"If Only" Counterfactual Thoughts about Cooperative and Uncooperative Decisions in Social Dilemmas

Stefania Pighin[1*], Ruth M. J. Byrne[2], & Katya Tentori[1]

[1] Center for Mind/Brain Sciences, University of Trento, Trento, Italy;

[2] School of Psychology and Institute of Neuroscience, Trinity College Dublin, University of Dublin, Ireland.

[*] Corresponding Author

Stefania Pighin
Center for Mind/Brain Sciences (CIMeC) - University of Trento
Corso Bettini, n. 31 38068 Rovereto (TN), Italy
email: stefania.pighin@unitn.it

Abstract

We examined how people think about how things could have turned out differently after they made a decision to cooperate or not in three social interactions: the Prisoner's dilemma (Experiment 1), the Stag Hunt dilemma (Experiment 2), and the Chicken game (Experiment 3). We found that participants who took part in the game imagined the outcome would have been different if a different decision had been made by the other player, not themselves; they did so whether the outcome was good or bad for them, their own choice had been to cooperate or not, and the other player's choice had been to cooperate or not. Participants who only read about a fictional protagonist's game imagined changes outside the protagonist's control (such as the other player's decision) after a good outcome but within the protagonist's control (such as the protagonist's decision) after a bad outcome. The implications for theories of counterfactual thinking and moral decision-making are discussed.

Key words: counterfactuals, cooperation, moral decisions, Prisoner's dilemma, Stag hunt dilemma, Chicken game

1. Introduction

In one of Puccini's most dramatic operas, Tosca, a singer living in Rome in the 1800's, begs Scarpia, a secret police General, to save the life of her lover, Cavaradossi. Scarpia offers Tosca a deal: if she yields herself to him, he will spare Cavaradossi by staging a mock execution; Tosca accepts. However, they both betray their agreement: Scarpia, assuming that Tosca would have made love to him before the execution, instructs the firing squad to use real bullets; Tosca, assuming that the mock execution had already been arranged, stabs Scarpia to death rather than submit to his advances. In the end, a despairing Tosca takes her own life. The opera has been described as revolving around a social dilemma that results in mutual defection: Tosca and Scarpia each experience a worse outcome than they would have if they had both respected their agreement (e.g. Dawes, 1980; Poundstone, 1992). Yet neither would have gained if their own decision only were changed: things would not have been better for Tosca if she had not betrayed Scarpia, nor would they have been better for Scarpia if he had not betrayed Tosca.

Like the characters of Puccini's opera, people frequently face situations in which they must weigh the common interest against than their own self-interest, and they do so for a range of social dilemmas, from personal relationships, to environmental and political policies, to economic and social competition (e.g., Capraro, Jordan, & Rand, 2014; see Rand & Nowak, 2013 for a review). Suppose you and another person, located in a different room, must make a decision – to cooperate or to defect – without knowing what the other will decide. Your outcome depends on your choice and on the other player's choice. If one of you cooperates and the other does not, the one who cooperates gets nothing and the one who does not cooperate gets the highest payoff (e.g., 5 chocolates); if both of you cooperate, you both receive an intermediate payoff (e.g., 3 chocolates); if both of you defect, you both receive a lower payoff (e.g., 1 chocolate). What would you choose, to cooperate or to defect? The

Prisoner's dilemma, and variations on it, is the most extensively studied game in the social sciences perhaps because it pits an individual interest against the collective interest (e.g., Embrey, Frechette, & Yuksel, 2017; see also Van Lange, Joireman, Parks, & Van Dijk, 2013; Sell & Reese 2014). For the individual, the rational choice is to defect – you will be better off than if you had chosen to cooperate, regardless of what the other player chooses – but, for the collective, the rational choice is to cooperate – you will both be better off than if you had both chosen to defect (for more details, see the Method section). In fact, most people chose to cooperate when they take part in the Prisoner's dilemma (Flood, 1952; Sanfey, 2007).

The Prisoner's dilemma implicates not only rationality, but also morality. It requires judgments about sacrificing some of your benefits for the sake of others, at least for those individuals who take into consideration the benefits that others might receive, and, at the end, pits individual selfishness against collective altruism. Moral decisions to carry out a self-sacrificial action for the benefit of others are difficult to make but they are encouraged by the moral uplift experienced by hearing about other such acts (Algoe & Haidt, 2009; Schnall, Roper & Fessler, 2010), and by knowing the outcome turned out well – a "moral hindsight" effect – and imagining it had not (Byrne & Timmons, 2018; Timmons, Gubbins, Almeida & Byrne, 2019). Our aim is to have a better understanding of how people make the decision to cooperate or not in social dilemmas, by examining how they interpret the decision outcomes. We focus on how they imagine things could have turned out differently. These thoughts provide a unique and previously unexplored window onto the factors people consider important in making their decision.

1.1 "If only" thoughts

Suppose you decide to cooperate, but the other person does not; you fail to get anything whereas they maximize their outcome. You might imagine how things could have turned out differently, "if only…". Would you think "If only I had chosen to defect" or "If

only the other person had chosen to cooperate"? When people experience a bad outcome, they tend to create an alternative to reality by focusing on aspects of a situation that appear most salient to them (Kahneman & Tversky, 1982; for a review see Byrne, 2016).

Imagining your own choice had been different changes something within your control; imagining the other player's choice had been different changes something outside your control. Our primary objective is to establish whether people imagine how things could have turned out differently after a decision to cooperate or to defect, by changing their own choice within their control, or by changing the other person's choice outside their control. People tend to change something within an individual's control when they read hypothetical scenarios (e.g., Girotto, Legrenzi, & Rizzo, 2001; Mandel & Lehman, 1996; McCloy & Byrne, 2000; McEleney & Byrne, 2006; Roese & Olson, 1995) and when they recall episodes from their own lives (e.g., Davis, Lehman, Silver, Wortman, & Ellard, 1996; Hammell & Chan, 2016; Mandel, 2003; Markman & Miller, 2006; Roese, Smallman, & Epstude, 2017). Curiously, they do not do so when they experience a situation first-hand. For example, participants read about a person who must choose between two envelopes, one containing an easy and the other a difficult multiplication problem, and she must try to solve it in 30 seconds. When the person failed, *readers* focused on things within her control, "if only she had chosen the other envelope". But when they experienced the situation, *actors* focused on things outside their control, such as, "if only I had had more time" (e.g., Ferrante, Girotto, Stragà, & Walsh, 2013; Girotto, Ferrante, Pighin, & Gonzalez, 2007; Mercier, Rolison, Stragà, Ferrante, Walsh, & Girotto, 2017; Pighin, Byrne, Ferrante, Gonzalez, & Girotto, 2011). Interestingly, *observers*, who witnessed the actor's attempt to solve the chosen problem, focused on things outside the actor's control, just like actors - and unlike readers (Pighin et al., 2011). This difference between observers and readers indicates that the *actor-reader* effect in counterfactual thinking cannot be explained as an attribution effect (see

Elster, 1999; Jones & Nisbett, 1972). Indeed, an attributional explanation predicts that actors will prefer situation-based counterfactuals (e.g., to avoid self-blame) whereas observers and readers will show a preference for person-based counterfactuals, but the data show otherwise (see Pighin et al., 2011 for further discussion). The asymmetry between actors and readers is also reflected in the activation of different brain regions involved in the simulation of personal and impersonal counterfactual thoughts (De Brigard & Parikh, 2018; De Brigard, Spreng, Mitchell, & Schacter, 2015). Specifically, the generation of personal counterfactuals mainly recruits autobiographical details and therefore relies preferentially on episodic memory, whereas the generation of counterfactuals featuring unfamiliar characters (as in the case of hypothetical scenarios) mainly recruits stereotypical social knowledge and therefore relies preferentially on semantic memory (see also De Brigard & Parikh, 2018).

Counterfactuals that identify what an individual could have done differently can help people to work out how to prevent similar bad outcomes in the future (e.g., Roese & Epstude, 2017). But the *actor-reader* and *observer-reader* differences suggest that not all counterfactuals have a preparatory function (Girotto et al., 2007; Pighin et al., 2011), since they focus on things the individual could not have changed to prevent a similar outcome in the future. Instead, some counterfactuals are intended to explain or justify past events (e.g., Markman, Mizoguchi, & McMullen, 2008; McCrae, 2008).

