Electronic Supplementary Materials for

What Motivates Direct and Indirect Punishment?

Extending the 'Intuitive Retributivism' Hypothesis

Catherine Molho[a*], Mathias Twardawski[b], & Lei Fan[c]

[a] *Institute for Advanced Study in Toulouse*

[b] *Ludwig-Maximilians Universität München*

[c] *Vrije Universiteit Amsterdam*

* Corresponding author | catherine.molho@iast.fr

Institute for Advanced Study in Toulouse

Université Toulouse 1 Capitole

1, Esplanade de l'Université

31080 Toulouse, France

Table of Contents

Materials **3**
Vignette Instructions 3
Moral Violation Vignettes 3
Manipulation Checks 6
Punishment Responses 6
Punishment Motives 7
Reputational Concern 8
Robustness of Main Analyses **10**
Full Sample 10
Manipulation Checks 10
Main Analyses 11
Robustness of Auxiliary Analyses **13**
Self-Reported Motives and Overall Punishment Tendencies 13
Self-Reported Motives and Direct versus Indirect Punishment 13
Additional Auxiliary Analyses **14**
Disentangling General Deterrence and Reputation Accounts 14
General Deterrence, Reputation Concern, and Overall Punishment Tendencies 15
General Deterrence, Reputation Concern, and Direct versus Indirect Punishment 16

Materials

Vignette Instructions

On the next page, you will read a scenario. We would like you to focus on how you would feel and react in response to the situation described in the scenario. Please read the scenario carefully and try to experience what is described in it as vividly as possible. After you read the scenario, we will ask you about how it made you feel and how you would react to it.

Moral Violation Vignettes

Vignette #1. Picture attending a party that is being hosted by a casual friend of yours. Some of your close friends are at the party, but most of the people there are just acquaintances. After you've been at the party for a while, you realize that you need to make a phone call. You go to the room where you and the other guests have left your coats to make the call. When you enter the room, you see that another guest – a man that you recognize, but whom you're not friends with – is smoking a cigarette and that he has been casually flicking ashes onto the top jacket on a pile of jackets.

[High severity] He looks at you and gives you a tight smile before he purposefully stubs out his cigarette on the jacket. You look closer and see that the jacket on the top of the pile has been badly damaged by the cigarette.

[Low severity] He looks at you and gives you a tight smile before flicking another bit of ash on the jacket. You look closer and see that the jacket on the top of the pile has been slightly stained by the ashes.

[High observability] Later that night, you bump into this same guest in the living room area. You look around and see that many other guests are still present.

[Low observability] Later that night, you bump into this same guest in the living room area. You look around and see that a few other guests are still present.

Vignette #2. Picture attending a party that is being hosted by a casual friend of yours. Some of your close friends are at the party, but most of the people there are just acquaintances. After you've been at the party for a while, you decide to step outside to get some air. When you walk outside, you see that another guest – a man that you recognize, but whom you're not friends with – is making a phone call. While talking on the phone, he is casually pulling the flowers and the leaves off the plants.

[High severity] He looks at you and gives you a tight smile before he purposefully tears off the blossoms of several flowers. You look closer and find that there are many scattered leaves and that the plants are likely destroyed.

[Low severity] He looks at you and gives you a tight smile before he continues to play with the plants. You look closer and find that there are a few scattered leaves and that the plants are only slightly damaged.

[High observability] Later that night, you bump into this same guest in the living room area. You look around and see that many other guests are still present.

[Low observability] Later that night, you bump into this same guest in the living room area. You look around and see that a few other guests are still present.

Vignette #3. Picture attending a party that is being hosted by a casual friend of yours. Some of your close friends are at the party, but most of the people there are just acquaintances. After you've been at the party for a while, you decide to step outside to get some air. When walking outside, you see another guest – a man that you recognize, but whom you're not friends with.

[High severity] He is putting various expensive silverware that you previously saw on the dining table into his backpack. He looks at you and gives you a tight smile before he walks away.

[Low severity] He is putting a cheap ashtray that you previously saw on the balcony into his backpack. He looks at you and gives you a tight smile before he walks away.

[High observability] Later that night, you bump into this same guest in the living room area. You look around and see that many other guests are still present.

[Low observability] Later that night, you bump into this same guest in the living room area. You look around and see that a few other guests are still present.

