Bad is Freer than Good: Positive–Negative Asymmetry in Attributions of Free Will

Gilad Feldman

Department of Work and Social Psychology

Maastricht University

Maastricht, 6200MD, The Netherlands

gilad.feldman@maastrichtuniversity.nl

Kin Fai Ellick Wong

Department of Management

Hong Kong University of Science and Technology

Clearwater Bay, Kowloon

Hong Kong SAR

mnewong@ust.hk

Roy F. Baumeister

Department of Psychology

Florida State University

Tallahassee, Florida, USA

baumeister@psy.fsu.edu

Conditionally accepted at Consciousness and Cognition
Word count. Abstract: 143; Article: 8950

Abstract

Recent findings support the idea that the belief in free will serves as the basis for moral responsibility, thus promoting the punishment of immoral agents. We theorized that free will extends beyond morality to serve as the basis for accountability and the capacity for change more broadly, not only for others but also for the self. Five experiments showed that people attributed higher freedom of will to negative than to positive valence, regardless of morality or intent, for both self and others. In recalling everyday life situations and in classical decision making paradigms, negative actions, negatives outcomes, and negative framing were attributed higher free will than positive ones. Free will attributions were mainly driven by action or outcome valence, but not intent. These findings show consistent support for the idea that free will underlies laypersons' sense-making for accountability and change under negative circumstances.

Keywords: accountability; action valence; attributions; free will; outcome valence

Bad is Freer than Good: Positive–Negative Asymmetry in Attributions of Free Will

The idea that people have the capacity to make free, autonomous, and responsible choices is one fundamental assumption of most, if not all, modern civilizations. Although most cultures operate based on some degree of belief in freedom of choice, people vary in how much they regard human beings, including themselves, as capable of making free choices (e.g., Baumeister, 2008; Paulhus & Carey, 2011). They also differ in their perceptions of how much free will they have compared with others (Gray, Knickman, & Wegner, 2011; Pronin & Kugler, 2010), and how much free will is exerted in certain situations (Helzer & Gilovich, 2012). Such fluctuations in free will beliefs and attributions of free will are far more than idle metaphysical speculations, having been shown to alter cognition and behaviors (e.g., Alquist, Ainsworth, & Baumeister, 2013; Vohs & Schooler, 2008) and with important legal and societal implications (Greene & Cohen, 2004; Roskies, 2006).

Over two millennia, philosophers have been debating what the concept of free will means, how it should be defined, and what purpose it serves. Despite the long-standing debate there has so far been very little convergence with several schools of thought and countless views conceptualizing free will in different—sometimes conflicting—ways. In recent years, a group of experimental philosophers and social-cognitive psychologists have begun to look beyond the academic and philosophical debate on the meaning of free will and have instead examined laypersons' beliefs, cognition, and the behavioral consequences related to the elusive concept of free will.

A growing body of literature examining free will beliefs and attributions has converged on the idea that in laypersons' minds the concept of free will is associated with moral

responsibility. The belief in free will has been shown to promote socially responsible and moral behavior, such as more honesty (Vohs & Schooler, 2008), better learning from emotional experiences (Stillman & Baumeister, 2010), and more prosocial behavior (Baumeister, Masicampo, & DeWall, 2009). The theory underlying these findings is grounded on the philosophical argument that free will is a prerequisite for holding people morally responsible for their actions (Kant, 1788/1997). The link between free will and moral responsibility is also reflected in the attributions people make to agents, with immoral agents perceived as having higher free will and higher perceived blameworthiness (Phillips & Knobe, 2009), resulting in more retributive behavior (Shariff et al., 2014), as well as activating the belief in free will so as to allow punishment of these agents (Clark et al., 2014).

However, studies on free will beliefs and cognition are not limited to morality and moral situations but have extended to broader behavior in everyday life to reflect a wider view of the self as an active agent who is free to choose actions and pursue goals. The laypersons' concept of free will is theorized as a core mechanism that enables the person to better pursue what he/she wants (Dennett, 2003; Hume, 1748; Edwards, 1754/1957). Those who believe in free will enjoy greater self-efficacy and less helplessness (Baumeister & Brewer, 2012), have higher levels of autonomy and more proactivity (Alquist et al., 2013), exhibit better academic performance (Feldman, Chandrashekar, & Wong, 2016) and job performance (Stillman et al., 2010), and have more positive attitudes and higher perceived capacity for decision making (Feldman, Baumeister, & Wong, 2014). These findings are not about morality but rather about agents capable of change to improve their own behavior and take responsibility for their actions. To exemplify that free will attributions are about more than just morality, it has been shown that free will attributions are affected by self-serving biases differentiating between perceptions of the self and of others.

Pronin and Kugler (2010) showed that people tend to perceive themselves as having more free will than others, perceiving their own behaviors as less predictable, and their own futures as less determined and more driven by intent. If free will were mainly about holding people responsible and punishing them for immoral actions, it would make little sense for people to attribute high free will to themselves, because that simply increases their own vulnerability to punishment.

We propose that the concept of free will extends beyond morality and punishment to encompass change and accountability more broadly. In our use of the term 'accountability' we refer to the acknowledgement and assumption of responsibility. Thus, if a behavior or an outcome deviates from the expected, then an accountable person accepts his or her own role and seeks to learn from mistakes and correct future action. Using this view, the popular notion of free will may have evolved to enable proactivity and learning by promoting people to see themselves as more accountable for their own actions (Baumeister, 2008). A judgment based in accountability (moral, legal, performance, learning, and otherwise) is essentially a decision about whether a person should have acted differently in a particular situation, especially if the outcome was undesirable or if it did not meet with expectations (Malle, Guglielmo, & Monroe, 2014). To assert that someone *should* have acted differently only makes sense if one assumes that the person *could* have chosen to act differently. This assertion implies that the actions and outcomes, although subject to many causes, were not fully coerced or predetermined, in the sense that there was no room left for agentic choice (Nichols, 2006). Choice has been shown to be an important factor in people's perception of agency and freedom (Barlas & Obhi, 2013; Bode et al., 2014; Feldman et al., 2014), and the assumption that a person could have chosen to act differently in the same situation is the essence of most laypersons' conception of free will (Monroe, Dillon, & Malle, 2014; Monroe & Malle, 2010; Stillman, Baumeister, & Mele, 2011).