Counterfactuals affect moral judgments by impacting blame and responsibility attributions (e.g., Byrne, 2005; Malle, Guglielmo, & Monroe, 2014; McCloy & Byrne, 2000; Migliore, Curcio, Mancini, & Cappa, 2014; Monroe & Malle, 2017; Phillips, Luguri, & Knobe, 2015). For instance, people are more lenient in their judgments of punishment for an individual who intended to harm someone and acted to do so but failed (e.g., failed to burn her partner's hand), compared to an individual with the same intentions and actions who succeeded (e.g., Lench, Domsky, Smallman, & Darbor, 2015). This leniency can be reduced

when people imagine that the perpetrator who failed had instead succeeded (Parkinson & Byrne, 2017a). Similarly, people judge that a morally good action (e.g., running into traffic to save a child who fell in front of an oncoming truck) should have been taken when they imagine that if the good action had not been taken, the outcome would have been worse (Byrne & Timmons, 2018). Hence people tend to think not only about what *could* have happened but what *should* have happened (e.g., Malle et al., 2014). They may access default possible actions by sampling those that are useful for future moral decisions (Phillips, Morris & Cushman, 2019). However, little is known about how people imagine how things could have turned out differently following acts of fairness or unfairness in social dilemmas. The only relevant study that examined the impact of retrospective thoughts on cooperative choices in a repeated social dilemma task showed that when participants are asked to think about how things could have been different, their generation of counterfactuals about how things could have been *better* encourages cooperation in future repetitions of the task, whereas counterfactuals about how things could have been *worse* inhibits it (Parks, Sanna, & Posye, 2003). Our question is different: we examined whether people create counterfactual thoughts that focus on actions within their control or outside their control after a decision turned out badly and they have to imagine how it could have turned out better, or after a decision turned out well and they have to imagine how it could have turned out worse.

We predict that when people act in a social dilemma, they will imagine that things would have turned out differently if their partner had acted differently rather than if they themselves had acted differently, that is, they will change things outside their control. We predict they will do so regardless of whether their decision turns out well or badly and so they imagine how it could have been worse or better. Our predictions derive from the observation that people create counterfactual thoughts that focus on things outside their control when they take part in an individual non-social game (as in the multiplication problem experiment

described above). Such a finding would have important consequences for understanding how people will react to the consequences of cooperation and defection in real situations. Moreover, we predict that people who read about a person who takes part in a social dilemma will focus instead on actions within the protagonist's control, that is, we expect to replicate an *actor-reader* effect in the context of social dilemmas.

1.2 The present experiments

To these aims, we employed three well-known one-shot strategic social dilemmas in which each player independently makes a single decision without knowledge of the other player's choice: The Prisoner's dilemma in Experiment 1, the Stag Hunt dilemma (also known as the Trust game) in Experiment 2, and the Chicken game (also known as the Hawk-Dove game) in Experiment 3. In the present study, all three dilemmas involve just two players who face a choice between two alternative options which are presented using the neutral labels "C" and "D" (that represent the typical "Cooperation" and "Defection" options, respectively). The payoffs associated with the combination of the players' choices (i.e., CC, CD, DC, and DD) determine the incentive structure of the dilemma, as shown in Figure 1. Listing the protagonist first and the other player second, the three 2 x 2 matrices (one for each of the social dilemmas) report the four possible outcomes. As the figure shows, the three dilemmas are symmetrical, that is the game "looks the same" to both players (i.e., if the order of the two players is reversed, the matrix does not change).

We investigated counterfactual thoughts in multiple social dilemmas to enhance the generalizability of our results, because, similarly to what happens in real-life situations, the reasons for cooperative or competitive behavior may change based on the specific payoff configuration. In particular, the use of the Prisoner's dilemma allowed us to study the focus of counterfactual thoughts in a social interaction where cooperation has a clear moral imperative: the decision to cooperate is motivated solely by the moral end of prioritizing the common

good pursued at the expense of individual benefits. The use of the Stag Hunt and the Chicken game allowed us to extend the findings to different payoff structures in which the decision to cooperate does not have a strict moral connotation. In both of these games, indeed, cooperation is both individually and collectively rational.

Our focus is on counterfactual thoughts and the use of these three different social dilemmas allowed us to investigate those created by participants who experienced (or read about) a wide range of situations. Importantly, the different structures made it possible to associate cooperative and competitive decisions with both positive and negative outcomes. Table 1 shows, for each social dilemma used, the complete set of combinations of factual choices and outcomes, as well as the main counterfactual modifications (of players' decisions) and corresponding counterfactual outcomes.

In order to control for possible confounds between the outcome of the counterfactual alternative (a better or a worse outcome) and the focus of the counterfactuals (a controllable or an uncontrollable action) in all three experiments, we examined the counterfactuals of participants who experienced only the best possible individual outcome (i.e., win 5 chocolates) or the worst (i.e., win 0 chocolates in Experiment 1 and 2, and win 1 chocolate in Experiment 3). As can be seen in Table 1, the best and the worst individual outcomes (in bold) are the only ones that allow the generation of all possible (i.e., controllable, uncontrollable, and mixed) counterfactual modifications. This is not the case for "intermediate" outcomes, which confound the controllability of the counterfactual modification with its direction (as better or worse). For example, consider the situation in which, in the Prisoner's dilemma, both players choose C and, consequently, the protagonist experiences an intermediate outcome (i.e., 3 chocolates). In such a situation, an *upward* counterfactual (i.e., a counterfactual about how things could have been better) is necessarily focused on a controllable modification (i.e., the actor/protagonist's choice is undone) whereas

a *downward* counterfactual (i.e., a counterfactual about how things could have been worse) is necessarily focused on uncontrollable (i.e., the other player's choice is undone) or mixed (i.e., both choices are undone) modifications. Note also that in the case of the Prisoner's dilemma, the best and the worst individual outcomes differ from the best and the worst collective outcomes, whereas in the Stag hunt and the Chicken game dilemmas the best and worst individual outcomes coincide exactly with the best and worst possible collective outcomes.

In all three experiments, we informed participants that they were paired with another participant, although the other partner was fictitious. We decided to use this deception based on careful consideration of standard protocols commonly accepted in the literature on social dilemmas and interactive games (e.g., Eimontaite, Schindler, De Marco, Duzzi, Venneri, & Goel, 2019; Muñoz-Reyes et al., 2020; Oren & Shamay-Tsoory, 2019). We determined that this deception was unlikely to cause pain or distress. Indeed, no emotional distress was reported when participants were fully debriefed immediately upon completion of their session: none indicated any concern or disappointment about the deception and/or asked for their data to be withdrawn. All participants confirmed that the manipulation was plausible, and fewer than 1% indicated that they suspected there was no real opponent. Our scientific rationale for using deception is that the only viable alternative, pairing each participant with another, has two serious drawbacks. The first is that it would necessitate the elimination of a large number of participants, since only the counterfactual thoughts of participants who experienced the best and the worst possible outcomes are examined. Hence, only two out of the four possible situations can be included in each experiment. For example, in Experiment 1, we can examine only the two situations in which different choices are made by the actor and the second player (i.e., the actor cooperates and the other player defects, or the actor defects and the other player cooperates), while we would have to eliminate the other two situations that can commonly arise (i.e., the actor and the other player make the same choice:

both cooperate, or both defect). So, for every pair of two real people, about half the data would have to be eliminated, with a waste of participants' time and experimental resources. The second drawback is that pairing real participants, even after elimination, would have resulted in uneven numbers of observations in each condition. For example, given the probability of cooperation versus defection, Experiment 2 would result in 3 times as many downward counterfactuals as upward ones. Such unevenness between conditions could undermine statistical reliability.

Finally, we used chocolate instead of money, an incentive that has been employed in the study of counterfactuals (e.g., Girotto et al. 2007; Pighin et al., 2011) and in other economic behavioral studies (e.g., Shen, Fishbach, & Hsee, 2015). Chocolates are an attractive incentive (for the majority of participants, students in their 20s), and a convenient one (in terms of storage, allocation, and administrative costs). The affective reactions measured at the end of the experiment showed that the use of chocolates was effective: Those who obtained the maximum payoff (5 chocolates) systematically expressed greater happiness than those who got fewer (e.g., 1) or no chocolates.

2. Experiment 1: Prisoner's dilemma

In Experiment 1, we employed the Prisoner's dilemma. As highlighted in Figure 1, in the Prisoner's dilemma there is only one pure *Nash equilibrium* – i.e., "[…] a steady state of the play of a strategic game in which each player holds the correct expectation about the other players' behavior and acts rationally" (Osborne & Rubinstein, 1994, p. 14). It occurs when both players defect: no matter what the other player does, defection always yields a higher payoff than cooperation (indeed, in case the other player cooperates, the outcome is 5 chocolates rather than 3; in case the other player defects, the outcome is 1 chocolate rather than 0). As mentioned earlier, cooperation instead prioritizes the collective outcome at the expense of individual gain. The *Pareto-optimal outcome* (i.e., the outcome that allows the

players as a whole to be better off and no individual player can improve their payoff without making at least one of the players worse off) occurs when both players cooperate.

In the Prisoner's dilemma, the worst individual outcome experienced by the actor/protagonist of the story follows when the actor/protagonist decides to cooperate and the other player decides to defect (i.e., CD in Table 1); in contrast, the best individual outcome experienced by the actor/protagonist of the story follows when the actor/protagonist decides to defect and the other player decides to cooperate (i.e., DC in Table 1).

2.1 Method

Participants were tested individually, and the materials were presented on a computer screen. Two groups of participants were involved in the study. One group (readers) read a story about a fictional protagonist who took part in the Prisoner's dilemma game, and one group (actors) took part in the game (paired with a fictitious player). Payoffs were constructed in terms of the number of chocolates (from 0 to 5) that the actor/protagonist could win. Actors received the payoff they won in their interaction with the other player, whereas readers were only informed about the payoff won by the protagonist.