Vignette #4. Picture attending a party that is being hosted by a casual friend of yours. Some of your close friends are at the party, but most of the people there are just acquaintances. After you've been at the party for a while, you begin to feel a little hungry. You go to the kitchen to get some food. When you enter the kitchen, you see another guest – a man that you recognize, but whom you're not friends with.

[High severity] He is hastily putting a bottle of expensive whiskey into his backpack. He looks at you and gives you a tight smile before walking away.

[Low severity] He is hastily putting a couple of beer cans into his backpack. He looks at you and gives you a tight smile before walking away.

[High observability] Later that night, you bump into this same guest in the living room area. You look around and see that many other guests are still present.

[Low observability] Later that night, you bump into this same guest in the living room area. You look around and see that a few other guests are still present.

Manipulation Checks

Next, we would like you to answer a few questions about the scenario you just read. '**The offender'** refers to the person who did something wrong in the scenario you read.

- How morally wrong do you think the offender's behavior was?

(1 = *not at all morally wrong*; 2 = *not morally wrong*; 3 = *neutral*; 4 = *morally wrong*; 5 = *extremely morally wrong*)

- How harmful do you think the offender's behavior was?

(1 = *not at all harmful*; 2 = *not harmful*; 3 = *neutral*; 4 = *harmful*; 5 = *extremely harmful*)

- How likely do you think it is that other guests will know your reaction to the offender's behavior?

- How likely do you think it is that only you will know your reaction to the offender's behavior?

(1 = *not at all likely*; 2 = *unlikely*; 3 = *neutral*; 4 = *likely*; 5 = *extremely likely*)

Punishment Responses

Punishment severity:

- To what extent do you think the offender should be punished?

(1 = *not at all*; 7 = *very much*)

On the next page, we will ask you to read some statements and rate how well each of them describes how you would act towards **the offender** (the person who did something wrong).

Direct punishment items:

- I would hit the offender.

- I would insult the offender to his face.

- I would shove the offender.

- I would get in the face of the offender.

- I would yell at or argue with the offender.

Indirect punishment items:

- I would spread negative information about what the offender did to other guests.
- I would mention something bad about what the offender did to others who know him.
- I would try to get other guests to dislike the offender based on what he did.
- I would express my disapproval about what the offender did to other guests.
- I would tell other guests what the offender did.

(1 = *not at all*; 7 = *very much*)

Punishment Motives

(adapted from the Sentencing Goals Inventory**;** McKee & Feather, 2008)

Listed below are a number of statements that describe attitudes that different people have about justice in the community. There are no right or wrong answers, only opinions. Read each item and decide whether you agree or disagree and to what extent.

Retribution items:

- Severe sentences are appropriate for offenders who commit serious offenses.
- Offenders should be punished in proportion to the seriousness of their crimes.
- Harm to the victim should be considered when setting the punishment for a given offense.
- Offenders should be made to bear full responsibility for their actions.
- Harsher crimes deserve harsher punishment.

Deterrence items:

- Every punishment should be well publicized.
- Emphasis should be placed on keeping potential other offenders from doing any harm.
- Offenders should be harshly punished as examples to others.

- If there would be tougher punishments against offenses, there wouldn't be so many offenders.

- Light punishments do not provide enough threat to deter people from offenses.

Importance items:

Next, we would like you to rate how important you think each goal is on the 1 (*not at all important*) to 7 (*very important*) scales provided.

- To make sure that the offender "pays" in some way for what they have done.
- To deter other potential offenders.
- To deter the offender from committing similar offenses in the future.

These items will be rated on 7-point Likert scales (1 = *not at all important*; 7 = *very important*)

Reputational Concern

(Jordan & Rand, 2019; adapted from the Brief Fear of Negative Evaluation Scale)

These items will be rated on 7-point Likert scales (1 = *not at all characteristic of me*; 7 = *very characteristic of me*)