The conceptual link between free will and accountability provides the basis for the proposition that attributions of free will would be higher for negative actions or outcomes because these attributions allow people to perceive that change is possible and accept their personal role in affecting such change. This would not be merely for the sake of punishment, but also to promote change and learning (Seligman, Railton, Baumeister, & Sripada, 2013). Negative outcomes are generally undesirable, and they motivate people to engage in counterfactual thoughts about what could have happened differently (Epstude & Roese, 2008; Roese, 1997), yet to allow for the possibility of controlled change, negative outcomes would more specifically trigger a search for what the agents involved could have chosen to do differently to prevent, correct, or overcome the negative outcome (Alquist, Ainsworth, Baumeister, Daly, & Stillman, 2015).

Our theoretical model is summarized in Figure 1. A perception of an action or an outcome as negative leads to an attribution of responsibility, which is dependent on the perceptions of the agent as having free will. On the basis of this link, we expect that the perception of a negative action or outcome would also directly trigger the attribution of free will as to allow for accountability and future change in behavior.

We expected that attributions of free will would be triggered in a wide array of negative circumstances, including those that are morally neutral or devoid of intent. From moral situations to everyday life learning, the perception of having free will (the capacity to have chosen otherwise) is an essential component of accepting responsibility for one's past, present, and future behavior so that negative outcomes can be corrected. In this view, the attribution of higher free will to agents would not be limited to the judgments of others, but will also extend to those

of the self, and would not be limited to specific actions, but would also extend to outcomes regardless of action.

The present studies

Five experiments were conducted to test our hypothesis that negative actions and outcomes, compared to positive ones, lead to higher attributions of free will to the enacting agent, for both self and for others. Experiment 1 employed a classic Asian disease scenario to examine the attributions following negative versus positive outcomes to a risky decision in which the action and the intent were held constant. Experiment 2 assessed the attributions to negative versus positive framings of identical outcomes in the same scenario, with both intent, action, and outcome held constant. Experiment 3 used a prisoner's dilemma game theory paradigm to examine free will attributions to defection versus cooperation, showing that the effect extends to social behavior. Experiment 4 was designed to extend the test to real-life events by having participants recall ordinary interactions they had with other people. Lastly, Experiment 5 contrasted action valence and outcome valence in free will attributions and their link to intent. Taken together, these experiments provide a comprehensive test for the hypothesized positive–negative asymmetries in general free will attributions.

Experiment 1

Experiment 1 provided the first test of the hypothesis, using a variation on the Asian disease scenario originally used by Tversky and Kahneman (1981). The scenario refers to an impending epidemic and a choice between two public health interventions, one of which may save everyone but may also save no one, whereas the other would guarantee saving some lives but not all. Tversky and Kahneman used this scenario to show that people’s preferences between those options shift as a function of whether the interventions are described in terms of lives saved

or lives lost. We adapted this scenario to hold the action constant while manipulating outcome. Participants were told to imagine that the decision had been made either by themselves or by someone else, and that the decision made was to pursue the high-risk high-payoff option (i.e., take a chance on saving everyone). Half were told that the decision had turned out well, thus all lives had been saved, whereas the others were told to imagine it had gone badly and no one was saved. Our hypothesis predicted that people would attribute more freedom to the decision-maker whose decision turned out badly rather than favorably.

The specific scenario was chosen as it allowed us to hold both the action and the intent constant. The two options presented in the scenario did not contrast between a positive or a negative action but rather an action that is either risk-seeking or risk-averse, thereby holding intent (to save lives) constant. Furthermore, in all scenarios the agent chose the risky option, as this option allowed us to hold the action constant while contrasting two different possible outcomes. Thus, in the contemplated scenario there is no question that the decision maker intends to minimize the number of lost lives, there are two viable alternatives and both can be argued as acceptable, and in our manipulation of the outcomes the agent has chosen the same action.

Holding action and intent constant served two main purposes in this experiment. First, we aimed to establish that free will attributions are different from previous effects about intent, such as Knobe's (2003) finding that people attributed greater intent when a decision maker chose an action that had bad rather than good side effects. Second, since the action is the same, there is no contrast between moral and immoral behavior, which would demonstrate that the effect is not merely about encouraging moral over immoral behavior (Clark et al., 2014; Phillips & Knobe,

2009; Shariff et al., 2014) but also extends to envisioning the possibility for change more broadly.

Method

Participants and design. A total of 204 participants (M_{age} = 30.15, SD_{age} = 8.99; 84 females) were recruited using Amazon Mechanical Turk, and each received US$0.05 for completing one of the four versions of the questionnaire (2×2, between-subject, self versus other, and negative versus positive, randomly assigned).

Manipulations. The participants were presented with the Asian disease problem (Tversky & Kahneman, 1981) in which a decision maker faces the dilemma of choosing between two types of medicine aimed to help in a situation in which 600,000 people are expected to die from an impending epidemic. There were two options. The riskier option offers a one-third chance of saving everyone and a two-thirds chance of not saving anyone. The safer option presented a certainty of saving one-third of the lives but consigning the other two-thirds to certain death. The participants were also told that the decision had been made to pursue the riskier option. Half were told to assume that they themselves had made that decision, whereas the rest were told to imagine that someone else had made the decision. The outcome was also manipulated: half were told to assume that the intervention had been successful and everyone had been saved, whereas the rest were told to imagine that it had failed and that all 600,000 lives had been lost.

The scenario was followed by two multiple-choice quiz questions that the participants had to answer correctly to proceed and a manipulation check. The participants were asked to indicate their level of agreement with free will statements.

Free will attributions. In order to avoid priming people's beliefs in free will and confounding answers to the different perspectives people hold regarding the loaded term of 'free will' we applied indirect measures of free will attributions and avoided using the term of free will (Pronin & Kugler, 2010), instead asking the relevant operative regarding the agency's capacity to choose differently (Chernyak & Kushnir, 2013; Nichols, 2004). We adopted the definition of free will reached by a recent combined effort of social psychologists, neuroscientists, and experimental philosophers as being the capacity to perform free actions (Haggard, Mele, O'Connor, & Vohs, 2010), meaning that the person could have acted otherwise in the availability of options and with the capacity to choose among those options without coercion (Baumeister, 2008; Kane, 2002; Wong & Cheng, 2013). Both sophisticated philosophical treatments (e.g., Kane, 2011) and layperson views (e.g., Feldman et al., 2014; Monroe & Malle, 2010) tend to regard the capacity for choice to act otherwise as an essential core of free will (Nichols, 2004). These measures were specifically meant to provide the clearest and simplest measure of free will attributions without addressing the issue of determinism, capturing the participant's own views of free will regardless of views on compatibilism (Nahmias, Morris, Nadelhoffer, & Turner, 2005).

