



TOURO UNIVERSITY
JACOB D. FUCHSBERG LAW CENTER
Where Knowledge and Values Meet

Touro Law Review

Volume 37 | Number 1

Article 12

2021

The Robber Wants to Be Punished

Uri Weiss

Follow this and additional works at: <https://digitalcommons.tourolaw.edu/lawreview>



Part of the [Criminal Law Commons](#), and the [Criminology and Criminal Justice Commons](#)

Recommended Citation

Weiss, Uri (2021) "The Robber Wants to Be Punished," *Touro Law Review*. Vol. 37: No. 1, Article 12.
Available at: <https://digitalcommons.tourolaw.edu/lawreview/vol37/iss1/12>

This Article is brought to you for free and open access by Digital Commons @ Touro Law Center. It has been accepted for inclusion in Touro Law Review by an authorized editor of Digital Commons @ Touro Law Center. For more information, please contact lross@tourolaw.edu.

THE ROBBER WANTS TO BE PUNISHED

*Uri Weiss**

ABSTRACT:

It is a commonly held intuition that increasing punishment leads to less crime. Let us move our glance from the punishment for the crime itself to the punishment for the attempt to commit a crime, or to the punishment for the threat to carry it out. We argue that the greater the punishment for the attempted robbery, i.e., for the threat, “give me your money or else,” the greater the number of robberies and threats there will be. The punishment for the threat makes the withdrawal from it more expensive for the criminal, making the relative cost of committing the crime lower. In other words, the punishment of the threat may turn an incredible threat into a credible one. Therefore, the robber has a strong interest in a legal system that increases the punishment of the threat.

* Dr. Uri Weiss, a Polonsky Fellow at Polonsky Academy – Van Leer Jerusalem Institute, uriw11@gmail.com, +972 5605238. This paper was part of my Doctoral Dissertation: “Issues in Law and Game Theory,” (The Hebrew University of Jerusalem) that was written under the supervision of Professor Robert J. Aumann and Professor Ehud Guttel. My gratitude is extended to them for their help, which has improved this paper significantly. I also wish to thank Joseph Agassi, Kenneth Arrow, Omri Ben-Shahar, Mike Borns, Shai Dothan, Ezra Einy, Osnat Jacobi, Alon Harel, Gil Kalai, Eric Maskin, Ariel Porat, Muli Safra, Alexander Stremitzer, Doron Teichman, Omri Yadlin, and Eyal Winter for our conversations about this article. In addition, I wish to thank the participants at the 2008 Israeli Law and Economics Conference in Tel Aviv University, 2009 Stony Brook Game Theory Festival, Bonn University Law and Economics Workshop, the Hebrew University Law and Economics Workshop, the Hebrew University Criminal Law Workshop, and the Center for the Study of Rationality annual retreat (2010) for their remarks.

THE ROBBER WANTS TO BE PUNISHED

I. INTRODUCTION

One of the main justifications for punishing people is deterrence.¹ According to deterrence theory, it is just to punish since more punishment leads to less crime.² The claim that punishment leads to fewer crimes reflects a commonly held intuition. This intuition assumes that human beings are rational and do not wish to be punished. However, we will show that, surprisingly, when people are rational and seek to avoid punishment, punishing them for threatening to commit a crime leads to more threats and more crimes, such as extortions, black-mails, incitement, and robberies. At the civilian level, “punishing,” i.e., imposing liability, e.g., for threatening to breach a contract or to bring frivolous lawsuits will lead to more such threats—and indeed to more breaches of contracts, and frivolous lawsuits. We also argue, that in international relations, sanctioning a nation for threatening to attack another nation will lead to more threats and more attacks. In other words, in the case of threats, there is incentive reversal³: increasing the punishment for illegal threats leads to more threats. The reason for this is that the punishment of the threat (which can be considered the same as the punishment of the attempt) lends credence to the threat. In other words, we argue that the punishment of the threat strengthens its validity.

Deterrence theory is used to justify the punishment of the threat. We will challenge this justification specifically, and we will also challenge the general claim that rational people will *always* commit less crime, and surely not more so, when the punishment is more severe. In short, we will challenge the generality of deterrence theory to justify punishments. To see this, we will distinguish between: (1) punishment for threats and punishment for crimes; (2) punishment for unrealized threats that were not heeded by the victim and punishment for unrealized threats that were heeded; (3) punishment for unrealized

¹ See CESARE BECCARIA, ON CRIMES AND PUNISHMENTS (Graeme R. Newman & Pietro Marongiu trans., Transaction Publishers 5th ed. 2016).

² See JEREMY BENTHAM, PRINCIPLES OF PENAL LAW 399 (1843); See Gary S. Becker, *The Economics of Crime*, 12 CROSS SECTIONS 8 (1995).

³ See Eyal Winter, *Incentive Reversal*, 1 AM. ECON. J.: MICROECONOMICS 133 (2009).

threats that were not heeded and punishment for realized threats; and (4) punishment for crimes and punishment for attempted crimes.

A. Review of Threat Games

Thomas Schelling contributed to our understanding of threats by developing the concept of commitment.⁴ In a bargaining situation it may be beneficial for a person to limit his set of actions.⁵ The qualities of being more intelligent, more skilled in debate, having more financial resources, more physical strength, more military might, or more ability to withstand losses are by no means universal advantages in bargaining situations; they often have a negative value.⁶ Furthermore, the threatener does have an incentive to commit to carrying out the threat, if he thinks that the threat will be successful, because then he will achieve his goal without paying the price for carrying out the threat.⁷

An additional contribution to the analysis of threat games is the model of the empire and the provoking state.⁸ In order to prevent the

⁴ Thomas C. Schelling, *An Essay on Bargaining*, 46 AM. ECON. REV. 281 (1956).

⁵ *Id.* at 283. See for example: "if the buyer can accept an irrevocable commitment, in a way that is unambiguously visible to the seller, he can squeeze the range of indeterminacy down to the point most favorable to him."

⁶ *Id.* at 282. See particularly his words: "Bargaining power," "bargaining strength," "bargaining skill" suggest that the advantage goes to the powerful, the strong, or the skillful. It does, of course, if those qualities are defined to mean only that negotiations are won by those who win. But if the terms imply that it is an advantage to be more intelligent or more skilled in debate, or to have more financial resources, more physical strength, more military potency, or more ability to withstand losses, then the term does a disservice. These qualities are by no means universal advantages in bargaining situations; they often have a contrary value."

⁷ *Id.* at 293. See *Particularly his worlds*

But more than communication is involved when one threatens an act that he would have no incentive to perform but that is designed to deter through its promise of mutual harm. To threaten massive retaliation against small encroachments is of this nature, as is the threat to bump a car that does not yield the right of way or to call a costly strike if the wage rate is not raised a few cents. The distinctive feature of this threat is that the threatener has no incentive to carry it out either before the event or after. He does have an incentive to bind himself to fulfill the threat, if he thinks the threat may be successful, because the threat and not its fulfillment gains the end; and fulfillment is not required if the threat succeeds. The more certain the contingent fulfillment is, the less likely is actual fulfillment.

⁸ See Stephen L. Quackenbush & Frank C. Zagare, *Game theory: Modeling Interstate Conflict*, in MAKING SENSE OF INTERNATIONAL RELATIONS THEORY, 98 (Jennifer Sterling-Folker ed., 2006).

provoking state from invading its neighbor, it may be efficient for the empire to “burn its bridges” by excluding the possibility of not responding.⁹ For example, Yugoslavia needs to decide whether to invade Bosnia. The best result for Yugoslavia is to invade and not to be punished by the North Atlantic Treaty Organization (NATO), and the worst result for Yugoslavia is to invade and to be punished by NATO. The best result for NATO is that Yugoslavia will not invade, and the worst result for NATO is that Yugoslavia will invade, and they will punish Yugoslavia. Thus, Yugoslavia will invade, if and only if, Yugoslavia believes NATO will not respond. If Yugoslavia invades, NATO can respond by punishing Yugoslavia, and this will be the worst result for both of them. NATO can also refrain from responding, and this will be the second best result for NATO. So, if Yugoslavia invades, NATO will not respond, and the result will be that Yugoslavia will invade and NATO will not respond. However, NATO can also limit its set of actions. NATO can burn its bridges. For example, NATO can decide in advance, and irreversibly, that if Yugoslavia invades, the chief commander of NATO must attack; then, the result will be that Yugoslavia will not invade, and NATO will not respond. Another mechanism of NATO may be to transfer part of its forces to a neighboring state and to bear the first 50 from the cost of 100. This time, if Yugoslavia invades Bosnia, NATO will respond since the remaining cost is only 50. As a result, Yugoslavia will not invade. In this case the “paying” of the first 50 was the commitment of NATO to respond.

An important paper that studies threat games in litigation is Bebachuk's analysis of the phenomena of “Negative Expected Value Suits.”¹⁰ Bebachuk tried to solve the puzzle why Negative Expected Value Suits, i.e., suits in which the expected litigation cost of the plaintiff is larger than the expected judgment, are submitted.¹¹ The explanation is, of course, that the plaintiff is interested in settlement.¹² However, why does he believe that the defendant will reach a settlement with him? If the plaintiff's litigation cost is bigger than his gain from the trial, does he have a credible threat to continue with the suit until judgment? Bebachuk's answer is that the capacity to divide the

⁹ *Id.*

¹⁰ See Lucian Arye Bebchuk, *A New Theory Concerning the Credibility and Success of Threats to Sue*, 25 J. LEGAL STUD. 1, 1 (1996).