1. I am afraid that others will not approve of me.

2. I am afraid that people will find fault with me.

3. I often hope that I will say or do the right things.

4. Sometimes I think I am too concerned with other people liking me.

5. I hope that people will view me favorably.

6. I frequently hope that other people will notice my positive attributes.

7. I worry about what other people will think of me even when I know it doesn't make a difference.

8. When I am talking to someone, I worry about what they may be thinking about me.

9. I am frequently afraid of other people noticing my shortcomings.

10. I often worry that I will say or do the wrong things.

11. I hope that other people will like me even when I know it doesn't make a difference.

12. I hope that others will approve of me.

13. When I am talking to someone, I hope that they will be thinking positive things about me.

14. I am usually worried about what kind of impression I make.

15. Sometimes I think I am too concerned with what other people think of me.

16. I am usually excited about the idea of making a good impression.

Robustness of Main Analyses

Full Sample

The full recruited sample (including potentially inattentive respondents) consisted of 350 participants (63.4% male; M_{age} = 46.5 years, SD_{age} = 12.18). In terms of educational attainment, eight participants had some high school education (2.3%); 89 had completed high school (25.4%); 105 had some college education (30.0%); 108 had obtained a bachelor's degree (30.9%); 35 a master's degree (10.0%); and 5 a doctoral degree (1.4%). In sum, our full sample was diverse in terms of age and educational background, though skewed toward including more male participants.

Manipulation Checks

We repeated the analyses reported in our main manuscript using the full sample of participants. Specifically, we conducted 2 × 2 ANOVAs testing the effects of the severity (*high* versus *low*) and the observability (*high* versus *low*) manipulations on the manipulation checks. Our manipulation of offense severity worked as intended. The severity manipulation had a main effect on the perceived severity aggregate, $F(1, 346) = 25.00$, $p < .001$, $\eta^2 = 0.07$, with participants in the high severity condition ($N = 169$) correctly perceiving offenses as more severe ($M = 5.51$, $SD = 1.16$) than participants in the low severity condition ($N = 181$; $M = 4.83$, $SD = 1.38$). There was no main effect of the observability condition ($p = .474$) and no severity condition × observability condition interaction ($p = .846$) predicting perceived severity.

When using the full sample, our manipulation of punishment observability again did not work as intended. Similar to the results reported in the main manuscript, there was no main effect of the observability manipulation on the perceived observability aggregate, $F(1, 346) = 0.71$, $p = .401$, $\eta^2 < .01$. Participants perceived punishment observability similarly irrespective of whether

they were in the high (N = 181; M = 3.36, SD = 1.38) or low (N = 169; M = 3.50; SD = 1.55) observability condition. In contrast, the severity manipulation had a main effect on perceived observability of punishment, $F(1, 346) = 11.28$, $p = .001$, $\eta^2 = 0.03$, with participants in the high severity condition perceiving punishment as somewhat more observable (M = 3.70, SD = 1.46), compared to participants in the low severity condition (M = 3.17, SD = 1.43). There was no severity condition × observability condition interaction (p = .652) predicting perceived observability.

Main Analyses

Overall punishment tendencies. As in the main manuscript, to test **H1** and **H2**, we conducted a 2 × 2 ANCOVA testing the effects of the severity manipulation (*high* versus *low*), the observability manipulation (*high* versus *low*), and their interaction on individuals' overall punishment tendencies (i.e., ratings of how much the offender should be punished)—this time using the full available sample. We included gender as a covariate, to control for sex differences in aggressive tendencies.

Results replicate the patterns reported in our main manuscript. Specifically, when using the full sample, there was no severity condition × observability condition interaction (p = .897) predicting participants' overall punishment tendencies. Consistent with the 'intuitive retributivism' account, the severity manipulation had a main effect on overall punishment ratings, $F(1, 345) = 30.22$, $p < .001$, $\eta^2 = 0.08$, such that participants thought the offender should be punished *more* when offense severity was high (M = 5.28, SD = 1.43) rather than low (M = 4.36, SD = 1.60). Consistent with a 'strong' version of the intuitive retributivism perspective (**H1b**), there was no main effect of the (unsuccessful) observability manipulation on overall

punishment ratings (p = .802). We also did not observe an effect of participant gender on overall punishment tendencies (p = .157).

Direct and indirect punishment. As in the main manuscript, to test **H3** and **H4**, we conducted a mixed 2 (between-subjects severity: *high* versus *low*) × 2 (between-subjects observability: *high* versus *low*) × 2 (within-subjects punishment type: *direct* versus *indirect*) ANCOVA. The focus of these analyses was on the severity × punishment type and the observability × punishment type interactions. However, we also tested for main effects of the severity and observability manipulations and included the three-way interaction between severity × observability × punishment type in our model (for the sake of completeness). We used the aggregates of direct punishment items and indirect punishment items as two levels of the within-subjects punishment type factor. Again, we tested for a main effect of participant gender, as well as the gender × punishment type interaction.