2.1.1 Participants

The minimum sample size needed was calculated as 30 participants per group, based on an *a priori* power analysis, computed by G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009), to detect an effect size of at least .44 (the effect size of the *actor-reader* effect in Experiment 1 of Girotto et al., 2007), assuming $\alpha = .05$ and $1 - \beta = .95$. Accordingly, data collection was stopped when at least 32 participants were included in each experimental condition.

The participants were 152 students from the University of Trento, Italy, 79 women and 73 men with an average age of 23 years ($SD = 3.27$). They took part in the experiment in exchange for course credits. All participants gave written consent and were debriefed at the

end of the experiment, and the experiment was carried out in accordance with the APA Ethical Guidelines.

The participants were randomly assigned to either the reader (n = 76) or the actor (n = 76) groups. In the reader group, participants were further randomly assigned to either the upward counterfactual (n = 37) or downward counterfactual (n = 39) conditions. In the actor group, participants who decided to cooperate (58%) were assigned to the upward counterfactual condition by telling them the other player had defected (n = 44), while participants who decided to defect (42%) were assigned to the downward counterfactual condition by telling them that the other player had cooperated (n = 32).

2.1.2 Materials

The materials for *readers* consisted of the following story (presented to participants in their native Italian):

> Anna, a student from your program, participates in a game in one of the university labs.
> In this game, two players play simultaneously: Anna and a second person who is in another lab and with whom Anna cannot communicate.
>
> Both players have to choose to press either the "C" key or the "D" key. Their answers will then be automatically linked by a software program.
>
> The outcome of the game (and, therefore, the number of chocolates won by each player) will depend on Anna's choice and on the other player's choice, as indicated in the following table:
> (*The upper panel of Figure 1 was displayed on screen*).

For readers in the upward counterfactual condition the story ended badly:

> Anna chose to press C and the other player chose to press D.
> *Outcome of the game:*
> Anna wins 0 chocolates, the other player wins 5 chocolates.

Once informed about the outcome of the game, readers in this condition were asked the question:

> How could things have been better for Anna? Please begin your answer with 'if only'.

For readers in the downward counterfactual condition, the story ended well:

> Anna chose to press D and the other player chose to press C.
> *Outcome of the game:*
> Anna wins 5 chocolates, the other player wins 0 chocolates.

Readers in this group were asked the question:

> How could things have been worse for Anna? Please begin your answer with 'if only'[1].

Female readers read about a female protagonist (Anna) and male readers about a male protagonist (Luca).

The materials for *actors* were similar. They were invited to take part in a game and received the following instructions:

> In this game, two players play simultaneously: you and a second person who is in another lab and with whom you cannot communicate.
> Each of you has to choose to press either the "C" key or the "D" key. Your answers will then be automatically linked by a software program.
> The outcome of the game (and, therefore, the number of chocolates won by each of you) will depend on your choice and on the other player's choice, as indicated in the following table:
> *(The upper panel of Figure 1 was displayed on screen).*

Actors made their decision and the software linked their response with the decision of a (fictitious) second player. After a few seconds, they were informed about the outcome of the game, in the same way as described above. When things ended badly for the actor (i.e., upward counterfactual condition), they were asked to answer the question:

> How could things have been better for you? Please begin your answer with 'if only'.

When things ended well for the actor (i.e., downward counterfactual condition), they were asked to answer the question:

> How could things have been worse for you? Please begin your answer with 'if only'.

The framing of the question is standard in the literature (see Byrne, 2016), and its neutral and general nature was intended to avoid orienting participants to focus on actions of one or other

[1] Note that in Italian, unlike English, the 'If only' ('Se solo') construction makes sense and is commonplace for both upward and downward counterfactuals. We can confirm that the use of 'Se solo' for downward counterfactuals was not confusing for our participants since they were explicitly asked to indicate how things could have been worse for them and their modifications were clearly focused on downward outcomes (i.e., outcomes that were worse than the factual one).

players, but instead to allow an unconstrained generation of modifications (whether controllable or uncontrollable). Such a general question allows a participant to choose to reflect on their own choice and justify it by blaming the other player or, equally, to reflect on their own choice and the other player's choice and prepare to avoid the same outcome again by imagining a change of strategy.

The payoff table appeared again on the screen when participants were constructing their counterfactual alternative to reduce memory load. The main dependent measure was the nature of the counterfactual participants constructed in the sentence stem completion task.

2.1.3 Affective reaction measures

Three additional measures were included in order to assess participants' emotional states and the items used are reported in Table 2. The first item served as a manipulation check for participants' happiness for the actual outcome (*happiness* item), the second item was aimed at assessing participants' emotional reaction to their own [the protagonist's] choice (*regret* item), and the third item was aimed at assessing participants' emotional reaction to the other player's choice (*disappointment* item).

2.2 Results

The datasets for this experiment and the other two are available at OSF, doi: 10.17605/OSF.IO/RDCXJ. Participants' completions of the "if only" sentence stem were coded into four categories: (a) modifications that undid aspects that were under the protagonist's control (e.g., "If only I /the protagonist had chosen differently" or "If only I/the protagonist had decided to defect/cooperate") were coded as "controllable"[2]; (b) modifications that undid aspects that were outside the protagonist's control (e.g., "If only the

[2] In line with previous literature, in the present study, we have used the label "controllable" to indicate a counterfactual thought focused on an element of the past event that the actor/protagonist could have acted upon and the label "uncontrollable" to indicate a counterfactual thought focused on an element of the past event that the actor/protagonist could have not acted upon.

other player had chosen differently", or "If only the other player had decided to defect/cooperate", or "If only we/the players could have communicated") were coded as "uncontrollable"; (c) modifications that undid controllable *and* uncontrollable aspects (e.g., "If I/the protagonist had chosen to defect and the other player had cooperated") were coded as a "mixed"; (d) ambiguous modifications or answers that indicated that the participant did not correctly understand the payoff structure of the game were coded as "other". When participants provided more counterfactuals (or more modifications connected with an "or" within a single "if only" sentence stem), only the first modification was considered. Two independent judges coded the answers, and their agreement was above 98%. Disagreements were solved via discussion. The frequency (and corresponding percentages) of the four types of modifications provided by participants are reported in Table 3.

We analyzed the data with a *G*-test (Sokal & Rohlf, 1981), a log likelihood ratio test that approaches a χ^2 distribution. Williams' correction was applied prior to calculation of the *G* statistic to improve the exactness of the probability estimates (Williams, 1976). No answers fell into the category "other". The four experimental conditions significantly differed in the tendency to produce the remaining three sorts of modifications, $G^2(6, N = 152) = 83.68$, $p < .001$, $w = .75$. As Table 3 shows, actors' and readers' counterfactuals focused on different sorts of actions, both in the upward condition, $G^2(2, N = 81) = 36.1$, $p < .001$, $w = .66$, and in the downward condition, $G^2(2, N = 71) = 10.53$, $p = .005$, $w = .38$.

In the upward condition, when the outcome was bad and participants thought about how things could have been better, the majority of actors (71%; $p = .002$, binomial test) focused on actions outside their control (always the other player's choice) rather on actions within their control, whereas the majority of readers (92%; $p < .001$, binomial test) did the opposite: they focused on actions within the protagonist's control (always the protagonist's choice) rather than actions outside it, as Figure 2 shows. Actors' and readers' counterfactuals

also focused on different sorts of actions in the downward condition, when the outcome was good and participants thought about how things could have been worse: The majority of actors (85%; $p < .001$, binomial test) again focused on actions outside their control, whereas readers focused roughly equally on actions outside the protagonist's control (49%), within his/her control (20%), or a combination of the two (31%). When mixed modifications are excluded, however, actors' and readers' modifications did not differ in the downward condition, $G^2(1, N = 57) = 3.43$, $p = .06$, $w = .32$. Hence, in every condition except one, participants modified actions outside their control (i.e., the other player's decision), as Figure 2 shows. The exception is the reader condition for upward counterfactuals, in which participants modified actions within the protagonist's control (i.e., the protagonist's own decision to cooperate).

Mean judgments of the three affective reaction measures are reported in Table 4. A multivariate analysis of variance (MANOVA) was carried out to examine the effect of the two independent variables (counterfactual: upward vs. downward; role: actor vs. reader) on the three emotion measures. In order to decompose interaction effects, we computed post-hoc tests with a Bonferroni corrected α of $p < .0127$ for the four pairwise comparisons. For happiness, there was a main effect of counterfactual, $F(1,148) = 280.2$, $p < .001$, $\eta^2 = .65$, a main effect of role, $F(1,148) = 7.59$, $p = .007$, $\eta^2 = .05$, and an interaction between the two, $F(1,148) = 7.54$, $p = .007$, $\eta^2 = .05$. Post-hoc comparisons showed that there was a significant difference in happiness judgments between the upward and downward condition both for readers, $t(74) = 14.76$, $p < .001$, $d = 3.39$, and actors, $t(74) = 9.30$, $p < .001$, $d = 2.16$. Readers' judgments of happiness, however, were higher than actors' ones in the upward condition, $t(79) = 4.69$, $p = .005$, $d = 1.05$, but not in the downward, $t(69) = .005$, $p = .095$.

