($p = 0.001$). It is, however, still greater than zero ($p = 0.000$).

We now turn to analyze the punishment that the 74 unbound subjects inflict on their respective opponents depending on whether the opponent happens to be bound or unbound.[1] Unbound subjects had 48 encounters with bound subjects (64.9 % of encounters). Twenty-nine of the unbound subjects who encountered a bound subject punished their opponent (60.4 %). The punishment level that they inflicted on the bound subjects is on average 17.00 ECU (sd = 28.86

[1] P-values reported are the results of two-sided Mann-Whitney U-tests.

ECU). This level is significantly different from zero ($p = 0.000$), which confirms Hypothesis 1.

Result 1: Spite. Unbound subjects punished bound subjects significantly.

Spite thus plays an important role for subjects' punishment behavior after competition. This is the case since fear of the opponent is precluded if the opponent chose to use a self-binding device. It is noteworthy that the level of punishment which unbound subjects inflict on bound subjects is not significantly lower than the overall punishment of 20.11 ECU in the baseline treatment in which all subjects were unbound ($p = 0.227$).

In comparison, the 74 unbound subjects had 26 encounters with other unbound subjects (35.1 % of encounters). Twenty of the 26 subjects who encountered other unbound subjects punished their opponent (76.9%). The punishment level that they inflicted on the other unbound subjects is 38.27 ECU (sd = 40.22 ECU). This level is significantly higher than the punishment inflicted on bound subjects ($p = 0.017$), which confirms Hypothesis 2.

Result 2: Preemptive Retaliation. Unbound subjects punished other unbound subjects more severely than they punished bound subjects.

Result 2 indicates the importance of the role that preemptive retaliation plays for punishment behavior. This leads to the question of whether the existence of self-binding signals escalates conflict between two unbound parties as compared to a situation where two unbound subjects encounter in a setting where no self-binding signals exist.

Indeed, the level of punishment between unbound subjects in the setting with signals is significantly higher than the respective punishment level of 20.11 ECU in the setting without signals ($p = 0.003$).

Result 3: Escalation Through Unused Signals. When two intentionally unbound subjects encountered each other, punishment was higher than in a situation in which no self-binding device existed.

While a used self-binding device does not substantially propitiate the unbound subject, an unused signal leads to an escalation of conflict.

5 Discussion

Our results show that aggression after competition is driven by two different motives: preemptive retaliation and spite. We find that subjects inflict substantial harm to others, even if they are certain that their counterpart constitutes no threat. Things get worse when the fear of retaliation is also present. In this case aggression increases dramatically.

In contrast to the findings of Abbink and de Haan (2014) as well as Simunovic et al. (2013), where almost no spiteful decisions could be identified, we did find spite to be a motivator for punishment in a competition-induced setting. Kessler et al. (unpublished) did find money burning if the fear of the counterpart one is ruled out qua design, but they used a veiling mechanism as in Abbink and Sadrieh (2009) and the percentage of subjects engaging in burning (15 %) was considerably lower than in our study. For this reason, further research is needed to better understand the mediating effects of competition in a setting where aggression against a former competitor is possible.

Against the background of these two motives, we explored the effect of credible self-binding signals on the level of aggression. We investigated how this device to voluntarily self-restrict one's option to aggress can influence the inclination to burn other's money. If spite is the sole motive self-binding is completely ineffective. Although we identify a significant level of spiteful behavior, self-binding reduces punishment on the aggregate level substantially. This is due to the fact that in some cases it can effectively rule out the motive of preemptive retaliation. But there is a downside to the introduction of such a device. Once the self-binding option is mutually dismissed and this is common knowledge, punishment can escalate to a higher level as in a setting where such a device never existed.

This indicates that self-binding devices need to be handled with care. In our conservative setting, parties could not gain monetarily from striking against an opponent. This was done to exclude any other motives than preemptive retaliation and spite. Here, the positive overall effects outweighed the negative ones due to the fact that the self-binding rate of two thirds was relatively high. But we have good reasons to assume that those high rates may erode under different circumstances, be it that self-binding becomes costly or that the strike against the other becomes profitable. This may often be the case outside the lab whenever attacks involve monetary gains or when deterrence plays a role. Further research should explore these aspects.

If the percentage of intentionally and knowingly unbound subjects crosses a certain threshold, the net effects of self-binding devices may

shift. Self-binding signals reduce uncertainty about the other's intentions. If selfish motives for aggression are also present, fear of the opponent may become a predominant motive. Under these circumstances, however, uncertainty might be a blessing.

Appendix: Instructions (translated from German)

[The instructions for the baseline and the self-binding treatment are identical, except for one paragraph which is added in the self-binding treatment. This paragraph is italized in this reprint.]

Welcome to this experiment and thank you very much for your participation!

At the beginning of the experiment every participant is assigned a letter from A to X. Every letter is assigned only once, participants keep their letters until the ending of the experiment. Letters are allocated randomly and remain anonymous.

This experiment consists of two parts. The first part consists of 19 calculation tasks, which you have to play out against another participant from this room. This participant is randomly assigned to you. In each round you are presented with a matrix. Every matrix consists of 16 cells. In each cell, there is a one-digit number with one decimal place. Per matrix there are exactly two pairs of numbers which add up to 10. All participants receive the same matrices in identical order. In each matrix presented, you have to click on ONE of the two pairs which add up to 10.