¹¹ *Id.* at 2.

¹² *Id.* at 2.

litigation cost explains the credibility of the threat to continue with Negative Expected Value Suits.¹³ If, for example, the expected judgment is 75, and the litigation cost is 100, then it will not be worthwhile to sue, and thus, the threat to “settle or be sued” is not credible. However, if it is possible to divide the litigation cost of the plaintiff, such that in the first stage he will pay 50, and in the second stage he will pay 50 (and we will assume that the defendant cannot divide his cost and that plaintiff and defendant have symmetrical bargaining power), then the plaintiff’s threat to sue becomes credible. The reason for this is that after he pays the first 50, it will be rational for the plaintiff to sue because he needs to pay only 50 for litigation costs, in order to gain 75—the judgment. Hence, it is worthwhile to sue. Therefore, even before spending a cent on the suit, the plaintiff’s threat is credible.

Bar-Gill and Ben-Shahar also study threat games.¹⁴ They discuss the situation of a person who threatens to breach a contract if the other side does not agree to change its conditions, and the other side capitulates to the threat.¹⁵ They discuss when the courts should enforce those modified contracts, and they conclude that sometimes the two sides will be better off if the law enforces such modifications, and then the court should enforce them.¹⁶ If the law does not enforce the modified contract, it may lead to a breach, a result that, for the promisee, would be worse than reaching a new contract. If the promisee accepts the new contract in a legal system that enforces modified contracts, this signals that he will gain from the new contract, unless there is an information problem that misleads him to believe in an incredible threat to breach. Hence, their recommendation is that the law will recognize the new contract if and only if the threat is credible.¹⁷ By contrast, in cases where the threat to breach is incredible, it is better for the threatened side if the law does not recognize the validity of the new

¹³ *Id.* at 4.

¹⁴ Oren Bar-Gill & Omri Ben-Shahar, *The Law of Duress and the Economics of Credible Threats*, 33 J. LEGAL STUD. 391 (2004).

¹⁵ *Id.* at 392. *In their words*: “The negotiation of a transaction often involves threats by one party to refrain from dealing unless a particular provision, favorable to the threatening party, is accepted.”

¹⁶ *Id.* at 391. *They Claimed*: “This paper argues that enforcement of an agreement, reached under a threat to refrain from dealing, should be conditioned solely on the threat’s credibility.”

¹⁷ Only in situations of credible threats will it be better for both sides to reach a new contract, since only in such case will the threatener breach the contract if he has no legal capacity to modify the contract.

contract and continues to recognize the validity of the old contract, since this ruling will prevent his coercion.

B. Road Map

The rest of the paper is organized as follows: Section 2 presents an example of robbery; Section 3 presents the general model; Section 4 provides the intuition for the conclusion that punishing for the threat makes the threat more credible; Section 5 proposes some applications; and Section 6 surveys the literature. In the appendix we modify the assumptions of the model and test its robustness.

II. EXAMPLE

Imagine two different legal systems: one lenient and the other severe. In the lenient system the expected punishment for attempted robbery (or the threat of it) is 1 year in prison, while in the severe system it is 7 years in prison. In both systems, the expected punishment for robbery is 8 years in prison, and for murder it is 15 years in prison. Thus, the robber needs to decide whether or not to attempt a robbery, which for him is “worth” 10 years in prison. In other words, he is indifferent between doing nothing and participating in a lottery with a 50 percent probability it yields the expected benefit from the robbery and with a 50 percent probability it yields the expected 10 years in prison. The sequence of events is as follows:

First, the robber needs to decide whether to attempt to rob—i.e., to threaten—or not.

If he chooses not to make the attempt, then that is the end of the game and the outcome is “no threat.” If he chooses to make the attempt, the victim decides whether to give in or not.

If the victim gives in, then that is the end of the game and the outcome is a “successful robbery.” If the victim does not give in, then the robber needs to decide whether to kill the victim and take the money.

If the robber decides to carry out his threat, then that’s the end of the game and the outcome is “murder and robbery.” If the robber decides not to carry out his threat, then that’s the end of the game and the outcome is “withdrawal from the threat.”

Hence, the decision tree in the lenient system is:

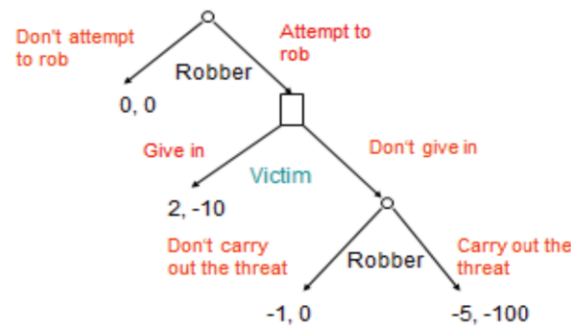


Figure 1. The decision tree in the lenient system.

We can see that in the lenient system, if the victim does not give in, the robber will withdraw (as -1 is greater than -5). As a result, the victim will not give in when threatened and the robber will not make the threat (as 0 is greater than -1).¹⁸ We can express this analysis by crossing out the alternatives not taken, as follows:

¹⁸ To solve the game, we use backward induction. In other words, we solve the game from the end to the beginning. In the last stage, if it is reached, the robber will need to decide whether to carry out the threat or to withdraw. In the lenient system, if the robber withdraws, he expects to be sentenced to one year in prison. On the other hand, if he murders and takes the asset, he expects to be sentenced to 15 years in prison, but the expected gain from the asset is worth 10 years in prison to him, so the expected payoff will be -5; therefore, the robber will prefer to withdraw. His payoff from withdrawing will be -1, while the payoff from the alternative of carrying out the threat will be -5. What will do the victim in the previous stage, if reached? Will the victim choose to give in or not? If the victim gives in, he will lose the asset and the payoff will be -10. On the other hand, if he does not give in, the robber will withdraw, and the payoff for the victim will be 0. Therefore, the victim will choose not to give in. What will do the robber in the first stage? Will he rob or not? As we have shown, if the robber chooses to attempt to rob, the victim will not give in, the robber will withdraw, and the payoff from attempting to rob will be -1, which is the expected punishment from attempting to rob. On the other hand, if he chooses not

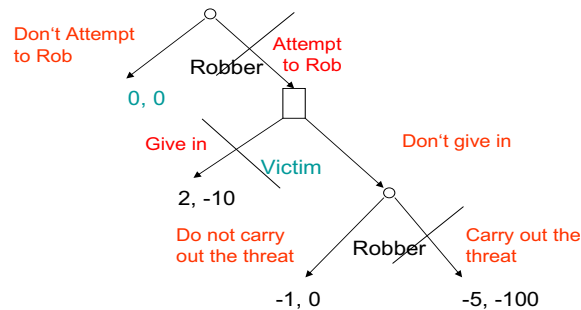


Figure 2. The solution of the decision tree in the lenient system.

However, in the severe system, the decision tree is:

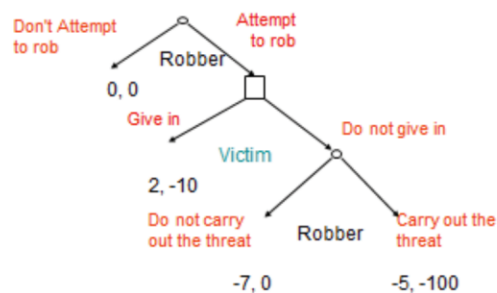


Figure 3. The decision tree in the severe system.

In the severe system, if the victim does not give in, the robber will carry out the threat (as -5 is greater than -7). Therefore, the victim

to attempt to rob, his payoff will be 0. Therefore, in the lenient system, the robber will choose not to attempt to rob.

will give in, and the robber will make the threat (as 2 is greater than 0).¹⁹ We can express this analysis as follows:

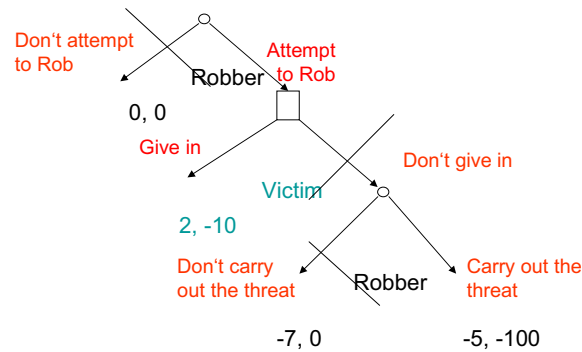


Figure 4. The solution of decision tree in the severe system.

The conclusion is that in our example reducing the punishment for attempted robbery will prevent the robbery and even the threat.

¹⁹ We also solve this game by backward induction. We ask what the robber will do in the last stage, if it is reached. Will he choose to carry out the threat or not? Also, this time his payoff from carrying out the threat is -5 (15 years in prison plus the asset that for him is worth 10 years), but this time his payoff from not carrying out the threat is -7, which is his punishment for attempting to rob in the severe legal system. Therefore, this time the robber will choose to carry out the threat, i.e., to murder and take the asset violently. What, then, would the victim do in the previous stage? The victim knows that if he does not give in, the robber will carry out the threat, and therefore the victim will choose to give in. He prefers losing the asset to losing his life. Therefore, if the second stage is reached, the result will be a successful robbery. It remains for us to ask what the robber would choose to do in the first stage. In the first stage, the robber knows that if he attempts to rob, the result will be a successful robbery. His payoff from a successful robbery is 2. He expects to be sentenced to 10 years in prison, but at the same time he will gain an asset that for him is worth 10 years. On the other hand, his payoff from not attempting to rob is 0. Therefore, in the severe legal system, the robber will choose not to rob.