Also for regret, there was a main effect of counterfactual, $F(1,148) = 24.67$, $p < .001$, $\eta^2 = .14$, a main effect of role, $F(1,148) = 19.75$, $p < .001$, $\eta^2 = .12$, and a significant

interaction between the two variables, $F(1,148) = 6.29$, $p < .013$, $\eta^2 = .04$. Readers' judgments of regret were higher than actors' ones in the upward condition, $t(79) = 5.08$, $p < .001$, $d = 1.13$, whereas there was no difference in the downward condition, $t(69) = 1.34$, $p = .188$. A significant difference in regret judgments between the upward and the downward condition was found for readers, $t(74) = 5.34$, $p < .001$, $d = 1.23$, but not for actors, $t(74) = 1.72$, $p = .089$.

For disappointment, there was a main effect of counterfactual, $F(1,148) = 153.35$, $p < .001$, $\eta^2 = .51$, as judgments of disappointment were higher in the upward condition than the downward one, but no main effect of role, $F(1,148) = 1.31$, $p = .254$, and no interaction, $F(1,148) = .39$, $p = .532$.

2.3 Post-hoc control condition

We introduced a post-hoc control condition for Experiment 1 to rule out a potential confound. The difference between actors and readers could have arisen because the former but not the latter received the payoff (the chocolates) corresponding to the outcome. More specifically, one might wonder whether in the upward condition (when the worse outcome was obtained and participants had to think about how things could have been better) actors have focused on actions outside their control because the corresponding counterfactual outcome would have represented a maximization of their payoff (3 chocolates rather than one) while readers focused on actions within the protagonist's control, because they lacked this motivation.

2.3.1 Participants

A new sample of 33 university students, consisting of 18 women and 15 men, with an average age of 22 years ($SD = 2.4$) was tested in the upward counterfactual reader condition, with the only difference that they received the same number of chocolates won by the

protagonist of the story, and were informed of this before being told the protagonist's and the other player's choices.

2.3.2 Results

Participants in the post-hoc control condition focused on controllable actions, just as readers in the upward counterfactual condition in Experiment 1 did, as Table 3 shows, and there was no difference between the two groups of readers, $G^2(2, N = 70) = 3.21$, $p = .20$. The readers in the post-hoc control condition created different counterfactuals from the actors in Experiment 1, $G^2(2, N = 77) = 20$, $p < .001$, $w = .51$. Thus, the results show that the responses of the readers in this post-hoc condition did not differ from those of standard readers (who could not receive any payoff), but did differ from those generated by actors. This indicates that the difference between actors' and readers' upward counterfactual thoughts observed in Experiment 1 cannot simply be ascribed to their different motivations at maximizing their payoffs.

3. Experiment 2: The Stag Hunt

In Experiment 2, we employed the Stag Hunt dilemma. In contrast with the Prisoner's dilemma, in the Stag Hunt there are two Nash equilibria (see Figure 1). One is Pareto-optimal and occurs when both players cooperate. This equilibrium is also referred to as *payoff dominant* since it corresponds to the highest payoffs for both players compared to the other equilibrium. The other equilibrium is Pareto-dominated and occurs when both players defect. This equilibrium is *risk dominant*, since it is less risky compared to the other equilibrium. Therefore, in this dilemma, cooperation is not motivated by moral ends, since it is both individually and collectively rational.

As in the Prisoner's dilemma, the worst possible individual outcome experienced by the actor/protagonist of the story follows when the actor/protagonist decides to cooperate while the other player decides to defect (i.e., CD in Table 1). In the Stag Hunt dilemma,

however, this outcome is also one of the two worst possible collective outcomes. Unlike in the Prisoner's dilemma, the best individual outcome (which corresponds also to the best possible collective outcome in this case) follows from mutual cooperation (i.e., CC in Table 1). Hence, the use of the Stag Hunt allowed us to investigate *downward* counterfactual thoughts about how things could have been worse, generated by participants whose cooperative decisions (rather than competitive decisions, as in the Prisoner's dilemma) led to a positive outcome. The Stag Hunt dilemma also allows us, more generally, to investigate counterfactual thoughts after outcomes that have a positive or a negative valence both individually and collectively.

3.1 Method

3.1.1 Participants

Participants in Experiment 2 were 156 students, 112 women and 44 men with an average age of 22 years (*SD* = 2.35). They were randomly assigned to the reader (*n* = 69) or actor (*n* = 87) group. In the reader group, participants were further randomly assigned either to the upward counterfactual (n = 34) or downward counterfactual (n = 35) conditions. In the actor group, most participants decided to cooperate (*n* = 65, 75%). They were randomly assigned to either the upward counterfactual condition by telling them the other player had defected (*n* = 33), or the downward counterfactual condition by telling them the other player had cooperated (*n* = 32). Participants who decided to defect (*n* = 22) could not experience the worst or best individual outcomes (they always received 3 chocolates, regardless of the other player's choice), and therefore we excluded them from the analyses.

3.1.2 Materials

The materials were similar to the previous experiment but the payoff matrix was the one depicted in the middle panel of Figure 1. For readers in the upward counterfactual condition the story ended as follows:

> Anna chose to press C and the other player chose to press D.
> *Outcome of the game:*
> Anna wins 0 chocolates, the other player wins 3 chocolates.

For readers in the downward counterfactual condition, the story ended as follows:

> Anna chose to press C and the other player chose to press C.
> *Outcome of the game:*
> Anna wins 5 chocolates, the other player wins 5 chocolates.

As in Experiment 1, the decisions and experienced outcomes of the actors mirrored those of the protagonists that readers followed.

Upward and downward counterfactual thoughts were elicited as in Experiment 1 for both readers and actors, and the same affective reaction measures were assessed.

3.2 Results

The counterfactual sentence stem "if only" completions were coded as in Experiment 1, as Table 3 shows, and the same set of statistical analyses was performed. The four experimental conditions differed significantly in the tendency to produce the four sorts of modifications, $G^2(9, N = 134) = 35.10$, $p < .001$, $w = .53$. Actors and readers' modifications were different in the upward condition, $G^2(3, N = 67) = 12.06$, $p = .007$, $w = .47$, but not in the downward condition, $G^2(3, N = 67) = 7.26$, $p = .064$, $w = .32$.

As in Experiment 1, in the upward condition, when the outcome was bad and participants thought about how things could have been better, the majority of actors (73%, $p = .004$, binomial test) focused on actions outside their control (all except one focused on the other player's choice) whereas the majority of readers (68%, $p = .006$, binomial test) did the opposite: they focused on actions within the protagonist's control (always the protagonist's choice). In the downward condition, when the outcome was good and participants thought about how things could have been worse, the majority of actors (75%, $p < .001$, binomial test) and readers (60%, $p = .02$, binomial test) focused on actions outside their/the protagonist's control (always the other player's choice) rather than actions within their/the protagonist's

control. Hence, once again, in all conditions but one, participants modified actions outside their control (the other player's decision), as Figure 2 shows. The exception is again the reader condition for upward counterfactuals, in which participants modified actions within the protagonist's control (the protagonist's own decision to cooperate).

As in Experiment 1, a MANOVA was carried out to examine the effect of the two independent variables (counterfactual: upward vs. downward; and role: actor vs. reader) on the three emotion measures. For happiness, there was a main effect of counterfactual, $F(1,130) = 346.4$, $p < .001$, $\eta^2 = .73$, as judgments of happiness were higher in the downward condition than the upward one, but no main effect of role, $F(1,130) = 1.16$, $p = .284$, and no interaction, $F(1,130) = 2.56$, $p = .617$.

For regret, there was a main effect of counterfactual, $F(1,130) = 58.14$, $p < .001$, $\eta^2 = .31$, a main effect of role, $F(1,130) = 18.5$, $p < .001$, $\eta^2 = .13$, and a significant interaction between the two factors, $F(1,130) = 7.03$, $p < .009$, $\eta^2 = .05$. Readers' judgments of regret were higher than actors' ones in the upward counterfactual condition, $t(65) = 5.07$, $p < .001$, $d = 1.23$, whereas there was no difference between readers and actors in the downward condition, $t(65) = 1.13$, $p = .261$. Both readers' and actors' judgments of regret were significantly higher in the upward than in the downward condition, $t(67) = 7.47$, $p < .001$, $d = 1.8$, and $t(67) = 3.42$, $p = .001$, $d = .85$ respectively.

For disappointment, there was a main effect of counterfactual, $F(1,130) = 279.98$, $p < .001$, $\eta^2 = .68$, as judgments of disappointment were higher in the upward condition than the downward one, but no main effect of role, $F(1,130) = .413$, $p = .521$, and no interaction, $F(1,130) = 3.62$, $p = .06$.