The participant, who is assigned to you, gets to see the same matrix simultaneously. He, too, has to find one pair as fast as possible. As soon as the faster one has marked two cells, the slower one can not click on the cells any longer and the round is over. If the chosen pair of numbers is right, the faster one scores. If the faster one chooses a wrong combination, the slower one scores. The scores of you and the participant assigned to you, remain displayed on top of the screen and constantly adopt to the state of play.

If one round is finished, after three seconds you are presented with the next matrix until all 19 rounds are completed. In case you or the other participant do not click on any cells at all, there is a time-out

after 45 seconds and the next matrix is presented. In case of a draw after 19 rounds, a drawing of lots with equal probabilities decides who the winner and who the loser is. Both scores are displayed on top of the screen. The first part is then completed.

In the second part of the experiment, each participant is financially endowed with 100 ECU (Experimental Currency Units). This endowment can be altered during the experiment. At the ending of the experiment your gainings are converted into Euros and disbursed to you in private. The exchange rate is as follows: 10 ECU = 70 Eurocents.

For the second part of the experiment each participant remains assigned to the participant, he played against in the first part.

Both participants simultaneously get the opportunity to reduce their counterpart's endowment of 100 ECU by any integer amount down to a minimum of 0 ECU. This reduction does not influence their own endowment. Likewise each participant can leave the counterpart's endowment unchanged. The possible reductions of the endowment by the assigned participant come into effect with different probabilities for winners and loser from the first part of the experiment:

As you can see, there are 4 black and 6 white balls in the winner's urn. In the loser's urn the condition is reversed: Herein contained are 6 black and only 4 white balls.

In case a black ball is drawn from a participant's urn, he gets hit by the decision of his counterpart. His endowment is then reduced or left untouched according to the decision of the counterpart. If, however, a white ball is drawn, the decision of the counterpart is ignored.

Consequently the participant's endowment is left untouched irrespective of the decision of the counterpart.

Winners bear a risk of 40 % to be hit by the counterpart's decision. Losers bear a risk of 60 % to be hit by the counterpart's decision. The drawings from both urns take place simultaneously. Hence both, one or no decision can come into effect.

Before the screen with the option to reduce the endowment of the respective counterpart is shown, you and your counterpart may unilaterally exclude this possibility. You and your counterpart make this choice simultaneously and both choices are communicated to you and your counterpart. Depending on the respective choices, you and your counterpart will then receive the option to reduce the endowment of the other or not.

Subsequently both participants receive their pay-off in Euro according to the decision of their counterpart and the result of the drawing from the urn. You will receive the individual pieces of information that are relevant to your decisions again during the experiment.

If you completed reading and understood the instructions, please confirm by clicking on the accordant button on your screen. The instructions are then read out aloud once more. If you have any questions, please stand at your place and raise your hand.

Acknowledgments

We are grateful to the Technical University of Munich for funding.

References

Abbink, K., & Sadrieh, A., 2009. The pleasure of being nasty. Economics Letters 105 (3), 306–308.

Abbink, K., & de Haan, T., 2014. Trust on the brink of Armageddon: The first-strike game. European Economic Review 67, 190–196.

Adachi, P. J. C., & Willoughby, T., 2011: The effect of video game competition and violence on aggressive behavior: Which characteristic has the greatest influence? Psychology of Violence 1 (4), 259–274.

Cushing, B. S., 1995. When a Hawk can Damage a Dove: An Extension of Game Theory. Journal of Theoretical Biology 175, 173–176.

Greiner, B., 2004. An online recruitment system for economic experiments. In: Kremer, K.; Macho, V. (Eds): Forschung und wissenschaftliches Rechnen, Ges. für wiss. Datenverabeitung, Göttingen, 79–93.

Harbring, C., Irlenbusch, B., 2008. How many winners are good to have? Journal of Economic Behavior & Organization 65 (3-4), 682–702.

Harbring, C.; Irlenbusch, B., 2011. Sabotage in Tournaments: Evidence from a Laboratory Experiment. Management Science 57 (4), 611–627.

Jauernig, J., Uhl, M., Luetge, C., unpublished results. Competition-Induced Punishment: Who Is the Target? Working Paper 2191364, MediaTUM, `https://mediatum.ub.tum.de/node?id=1221336`.

Kessler, E., Ruiz-Martos, M., Skuse, D., unpublished results. Destructor Game. Working Paper 2012-11, Economics Department, Universitat Jaume I, 1–9.

Muller, D., Bushman, B. J., Subra, B., & Ceaux, E., 2012. Are People More Aggressive When They Are Worse Off or Better Off Than Others? Social Psychological and Personality Science 3 (6), 754–759.

Schelling, T. C., 1980. The Strategy of Conflict. Cambridge, MA: Harvard University Press.

Simunovic, D., Mifune, N., Yamagishi, T. 2013. Preemptive strike: An experimental study of fear-based aggression. Journal of Experimental Psychology 49, 1120–1123.

Vohs, K. D., Baumeister, R. F., 2007. Feeling Duped: Emotional, Motivational, and Cognitive Aspects of Being Exploited by Others. Review of General Psychology 11 (2), 127–141.

Yamagishi, T., Sato, K., 1986. Motivational bases of the public goods problem. Journal of Personality and Social Psychology 50 (1), 67–73.