We therefore asked the participants about the ability of the decider to choose otherwise. The first question asked whether the person (self or other) "could have chosen to act differently in that situation". To go beyond the specific context of the situation and examine the implications for possible learning and change in the future, we also asked about future situations of what may be, so the second question asked that if the two people "were to face the same situation again" whether the actor "would be able to choose a different course of action that would lead to a different outcome" (1 = *strongly disagree*, 5 = *strongly agree*). These two items were designed to capture the dimensions of both past reflection and future prospection.

Clearly, in the scenario described agents had an alternative option. Therefore, any differences in attributions between the manipulation conditions would reflect a bias in the perception of the agent's capacity to be able to choose differently based on who decided and what the outcome was.

Results and discussion

Means and standard deviations for the manipulation check and the free will attributions are reported in Table 1. The manipulation of valence was successful. The participants in the negative conditions rated the outcome as more negative than those in the positive conditions ($F(1, 200) = 303.97$, $p < .001$, $\eta_p^2 = .60$).

Contemplating the bad outcome led people to attribute greater freedom to the decision maker than contemplating the good outcome. Two-way between-subject analysis of variance (ANOVA) revealed the main effects of outcome valence on both free will attribution measures (findings are plotted in Figure 2). That is, participants rated the bad-outcome decision as freer than that of the good-outcome ($F(1, 200) = 16.45$, $p < .001$, $\eta_p^2 = .08$), and likewise they rated the decision leading to a bad-outcome as freer with regard to possible future decisions in similar situations ($F(1, 200) = 30.76$, $p < .001$, $\eta_p^2 = .13$). The main effects of the self/other variable were not significant on either measure ($F < .63$, $p > .427$). The interactions between outcome valence and self-other were likewise not significant ($F < .24$, $p > .627$).

Experiment 1 showed that people attributed more free will to the authors of an action which resulted in a negative outcome than to the authors of the same action taken by the same author leading to a positive outcome. The pattern was the same regardless of whether self or another made the decision. This provided the first support for our hypotheses about a positive-negative asymmetry in attributions of free will. The action and the intent were controlled for, and

the effects consistent for when the contemplated enacting agent was the self or someone else and for the current situation and future similar situations, lending support for the theory that free will attributions are not merely about the contrasts between moral and immoral behavior and the punishment of misbehaving agents, but rather a broader bias triggering the perceived capacity for change when well-intended actions turn out badly.

Experiment 2

Experiment 2 was a companion to Experiment 1. It again used the Asian disease scenario, but this time we went further by holding both the action and the outcome constant, varying only the framing of the outcome. In Experiment 1, we manipulated the outcome of a risky decision that turned out either positively or negatively. However, in Experiment 2, the actions and outcomes were the same across conditions, and we only manipulated whether the outcome was described in terms of lives saved (positive) or lives lost (negative).

If the effect found in Experiment 1 is replicated in Experiment 2 when both the action, intent, and outcome are the same, then this would show a fundamental bias in which simply focusing on the negative rather than the positive in any given situation would trigger higher free will attributions and the envisioned capacity to act differently and obtain different outcomes in future situations.

Method

Participants and design. A total of 221 participants (M_{age} = 29.98, SD_{age} = 9.11; 93 females) on Amazon Mechanical Turk received US$0.05 for completing one of the four versions of the Asian disease scenario (2×2, between-subject, self versus other, and negative versus positive frame, randomly assigned).

Procedure and materials. The Asian disease scenario in Experiment 1 was adapted for Experiment 2. As in the previous study, it described an impending epidemic and a choice of interventions made by either the self or another person. This time, however, participants were instructed to visualize that the non-risky option rather than the risky option had been chosen, thus saving a third of the at-risk lives (or, framed negatively, killing two thirds). For half of the participants, this outcome was described in terms of the lives that were saved. For the rest, the outcome was described in terms of the number of deaths. This differential framing was similar to what was originally used by Tversky and Kahneman (1981) with this scenario.

The scenario was followed by two multiple-choice quiz questions that the participants had to answer correctly in order to proceed (again, intended as checks on attention and understanding) and a manipulation check. The participants were then asked to rate their perceptions of whether the person could have chosen to act differently, using the same two items as in Experiment 1.

Results and discussion

The means and standard deviations for the manipulation check and the free will attributions are reported in Table 2. The participants in the positive framing conditions rated the outcome as more positive than those in the negative framing conditions (F (1,217) = 44.93, p < .001, η_p^2 = .17), even though the outcome was the same — thus indicating a successful manipulation of the framing valence.

A two-way between-subject ANOVA revealed the significant main effects of framing on both free will attribution measures (findings are plotted in Figure 3). Regarding the decision itself, the participants perceived more scope for the decision maker to have chosen otherwise when the outcome was framed in terms of losses than gains (F(1, 217) = 16.54, p < .001, η_p^2 =

.07). They also attributed higher freedom to the decision maker in similar future situations following the death frame than the lives-saved frame ($F(1, 217) = 7.11$, $p = .008$, $\eta_p^2 = .03$).

As in Experiment 1, the main effects of the decision maker (self versus other) were not significant on either measure ($F < .89$, $p > .347$). The interactions were also not significant ($F < 1.08$, $p > .300$).

Thus, even when the decision outcome was objectively the same (i.e., 400,000 deaths and 200,000 lives saved) and only the framing was manipulated, the attributions of free will varied as a function of valence. Contemplating the outcome in terms of lives lost made people attribute more freedom to the decision maker than contemplating it in terms of lives saved. As in the preceding experiment, the same effect was found for judging the specific decision as for evaluating future possible similar actions, and for contemplating the action taken by both the self and others.

Experiment 2 extended Experiment 1 to show an even broader bias triggering the perceived capacity for change not only when well-intended actions turn out badly but even when simply thinking about outcomes as negative. In both experiments, the action and the intent were controlled for, and the effect was observed when contemplating both self and others for either the current situation or other similar situations in the future.

Experiment 3

Experiment 3 sought to extend Experiments 1 and 2 using a more specific fixed interaction between the participant and another person to capture a social situation involving the participant with consequences for the participant in a more realistic situation. That is, the experiment measured free will attributions in a two-person “prisoner’s dilemma” interaction (Rapoport & Chammah, 1965) in which the actions of a person hold direct and clear

consequences for the other party and one of the parties is the participant. The decision is either prosocial or selfish, and it is made either by the participant toward a friend or by a friend toward the participant. Defection in this game is not an immoral act but rather a selfish act, as there is no cheating, reciprocation or lack of, and both options of cooperation or defection are within the set rules of the game[1].