Contrary to our expectations, we did not observe a severity manipulation × punishment type interaction (**H3**; p = .667), nor an observability manipulation × punishment type interaction (**H4**; p = .583) predicting punishment endorsements. Instead, consistent with our analyses on overall punishment tendencies, we observed a main effect of the severity manipulation on punishment, $F(1, 345) = 21.13$, $p < .001$, $\eta^2 = 0.06$, such that both direct and indirect punishment were stronger when the offense severity was high (*direct*: $M = 2.91$, $SD = 1.50$; *indirect*: $M = 4.12$, $SD = 1.43$) as compared to low (*direct*: $M = 2.20$, $SD = 1.36$; *indirect*: $M = 3.49$, $SD = 1.44$). There was no main effect of the observability manipulation on punishment (p = .500), nor a severity × observability interaction predicting punishment (p = .145). Consistent with results in our main manuscript, there was a substantial difference in the overall endorsement of direct versus indirect punishment, $F(1, 345) = 329.26$, $p < .001$, $\eta^2 = 0.49$, with participants more

strongly inclined to punish indirectly (M = 3.80, SD = 1.46) than directly (M = 2.54, SD = 1.47). Consistent with previous work, we observed that gender had a main effect on punishment tendencies, $F(1, 345) = 10.05$, $p = .002$, $\eta^2 = 0.03$, with men reporting stronger punishment tendencies than women—and this effect not qualified by punishment type ($p = .227$).

Robustness of Auxiliary Analyses

Self-Reported Motives and Overall Punishment Tendencies

As in the main manuscript, we conducted auxiliary analyses to examine the relations of self-reported retribution and deterrence motives with overall punishment tendencies. First, we run a general linear model testing the effects of retribution motives, deterrence motives, and their interaction predicting participants' overall punishment tendencies (while controlling for participant gender). Results showed that self-reported retribution motives had no main effect on overall punishment ratings, $F(1, 345) = 0.76$, $p = .383$, $\eta^2 < .01$. Instead, we observed a positive main effect of deterrence motives on overall punishment, $b = 0.95$, $F(1, 345) = 9.54$, $p = .002$, $\eta^2 = 0.03$. There was no interaction of retribution × deterrence motives ($p = .141$), nor a main effect of gender predicting overall punishment tendencies ($p = .087$).

Self-Reported Motives and Direct versus Indirect Punishment

As in the main manuscript, we then run a general linear model testing the effects of retribution and deterrence motives on endorsements of direct versus indirect punishment tendencies (as two levels of a within-subjects punishment type factor). We focused on the retribution motives × punishment type and the deterrence motives × punishment type interactions, but also tested for main effects of retribution and deterrence motives across punishment types. Again, we controlled for participant gender, and tested for the gender × punishment type interaction.

Pertaining to our question of whether retribution and deterrence motives differentially relate to direct versus indirect punishment, we again found no evidence to support this prediction. When using the full sample, we found no evidence of a retribution motives × punishment type interaction ($p = .120$), nor a deterrence motives × punishment type interaction ($p = .136$) predicting punishment tendencies. Instead, consistent with our results on overall punishment tendencies, we observed a positive, main effect of deterrence motives on punishment, $F(1, 345) = 24.18$, $p < .001$, $\eta^2 = 0.07$, which held both for direct punishment tendencies ($b = 1.40$, $p < .001$, $\eta^2 = 0.07$) *and* indirect punishment tendencies ($b = 1.03$, $p < .001$, $\eta^2 = 0.04$). In contrast, we did not observe a main effect of retribution motives on punishment, $F(1, 345) = 0.47$, $p = .492$, $\eta^2 < 0.01$. Finally, there was a main effect of gender on punishment endorsements, $F(1, 345) = 12.03$, $p = .001$, $\eta^2 = 0.03$, with men reporting stronger punishment tendencies than women—this effect was not qualified by punishment type ($p = .248$).

Additional Auxiliary Analyses

Disentangling General Deterrence and Reputation Accounts

To disentangle two potential mechanisms by which observability of punishment may influence individuals' punishment tendencies, we conducted auxiliary analyses including measures of deterrence concerns and reputational concerns as additional predictors. According to a general deterrence account, we would expect the effect of observability on punishment tendencies to be stronger among individuals with higher self-reported deterrence motives. In contrast, according to a reputation account, we would expect the effect of observability on punishment tendencies to be stronger among individuals with higher self-reported reputational concerns.