III. THE MODEL

m – The money

q_r – The probability that the robber will be caught in case of a robbery (and then punished and the money taken from him)

q_t – The probability that the robber will be caught in case of an attempt (and then punished for the attempt)

q_m – The probability that the robber will be caught in case of a murder (and then punished for murder and the money taken from him)

p_r – The punishment for successful robbery

p_t – The punishment for the threat (the attempt to rob)

p_m – The punishment for murder

$P_r = q_r p_r$ – The expected punishment for robbery

$P_t = q_t p_t$ – The expected punishment for the threat to commit robbery

$P_m = q_m p_m$ – The expected punishment for murder

$M = (1 - q_r)m$ – The expected benefit from a successful robbery

$M' = (1 - q_m)m$ – The expected benefit from robbery and murder

We set -100 to be the worth for the victim of being murdered. The sequence of events is as described above. Hence, the decision tree of the model is:

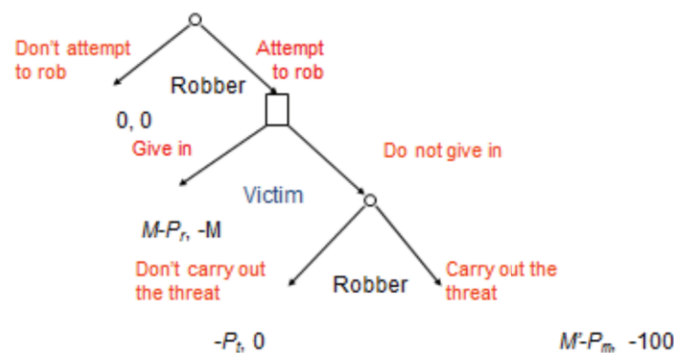


Figure 5. The decision tree in the model.

When will the robber attempt the robbery? He will do so if, and only if, the following conditions are met:

1. The expected net cost of carrying out the threat (i.e., expected cost minus expected benefit) is less than the expected cost of withdrawing, and
2. The expected punishment for a successful robbery is less than the expected benefit from it.

In other words, the robber will rob if and only if:

1. $P_m - M' < P_i$; and
2. $M > P_r$

The conclusion is that in the described game, the more severe the expected punishment is for attempted robbery (i.e., for uttering the threat “give me your money or else”), the more robberies and threats there will be.²⁰

²⁰ Sometimes there is a higher probability of catching the threatener if he carries out his threat than if he does not. For example, in the case of robbery, the body of the victim will stimulate the police to investigate, which increases the probability of the robber being caught. However, sometimes there is a higher probability of catching the threatener if he does not carry out his threat than if he does. One example of this is sexual harassment of an employee by her employer. In this case, the legal system enables the employer to effectively deny that he fired the worker for refusing to give in to his sexual demands.

Another conclusion from the model is that contrary to the canonical thesis of Gary Becker that in the case of a risk-neutral criminal, multiplying the probability of being caught by X is *not* equivalent to multiplying the punishment by X . When $X > 1$, multiplying the probability of being caught by X deters *more* than does multiplying the punishment by X , since when the criminal is caught, he needs to give the asset back. This also challenges the recommendation of Becker to reduce the probability of being caught and compensating for it by increasing the punishment. Becker’s recommendation is intended to reduce the cost of the police. Becker dismisses the claim that capture deters, since it also takes the robbed asset from the criminal.

IV. DISCUSSION²¹

A. Intuitive Explanation of the Model

The above example and model show that increasing the punishment for threatening—i.e., attempting—to rob leads to a higher incidence of threats and robberies. We now provide an intuitive explanation of this. We argue that increasing the punishment for threatening to rob leads to more threats and robberies since it makes incredible threats credible ones. That is, it increases the cost of withdrawing from the threat, and hence makes carrying out the threat more attractive relative to the alternative of withdrawing from it; this in turn makes the *relative* cost of carrying out the threat less expensive. We argue that the punishment for carrying out the threat becomes a “sunk cost” after the threat is made. The robber will bear the expected punishment of the threat regardless of whether he carries out the threat or not. Therefore, the expected punishment borne by the robber for carrying out the threat becomes less expensive. In a regime that does not punish for threatening to carry out a crime, the robber who must decide whether to carry out the threat by taking the asset violently expects the full punishment for murder. In contrast, in a regime that does punish for threatening to carry out a crime, the robber expects a less severe punishment. The punishment he expects is the punishment for murder minus the punishment for the threat (the attempt). The robber will bear the punishment for the attempt regardless of his decision whether to murder. The risk that will be added for the murder is the expected punishment for murder minus the expected punishment for the threat. Hence, when we increase the punishment for the threat, there will be more cases in which the robber will respond to the refusal of the victim by carrying out the threat. Thus, the threat will be credible in more cases. Therefore, increasing the punishment for threats results in more crimes, like threatening, robbery, extortion, blackmail, etc.

B. How Can It be Beneficial for a Person to be Punished?

²¹ “The first (and only one, to the best of our knowledge) to point to what we call the credibility effect is Uri Weiss, *The Robber Wants to be Punished* (Hebrew Univ. of Jerusalem, Federman Ctr. for the Study of Rationality, Discussion Paper No. 685, 2015).” Saul Levmore & Ariel Porat, *Threats and Criminal Deterrence in Several Dimensions*, 2017 U. ILL. L. REV., 1333, 1351 n.45 (2017). In a former version of their paper, they described this work by these words: “demonstrating that law’s punishing the attempted robbery can help the robber by making his threat more credible.”

In interactive situations like negotiation, it may be worthwhile for a person to limit his set of actions or to make some of them more expensive to carry out.²² This is the case in the game of Chicken. The law may be used as a mechanism to limit the set of actions of a player. If we modify the chicken game, such that a particular player is subject to the rule that if he moves, he will be executed, then he will “win” the game. In other words, he will benefit from this limitation of his power. Similarly, in the Battle of the Sexes game, in which each player chooses between two places and the worst result for any player is that they choose different places but each player prefer another place, a player will benefit if they are banned from going to the place preferred by the other, because it incentivizes the other to choose the place the player prefers (and not the one the other prefers). We argue that this also holds true in situations of threats, such as robbery, extortion, and blackmail. These are situations of negotiation in the shadow of the law. Capitulating to the threat is *equivalent to a settlement*. The robber/threatener is equivalent to the plaintiff. Sometimes the robber/threatener is equivalent to a plaintiff whose lawsuit has negative expected value. If the law punishes the plaintiff (say by imposing a court fee) for cancelling the suit, then his threat to continue with the suit will become more credible, which will promote his interest. Equivalently, the interest of the robber/threatener is promoted when the law punishes him for withdrawing from the threat. The punishment for withdrawing from the threat actually makes the threat credible.

V. APPLICATIONS

A. General Conclusion regarding the Need to Adapt Punishment Theory to Interactive Situations

We argue that when it comes to interactive situations, punishment theories need to be specially adapted. A person in an interactive situation does not respond to incentives in the same way that Robinson Crusoe would.²³ When an isolated decision maker, “Robinson Crusoe,” is threatened with a penalty if he performs a specific action, he

²² See THOMAS C. SCHELLING, *THE STRATEGY OF CONFLICT* (Harvard Univ. Press rev. ed., 1980) (1960).

²³ This should be the moral from the following book: THOMAS C. SCHELLING, *THE STRATEGY OF CONFLICT* (1960).

will refrain from doing so. In standard punishment theories, like that of Gary Becker,²⁴ it is assumed that the potential criminal will behave like Robinson Crusoe. Hence, these theories reach false conclusions regarding crimes committed in an interactive context. In an interactive context it may be beneficial for a person to be limited from doing X; therefore, it may be beneficial for him to be penalized for doing X. Therefore, increasing the punishment for doing X may lead to X being done more often.

In this paper we have focused on penalties for robbery, but the conclusion is of much wider application. For example, the analysis of robbery is also valid in the case of hijacking an airplane for a bargaining goal. In this case, it is better for the state whose airplane was hijacked to concede to the terrorist that if he withdraws from the hijacking, i.e., releases the hostages, he will be exempted from a full punishment. Furthermore, even rational bankers may want to be exposed to the threat to be punished. If the law exempts women from criminal liability for Bank fraud, women will not get Bank jobs.