This specific paradigm was chosen as a widely used and simplified classical representation of a social interaction. Based on the findings in the first two experiments, the prediction was that participants would perceive more free will when it is perceived that the person acted negatively towards the other person than when the person is perceived as having acted positively towards the other person.

Method

A total of 208 participants on Amazon Mechanical Turk received US$0.05 for completing the study. They were randomly assigned among four conditions (2×2, between-subject, self versus other, and negative versus positive action).

The participants were presented with the prisoner's dilemma scenario, in which they were asked to imagine playing together with a friend to win a possible prize and in which both the self and the friend have the option to cooperate or defect. The unilateral defection brought significant gains for the defector (US$75), whereas the cooperating partner received nothing. Mutual cooperation brought both parties a good outcome (US$45). Mutual defection resulted in a small benefit (US$15) to both. Four multiple-choice questions were administered to ascertain that the

[1] Prosocial and moral behavior are not the same construct, as – for example – two players can act cooperate to act unethically (Feldman, Chao, Farh, & Bardi, 2015; for an example using game theory defection and cooperation see Weisel & Shalvi, 2015).

participant understood the game and the instructions. These had to be answered correctly before proceeding.

Next, the participants were told to imagine that the game had been played. They were told to imagine a particular response either by themselves or by the friend (other player). Thus, four conditions existed: the friend cooperated, the friend defected, the participant cooperated, or the participant defected. The scenario only described a single action taken by one of the actors without indicating the other player's decision in order to minimize the possible confound of reciprocity. The perceptions of the outcome valence in game theory scenarios can vary considerably across participants based on many factors, and we therefore administered outcome comprehension check questions also serving as manipulation checks (1 - "indicate whether [your/your friend's] choice has positive or negative implications for [your friend/you]" with an answer of either positive or negative, and 2 - "on a scale of -100, most negative, to +100, most positive, how would you rate the implications of [your/your friend's] decision for [your friend/you]"). Correct answers in the manipulation checks were a prerequisite to inclusion in the analyses (above or below zero for the second manipulation check), because only participants who properly answered the manipulation can be considered as a test of the hypotheses regarding valence. After excluding those who gave wrong answers about the manipulation, we were left with a sample of 137 (M_{age} = 30.36, SD_{age} = 9.56; 63 females). Nonetheless, the pattern of results for the full sample was similar to the findings reported below.

Finally, the participants were asked about their attributions of free will. Two items were similar to the items used in the preceding studies: the choice to act differently in the current situation and choice to act differently in similar future situations. Based on Pronin and Kugler's (2010, Experiment 1) conceptualization of free will attributions, we also added an item which

measured free will attributions by an indirect measure of predictability (more predictable indicative of lower capacity for free will) - “I could have predicted the other person’s behavior in that situation even before it happened” or "the other person could have predicted my behavior in that situation even before it happened" (1 = *strongly disagree*, 5 = *strongly agree*; reversed).

Results and discussion

The means and standard deviations for the free will attributions are reported in Table 3. The results of a two-way between-subject ANOVA are plotted in Figure 4. For the attributions of the current decision made, the participants perceived a negative action toward another person to indicate higher free will than a positive action ($F(1, 133) = 4.03$, $p = .047$, $\eta_p^2 = .03$). If the same situation were to arise again in the future, the participants perceived that an agent who behaved negatively would have a higher capacity to behave differently in the future than an agent who acted positively ($F(1, 133) = 5.75$, $p = .018$, $\eta_p^2 = .04$). The predictability measure in this study followed the overall expected pattern, and negative behaviors were rated as less predictable indicating higher free will than positive behaviors ($F(1, 133) = 7.22$, $p = .008$, $\eta_p^2 = .05$). There was a marginal interaction for attributions to the current situation ($F = 3.69$, $p = .057$), indicating that the effect was stronger for attributions to others. However, no other significant self-other main effects or interactions were found ($F < 1.19$, $p > .278$).

Experiment 3 tested the hypotheses using a game theory of prisoner's dilemma simulation of an everyday life social interaction. Overall, free will was again perceived more strongly in connection with the negative than with the positive valence, which in this study took the form of making selfish moves rather than cooperation.

Although the prisoner’s dilemma paradigm resembled a realistic social interaction context, a limitation of this paradigm is that the outcome of either cooperation or defection could

be interpreted to be both positive and negative, depending on the perspective taken, an issue which we attempted to control for using the manipulation checks. Another limitation of the game theory paradigm has to do with the ambiguous intent, as it is not clear whether the action taken by either side was intended as positive or negative toward the other party or merely as a reaction to an anticipated decision by the other party.

Notwithstanding these limitations, Experiment 3 extended the findings of Experiments 1 and 2 to demonstrate the effect in a more realistic scenario involving the participant contrasting between selfish and prosocial actions.

Experiment 4

In Experiment 4, we sought to extend the findings from Experiments 1 to 3 to actual behaviors in everyday life rather than hypothetical vignettes (for the importance of evaluating an effect from both reader and observer perspectives, see Girotto, Ferrante, Pighin, & Gonzalez, 2007). The participants rated their perceptions of free will after recalling either a good or a bad action by themselves or by someone else. The main prediction was, again, that people would attribute more free will for the bad than the good actions, regardless of whether the actions were performed by self or others.

To complement the attribution questions common in the literature and demonstrate that negative valence triggers perceiving higher capacity for agents to do otherwise, we also measured agentic counterfactuals. Counterfactual thinking involves the tendency to think of all possible alternative realities that could have taken place, both in general circumstances or for the person, upwards or downwards. Past work has shown that negative actions and outcomes generally elicit more upward counterfactuals than good ones, with regard to what could have happened differently to produce a better outcome (Boninger, Gleicher, & Strathman, 1994;

Roese, 1997; Roese & Olson, 1997). We hypothesized that the concept of accountability underlying free will and the need for responsibility are likely to elicit a very specific set of counterfactuals, one that mainly focuses on what the agent could have chosen to do differently[2].

Method

Participants and design. Undergraduate students (N = 212, 112 females; M_{age} = 19.17, SD_{age} = .97) received course credit for completing one of the four versions of a survey questionnaire, assigned at random. The design was a 2x2 between-subject factorial, varying whether the self or the other person was the responsible agent and whether the action was negative versus positive.

Procedure and materials. The participants were instructed to recall and describe in writing a recent interaction with another person, in which one person did something that affected the other. Half of the participants were randomly assigned to write about them doing something that affected another person, and the rest were assigned to write about the other person doing something that affected them. Cross-cutting this, half were randomly assigned to write about positive actions, and the rest wrote about negative ones.