4. Experiment 3: The Chicken game

In Experiment 3, we employed the Chicken game. As in the case of the Stag Hunt, in the Chicken game there is no dominant strategy and there are two pure Nash equilibria. These

correspond to the Pareto-optimal outcomes and occur when the two players play opposite strategies (i.e., one cooperates and the other defects, see the third panel of Figure 1). This means that players benefit from playing different strategies both at the individual and the collective level. Of course, for each player the equilibrium in which they defect and the other player cooperates is preferable.

Importantly, for the purposes of this study, unlike the Prisoner's and Stag Hunt dilemmas, in the Chicken game, the worst possible individual outcome experienced by the actor/protagonist of the story (which is also the worst collective outcome) follows from mutual defection (i.e., DD in Table 1). Hence this game allowed us to investigate *upward* counterfactual thoughts about how things could have been better, generated by participants whose competitive decision (rather than cooperative decision, as in the Prisoner's and Stag hunt dilemmas) led to a negative outcome. As in the Prisoner's dilemma (but unlike the Stag Hunt), the best possible individual outcome for the actor/protagonist of the story (which is also one of the two best collective outcomes) follows when the actor/protagonist decides to defect while the other player decides to cooperate (i.e., DC in Table 1).

4.1. Method

4.1.1 Participants

Participants in Experiment 3 were 243 students, 184 women and 59 men with an average age of 21 years (SD = 3.37). They were randomly assigned to the reader (n = 64) or actor (n = 179) group. In the reader group, participants were further randomly assigned either to the upward counterfactual (n = 32) or downward counterfactual (n = 32) conditions. In the actor group, only 68 (38%) participants decided to defect. They were randomly assigned to either the upward counterfactual condition by telling them the other player had also defected (n = 33), or the downward counterfactual condition by telling them the other player had cooperated (n = 35). Participants who decided to cooperate (n = 111, 62%) could not

experience the worst or best individual outcomes (i.e., they always received 2 chocolates, regardless of the other player's choice), and therefore they were excluded from the analyses.

4.1.2 Materials

The materials were the same as the previous experiments but the payoff matrix was the one depicted in the lower panel of Figure 1. For readers in the upward counterfactual condition the story ended as follows:

> Anna chose to press D and the other player chose to press D.
> *Outcome of the game:*
> Anna wins 1 chocolate, the other player wins 1 chocolate.

For readers in the downward counterfactual condition, the story ended as follows:

> Anna chose to press D and the other player chose to press C.
> *Outcome of the game:*
> Anna wins 5 chocolates, the other player wins 0 chocolates.

Yet again, the decisions and experienced outcomes of the actors mirrored those of the protagonists that readers followed.

Upward and downward counterfactual thoughts were elicited as in Experiment 1 and 2 for both readers and actors, and the same affective reaction measures were assessed.

4.2 Results

The counterfactual sentence stem "if only" completions were coded as in Experiment 1 and 2, and the same set of statistical analysis was performed. The four experimental conditions differed significantly in the tendency to produce the four sorts of modifications, $G^2(9, N = 132) = 38.46$, $p < .001$, $w = .55$. Actors and readers' modifications were different in the upward condition, $G^2(3, N = 65) = 17.91$, $p < .001$, $w = .55$, but not the downward condition, $G^2(3, N = 67) = 1.40$, $p = .706$.

As in the previous two experiments, in the upward condition, when the outcome was bad and participants thought about how things could have been better, the majority of actors (76%, $p < .001$, binomial test) focused on actions outside their control (always the other

player's choice) whereas the majority of readers (66%, $p = .008$, binomial test) did the opposite: they focused on actions within the protagonist's control (always the protagonist's choice). In the downward condition, when the outcome was good and participants thought about how things could have been worse, the majority of both actors (66%, $p < .001$, binomial test) and readers (72%, $p < .001$, binomial test) focused on actions outside their/the protagonist's control (always the other player's choice).

Finally, the MANOVA showed that for happiness there was a main effect of counterfactual, $F(1,128) = 178.7$, $p < .001$, $\eta^2 = .58$, as judgments of happiness were higher in the downward condition than in the upward one, and no main effect of role, $F(1,128) = .91$, $p = .342$. The analysis revealed a significant interaction between counterfactual and role, $F(1,128) = 9.23$, $p = .003$, $\eta^2 = .07$: in the upward condition, readers' judgments of happiness were lower than those of actors, $t(63) = 2.93$, $p = .005$, $d = .73$, but readers' and actors' judgments did not differ in the downward condition, $t(65) = -1.43$, $p = .158$.

For regret, there was a main effect of counterfactual, $F(1,128) = 25.09$, $p < .001$, $\eta^2 = .316$, as participants' judgments were higher in the upward than in the downward condition, a main effect of role, $F(1,128) = 8.05$, $p = .005$, $\eta^2 = .06$, as actors judgments were significantly lower than readers' ones, and no significant interaction between the two factors, $F(1,128) = 1.83$, $p = .178$.

For disappointment, there was a main effect of counterfactual, $F(1,128) = 152.69$, $p < .001$, $\eta^2 = .16$, as judgments of disappointment were higher in the upward condition than the downward one, but no main effect of role, $F(1,128) = 2.80$, $p = .097$, and no interaction, $F(1,128) = .247$, $p = .62$.

5. Pooled analysis across experiments

An overall logistic regression analysis across the three experiments was performed to ascertain the effects on controllable and uncontrollable modifications of the payoff structures

outlined in the Prisoner's dilemma, Stag Hunt, Chicken game, the counterfactual condition, the role, and the interaction between counterfactual condition and role. The three predictors were included in the logistic regression analysis as categorical independent variables. The logistic regression model was statistically significant, $\chi^2(5) = 116.9$, $p < .001$. The model explained 37% (Nagelkerke R^2) of the variance in counterfactual provided. As shown in Table 5, the counterfactual condition ($p < .001$), the role ($p = .015$), and the interaction between counterfactual condition and role ($p = .006$) significantly predicted the type of counterfactual generated, while the payoff structure did not ($p > .05$). Across the three experiments, the probability of producing counterfactual modifications that focused on an uncontrollable event was greater for actors than readers (OR = 2.77, 95% CI 1.22 to 6.27), and in the downward than in the upward condition (OR = .09, 95% CI .04 to .18). The interaction effect indicates that role (actor vs. reader) was a significant predictor of counterfactual type in the upward but not in the downward condition.

6. Discussion and conclusions

When participants cooperated in a social interaction that turned out badly because unbeknownst to them their partner betrayed them, they tended to imagine how things could have turned out better by wishing the other person had not defected, rather than by wishing that they themselves had not cooperated. Our first discovery is that people do not respond to betrayal by thinking about changing their own choice, instead they wish the other person had acted differently. This result occurs not only when cooperation is collectively rational as in the case of the Prisoner's dilemma (Experiment 1) but also when it is both collectively and individually rational as in the case of the Stag Hunt (Experiment 2). The very same result was obtained also when the social interaction turned out badly after participants' decision to defect, as in the case of the Chicken game (Experiment 3). In such a situation, participants tended to imagine how things could have turned out better by once again wishing the other

person had not defected, rather than by wishing that they themselves had made a different choice.

Strikingly, all three experiments also show that when participants read a story, they tended to imagine that things could have turned out better for the protagonist by undoing the protagonist's decision to cooperate (Experiments 1 and 2) or defect (Experiment 3). This second finding confirms and extends the *actor-reader* effect in upward counterfactuals to the domain of social interaction: actors who actually experience a social dilemma imagined an action outside their control, whereas readers who read about the social dilemma imagined an action within the protagonist's control. The results for the Prisoner's dilemma, for which cooperation has a moral connotation, imply we sometimes think about the morality of other people's choices quite differently from the way we think about our own (see also Parkinson & Byrne, 2017b; Goodwin, 2015). The results for the Stag Hunt and Chicken game, for which cooperation is both individually and collectively rational, imply we also sometimes think about other people's rationality differently from the way we think about our own.

When the social interaction turned out well, no difference was observed between actors and readers' counterfactuals: participants tended to think that things could have turned out worse by imagining that the other person decided differently, rather than by imagining that they themselves or the protagonist of the story had made a different choice. This effect occurred not only when participants betrayed their partners and the social interaction turned out to benefit the participant, as in case of the Prisoner's dilemma (Experiment 1), but also when the social interaction turned out to benefit both players, as in the case of mutual cooperation in the Stag Hunt (Experiment 2), and when participants opted for different strategies as in the case of the Chicken game (Experiment 3). Our third result implies therefore that people do not always question their choices when the outcome turns out well, even when such choices have a moral connotation (see also Timmons & Byrne, 2018; Merritt,

Effron, & Monin, 2010). This inclination towards "absolution" occurred not only for choices that participants made themselves, but also for choices made by a fictional protagonist. Hence, as Experiment 1 showed, when an individual's betrayal of others turned out well for them, both readers and actors imagined things could have turned out differently by focusing on actions outside their control, that is, the other player's choice.