We re-run the ANCOVA models from our main analyses (reported in the main manuscript), but this time including deterrence concerns, reputation concerns, and the deterrence concern × observability and reputation concern × observability interactions. If observability influences punishment via a deterrence mechanism, we would expect to see a statistically significant deterrence concern × observability interaction. More specifically, we would expect observability to have a stronger effect on punishment among individuals with high (rather than low) deterrence concerns. In contrast, if observability influences punishment via a reputation mechanism, we would expect to see a statistically significant reputation concern × observability interaction. More specifically, we then expect observability to have a stronger effect on punishment among individuals with high (rather than low) reputation concern.

General Deterrence, Reputation Concern, and Overall Punishment Tendencies

We conducted a 2 × 2 ANCOVA testing the effects of the severity manipulation (*high* versus *low*), the observability manipulation (*high* versus *low*), and their interaction on individuals' overall punishment tendencies. In these analyses, we included participant gender as a covariate, and we also tested for effects of general deterrence concern, reputation concern, and the general deterrence × observability condition and reputation concern × observability condition interactions predicting overall punishment ratings.

The pattern of results replicates the findings reported in our main manuscript. We observed a main effect of the severity manipulation on overall punishment tendencies, $F(1, 299) = 30.63$, $p < .001$, $\eta^2 = 0.09$ (*full sample*: $F(1, 341) = 36.46$, $p < .001$, $\eta^2 = 0.10$), with participants in the high severity condition indicating that the offender should be punished more ($M = 5.28$, $SD = 1.46$; *full sample*: $M = 5.28$, $SD = 1.43$) compared to participants in the low severity condition ($M = 4.35$, $SD = 1.62$; *full sample:* $M = 4.36$, $SD = 1.60$). We also observed a

positive main effect of self-reported deterrence motives on overall punishment ratings, $F(1, 299) = 34.37$, $p < .001$, $\eta^2 = 0.10$ (*full sample:* $F(1, 341) = 41.47$, $p < .001$, $\eta^2 = 0.11$). There was no main effect of the observability condition on overall punishment tendencies ($p = .597$; *full sample:* $p = .258$) and no main effect of reputational concern on overall punishment tendencies ($p = .170$; *full sample:* $p = .105$). Further, results showed no evidence for the general deterrence × observability condition ($p = .331$; *full sample:* $p = .143$) and reputation concern × observability condition ($p = .612$; *full sample:* $p = .813$) interactions predicting punishment ratings. No other effects were statistically significant ($p \geq .100$; *full sample:* $p > .090$).

General Deterrence, Reputation Concern, and Direct versus Indirect Punishment

We conducted a mixed 2 (between-subjects severity: *high* versus *low*) × 2 (between-subjects observability: *high* versus *low*) × 2 (within-subjects punishment type: *direct* versus *indirect*) ANCOVA. We included participant gender as a covariate and tested for the gender × punishment type interaction. In these analyses, we also tested for effects of general deterrence concern, reputation concern, and the general deterrence × observability condition and reputation concern × observability condition interactions predicting direct versus indirect punishment.

Results were again consistent with the findings reported in our main manuscript but provided no support for either the general deterrence or the reputation concern accounts laid out above. Specifically, none of the predictors included in our model differentially related with direct versus indirect punishment tendencies ($p > .089$ for all within-subjects effects; *full model*: all ps $> .130$, with the exception of a statistically significant deterrence motives × punishment type interaction, $p = .041$). Consistent with the results above, we observed a main effect of the severity manipulation on punishment tendencies, $F(1, 299) = 20.71$, $p < .001$, $\eta^2 = 0.06$ (*full sample*: $F(1,341) = 24.36$, $p < .001$, $\eta^2 = 0.07$) and a main effect of self-reported deterrence

concerns on punishment tendencies, $F(1, 299) = 27.53$, $p < .001$, $\eta^2 = 0.08$ (*full sample:* $F(1,341) = 23.60$, $p < .001$, $\eta^2 = 0.06$). There was also a main effect of gender on overall punishment tendencies, $F(1, 299) = 6.00$, $p = .015$, $\eta^2 = 0.02$ (*full sample:* $F(1,341) = 11.98$, $p = .001$, $\eta^2 = 0.03$), but no other predictors were statistically significant (all *p*s > .120; *full model*: *p*s > .090).