B. A General Conclusion Regarding the Rationality of Masochism

The proposed reasoning may also help us to understand social phenomena that are difficult to explain, especially within the framework of rational choice. The proposed reasoning of this paper may provide us with the key to understanding masochistic behavior, which is perceived to be irrational and even to challenge the validity of rational choice theory. Our explanation is that the reputation for self-punishment is used as a mechanism for self-commitment, which may be beneficial to the masochist in interactive situations. For example, if the threatening person develops a reputation for harming himself when his threats are not heeded, then his threats will be more credible. This proposition is valid not only for robberies, but also for legitimate negotiations: a person will improve his bargaining position by developing a reputation for self-punishment when he gets less than x of the surplus. Self-punishment may function as a mechanism that helps the

²⁴ Gary S. Becker, *Crime and Punishment: An Economic Approach*, 76 J. POL. ECON. 169 (1968).

individual, in much the same way as insult does.²⁵ Furthermore, a reputation for self-punishment may help not only the person making the threat, but also the person facing the threat. If a person develops a reputation for punishing himself when submitting to a threat (where the cost of submitting to the threat—including the threat of self-punishment—is greater than the benefit from it), he will be threatened less often. The insight that in some situations the masochist stands to benefit more than the non-masochist can help us to understand why masochism has survived natural selection.²⁶

C. Sunk-Cost Bias as a Commitment

Our thesis may rationalize the bias of sunk cost. We conjecture that sunk-cost bias is a rational mechanism (or rational *rule*) that provides the biased player with a commitment to fight. A player with sunk-cost bias sees his exit from a conflict, such as war or litigation, as more expensive, relative to a player without sunk-cost bias. This is because he needs to bear his regret. The regret (or loss aversion) is the self-punishment that may make his threat credible.

Let us proceed with an example. Anne sues Bob. She invests 50 in litigation. If she continues with the suit, she will need to invest an additional 100, but she will gain 60 in court. So, the “rational Anne” has no credible threat to continue with the litigation. However, the “irrational Anne,” the biased one, may see the result of investing 50 and withdrawing her claim as inferior to investing 150 and getting 60 from the defendant—or at least she may act as if this is her preference. In this case she gains a commitment to go on with the lawsuit, which leads Bob to settle with her even before she invests the first 50. Her irrationality will give her a credible threat, which rationalizes the “irrational” emotion.

²⁵ Robert J. Aumann, *Rule-Rationality versus Act-Rationality*, 12 (Federmann Ctr. for Study of Rationality Discussion Paper, Paper No. 497, 2008), <http://www.ratio.huji.ac.il/sites/default/files/publications/dp497.pdf>.

²⁶ THOMAS C. SCHELLING, *STRATEGY OF CONFLICT* 17 (1960) (“I am told that inmates of mental hospitals often seem to cultivate, deliberately or instinctively, value systems that make them less susceptible to disciplinary threats and more capable of exercising coercion themselves. A careless or even self-destructive attitude toward injury—‘I’ll cut a vein in my arm if you don’t let me . . .’—can be a genuine strategic advantage.”).

D. The Regressive Effect of Severe Punishment for Threats

A severe punishment for threats also has the effect of shifting threats away from rich people and onto poor people. To see why, consider a robber who is wondering whether to rob a poor person or a rich one. If the cost of the robbery is the same (or if it is less expensive to rob a rich person), the robber will choose to rob the rich person. However, what happens when it is less expensive to rob the poor person, when, for example, the poor person lives closer to the robber. In this case, if the robber poses a credible threat to both the poor person and the rich person, he may choose to rob the poor person. If the asset is of little worth, the robber may lack a credible threat to take it by main force. However, if the punishment for the threat is severe enough, he will have a credible threat to rob even assets of little worth, i.e., to rob even poor people. Hence, increasing the punishment for threats shifts threats away from rich people and onto poor people.

We can show this using the main example of this paper: the expected punishment for robbery is 8 years in prison, and the expected punishment for murder is 15 years in prison. Therefore, in a regime in which the expected punishment for the threat is 7 years in prison, the robber has a credible threat if and only if the asset has a worth for him of more than 8 years in prison. However, if the punishment for the threat is 3 years in prison, then the robber has a credible threat if, and only if, the asset has a worth for him of more than 12 years in prison. Anne has an asset that for the robber is worth 11 years in prison, while Bob has an asset that has for the robber a worth of 13 years in prison. Additionally, in order to rob Bob, the robber needs to bear a transition cost that is equivalent to 2.5 years in prison. How will the robber behave? In the severe system, he will prefer to rob poor Anne rather than rich Bob.

The robber poses a credible threat to both of them but robbing Anne will give him a total benefit that for him is worth 3 years in prison, while robbing Bob will give him a total benefit that for him is worth 2.5 years in prison. By contrast, in the lenient legal system he will prefer to rob rich Bob rather than poor Anne, since he poses a credible threat to Bob, but not to Anne. The conclusion is that in moving from the severe legal system to the lenient one, we see that more

crimes are committed against the rich than against the poor. Punishing severely for threats thus has a regressive effect.²⁷

E. Lenient Punishment for Robbery of Expensive Assets

We will claim that if the asset that the robber attempted to rob is expensive, it is important to impose a lenient punishment. And the reason is this: when the benefit from a successful robbery (M) is greater than the expected punishment for robbery (P_r), then in order to prevent the threat, the expected punishment for the threat should be lower than the difference between the punishment for carrying out the threat and the benefit from carrying it out ($P_m - M$). If, for example, the expected punishment for robbery is 8 years in prison, and the expected punishment for murder is 15 years in prison, then if the robbed asset is worth 13 years in prison in order to prevent the robbery the expected punishment for the threat should be less than 2 years. However, if the robbed asset is worth 9 years in prison, then in order to prevent the robbery the expected punishment for the threat should be less than 6 years.

F. The Case of Renegotiation in Contract Law

In contracts one party may threaten the other: “I will breach the contract, unless you adjust the price.” However, whereas criminal law punishes for the threat itself, contract law does not impose damages for the threat themselves. The promisee cannot sue for unrealized threats to breach the contract that were not heeded by him (in those cases there is no damage). Furthermore, in the case of a realized threat to breach the contract, the remedy is the same (damages) as in the case of a

²⁷ Let us now provide a general model explaining why punishment for threats is regressive. We denote by c_i the cost of robbing a person. A robber must choose whether to rob J, the rich person, or K, the poor person, or neither of them. He will rob the rich person if and only if:

1. $M_j > P_r$ (a successful robbery of the rich person is worthwhile); and
2. $P_r > P_m - M'_j$ (the robber poses a credible threat to the rich person); and also
- 3a. $M_j - P_r - c_j > M_k - P_r - c_k$ (it is more worthwhile to rob the rich person than the poor person), which means $M_j - c_j > M_k - c_k$; or
- 3b. $P_r < P_m - M'_k$ (the robber poses no credible threat to the poor person).

We can see that when we increase the punishment for threat P_t beyond $P_m - M'_k$ in cases in which (1) $P_r > P_m - M'_k$, (2) $(1 - q_r)M_j > q_r P_r$, and (3) $M_j - c_j < M_k - c_k$, then we increase the likelihood that the rich person will be robbed.

breach that was not accompanied by any threat. The threat is of no consequence when determining the damages. We wish to point out that those rules are also efficient.

If the law imposes damages in a case of an unrealized threat to breach that has not heeded, then it effectively turns incredible threats into credible ones. Moreover, in a legal regime of expectation damages with no problem of under-compensation, then the threat to breach will be credible if, and only if, the cost of performance is greater than the value to the promisee from performing; this means if, and only if, the performance is inefficient.²⁸ However, if we impose a compensation also for the unrealized threat to breach that has not been heeded, then the threat to breach will be credible also regarding efficient performance.²⁹ For example, if the value for the lawsuit to the customer is 100, and the cost of its production is 90 and the consented price is 95; then there is no credible threat to breach a contract to produce it. However, if there is a sanction of 11 for the threat, then there will be a credible threat to breach, since in a case of breaching the tailor needs to pay damages of 5, so his payoff will be -5; whereas in the case of production, his payoff will be the price minus the cost of production minus the sanction, i.e., $100 - 95 - 11 = -6$.

Furthermore, if this is an efficient contract, but we have a moderate problem of under-compensation, such that the value to the promisee minus the price is bigger than the damages that will be awarded in a case of breach; and the damages are bigger than the cost of performance minus the price; then the threat to breach under the rule of just compensation is incredible. However, if there is a high enough sanction (punitive damages) for the threat to breach, the threat to breach will be credible in this case.³⁰ This may, *ex ante*, block efficient

²⁸ I would like to thank Osnat Jacobi for showing me this point.

²⁹ The threat to breach the contract is credible, if and only if in the case the promisee does not give in to the threat; the threatener gains more from breaching the contract than from respecting the contract. In other words, the threat to breach the contract is credible, if and only if the cost of respecting the contract (c) plus the sanction for the threat (s) minus the price (p) is greater than the value to the promisee from respecting the contract (v) minus the price (p); i.e., if and only if $c + s - p > v - p$; i.e., if and only if $c + s > v$. When $s=0$, the threat of breach is credible if and only if $c > v$, i.e., if and only if the contract is efficient (the contract is efficient when $c - p > v - p$, i.e., when $v > p$).