Participants' affective reactions were consistent across the three experiments. As expected, positive outcomes elicited greater happiness evaluations than negative ones, regardless of whether participants actually experienced the social interaction or read about it. Overall, readers tended to evaluate more negatively bad outcomes than actors, but this difference was statistically significant only in the first experiment. In all three experiments, participants' responses about regret were in line with their upward counterfactual modifications: actors reported a lower level of regret than the one expected by readers, and, accordingly, generated a lower number of controllable modifications. This could suggest that actors readily rationalize negative outcomes by avoiding regret for their own choices and by moving away from possible self-blame (see also Gilbert, Morewedge, Risen, & Wilson, 2007; Zeelenberg & Pieters, 2007) better than expected. Actors' and readers' disappointment in the choice of the other player did not differ but, unsurprisingly, was higher when the social interaction went badly, rather than well.

As the descriptive statistics show (see Table 3), the pattern of results is robust and clear-cut, and the effect sizes are large. Our findings indicate that, when a social interaction goes awry, actors, unlike readers, create counterfactuals that focus on actions outside their control. These results extend previous findings by showing an actor-reader effect in upward counterfactual thinking in the domain of a social interaction (Ferrante et al., 2013; Girotto et al., 2007; Mercier et al., 2017; Pighin et al., 2011). What gives rise to the reader-actor effect in upward counterfactual thinking? One potential explanation is that this effect arises from a

difference in the *goals* that drive the generation of counterfactual thoughts in actors and readers. Counterfactual thinking is recognized as the explicit manifestation of an underlying process of causal analysis through which a past event is scrutinized to uncover the chain of causes and effects that, potentially, could have brought about a different outcome from the one that happened (e.g., Roese & Epstude, 2017). Since this causal analysis is viewed as goal oriented, the discrepancy in counterfactual thoughts produced by actors and readers could reflect a difference in their goals. Within this perspective, the actor-reader effect could be interpreted as a special case of the well-known actor-observer attributional effect (Jones & Nisbett, 1972), such that actors prefer situation-based counterfactuals (say, to avoid self-blame) whereas readers show a preference for person-based counterfactuals. However, previous experiments have revealed that observers (i.e., individuals who witnessed the actor's performance) tend to generate counterfactuals focused on things outside the actor's control, just as actors do, but unlike readers (Pighin et al., 2011). Such an observer-reader difference is incompatible with the explanation that the actor-reader effect depends on goals.

A second potential explanation relies on the well-known *temporal order* effect in counterfactual thinking. Previous studies have found that participants tend to modify the most recent event in a temporal sequence of two independent events (Miller & Gunasegaram, 1990; Byrne et al., 2000; Segura, Fernandez-Barrocal, & Byrne, 2002; Walsh & Byrne, 2004). In our experiments, players' choices were intended to be interpreted as synchronous, but actors experienced them sequentially: they made their own choice and only later were they informed about the other player's choice. Hence, actors' modifications of the other player's choice are consistent with a temporal order effect. However, this explanation cannot explain readers' choices in our experiments. Indeed, readers systematically altered the protagonist's choice, which was the first event presented in a sequence of two events: readers read about the two events in a single sentence, with the protagonist's choice always in the first position. Readers'

tendency to modify the first element in the sequence has also been observed in previous studies of the actor-reader effect (e.g., Girotto et al., 2007; Pighin et al., 2011). Accordingly, the temporal order explanation of the actor-reader effect in our experiments is incompatible with the asymmetry in modifications as a result of a participant's role.

A third potential explanation is that the differences arise because of the impact of the *different information* available to actors and readers. It is plausible to assume that the participants in the two roles construct different mental representations of the situation given that they have access to different semantic and pragmatic knowledge. Actors, for example, have direct access to the reasons for their own (cooperative or competitive) behavior and know what outcome they expect to achieve in the interaction. In contrast, readers have little access to the two players' reasons for their actions and their expectations of outcomes. The difference in the explicit representation of information about the facts of the situation could then activate different information about the alternative counterfactual possibilities (see Byrne, 2016). Previous studies have shown that when actors have an impoverished experience of a situation (e.g., when they simulate an unsuccessful attempt to solve a problem rather than actually experience it) their tendency to construct counterfactuals that focus on features of the situation is reduced (Girotto et al., 2007, Exp. 5). Hence, the difference in the information included in actors and readers mental representations of the factual situation may explain the difference in their counterfactual thoughts about it.

The three explanations we have considered are not mutually exclusive, and are worth exploring in future studies. Experiments that systematically manipulate participants' representation of a situation, e.g. by manipulating the quantity and quality of information available to actors and readers, may provide further insights into the nature of the actor-reader effect.

The present findings also question whether actors' counterfactuals have a preparatory function for future social interactions. Counterfactuals that focus on actions within one's own control enable people to work out how to avoid unwanted outcomes in the future (Roese & Epstude, 2017). Actors' focus on uncontrollable actions has been typically interpreted as indicating that counterfactuals do not always have a preparatory purpose (e.g., Girotto et al., 2007; Mercier et al., 2017 but see also Ferrante, Stragà, Walsh, & Girotto, 2016; Hammell & Chan, 2016). Of course, another function of counterfactuals is to enable people to explain or justify past actions (e.g., Markman et al., 2008; McCrae, 2008), and it may best be served by focusing on actions outside one's control. Nonetheless, in social dilemmas, counterfactuals focusing on uncontrollable actions could in fact enable individuals to prepare for future social interactions. Focusing on the other player's decision after a bad outcome could allow actors to reappraise their expectations of the chances of encountering cooperating versus non-cooperating players, which could lead them to revise their strategy in the future (e.g., to defect rather than to cooperate). Future studies that examine the effects of counterfactuals in repeated social dilemmas could provide insight about their preparatory role.

In the present study, we focused on social interactions in which the actor or protagonist in a story experienced the best or the worst possible individual outcome (which also corresponded to the best and worst possible collective outcome in Experiments 2 and 3). This focus was motivated by the attempt to avoid possible confounds between the direction (upward vs. downward) and the focus (controllable vs. uncontrollable) of counterfactual thoughts. In future research it would be fruitful to investigate spontaneous (rather than elicited) counterfactual thoughts, as well as to examine intermediate situations, in which participants can construct either type of counterfactual, to imagine a better or worse outcome. For example, in the Prisoner's dilemma when both players defect and it turns out fairly well for them, would participants imagine how things could have been better (if their partner had

not defected, or neither of them had), or worst (if one them had cooperated)? The answer to this question may help clarify how people develop their cooperative behavior in social situations.

Author Contributions

All authors developed the study concept and contributed to the study design. Data collection and data analysis were performed by the first author. All authors collaborated in the interpretation of the results. All authors drafted the manuscript and approved the final version of the manuscript for submission.

Open Practices

De-identified data for the three experiments are publicly available at doi 10.17605/OSF.IO/RDCXJ

Conflicts of interest

The authors declared no conflicts of interest with respect to the authorship or the publication of this article.

References

Algoe, S. B., & Haidt, J. (2009). Witnessing excellence in action: The 'other-praising'emotions of elevation, gratitude, and admiration. *The Journal of Positive Psychology*, *4*(2), 105-127.

Byrne, R. M. (2005). *The rational imagination: How people create alternatives to reality*. Cambridge: MIT Press.

Byrne, R. M. (2016). Counterfactual thought. *Annual Review of Psychology*, *67*, 135–157.

Byrne, R. M., Segura, S., Culhane, R., Tasso, A., & Berrocal, P. (2000). The temporality effect in counterfactual thinking about what might have been. *Memory & Cognition*, *28*(2), 264-281.

Byrne, R.M. & Timmons, S. (2018). Moral hindsight for good actions and the effects of imagined alternatives to reality. *Cognition*. 178, 82-91.

Capraro, V., Jordan, J. J., & Rand, D. G. (2014). Heuristics guide the implementation of social preferences in one-shot Prisoner's Dilemma experiments. *Scientific Reports*, *4*, 6790.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, *31*, 169–193.

Davis, C. G., Lehman, D. R., Silver, R. C., Wortman, C. B., & Ellard, J. H. (1996). Self blame following a traumatic event: The role of perceived avoidability. *Personality and Social Psychology Bulletin*, *22*, 557–567.

De Brigard, F., Spreng, R. N., Mitchell, J. P., & Schacter, D. L. (2015). Neural activity associated with self, other, and object-based counterfactual thinking. *Neuroimage*, *109*, 12–26.

De Brigard, F., & Parikh, N. (2018). Episodic Counterfactual Thinking. *Current Directions in Psychological Science*, *28*, 59–66.

Eimontaite, I., Schindler, I., De Marco, M., Duzzi, D., Venneri, A., & Goel, V. (2019). Left amygdala and putamen activation modulate emotion driven decisions in the iterated Prisoner's Dilemma game. *Frontiers in Neuroscience*, *13*, 741.

Elster, J. (1999). *Alchemies of mind*. Cambridge, UK: Cambridge University Press.

Embrey, M., Fréchette, G. R., & Yuksel, S. (2017). Cooperation in the finitely repeated prisoner's dilemma. *The Quarterly Journal of Economics*, *133*(1), 509-551.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 11491160.