The short essay was followed by the measuring of free will attributions in the recalled situation and similar future situations (Experiments 1 and 2) and predictability (Experiment 3). Because we examined real life complex interactions rather than a fixed scenario, this allowed us

[2] To clarify, free will attributions cannot be reduced to counterfactual thinking. Although the two are related, there are important differences between the two constructs. Counterfactuals are broader and include all that could have happened differently leading to a different outcome, while free will attributions focus more specifically on what the agent could have *chosen* to do differently, free from internal and external constraints (for a more detailed review see Alicke, Buckingham, Zell, & Davis, 2008; Alquist, Ainsworth, Baumeister, Daly, & Stillman, 2015; Baumeister Crescioni & Alquist, 2011). For example, counterfactual thinking may trigger many types of alternative realities that lack free will, such as external constraints that confound free will: luck ('if I only had luck on my side'), nature ('if only it did not rain'), fate ('if only I my astrological sign were different'), and laws of physics ('if only the sun did not rise this morning'), or internal constraints that confound free will, such as personality ('if only I were an extravert/'), background ('if only I had been born rich'), genes ('if only I were taller') or counterfactuals that do not involve a deliberate choice ('if only I were not so tired').

to add an additional measure. The fourth measure was adapted from the Pronin and Kugler measure of free will as alternatives to action (2010, Studies 2 and 3) and asked about agentic counterfactuals: how the self or the other person could have acted differently in the specific recalled situation. The participants were asked the following: “Looking back, how — if at all — could [you/the other person] have acted differently? Please provide as many options as you can about how you think [you/the other person] could have acted differently in that situation” and were further instructed to "write down 'no other possible actions' if and only if you think [you/the other person] could not have acted differently in any way." The responses were coded for the number of alternatives mentioned.

Results and discussion

The means and standard deviations for free will attributions are reported in Table 4. Bad actions were rated as freer than the good ones, as indicated by a two-way between-subject ANOVA (see Figure 5 for plots).

A significant main effect was found for the valence of the action. The participants gave higher free will ratings to the actor who performed the bad action than to the actor who did something good, both for the specific action they wrote about (F (1, 208) = 36.16, $p < .001$, $\eta_p^2 = .15$), and for possible similar situations in the future (F(1, 208) = 15.54, $p < .001$, $\eta_p^2 = .07$). Participants also wrote down more agentic counterfactuals in the negative conditions than in the positive condition (F (1, 208) = 4.86, $p = .029$, $\eta_p^2 = .02$). Lastly, a main effect for valence indicated lower predictability (higher free will) for negative actions (F (1, 208) = 6.14, $p = .014$, $\eta_p^2 = .03$). We examined the situations recalled by participants in terms of morality, by coding whether the situations recalled involved blatant unethical actions that intentionally harmed others, or clearly violated laws or regulations. We found that none of the recalled situations

involved unethical behavior or a moral dilemma, and that participants recalled every-day life situations (for example, negative behaviors recalled were a professor giving a bad grade, roommates making noise, someone chewing gum in class, romantic disappointments, etc.).

Unlike in the previous experiments, there was also a significant main effect indicating that more free will was attributed to the other person than to the self using three measures (recalled situation: $F(1, 208) = 17.40$, $p < .001$, $\eta_p^2 = .08$; predictability (F (1, 208) = 3.53, $p = .062$, $\eta_p^2 = .02$; agentic counterfactuals ($F(1, 208) = 15.23$, $p = .015$, $\eta_p^2 = .03$). A significant interaction also emerged in two measures (recalled situations: F (1, 208) = 5.46, $p = .02$, $\eta_p^2 = .03$; predictability: $F(1, 208) = 13.30$, $p < .001$, $\eta_p^2 = .06$) indicating that differences between self and other were larger with regard to the positive action than the negative one.

Across the four measures we found that negative valence was associated with a higher degree of free will than positive valence. The coding of the recalled situations showed that they did not involve any moral dilemmas or morally valenced actions but rather represented simple interactions between people in their everyday lives. Together with the previous three experiments, we conclude a consistent positive–negative asymmetry bias for free will attributions.

Some differences in attributions to self versus others were found in this study unlike in the previous experiments. High free will was attributed to others than to the self in two out of four measures for both positive and negative actions. The findings that positive and negative actions were both perceived as higher free will for others might be due to the easier recall of circumstances and constraints for their own actions while being typically unaware of the circumstances and constraints of others, thus possibly leading to more perceived freedom of action (Malle, Knobe, & Nelson, 2007). In Experiments 1-3, the alternatives to the action were

predefined and controlled. That is, the complexity of real-life situations may have introduced additional biases to free will attributions.

Experiment 5

Experiment 1 manipulated outcome, Experiment 2 manipulated outcome framing, and Experiments 3 and 4 manipulated action. In Experiment 5, we sought to manipulate both action and outcome.

To address previous challenges about the role of desire for blameworthiness and intent confounds in the Knobe Effect (Guglielmo & Malle, 2010), we also directly manipulated intent. That enabled us to examine and contrast all three in their effect on perceived free will.

Method

Participants and procedure. A total of 301 participants (M_{age} = 35.08, SD_{age} = 11.58; 169 females) were recruited from Amazon Mechanical Turk in return for US$0.15. The participants were presented with a scenario based on a design by Cushman (2008) and adapted from a scenario in Phillips and Knobe (2009), in which a doctor was ordered by the chief of surgery to prescribe medicine to a patient (each of the brackets below represents a single manipulation):

> At a certain hospital, there were very specific rules about the procedures doctors had to follow. The rules said that doctors have to follow the orders of the chief of surgery. One day, the chief of surgery went to a doctor and said: 'I don't care what you think about how this patient should be treated. I am ordering you to prescribe the drug Accuphine for her!'.

> The doctor had always [liked this patient and actually wanted the patient/disliked this patient and actually did not want the patient] to be cured.
>
> The doctor knew that giving this patient Accuphine would result in an immediate [recovery/death].
>
> The doctor went ahead and prescribed Accuphine.
>
> As a result of the medicine, the patient [recovered immediately/died shortly after].

The scenario manipulated the valence of three factors. First, intent was manipulated by whether the doctor had positive or negative attitudes toward the patient and wanted to see the patient helped or harmed. Second, the action varied, as in the doctor knew the outcome of the action taken would be either positive or negative. Third, the outcome was positive or negative. The design was therefore 2x2x2 for intent, action, and outcome as either positive or negative.