³⁰ Let d denote damages. We speak about a case in which $v - p > d > c - p$. There is a credible threat if and only if $c - p < s + d - p$, i.e., if $s > c - d$.

transactions, if the courts in this legal system do not fully compensate the promisee who agrees to adjust the price after the threat.³¹

However, what is the efficient law when one party threatens to breach the contract and carries out his threat? If the law imposes punitive damages in this case, then the incentive not to threaten will be strengthened. Such a rule cannot encourage such threats and, indeed, it can prevent them. However, the price is that threats will be prevented even when the threat is beneficial for both sides. When the damages are lower than the cost of carrying out the threat, and the cost of carrying it out is lower than the value of the promise, then it is in the best interest of both sides to modify the contract. Otherwise, the result will be an inefficient breach. However, if the law imposes punitive damages for the threat, the promisor will not threaten. This will

³¹ Let us now investigate the efficiency of this rule in situations of asymmetric information in which the threatened does not know if the threat is credible. If there is a regime of non-enforcement of modifications, then the uninformed threatened side may say to himself: If the threatener respects the new contract, knowing that the court will cancel the effect of the modification, he will also respect the old contract. This is so since, for the threatener, respecting the new contract will have the same result as respecting the old one. Hence, accepting the demand to modify the contract will protect the threatened only in situations in which he does not need protection. On the other hand, imposing punitive damages for the threats will turn incredible threats into credible ones and motivate the threatened to give up. The conclusion is that in the regime of non-enforcement, the current rule of non-compensation for the unrealized threat that was not heeded is efficient also in situations of asymmetric information regarding the credibility of the threat to breach.

Let us now investigate what is the efficient rule in a regime of enforcement of every modification. In such a regime, imposing punitive damages for the unrealized threat that was not heeded will have two effects: one is to incentivize the threatener to realize the threat, i.e., to breach the contract, which will turn incredible threats into credible ones. The second is to discourage the potential threatener from threatening to breach in case he thinks that there is a low probability that the other party will believe his threat. If the punitive damages are high enough, every threat to breach will be credible. This is, for example, the case when the threatener knows that the punitive damages for the threat are greater than the damages for the breach. Actually, if the threatened believes that the lower bound of the cost of respecting the contract is \underline{c} , then he believes that every threat in which the punitive damages for the threat (s) are greater than the damages for the breach (d) minus the lower bound of the cost of respecting the contract is credible. This means that if $s > d - \underline{c}$, then the threatened will give in to every threat. We do not investigate the effect of punitive damages for the threats themselves in a regime of enforcement of a contract that was modified only after credible threats, since this regime loses its justification when the legal system imposes damages for the threats themselves. In such a scenario, the threatened will no longer be better off (ex ante) from enforcement of modified contracts that were achieved after every credible threat.

increase the transaction cost for agreeing to a new contract, and the renegotiation may be blocked, and we may come to a result of inefficient breach instead of respecting the contract after its renegotiation.³²

G. How to Respond to Threats in International Relations

We argue that if the international community severely punishes a state that threatens to use force against another state, it renders the threat credible, and hence encourages the use of threats in international relations.

However, the situation in international relations is more complicated since we are much nearer to “the state of nature,” i.e. the pre-political situation: there is no effective enforcement mechanism for international law, especially before the International Criminal Court (ICC), whose influence even today is still ambiguous, was established. So, in the arena of international relations there may be a dilemma for how a particular state might respond to a threat against it. We can conclude from our thesis that when a state punishes another state for the threat itself, it makes the threat more credible, and by ignoring the threats it makes them less credible. For example, a state is threatening to attack another country. The cost of the attack is 120, that is, 60 from losing the opportunity to trade with the other country and 60 from losses incurred during the attack, while the benefit from the attack is 100. Therefore, the threatening country has no credible threat. However, if the threatened country responds to the threat by severing the trade relationship, then the threat of attack becomes credible.

H. How to Respond to Legitimate Threats against You

Our thesis is also applicable in advising a player how to respond to noncriminal threats. Would it be beneficial for him to respond by revenge, shaming, etc.? This time we will approach the problem not from the point of view of the social planner, but from that of the individual player.

People threaten each other a lot: if you do not pay me, I will sue you; if you do not give me a discount, I will not buy your product; if you do not give me a raise, I will leave the job; if you do not give

³² When we have a situation of asymmetric information, the problem increases since the promisee will not know that the rational thing for him to do is to propose a modification. Accordingly, he pays more.

our group a raise, we will go on strike; if you do not appoint our leader to this ministry, we, his party, will not join your coalition; if you do something I dislike, I will break off our relationship. If the other side responds to the threat by punishing the threatener instead of ignoring him, he will validate the threat and also have to bear the cost of punishing the threatener.

If a person responds to a threat of being sued by breaking off the personal or commercial relationship with the threatener, he may, in effect, make the threat credible. For example, the cost of litigation is 100, the additional cost of losing the personal/commercial relationship is 50, and the expected judgment is 130. If you respond to the threat to be sued by cutting off the relationship, you make the threat credible. If the other party predicts your response, it will probably encourage him to threaten you: he knows that after he makes the threat, you need to settle with him, which will also restore the relationship. This is applicable to situations of conflict between neighbors, siblings in an inheritance dispute, couples in divorce proceedings, and commercial partners. Therefore, it is not beneficial to have a reputation of being someone who punishes those who threaten to sue him.

If a bank director has a policy of making life hard for any worker who threatens to quit his job if he is not promoted, then he makes the threat of the worker credible. Similarly, if a firm has a policy that every threat by a worker to quit if his wage is not raised is answered either by raising his wage or by firing him, then it makes every threat of this kind credible. In a similar scenario, a worker who is an expert in a certain project threatens to quit in the middle of the project. If he leaves, the boss will lose \$50,000 per month. On the other hand, if he leaves, he will earn \$5,000 per month in his new job, whereas if he does not leave, he will continue to earn \$6,000 per month. Hence, his threat to leave is not credible; however, if the boss punishes him by reducing his wage by \$2,000, then his threat is credible. This is also the case if the prime minister makes fun of his political partner who does not carry out his threat to leave the coalition.

I. The Criminal Defense of Abandonment and Withdrawal

Our thesis that the “robber wants to be punished” can justify the controversial doctrine of criminal defense of abandonment and

withdrawal in a novel way.³³ The common justification for this defense is that we wish to discourage the criminal from carrying out a crime, after he has begun to regret his threat. We provide him with the opportunity to change his mind and to rejoin civil society. However, our justification is very different: a great advantage of the defense of abandonment and withdrawal is that it makes the criminal's threat *a priori* less credible by making the withdrawal of the criminal less expensive. We should grant the criminal the opportunity to withdraw without being punished in order to discourage him from undertaking his crime. This defense also applies as much to the very determined robber as to the hesitating type.

Our rationale for the defense of abandonment and withdrawal is valid in broader situations than the current doctrine covers and may push its boundaries. The current defense protects only criminals who felt regret prior to a change in circumstances. In contrast, our rationale is valid up to the moment that the criminal loses his chance to carry out his threat.

The price of the current doctrine is that it turns incredible threats into credible ones. In cases where circumstances have changed, there are threats that are incredible in a system in which you can refrain from carrying out your threat without being punished, but credible in a system in which you cannot do so. However, a doctrine that allows for regret at any time also has a price: in the severe system the criminal takes another risk when he decides whether to rob. If reality changes to such an extent that it becomes irrational for you to continue robbing, you still bear the punishment of refraining, and sometimes you will continue robbing even though the expected punishment from the robbery will be greater than the expected benefit. One doctrine may be that the criminal will not be exempted from punishment for the attempt if, and only if, the change in circumstances is of such significance that even if the criminal is punished for the threat, he is better off refraining. The result will be that the punishment for the threat will not give credence to his threat and will discourage him from attempting to rob. The result of the more severe doctrine will be that the robber will carry out his threat, which gives credence to his threat. However, such a

³³ "To establish the common-law defense of withdrawal from the crime of premeditated murder, a defendant must show that he abandoned and renounced his intention to kill the victim and that he clearly communicated his renunciation to his accomplices in sufficient time for them to consider abandoning the criminal plan." *See Hariman v. State*, 174 So. 3d 1044, 1047 (Fla. Dist. Ct. App. 2015).

doctrine also has an effect of deterring him from making the threat, since he knows that he has no exit if it becomes detrimental to him to carry out the crime, given the updated expected punishments. We leave it to future research to determine the precise limit of this defense, but what is already clear is that if there is a slight change in the circumstances the legal system should preserve the right of the threatener to refrain from carrying out a crime without being punished.

J. In Favor of a Pardon for War Criminals

Another application of the logic of this paper is that by providing war criminals with the possibility of a pardon, we may make their commitment to fight to the death less credible. By leaving room for pardon, we may make their commitment to fight less credible and hence discourage them and their supporters from fighting. This consideration should be taken into account in the dilemma faced by the international community regarding how to respond to war criminals like Bashar Hafez al-Assad and Muammar Gaddafi.

This logic may be used to criticize the 1970 United Nations Security Council Resolution. On February 26, 2011, the Security Council decided “to refer the situation in the Libyan Arab Jamahiriya since 15 February 2011 to the Prosecutor of the International Criminal Court.”³⁴ We argue, that this decision of the Security Council gave the ICC *retroactive* jurisdiction, which motivated Gaddafi to preserve his power in order to prevent his prosecution and strengthened his commitment to continue fighting until achieving his goals, rather than motivating him to abdicate power.

VI. LITERATURE REVIEW

A. Review of the Discussion on the Influence of Punishment on Criminal Activity

The common intuition is that the bigger the punishment for a particular act, the less this act will be done. This thesis is subject to two kinds of criticism. The first criticism is that the thesis assumes that human beings are deterred from punishment, but the truth is that they are not: they may be irrational,³⁵ or the risk of being punished may

³⁴ S.C. Res. 1970, ¶ 4 (Feb. 26, 2011).