Ferrante, D., Girotto, V., Stragà, M., & Walsh, C. (2013). Improving the past and the future: A temporal asymmetry in hypothetical thinking. *Journal of Experimental Psychology: General*, *143*, 23–27.

Ferrante, D., Stragà, M., Walsh, C., & Girotto, V. (2016). What could I have done or what can I do? The effect of counterfactual and prefactual thinking on predictions and intentions. In Paper presented at the EASP Small Group Meeting. "Counterfactual thinking in causality, emotion, communication, and behavior", Aix-en-Provence, France.

Flood, M.M. (1952). Some experimental games. Research Memorandum RM-789-1,The RAND Corporation, Santa-Monica, CA, USA.

Jones, E. E., & Nisbett. R. E. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. E. Kanhouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 79–94). Morristown, NJ: General Learning Press.

Gilbert, D.T., Morewedge, C.K., Risen, J.L., & Wilson, T.D. (2004). Looking forward to looking backward: The misprediction of regret. *Psychological Science*, *15*, 346–

350.

Girotto, V., Legrenzi, P., & Rizzo, A. (1991). Counterfactual thinking: The role of events controllability. *Acta Psychologica*, *78*, 111–133.

Girotto, V., Ferrante, D., Pighin, S., & Gonzalez, M. (2007). Postdecisional counterfactual thinking by actors and readers. *Psychological Science*, *18*, 510–515.

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science*, *24*(1), 38-44.

Hammell, C., & Chan, A. Y. C. (2016). Improving physical task performance with counterfactual and prefactual thinking. *PLoS One*, *11*, e0168181.

Kahneman, D., & Tversky, A. (1982b). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–211). Cambridge, England: Cambridge University Press.

Lench, H. C., Domsky, D., Smallman, R., & Darbor, K. E. (2015). Beliefs in moral luck: When and why blame hinges on luck. *British Journal of Psychology*, *106*(2), 272-287.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186.

Mandel, D. R. (2003). Counterfactuals, emotions, and context. *Cognition and Emotion*, *17*, 139–159.

Mandel, D. R., & Lehman, D. R. (1996). Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality and Social Psychology*, *71*, 450–463.

Markman, K. D., & Miller, A. K. (2006). Depression, control, and counterfactual thinking: Functional for whom? *Journal of Social and Clinical Psychology*, *25*, 210–227.

Markman, K. D., Mizoguchi, N., & McMullen, M. N. (2008). “It would have been worse under Saddam”: Implications of counterfactual thinking for beliefs regarding the

ethical treatment of prisoners of war. *Journal of Experimental Social Psychology*, 44, 650–654.

McCloy, R., & Byrne, R. M. J. (2000). Counterfactual thinking about controllable events. *Memory & Cognition*, *28*, 1071–1078.

McCrea, S.M. (2008). Self-handicapping, excuse making, and counterfactual thinking: consequences for self esteem and future motivation. *Journal of Personality and Social Psychology,* 95:274–92

McEleney, A., & Byrne, R. M. J. (2006). Spontaneous causal and counterfactual thoughts. *Thinking & Reasoning*, *12*, 235–255.

Mercier, H., Rolison, J. J., Stragà, M., Ferrante, D., Walsh, C. R., & Girotto, V. (2017). Questioning the preparatory function of counterfactual thinking. *Memory & Cognition*, *45*, 261–269.

Migliore, S., Curcio, G., Mancini, F., & Cappa, S.F. (2014). Counterfactual thinking in moral judgment: an experimental study. *Frontiers in Psychology*, 5:451.

Miller, D. T., & Gunasegaram, S. (1990). Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of personality and social psychology*, *59*(6), 1111.

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and Personality Psychology Compass*, *4*(5), 344-357.

Monroe, A. E., & Malle, B. F. (2017). Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General*, 146(1), 123.

Muñoz-Reyes, J. A., Polo, P., Valenzuela, N., et al. (2020). the Male Warrior Hypothesis: testosterone-related cooperation and Aggression in the context of Intergroup Conflict. *Scientific Reports*, *10*(1), 1-12.

Oren, C., & Shamay-Tsoory, S. G. (2019). Women's fertility cues affect cooperative behavior: Evidence for the role of the human putative chemosignal estratetraenol. *Psychoneuroendocrinology*, *101*, 50-59.

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. Cambridge: MIT press.

Parkinson, M. & Byrne, R.M.J. (2017a). Counterfactual and semifactual thoughts in moral judgments about failed attempts to harm. *Thinking & Reasoning.* 23, 4, 409-448.

Parkinson, M. & Byrne, R.M.J. (2017b). Moral judgments of risky choices: a moral echoing effect. *Judgment & Decision Making,* 12, 3, 236-252.

Parks, C. D., Sanna, L. J., & Posey, D. C. (2003). Retrospection in social dilemmas: How thinking about the past affects future cooperation. *Journal of Personality and Social Psychology*, *84*, 988–996.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30-42.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in Cognitive Sciences*, *23,* 1026–1040.

Pighin, S., Byrne, R. M., Ferrante, D., Gonzalez, M., & Girotto, V. (2011). Counterfactual thoughts by experienced, observed and narrated events. *Thinking & Reasoning*, *17*, 197–211.

Poundstone, W. (1992). *Prisoner's Dilemma: John Von Neumann, Game Theory and the Puzzle of the Bomb*. New York: Doubleday.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413–425.

Roese, N. J., & Epstude, K. (2017). The functional theory of counterfactual thinking:

New evidence, new challenges, new insights. *Advances in Experimental Social Psychology*, *56*, 1–79.

Roese, N. J., & Olson, J. M. (1995). Outcome controllability and counterfactual thinking. *Personality and Social Psychology Bulletin*, *21*, 620-628.

Roese, N. J., Smallman, R., & Epstude, K. (2017). Do episodic counterfactual thoughts focus on controllable action?: The role of self-initiation. *Journal of Experimental Social Psychology*, *73*, 14–23.

Sanfey, A. G. (2007). Social decision-making: insights from game theory and neuroscience. *Science*, *318*(5850), 598-602.

Schnall, S., Roper, J., & Fessler, D. M. (2010). Elevation leads to altruistic behavior. *Psychological Science*, *21*(3), 315-320.

Segura, S., Fernandez-Berrocal, P., & Byrne, R. M. (2002). Temporal and causal order effects in thinking about what might have been. *The Quarterly Journal of Experimental Psychology*, *55*(4), 1295-1305.

Sell, J., & Reese, B. (2014). Social Dilemma Experiments in Sociology, Psychology, Political Science, and Economics. In M. Webster, & J. Sell (Eds.), *Laboratory Experiments in the Social Sciences* (pp. 225-245). New York: Elsevier.

Shen, L., Fishbach, A., & Hsee, C. K. (2015). The motivating-uncertainty effect: Uncertainty increases resource investment in the process of reward pursuit. *Journal of Consumer Research*, *41*(5), 1301-1315.

Sokal, R. R., & Rohlf, F. J. (1995). *Biometry*. New York: Freeman.

Timmons, S. & Byrne, R.M.J. (2018). Moral Fatigue: The Effects of Cognitive Fatigue on Moral Reasoning. *Quarterly Journal of Experimental Psychology,* 72, 4, 943-954.

Timmons, S., Gubbins, E., Almeida, T., & Byrne, R. M. (2019). Imagined alternatives to episodic memories of morally good acts. *The Journal of Positive Psychology*, doi.org/10.1080/17439760.2019.1689410.

Van Lange, P. A., Joireman, J., Parks, C. D., & Van Dijk, E. (2013). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, 120(2), 125-141.

Walsh, C. R., & Byrne, R. M. (2004). Counterfactual thinking: The temporal order effect. *Memory & Cognition*, *32*(3), 369-378.

Williams, D. A. (1976). Improved likelihood ratio tests for complete contingency tables. *Biometrika*, *63*, 33–37.

Zeelenberg, M., & Pieters, R. (2007). A theory of regret regulation 1.0. *Journal of Consumer Psychology*, *17*, 3–18.

Table 1: The complete set of combinations of factual choices, factual outcomes, counterfactual modifications (of players' decisions), and counterfactual outcomes in the social dilemmas used in the three experiments. In bold are the cells that represent the best and the worst possible individual outcomes experienced by the actor/protagonist, which were used in the experiments.