The participants were then presented with three manipulation checks in which they were asked to indicate the valence of the doctor's intent, the action taken by the doctor, and the outcome on a scale of –100 (*very bad*) to 100 (*very good*).

The participants were then asked about the doctor's perceived capacity to have chosen not to prescribe the medicine as a measure or perceived free will (0 = *No choice - had to prescribe*; 100 = *Had choice - could have chosen NOT to prescribe*).

Results and discussion

The correlations between the manipulation checks and the dependent variables are detailed in Table 5. Free will attributions were negatively correlated with both action valence (r = -.27, p < .001) and outcome valence (r = -.17, p = .004) and only marginally correlated with

intent valence (r = -.10, p = .07ns). The strong correlations between intent and action and between action and outcome perceived valence indicate that perceptions of intent, action, and outcome valence may be linked to one another regardless of the manipulations.

The manipulations were successful. Participants indicated more negative intent when the doctor disliked and wanted to harm the patient (N = 153, M = -57.58, SD = 61.73) than when the doctor liked and wanted to help the patient (N = 148, M = 40.97, SD = 62.34; t(299) = 13.04, p < .001). Likewise, they perceived the action as more negative when the doctor thought that the medicine would be harmful (N = 150, M = -46.57, SD = 69.93) than when the doctor thought that the medicine would be helpful (N = 151, M = 31.10, SD = 74.38; t(299) = 9.33, p < .001). Last, and unsurprisingly, they rated the outcome as more negative when the patient died (N = 152, M = -87.52, SD = 37.97) than when the patient recovered (N = 149, M = 88.62, SD = 29.17; t(299) = 45.07, p < .001). However, the action valence manipulation also affected perceived intent (t(299) = 8.47, p < .001) and the outcome valence manipulation also affected perceived action (t(299) = 10.536, p < .001) and perceived intent (t(299) = 2.06, p = .04). Therefore, the analyses below were supplemented the regression analyses using manipulation checks.

A three-way ANOVA of the three manipulations examined the free will attributions and revealed significant effects for action (F(1, 293) = 20.57, p < .001, η_p^2 = .07) and outcome (F(1, 293) = 7.57, p = .006, η_p^2 = .03), but not for intent (F(1,293) = .89, p = .348ns), on free will attributions. No interactions were significant (F < 1.65). The participants in the negative action condition perceived higher free will (M = 81.47, SD = 27.53) than those in the positive action condition (M = 65.11, SD = 34.75; t(299) = 4.53, p < .001). Those in the negative outcome condition perceived higher free will (M = 78.16, SD = 30.83) than those in the positive outcome condition (M = 68.28, SD = 33.21; t(299) = 2.67, p = .008). A step-wise regression using the

manipulation checks showed that when put together in a regression, only action emerged as a significant predictor of free will attributions ($F(1, 299) = 23.91$, $p < .001$, $\beta = -.27$, $p < .001$, $\Delta R^2 = .07$), with no other effects found.

In summary, free will attributions were affected by both action valence and outcome valence, but were not affected by intent manipulation or associated with intent attributions. Thus, we conclude that the findings differ from the pattern shown in Knobe's (2003) findings about intention to produce unwanted side effects.

General Discussion

The primary finding of this investigation was that bad was perceived as freer than good for both actions and outcomes and regardless of the agent. The findings are summarized in Table 6. Higher attributions of free will for bad than for good actions and outcomes were consistent across multiple methods, including the hypothetical Asian disease scenario (Experiments 1 and 2), a two-person social interaction in a prisoner's dilemma game theory scenario (Experiment 3), and autobiographical experiences from participants' lives (Experiment 4). We showed that the valence effect on free will attributions generalized for outcomes (Experiment 1), the mere framing of an outcome (Experiment 2), and actions taken (Experiments 3 and 4), and using several measures, including free will attributions to current or recalled situation and a similar situation taking place in the future (Experiments 1 to 4), predictability (Experiments 3 and 4), and agentic counterfactuals (Experiment 4).

Free will as the capacity for change

The present results indicate that people perceive free will more strongly in connection with bad than good. The perception that someone is free and able to do otherwise is personally and socially useful, insofar as it emphasizes accountability and facilitates learning and possible

change. The present findings make sense in that context: It is seemingly most useful to perceive freedom of action when something goes wrong, because learning and change are most urgently desirable then. The very undesirability of negative actions or outcomes raises attention to the fact that the person really could have (and probably should have) done otherwise. That conclusion dovetails well with the research on counterfactual thinking: People engage in counterfactual thinking by reflecting on bad actions and bad outcomes, and the benefits of such thinking are derived from thinking about how one could have acted differently (Roese, 1997). The unique form of action control that humans exercise, which in layperson perspective corresponds to free will, may well be an adaptation to enable people to improve themselves so as to function better in society and, in turn, enable society to function better, thereby benefiting the group (e.g., Baumeister, 2008). People may only reflect on it when faced with subpar or negative outcomes, or when an agent behaves badly or makes bad decisions — because those are the cases in which it is most obvious that by acting otherwise, the person could benefit self and society.

Attributions of free will versus intent

Our findings extend previous literature regarding asymmetries in attributions. Knobe (2003) showed that people attribute more intent to a decision maker whose choices produced a bad than a good side effect that he had explicitly said was not his intention (he expressed complete indifference). Experiments 1 and 2 held intent and action constant while manipulating outcome or framing of an outcome. Experiment 5 directly addressed the question of the difference between intent and free will. The relationship between free will and intent attributions was not significant and negative. Thus, previous findings about attributions of intent do not explain the present findings.

Attributions of intent and free will are conceptually different. Intention is a mental representation of purposive action, and as such it could exist without free will. Meanwhile, some concepts of free will (e.g., random action; see Brembs, 2011) could operate without intention, simply by enabling the agent to make a different choice in the moment. Many concepts of free will also invoke the absence of external coercion, or even opposition to external pressure, which is largely irrelevant to intention. In the Knobe (2003) dilemma, the decision maker was presented as being fully able to do what he chose, and so his level of free will was conceptually the same across conditions, whereas people judged his intention quite differently depending on outcomes.

Furthermore, intent is internal and people know whether they intended for something to happen or not, and therefore the Knobe Effect is not meaningful for self attributions. The effect regarding the self can only be interpreted as the capacity for choice, not as intent. This is best exemplified in the Experiment 4 recall task, as it relates to real everyday life actions taken by the self, and the intent by the self is known.