³⁵ Russell claimed:

make the deed more attractive,³⁶ e.g., when the risk changes the potential criminal's preferences, or when the potential criminal is a masochist.³⁷ The second kind of criticism accepts the view that human beings are deterred from punishing—i.e., human beings, including potential criminals, are rational, and that their preferences are changed dramatically by the threat of punishment and they are not masochists—but it rejects the view that punishment *always* reduces criminal activity. One criticism of this kind is marginal deterrence theory. It assumes that there is an upper bound to the capacity to punish, and therefore increasing punishment may deter the potential criminal from undertaking a first criminal act. But, if he has already undertaken a first crime, the severe punishment may motivate him to undertake a second and third criminal act, as well as other criminal deeds that will help him to escape being detected and prosecuted.³⁸ Let us compare our theory to marginal deterrence theory. According to marginal deterrence theory, we should be lenient in punishing the first crime, so as to prevent *second and third crimes*. By contrast, according to the explanation in this paper, we should be lenient in punishing the first crime, i.e., the threat, so as to prevent the *first crime*, i.e., the threat.

Furthermore, the conclusion of marginal deterrence theory is that the maximum punishment is not necessarily the optimal one, whereas our conclusion is that in a world in which all players are rational and informed, the optimal punishment for the threat is epsilon. Another difference is that in marginal deterrence theory the criminal

The defect of punishment, as a means of dealing with impulses which the community wishes to discourage, is that it does nothing to prevent the existence of the impulses, but merely endeavors to check their indulgence by an appeal of self-interest. This method, since it does not eradicate the impulses, probably only drives them to find other outlets even when it is successful in its immediate object; and if the impulses are strong, mere self-interest is not likely to curb them effectually, since it is not a very powerful motive except with unusually reasonable and rather passionless people. It is thought to be a stronger motive than it is, because our moods make us decisive ourselves as to our interest and lead us to believe that it is consistent with the actions to which we are promoted by desire or impulse.

BERTRAND RUSSELL, *WHY MEN FIGHT* 38 (Routledge Classics, 1st ed. 2009) (1916).

³⁶ See *Proverbs* 9:17 (New King James) (“Stolen water is sweet, And bread eaten in secret is pleasant.”).

³⁷ See ALBERT CAMUS, *REFLECTIONS ON THE GUILLOTINE* (1957), reprinted in *RESISTANCE, REBELLION, AND DEATH* 175 (Justin O’Brien trans., 1966).

³⁸ George J. Stigler, *The Optimum Enforcement of Laws*, 78 J. POL. ECON. 526 (1970).

does not wish to be punished (not even for the first crime), whereas in our paper the robber wishes to be punished for the threat. Moreover, our theory points out that the punishment of the attempt turns the attempt into a commitment by the robber. Therefore, marginal deterrence theory actually belongs to decision theory, whereas our theory is game-theoretic.

The difference between the theories may be summarized as follows. Posner discusses a situation of robbery from the perspective of marginal deterrence theory:

[t]he attempter in our example will not, however, be punished so severely as if he had actually robbed the bank. There are two economic reasons for this: to give offenders an incentive to change their minds at the last moment (a form of marginal deterrence) and to minimize the costs of error, since there is a higher probability that an attempter really is harmless than that a person punished for an actual robbery has really done nothing.³⁹

According to Posner, we are lenient in punishing a threat in order to encourage the robber who has *already threatened* to withdraw. In contrast, according to the explanation in this paper we are lenient in punishing the threat in order to make the threat incredible, which means that we wish to encourage the *potential* robber not to make the threat.

Another criticism of the claim that increasing punishment reduces criminal activity is that what the criminal chooses is not only criminal activity but a package of criminal activity and avoidance activity that reduces the probability or magnitude of punishment.⁴⁰ Increasing punishment may lead the criminal to choose a package of more avoidance activity and more criminal activity. Another criticism is that increasing the usage of shaming penalties causes them to lose their effectiveness, since a large community of shamed people will be created.⁴¹ Polinsky and Shavell claim that punishing for extortion

³⁹ Richard A. Posner, *An Economic Theory of The Criminal Law*, 85 COLUM. L. REV. 1193, 1217-18 (1985).

⁴⁰ See Jacob Nussim & Avraham D. Tabbach, *Deterrence and Avoidance*, 29 INT'L REV. L. & ECON. 314 (2009).

⁴¹ See Alon Harel & Alon Klement, *The Economics of Stigma: Why More Detection of Crime May Result in Less Stigmatization*, 36 J. LEGAL STUD. 355 (2007).

indeed leads to fewer cases of extortion but leads to more framing of innocent individuals so as to extort money from them.⁴²

The kernel of the reason is that sanctioning extortion leads to one of two detrimental consequences: it will either fail to deter extortion and result in higher costs to innocent individuals (the sum of their expected extortion payment and the expected fine on them for paying extortion); or else it will cause enforcers to switch from extorting money from innocent individuals to framing them, which is socially worse.⁴³

Therefore, the authors claim that the optimal deterrence will be achieved by cancelling the punishment for extortion. This thesis has many problems, e.g., some policemen extort, but are not punished. Therefore, the policemen who survive in the police force are the corrupt ones. Moreover, there is also the problem of not taking into account that the credibility of a threat is a function of the punishment for the threat.⁴⁴

⁴² A. Mitchell Polinsky & Steven Shavell, *Corruption and Optimal Law Enforcement*, 81 J. PUB. ECON. 1, 3 (2001) (available earlier as Harvard Law and Economics Discussion Paper No. 288 (2000); John M. Olin Program in Law and Economics Working Paper No. 171 (1999)).

⁴³ *Id.* at 3-4.

⁴⁴ The thesis proposed in this paper is not subject to the problems challenging the thesis of Polinsky and Shavell. Firstly, even if we punish very leniently for the threat, i.e., for the attempt to extort, the result will not be that there will be people who threaten but are not punished, in contrast to Polinsky and Shavell's recommended rule not to punish for extortion. This is so since those whose threats are credible will bear the penalty for the extortion agreement (the robbery); and those whose threats are not credible will not threaten at all. An additional difference between our paper and Polinsky and Shavell lies at the heart of our thesis: we propose to examine the credibility of threats as a function of the punishment for them: punishment for threats turns incredible threats into credible ones and leads to more threats and crimes. Polinsky and Shavell ignored that even in cases where it is not beneficial to carry out the threat *before* it is made, it may be beneficial to carry it out after the threat is made because of the punishment for the threat. This means that the threat is credible not when the gain from carrying it out is greater than the cost of it, but rather when the gain from carrying it out is greater than the cost of it minus the punishment for threat! A third difference is that unlike in their model, in ours there is "incentive reversal." A fourth difference is that in their model both the criminal and the victim gain from cancelling the punishment for extortion, while in our model the victim gains from cancelling the punishment for extortion, but the robber wishes to be punished for the attempt to extort. This means that in our model, increasing the punishment promotes

VII. CONCLUSION

When the punishment for the threat is more severe, there will be more threats. The reason is that the punishment for threats turns incredible threats into credible ones. The punishment makes threats credible, since it makes withdrawing from them more expensive. The conclusion is that the punishment is a commitment device for the robber. Our conclusions are also stable in a world of irrational or informed players with the important exception that when a significant number of victims give in to an incredible threat, the optimal punishment for the threat is a moderate one.

VIII. APPENDIX – MODIFICATIONS OF THE ASSUMPTIONS

A. Asymmetric Information and Irrationality

This Section examines the robustness of the model. In a world in which all players are rational and informed, the punishment for a threat turns incredible threats into credible ones. Thus, there is an incentive reversal regarding the punishment of threat; and the punishment that minimizes the cases of threats is epsilon. What happens in a world in which not all the players are informed, or not all the players are rational?

First, we will discuss a world in which not all the players are informed and show that if the expected punishment for the threat is greater than a particular critical point, then it will be beneficial for the robbers to carry out all their threats, and it will be rational for the victims to give in to any threat. In this legal system the victim will not need any information regarding the preferences of the robber, or to make any calculation. There will be a rational rule to “give in to any threat,” and this rule will lead to a rational act in every case.

Next, we will discuss different situations of irrationality and asymmetric information and show that, in all such situations, the conclusion that the punishment for the threat turns incredible threats into credible ones is stable. Additionally, we will show that the phenomena

the interest of the criminal! A fifth difference is that we are interested in the punishment for the threat, while they are interested in the punishment for extortion, which is equivalent to the punishment for successful robbery.

of incentive reversal also exists in the “irrational worlds” of irrational victims or irrational threateners and in situations of asymmetric information. The moderate punishment for a threat will minimize threats, extortions, robberies, and blackmail in those cases where some victims give in to an incredible threat because they are irrational or uninformed and the robber cannot identify them as incredible. That is, too lenient a punishment for threats encourages the criminal to threaten in the event that the victim will be irrational, i.e., he will heed even an incredible threat. However, in situations of risk to irrational robbers or victims who give in to incredible threats, the punishment that will minimize robberies and threats is epsilon.

i. Asymmetric Information: When the Victim Does not Know If the Threat Is Credible

Let us investigate the case in which the victim does not know if the threat is credible. We will show that if the sanction for the threat is severe enough, the victim will give in to any threat, and this will motivate the robber to threaten and to ignore the risk of being punished for the threat. Let us assume that the sanction for a threat that is not heeded and not realized is capital punishment. In this legal system every threat will be credible, and the victim will give in to every threat, and the robber will not be punished for the threat (but for the robbery). This will happen even if the expected punishment for the threat is greater than the expected punishment for its being carried out. We wish to claim that the critical point that makes every threat credible is much lower. In order to discover this critical point, we will model the situation as follows:

We assumed that the victim knows how many years in prison the asset is worth for the robber. Let us now assume that the victim does not know how much the asset is worth for the robber.