	Other player chose C				Other player chose D			
	Factual		Counterfactual		Factual		Counterfactual	
Prisoner's dilemma Experiment 1	Choice	Outcome	Modification	Outcome	Choice	Outcome	Modification	Outcome
You/Protagonist chose C	CC	3:3	CD uncontrollable	0:5 worse	**CD**	**0:5**	**CC uncontrollable**	**3:3 better**
			DC controllable	5:0 better			**DD controllable**	**1:1 better**
			DD mixed	1:1 worse			**DC mixed**	**5:0 better**
You/Protagonist chose D	**DC**	**5:0**	**DD uncontrollable**	**1:1 worse**	DD	1:1	DC uncontrollable	3:0 better
			CC controllable	**3:3 worse**			CD controllable	0:3 worse
			CD mixed	**0:5 worse**			CC mixed	5:5 better
Stag hunt Experiment 2								
You/Protagonist chose C	**CC**	**5:5**	**CD uncontrollable**	**0:3 worse**	**CD**	**0:3**	**CC uncontrollable**	**5:5 better**
			DC controllable	**3:0 worse**			**DD controllable**	**3:3 better**
			DD mixed	**3:3 worse**			**DC mixed**	**3:0 better**
You/Protagonist chose D	DC	3:0	DD uncontrollable	3:3 same	DD	3:3	DC uncontrollable	3:0 same
			CC controllable	5:5 better			CD controllable	0:3 worse
			CD mixed	0:3 worse			CC mixed	5:5 better
Chicken game Experiment 3								
You/Protagonist chose C	CC	2:2	CD uncontrollable	2:5 same	CD	2:5	CC uncontrollable	2:2 same
			DC controllable	5:2 better			DD controllable	1:1 worse
			DD mixed	1:1 worse			DC mixed	5:2 better
You/Protagonist chose D	DC	**5:2**	**DD uncontrollable**	**1:1 worse**	**DD**	**1:1**	**DC uncontrollable**	**5:2 better**
			CC controllable	**2:2 worse**			**CD controllable**	**2:5 better**
			CD mixed	**2:5 worse**			**CC mixed**	**2:2 better**

Key: The "Modification" column specifies whether the possible counterfactual modification is controllable (i.e., only the actor's/protagonist's choice is undone), uncontrollable (i.e., only the other player's choice is undone), or mixed (i.e., both choices are undone). The counterfactual "Outcome" column specifies whether the modification leads to a better or worse counterfactual outcome for the actor/protagonist.

Table 2: Items used to measure happiness, regret, and disappointment in the three experiment for the reader and the actor groups. All materials are translated from Italian.

Reader group

[Happiness] In light of the outcome of the game, how do you think that Anna/Luca feels?

Extremely unhappy										Extremely happy
□	□	□	□	□	□	□	□	□	□	□
-5	-4	-3	-2	-1	0	1	2	3	4	5

[Regret] In light of the outcome of the game, how do you think that Anna/Luca feels about her/his decision?

S/he does not regret her/his decision at all										S/he regrets her/his decision
□	□	□	□	□	□	□	□	□	□	□
-5	-4	-3	-2	-1	0	1	2	3	4	5

[Disappointment] In light of the outcome of the game, how do you think that Anna/Luca feels about the other player's decision?

Deeply disappointed										Fully satisfied
□	□	□	□	□	□	□	□	□	□	□
-5	-4	-3	-2	-1	0	1	2	3	4	5

Actor group

[Happiness] In light of the outcome of the game, how do you feel?

Extremely unhappy										Extremely happy
□	□	□	□	□	□	□	□	□	□	□
-5	-4	-3	-2	-1	0	1	2	3	4	5

[Regret] In light of the outcome of the game, how do you feel about your decision?

I do not regret my decision at all										I fully regret my decision
□	□	□	□	□	□	□	□	□	□	□
-5	-4	-3	-2	-1	0	1	2	3	4	5

[Disappointment] In light of the outcome of the game, how do you feel about the other player's decision?

Deeply disappointed										Fully satisfied
□	□	□	□	□	□	□	□	□	□	□
-5	-4	-3	-2	-1	0	1	2	3	4	5

Table 3: The frequency (with percentages in parentheses) of the four sorts of counterfactual by participants in Experiments 1-3 (and the additional post-hoc control condition of Experiment 1).

Modifications	Experiment 1 - Prisoner's dilemma					Experiment 2 - Stag hunt				Experiment 3- Chicken game			
	Upward			Downward		Upward		Downward		Upward		Downward	
	Reader *n* = 37	Actor *n* = 44	Reader* *n* = 33	Reader *n* = 39	Actor *n* = 32	Reader *n* = 34	Actor *n* = 33	Reader *n* = 35	Actor *n* = 32	Reader *n* = 32	Actor *n* = 33	Reader *n* = 32	Actor *n* = 35
Controllable	34 (92%)	12 (27%)	25 (76%)	8 (20%)	3 (9%)	23 (68%)	9 (27%)	10 (28%)	3 (9%)	21 (66%)	7 (21%)	5 (16%)	4 (11%)
Uncontrollable	2 (5%)	31 (71%)	6 (18%)	19 (49%)	27 (85%)	9 (26%)	24 (73%)	21 (60%)	24 (75%)	8 (25%)	25 (76%)	23 (72%)	23 (66%)
Mixed	1 (3%)	1 (2%)	2 (6%)	12 (31%)	2 (6%)	1 (3%)	0 (0%)	2 (6%)	5 (16%)	3 (9%)	0 (0%)	2 (6%)	3 (9%)
Other	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (3%)	0 (0%)	2 (6%)	0 (0%)	0 (0%)	1 (3%)	2 (6%)	5 (14%)

*Additional post-hoc control condition

Table 4. Mean judgments of happiness, regret and disappointment (with *SD* in parenthesis) in the three experiments.

Emotion Measures	Experiment 1 - Prisoner's dilemma					Experiment 2 - Stag hunt				Experiment 3 – Chicken game			
	Upward			Downward		Upward		Downward		Upward		Downward	
	Reader	Actor	Reader[2]	Reader	Actor	Reader	Actor	Reader	Actor	Reader	Actor	Reader	Actor
Happiness	-2.92 (1.09)	-1.2 (1.98)	-2.82 (2.20)	3.15 (2.27)	3.16 (2.1)	-2.62 (1.97)	-2.51 (2.03)	3.67 (2.69)	4.31 (1.03)	-1.91 (1.63)	-.52 (2.15)	3.81 (1.94)	3.09 (2.20)
Regret	1.35 (2.70)	-1.91 (3.02)	1.91 (2.27)	-2.15 (3.0)	-3.06 (2.69)	2.06 (2.45)	-1.36 (3.04)	-3.00 (3.12)	-3.81 (2.7)	.97 (2.42)	-1.15 (2.68)	-2.25 (3.65)	-3,00 (2.73)
Disappointment[1]	2.16 (1.76)	1.98 (2.20)	1.97 (2.26)	-2.02 (2.71)	-2.66 (1.84)	1.94 (1.89)	2.42 (2.7)	-3.74 (2.80)	-4.72 (.81)	1.91 (1.49)	1.15 (2.14)	-2.56 (2.28)	-2.97 (1.98)

(1) The scoring was reversed so that higher disappointment had higher scores.
(2) Readers in the post-hoc condition

Table 5.

Results of the Logistic Regression on the controllable vs. uncontrollable counterfactual modifications across the three experiments.

Predictor	*B*	*SE b*	Wald	*df*	*P*	Exp(*b*)	95% confidence interval
Payoff structure			3.170	2	.205		
Payoff structure (2)	-.560	.317	3.108	1	.078	.571	[.31-1.07]
Payoff structure (3)	-.252	.328	.591	1	.442	.777	[.41-1.48]
Role	1.017	.418	5.93	1	.015	2.77	[1.22-6.27]
Direction	-2.423	.355	46.49	1	.000	.089	[.04-.18]
Role x Direction	1.483	.538	7.607	1	.006	4.407	[1.54-12.65]

Figure 1.

The payoff structure of the three social dilemma games employed in Experiments 1, 2, and 3. The Nash equilibria are indicated in bold, and the Pareto-optimal outcomes are in italics.

Prisoner's Dilemma Experiment 1		Other player chooses C	Other player chooses D
You/Protagonist choose(s)	C	*You/Protagonist: win(s) 3 chocolates* *Other player: wins 3 chocolates*	You /Protagonist: win(s) 0 chocolates Other player: wins 5 chocolates
You /Protagonist choose(s)	D	You/Protagonist: win(s) 5 chocolates Other player: wins 0 chocolates	**You/Protagonist: win(s) 1 chocolate** **Other player: wins 1 chocolate**
Stag Hunt Experiment 2		Other player chooses C	Other player chooses D
You/Protagonist choose(s)	C	***You /Protagonist: win(s) 5 chocolates*** ***Other player: wins 5 chocolates***	You /Protagonist: win(s) 0 chocolates Other player: wins 3 chocolates
You/Protagonist choose(s)	D	You/Protagonist: win(s) 3 chocolates Other player: wins 0 chocolates	**You/Protagonist: win(s) 3 chocolates** **Other player: wins 3 chocolates**
Chicken game Experiment 3		Other player chooses C	Other player chooses D
You/Protagonist choose(s)	C	You /Protagonist: win(s) 2 chocolates Other player: wins 2 chocolates	***You /Protagonist: win(s) 2 chocolates*** ***Other player: wins 5 chocolates***
You/Protagonist choose(s)	D	***You/Protagonist: win(s) 5 chocolates*** ***Other player: wins 2 chocolates***	You/Protagonist: win(s) 1 chocolate Other player: wins 1 chocolate

Figure 2: The percentages of controllable, uncontrollable, and mixed counterfactual modifications in the reader and actor conditions in the three experiments (the remainder were Other responses).