We therefore conclude that intent and free will attributions are different, and that the valence asymmetry effect found is unique, together allowing for fuller understanding of intentionality (Alicke, 2000; Malle & Knobe, 1997) with possible implications for recent theories about the attribution of blame and responsibility (Malle et al., 2014).

Free will extends beyond morality

We theorized that higher free will attributions would be elicited in a wide array of situations. Recent findings have linked the concept of free will serves mainly to moral responsibility and to the need to punish wrongdoers (Clark et al., 2014; Shariff et al., 2014), and we readily acknowledge the importance of free will in moral judgment. However, there is more

to free will than moral choice. Our findings extend beyond morality and punishment in multiple ways.

First, the actions assessed in Experiments 1 to 4 were morally neutral. In Experiment 1, the action was the same across conditions and therefore did not involve a contrast between moral and immoral, and indeed if anything the moral goal of saving lives was held constant across conditions. Experiment 2 went even further in minimizing moral variation, insofar as the both the action and the outcome were the same across conditions, only varying the framing of the outcome. Furthermore, in both experiments the two possible decisions were not between a moral and an immoral decision but rather a decision between a risk-seeking versus risk-averse decision — both undertaken in the service of the morally commendable goal of saving lives. In Experiment 3 the decision made was between prosocial versus selfish behavior, both morally acceptable within the set rules of the game. Experiment 4 used a sample of autobiographical memories from everyday life, and none of the stories involved any sort of blatantly immoral behavior.

Second, we examined free will attributions for both self and others. The hypothesis that people attribute free will to negative actions and outcomes may not be easily extended to cases in which the agent is the self., which might reflect a potential bias in attributions to self in order to try and reduce own responsibility and avoid punishment. If a key factor driving free will beliefs and attributions is the need to punish wrongdoers, then we would expect people to act in accordance with self-serving bias (e.g., Kunda, 1987; Zuckerman, 1979), by which people seek to take credit for success but deny blame for failure. Similarly, negative agency bias (Baumeister, Stillwell, & Wotman, 1990; Morewedge, 2009) argues that people tend to attribute success to internal factors and failures to external causes or to an agent. If so, positive actions

taken by the self might elicit higher attributions of free will to elicit praise and receive credit ("I could have done bad, but I made the decision to do good") and negative actions to lower attributions of free will in order to reduce feelings of guilt or possible social punishment (external: "I was forced to do it", or internal: "I was drunk/mentally insane"). On that basis, one would predict that people would attribute high free will to others who caused negative outcomes but not to themselves for producing negative outcomes. However, we found no evidence indicating such a bias. In fact, the attributions were for the most part similar for the self and other (Experiments 1 to 3). While this null difference cannot rule out the possibility of the above prediction, the best available evidence suggests that the core of free will attributions and belief is more about the assumption of responsibility as to envision change and enable learning. Punishment and retribution can be seen as one of several possible mechanisms to facilitate such learning, which are rendered ineffective when one does not perceive the capacity for change.

Implications and Future Directions

The present investigation focused on cognitive biases in attributions of free will, and it is plausible that these attributions may interact with beliefs, motivations, and affect. For example, Nichols and Knobe (2007) have shown that contemplating moral responsibility of emotionally valenced crimes led to more compatibilist attributions (attributing responsibility in a deterministic universe). Possibly, contemplating more affective situations would lead to an even stronger positive-negative bias in attributions of free will. Future studies may examine impact of affect for free will attributions and whether generalized beliefs in free will would moderate the positive-negative asymmetry bias.

The positive-negative asymmetry in free will attributions may also be related to lay-assumptions regarding human nature, specifically whether people are inherently good or

inherently bad. The predictability measures in Experiments 3 and 4 showed that people find bad actions to be more unpredictable than good ones, which according to some scholars (Pronin & Kugler, 2010; Brembs, 2011) is a measure indicative of more free will, because it suggests that people expect good behavior and are surprised by bad actions or bad outcomes. Future research may examine the interaction of these implicit lay-beliefs of agency and human nature against one another to see the effect those may have over attributions, teasing apart the two effects. It is possible that the two lay-beliefs may interact so that free will would be attributed to negative situations when assumptions are that human nature is good, but that this effect might be reversed if the assumption is that human nature is bad.

In Experiment 4 using a free recall of everyday real-life situations has also revealed that people tend to attribute higher free will to others than to themselves. Experiments 1-3 in which the alternative actions were controlled did not show a similar effect, which suggests that complex situations may involve additional biases in free will attributions. These findings also seem to counter Pronin and Kugler's (2010) finding that people believe they have more free will than other people. Pronin and Kugler acknowledged an alternative account for their findings regarding the self/other main effect: People may see predictability as undesirable and therefore may try to protect their self-image by rejecting being predictable. The interaction found between valence and agent using the predictability measure in Experiment 4 may suggest that participants were indeed acting to maintain their self image, but that in this case they emphasized wanting to be seen as being inherently good, as they report their negative actions to be less predictable than their good actions. We note, however, that Pronin and Kugler's research differs from this study in several key respects. In particular, they examined neutral situations in life rather than explicitly good or bad actions (e.g., school, career, romance, social life, and everyday life). Moreover, the

actions in their study held no clear implications. In contrast, the actions in the present study involve direct consequences for the other party. Quite possibly, people may perceive their actions as having more free will than others' for neutral everyday actions, which would bolster their sense of being capable or deserving — but when actions involve valence and other parties, the effect is weakened or even reversed. Future studies may more closely examine moderating factors for the self-–other bias in attributions of free will and the role of the positive–negative asymmetry.

Experiments 1 and 2 may also offer an insight into the classical paradigms of the framing effect (Tversky & Kahneman, 1981), potentially shedding light over an unexplored factor involved in decision making. A possible interpretation of our findings regarding the framing effect could be that the tendency to undertake riskier decisions under negative framing may be related to the perception that negative situations involve a greater ability to see other options or to choose non-conforming or unexpected options. A negative context may be a cognitive trigger to perceived free will, thereby leading one to consider taking more risks (Hills, Noguchi, & Gibbert, 2013). Therefore, the association between risk seeking or avoidance behaviors and free will deserves further exploration. Future studies may attempt to examine free will attributions for different negative outcomes.

Considerable evidence indicates that bad actions and bad events have a stronger impact than good ones (for reviews, see Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001). The present findings extend that work to show that bad is also seen as freer than good ones. However, there are important differences between these two effects, as 'stronger' situations usually imply more rather than less constraints, and hence less capacity for free choice. If the implications of a bad action are perceived as more impactful and with higher anticipated

affect, then this suggests that people would feel more constrained and hesitant to engage, therefore suggesting the perception of less free will to act. Future research could examine the interaction of these two effects and their possible implications.