However, in the case of robbery, the victim can conclude from the revealed preference of the robber to attempt to rob that the robbed asset is worth more for the robber than the expected punishment for robbery ($M > P_r$).

The victim will give in if he believes that the threat is credible, i.e., $P_m - M' < P_t$.

Let us denote: $g = M'/M = m(1-q_m)/m(1-q_r) = (1-q_m)/(1-q_r)$. This is the ratio between the probability of not being caught in the case of executing the threat, and the probability of not being caught in the case of a successful robbery.

If $P_m - gP_r < P_t$ (i.e., if the punishment for threat is great enough), then $P_m - M' < P_t$ (i.e., the threat is credible).⁴⁵

Therefore, in a legal system in which $P_t > P_m - gP_r$, i.e., the punishment for threat is great enough, every criminal threat is credible.

Therefore, even when the victim does not know how many years in prison the asset is worth for the victim, a severe enough punishment for the threat will lead to a result of “successful robbery.”

The interpretation of the result is that the law may make every threat that is made credible. The law may supply the robber with incentives to make only credible threats, such that the victim will get a signal that every threat with which he is threatened is credible. Let us explain this: from the choice of the robber to rob it can be concluded that the robbed asset is worth for him at least the expected punishment for robbery. The victim can know the lower bound of the robber's expected benefit from a successful robbery is; it is the expected punishment for robbery (P_r). Thus, he can also know the lower bound of the expected benefit from taking the asset violently.

The law can prevent the threat to take the asset violently from being carried out only if the value of the assets is low enough. Specifically, the value of these assets, when taken violently, is less than the expected punishment for carrying out the threat minus the expected punishment for making the threat ($P_m - P_t$). The point is that if the lower bound of the value of the asset that is worth robbing is greater than this size ($P_m - P_t$), then the threat becomes credible for every asset. When the punishment for the threat is severe enough, then it is beneficial to execute every threat that has already been made. One who lives in such a society (legal system) will benefit from giving in to every threat. In such a society a rule of “giving in to every threat” is not only

⁴⁵ This is so since $M > P_r$, which implies that $MM' > P_r M'$, which implies that $M' > P_r M'/M$, which implies that $M' > gP_r$, which implies that if $P_m - gP_r < P_t$, then $P_m - M' < P_t$.

a rational rule, but also a rational act. In other words, one who acts according to the rule of “giving up to every threat” will behave exactly like a person with perfect calculation capacity and perfect information. This rule will lead him to the right deed in every case. In such a society the evolutionarily stable strategy will be that all the robbers that made criminal threats will execute them whenever the victim refuses to give in, and all the victims will give in (this is also a Nash equilibrium,⁴⁶ which means a stable social norm⁴⁷). The victim need not act in a planned, rational way in order to do the rational thing, namely, to give in. It will be beneficial for all victims to give in, and they need not make any calculation or know the value of the asset. All they need to know is that in their society it is rational to give in to every criminal threat.

Moreover, let us investigate legal systems in which the punishment for a threat is not severe enough to make every threat credible. We can imagine a careful type of victim who refuses to capitulate only in “safe cases,” i.e., if and only if the upper bound of the robber's expected benefit from taking the money violently is less than the lower bound of the expected punishment for taking the money violently minus the upper bound of the punishment for the threat. If the victim knows that the upper bound of the expected benefit from taking the money violently is less than the lower bound of the punishment for taking the money violently, then he will not capitulate in a regime that does not punish for the threat. However, if the upper bound of the expected punishment for the threat is greater than the lower bound of the expected punishment for murder minus the upper bound of the expected benefit from taking the asset violently, then he will capitulate. Alternatively, we can imagine a careless type of victim who gives in if and only if he is sure the threat will be realized. He will give in if and only if the lower bound of M' is greater than the upper bound of the punishment for taking the money violently minus the lower bound for the punishment for the threat. We can see that increasing the punishment for the threat will influence him too. Actually, we can imagine that every type of victim across the spectrum has his critical point where if the probability he attributes to the credibility of the threat

⁴⁶ John F. Nash, Jr., *Equilibrium Points in n -Person Games*, 36 PROC. NAT'L ACAD. SCIS. 48 (1950).

⁴⁷ See Robert J. Aumann & Jacques H. Dreze, *Rational Expectations in Games*, 98 AM. ECON. REV. 72 (2008); See Uri Weiss & Joseph Agassi, *Game Theory for International Accords*, 16 S.C. J. INT'L L. & BUS. 1 (2020).

being carried out is greater than that critical point, then he will give in. The individual critical point for every rational victim is positively correlated with the sanction for the threat. The conclusion is that the punishment incentivizes the victim to give in, when he does not know whether the threat is credible.

ii. What Happens When the Victim Does not Know Whether the Robber is Rational?

We have assumed that the victim knows that the robber is rational, which means that he will not commit threats that are not worthy for him to be carried out. Let us assume now that some of the robbers are irrational, in the sense that they will execute the threat even when it is irrational for them to do so. The victim cannot distinguish between rational and irrational threateners.

* Let us denote by b the probability that the robber is irrational, i.e., the probability that he will commit the threat, regardless of whether it is beneficial for him to do so or not.

* Let us denote by L the loss that will be borne by the victim from the execution of the threat by the robber.

* The victim will give in to the robber either if the robber has a credible threat ($P_t > M' - P_m$), or if the expected damage from an irrational robber is greater than the expected damage from giving in, i.e., $Lb > M$.

* Thus, the analysis has not changed qualitatively: when the punishment is bigger, then there will be more threats and robberies.

Another interesting conclusion is that irrational robbers attract also rational robbers to join the marketplace of crime. We have a critical point that if the proportion of irrational robbers in the population (the marketplace of crime) is greater than that point, then new rational people will join the criminal population until the proportion of irrational robbers is not greater than that critical point. Thus, society benefits more when we deter irrational criminals than when we deter rational ones.

iii. What Happens If Some Victims Give in Even to Incredible Threats?

Let us assume that some victims will give in even to incredible threats. This may be either because the victim is irrational or uninformed (including about the rationality of the robber). We assume that the robber does not know if the specific victim will give in to his incredible threat.

Let us begin with an example: 50% of the victims give in to any threat, the value of the asset (when it is robbed or taken violently) is 10 months in prison, the expected punishment for robbery is 8 months in prison, and the expected punishment for taking the asset violently is 16 months in prison. The result is that the robber will attempt to rob if the punishment for the threat is less than 2 months in prison. The robber will have a 50 percent probability of gaining 2 (in the case of an “irrational victim”) and a 50 percent probability of losing 2 (in the case of a “rational victim”). However, he will also attempt to rob if the punishment is greater than 6 months, since then he will have a credible threat. Thus, he will not attempt to rob if and only if the punishment is greater than 2 months, but less than 6 months.

Let us now introduce some modifications to our example. First, if we modify the example such that $\frac{1}{3}$ of the victims give in to every threat, then in order to prevent the robbery of such an asset the punishment for the threat should be greater than 1 month (then the robber has a chance of $\frac{1}{3}$ to gain 2 months versus a risk of $\frac{2}{3}$ to lose more than 1 month) but less than 6 months (otherwise the threat will be credible).

Second, if we modify the example such that more than 0.75 of the victims give in to any credible threat, there is no punishment for a threat that will prevent the robbery: if it is more than 6 months, then every threat becomes credible, and it is worthwhile to rob. However, if it is less than 6 months, then it is worthwhile to attempt to rob in the event that the victim is one who gives in to any threat. Hence, even if the judge does not know the proportion of victims who give in to any threat, she can conclude that a punishment of more than 6 months for attempting to rob such an asset cannot prevent the robbery. More generally, even if the judge does not know the proportion of victims who give in to any threat, she should not impose a punishment greater than the critical point that turns incredible threat into credible ones, which is independent of the number of such victims.

Third, let us now modify the value of the asset such that the value of an asset robbed with violence is 15 months in prison. Fifty percent of the victims give in to any threat, and the punishment for robbery and taking the asset violently are as above. Thus, there is no punishment for the threat that can prevent the threat. If the punishment for the attempt is less than 7 months, then it is rational to rob in the event that the victim is irrational. However, if the punishment for the threat is greater than 1 month, then every threat is credible.

Furthermore, if the punishment for a successful robbery and the punishment for taking money violently are independent of the value of the asset, then in our example the upper bound of the value of the asset whose robbery the law can prevent is 12 months in prison. If the asset is more expensive than that, then a punishment of 4 or more months for the threat makes the threat credible, but a punishment of 4 or less months makes the attempt worthwhile in the event that the victim will give in to any threat. In this case, the robber will have a fifty percent probability of gaining 4 and a fifty percent probability of losing less than 4. The robbery of an asset with a value of 12 will be prevented by a punishment of 4 months for the attempt. This punishment will prevent even the robbery of any cheaper asset. Hence, this may be considered to be the optimal punishment.

Let us now investigate the general case. We assume that the robber attributes a probability e that the victim will give in to his incredible threat (e may be a function of the punishment for the threat (P_t)). The victim may give in because he is irrational or because he is rational but uninformed.

The robber will not rob if his expected benefit from a successful robbery is less than his expected punishment for robbery ($M < P_r$). However, if his expected benefit from a successful robbery is greater than his expected punishment for robbery ($M > P_r$), he may rob. In that case he will rob if and only if he has a credible threat *or* he has no credible threat, but it is still worthwhile to rob since the victim may be one who gives in to every threat. If the robber makes an incredible threat, his benefit will be eM (the victim will give him the money with probability e), and his cost will be $eP_r + (1-e)P_t$ (with probability e , he will be punished for robbery, and with probability $1-e$ he will be punished for the threat). Thus, when the robber's threat is incredible, he will attempt to rob if and only if $eM > eP_r + (1-e)P_t$, which means if and only if $P_t < (e/1-e)[M-P_r]$.

Thus, the robber will not rob if and only if:

1. $(e/1-e)(M-P_r) < P_t < P_m - M'$ (his threat is not credible, and, in addition, in the case of an incredible threat the punishment for the attempt makes it not worthwhile to attempt it in order to have the opportunity to meet an irrational victim); or
2. $M < P_r$ (even a successful robbery is not worthwhile).

Hence, we should have a moderate punishment for the attempt. It should be lenient enough to make the threat incredible and severe enough to deter the robber from attempting to rob in the event that the victim is either irrational or careful but uninformed. What is special in this case is that the robber may bear the punishment for the threat, since he may make threats he does not intend to realize.⁴⁸

Let us investigate now what the optimal punishment for the attempt is when the punishments for robbery and for carrying out the threat are independent of the value of the asset. In order to prevent the robbery, the punishment for the attempt should be: $(e/1-e)(M-P_r) < P_t < P_m - M'$. Thus, there is no punishment for the attempt that prevents the robbery, when $P_m - M' < (e/1-e)(M-P_r)$. In these cases, it is sufficient that the value of the robbed asset be greater than the punishment for robbery, in order that for the robber to attempt.⁴⁹ Let us now ask what is the most expensive asset (M^*) whose robbery can be prevented by the punishment for the threat. We will find the answer to this question by solving the following equation⁵⁰: $(e/1-e)(M^*-P_r) = P_m - M'^*$,

⁴⁸ However, when $e/1-e$ is a function of P_t , and in addition $\Delta(e/1-e)/\Delta P_t > 1$, then in order to make P_t greater than $(e/1-e)(M-P_r)$, we may need to *decrease* P_t . More specifically, when $\Delta(e/1-e)/\Delta P_t > 1$, i.e., when the change in punishment changes the ratio between the number of victims who give in to an incredible threat and the number of victims who do not give in, more quickly than the change in punishment; we need to *decrease* the punishment for the threat in order to discourage the robber from making an *incredible* threat in the event that the victim will give in to any threat. Of course, the conclusion, that we need to decrease the punishment for the threat in order to make threats incredible ones, is stable in those cases too. Furthermore, if $\Delta(e/1-e)/\Delta P_t > 1$ for every P_t , then the optimal punishment is epsilon.

⁴⁹ This may be a good reason why the punishment for robbery, or for carrying out the threat to take the money violently, should depend on the value of the asset.

⁵⁰ We implicitly assume that if the punishment is severe enough to prevent the making of an *incredible* threat to take violently m , it is sufficient to prevent the making of *incredible* threat to take violently any sum of money greater than m . In other words, we assume that $(e/1-e)(M-P_r)$ is always greater when M is greater. In other words, we assume that $\Delta[(e/1-e)(M-P_r)]/\Delta M > 0$.

which means $(e/1-e)(M^*-P_r)=P_m-gM^*$. Hence $M^*=[eP_r+(1-e)P_m]/(e+g-eg)$.

Now, let us ask which punishment for the threat (P_t^*) is needed in order to prevent the robbery of the asset (M^*). It is when $P_t^*=P_m-M^*$ (this is the biggest punishment for the threat that will not make the threat to rob such an asset credible). Thus, $P_t^*=P_m-g[eP_r+(1-e)P_m]/(e+g-eg)=\{P_m(e+g-eg)-g[eP_r+(1-e)P_m]\}/(e+g-eg)=(eP_m+gP_m-egP_m-egP_r-gP_m+egP_m)/(e+g-eg)=(eP_m-egP_r)/(e+g-eg)=e(P_m-P_r)/(e+g-eg)$.

Let us now investigate what is the optimal punishment when $M=M'$, i.e., when the value of the robbed asset is equal to the value of the asset taken violently, i.e., when $g=1$. In this case, we get the elegant result that the optimal punishment is $P_t^*=e(P_m-P_r)$. This means that when the gap between the robbed asset and the asset that is taken violently is negligible, the optimal punishment for the threat is the proportion of victims who give in to any threat multiplied by the gap between the punishment for robbery and the punishment for the attempt. Such a punishment for an attempt will prevent the attempt to rob the most expensive asset that the law can prevent, which is $M^*=(1-e)P_m+eP_r$, and generally⁵¹ any cheaper asset.

We can conclude that when some victims give in to any threat, the punishment for a threat may be productive. It should be severe enough to make it not worthwhile for the robber to attempt, and lenient enough not to make the threat credible. When the number of victims who give in to any threat is bigger, then the punishment should be bigger, but still low enough. When the number of victims who give in to any threat is big enough, the law may lose its capacity to prevent the robbery by punishing for the attempt. Similar to the results of Bar-Gill and Ben-Shahar,⁵² we need the intervention of the law especially in cases where the threatened *may* give in to incredible threats. However, if the punishment for the attempt in such cases is severe enough, then

Actually it is sufficient to assume that if $P_m-M'<(e/1-e)(M-P_r)$ for $M=x$, then $P_m-M'<(e/1-e)(M-P_r)$ for any $M>x$.

⁵¹ The qualifier 'generally' applies here since e may be negatively correlated with M , and if $-\Delta e/\Delta M$ is big enough, a punishment that deters the robber from making an incredible threat regarding an expensive asset will not deter him from making an incredible threat regarding a cheap asset, and in that case the punishment that prevents the robbing of the most expensive asset the law can, will not necessarily prevent the robbing of the cheaper one.

⁵² Oren Bar-Gill & Omri Ben-Shahar, *The Law of Duress and the Economics of Credible Threats*, 33 J. LEGAL STUD. 391, 422-24 (2004).

the threat will become credible and encourage robberies. Even if the judge or jury does not have all that data, it is important that in the case of attempted robbery they impose a punishment such that if the criminal knows its size in advance, he will carry out his threat.

iv. What Happens If the Robber Does Know That the Victim Will Give in to His Incredible Threat?

Let us now investigate a scenario in which the robber's threat is credible but he fears that the victim will not give in. This may be because the victim is irrational (he does not give in to credible threats) or because he is a risk-loving, uninformed type. We will denote by h the probability that the robber believes that the victim will not give in to his credible threat.

This time if the robber's threat is credible ($P_t > P_m - M'$), there is a probability of $1-h$ that he will obtain the payoff for a successful robbery ($M - P_r$) and a probability of h that he will obtain the payoff for taking the asset violently ($M' - P_t$). Hence, given that his threat is credible, he will attempt to rob if and only if $(M - P_r)(1-h) + h(M' - P_t) > 0$. We can see that in this case, our conclusions are stable. The explanation is that the robber knows that in any possible situation he will not pay the price for the threat. If the victim gives in to his threat, he will pay for the robbery, and if he does not give in to his threat, he will pay for taking the asset violently. Hence, the robber is not deterred by the punishment for the threat.

B. The Case of the Repeat Robber

Until now we have assumed that the robber is a one-time offender. Now, let us assume that he is a repeat player. Unlike a one-time offender, a repeat robber accumulates a reputation. Therefore, when this robber carries out his threat, he also gains a reputation (r). As a repeat player his benefit from a successful robbery or from taking the asset violently is actually increased, and the other parameters are not changed. Thus, we can see that the repeat robber, more than the one-time robber, will evaluate the expected benefit from carrying out the threat. As before, if the expected punishment for the threat is big enough, the threat will be credible. In order for the threat to be credible, the punishment for the threat (P_t) needs to be bigger than the difference between the expected punishment for carrying out the threat

(P_m) and the expected benefit from carrying out the threat, which in this case includes also the reputational gain ($M' + r$).⁵³ Therefore, the critical expected punishment for the threat beyond which the threats are credible is now lower. However, the conclusion is the same as before: the threat is credible if and only if the punishment for the threat is bigger than the critical punishment that makes the threat credible. When the punishment for the threat is greater, then the set of assets that the criminal can credibly threaten to take by violence becomes greater. Therefore, when the punishment for the threat is bigger, there will be more robberies (or extortions, blackmail, etc.). The conclusion is that in the case of a repeat robber, the threat will be credible when the punishment for it is big enough. We have shown that by moving to a game with a repeat robber, the conclusion of the paper is qualitatively unchanged. There is still incentive reversal and the punishment for the threats themselves makes them credible and encourages criminals to make them.

⁵³ The threat is credible if and only if $P_t > P_m - (M' + r)$.