Conclusion

Although it may at first seem disappointing that people seem to associate freedom of action with negative circumstance, that perception may actually serve an important role for both self and for societal functioning. Our findings suggest that invoking the concept of free will in connection with negative actions or outcomes facilitates the contemplation of what went wrong, the evaluation of what could have been done differently, and the perception that things could be done differently in the future. Ultimately, free will highlights the importance of being able to learn, evolve, and act differently.

Acknowledgements

The research was supported by the RGC General Research Fund (HKUST 644312) and UGC Infra-Structure Grant (SBI14BM23) awarded to Kin Fai Ellick Wong.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*(4), 556-574. doi:10.1037/0033-2909.126.4.556

Alquist, J. L., Ainsworth, S. E., & Baumeister, R. F. (2013). Determined to conform: Disbelief in free will increases conformity. *Journal of Experimental Social Psychology, 49*(1), 80-86. doi:10.1016/j.jesp.2012.08.015

Alquist, J. L., Ainsworth, S. E., Baumeister, R. F., Daly, M., & Stillman, T. F. (2015). The making of might-have-beens: Effects of free will belief on counterfactual thinking.

Personality and Social Psychology Bulletin, 41(2), 268-283. doi:10.1177/0146167214563673

Barlas, Z., & Obhi, S. S. (2013). Freedom, choice, and the sense of agency. *Frontiers in human neuroscience, 7*.

Baumeister, R. F. (2008). Free will in scientific psychology. *Perspectives on Psychological Science, 3*(1), 14-19.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323.

Baumeister, R. F., & Brewer, L. E. (2012). Believing versus disbelieving in free will: Correlates and consequences. *Social and Personality Psychology Compass, 6*(10), 736-745.

Baumeister, R. F., Masicampo, E., & DeWall, C. N. (2009). Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin, 35*(2), 260-268.

Baumeister, R. F., Stillwell, A., & Wotman, S. R. (1990). Victim and perpetrator accounts of interpersonal conflict: Autobiographical narratives about anger. *Journal of Personality and Social Psychology, 59*(5), 994.

Bode, S., Murawski, C., Soon, C. S., Bode, P., Stahl, J., & Smith, P. L. (2014). Demystifying "free will": the role of contextual information and evidence accumulation for predictive brain activity. *Neuroscience & Biobehavioral Reviews, 47*, 636-645.

Boninger, D. S., Gleicher, F., & Strathman, A. (1994). Counterfactual thinking: From what might have been to what may be. *Journal of Personality and Social Psychology, 67*(2), 297-307. doi:10.1037/0022-3514.67.2.297

Brembs, B. (2011). Towards a scientific concept of free will as a biological trait: Spontaneous actions and decision-making in invertebrates. *Proceedings of the Royal Society B: Biological Sciences, 278*(1707), 930-939. doi:10.1098/rspb.2010.2325

Chernyak, N., & Kushnir, T. (2013). The self as a moral agent: Preschoolers behave morally, but believe in the freedom to do otherwise. *Journal of Cognition and Development*, doi:10.1080/15248372.2013.777843

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F., & Baumeister, R. F. (2014). Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology, 106*(4), 501-513. doi:10.1037/a0035880

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353-380. doi:http://dx.doi.org/10.1016/j.cognition.2008.03.006

Dennett, D. C. (2003). The self as a responding—and responsible—artifact. *Annals of the New York Academy of Sciences, 1001*(1), 39-50.

Edwards, J. (1754/1957). *Freedom of the will*.

Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review, 12*(2), 168-192. doi:10.1177/1088868308316091

Feldman, G., Baumeister, R. F., & Wong, K. F. E. (2014). Free will is about choosing: The link between choice and the belief in free will. *Journal of Experimental Social Psychology, 55*(0), 239-245. doi:http://dx.doi.org/10.1016/j.jesp.2014.07.012

Feldman, G., Chandrashekar, S. P., & Wong, K. F. E. (2016). The freedom to excel: Belief in free will predicts better academic performance. *Personality and Individual Differences, 90*, 377-383.

Feldman, G., Chao, M. M., Farh, J. L., & Bardi, A. (2015). The motivation and inhibition of breaking the rules: Personal values structures predict unethicality. *Journal of Research in Personality*, *59*, 69-80.

Girotto, V., Ferrante, D., Pighin, S., & Gonzalez, M. (2007). Postdecisional counterfactual thinking by actors and readers. *Psychological Science*, *18*(6), 510-515. doi:10.1111/j.1467-9280.2007.01931.x

Gray, K., Knickman, T. A., & Wegner, D. M. (2011). More dead than dead: Perceptions of persons in the persistent vegetative state. Cognition, 121(2), 275-280.

Greene, J., & Cohen, J. (2004). For the Law, Neuroscience Changes Nothing and Everything. Philosophical Transactions: Biological Sciences, 1775-1785.

Guglielmo, S., & Malle, B. F. (2010). Enough skill to kill: Intentionality judgments and the moral valence of action. *Cognition, 117*(2), 139-150. doi:http://dx.doi.org/10.1016/j.cognition.2010.08.002

Haggard, P., Mele, A., O'Connor, T., & Vohs, K. D. (2010). "Lexicon of key terms." big questions in free will project. Retrieved from http://www.freewillandscience.com

Helzer, E. G., & Gilovich, T. (2012). Whatever is willed will be A temporal asymmetry in attributions to will. *Personality and Social Psychology Bulletin, 38*(10), 1235-1246.

Hills, T. T., Noguchi, T., & Gibbert, M. (2013). Information overload or search-amplified risk? set size and order effects on decisions from experience *Psychonomic Bulletin & Review*, *20*(5), 1023-1031. doi:10.3758/s13423-013-0422-3; 10.3758/s13423-013-0422-3

Hume, D. (1748). *An enquiry concerning human understanding*.

Kane, R. (2002). *Free will: New directions for an ancient problem*. Blackwell Publishers Malden.

Kane, R. (2011). *The oxford handbook of free will*. Oxford University Press.

Kant, I. (1788/1997). Critique of practical reason. Translated and edited by Mary J. Gregor.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*(279), 190-194.

Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology, 53*(4), 636.

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

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*(2), 101-121. doi:http://dx.doi.org/10.1006/jesp.1996.1314

Malle, B. F., Knobe, J. M., & Nelson, S. E. (2007). Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology, 93*(4), 491-514. doi:10.1037/0022-3514.93.4.491

Monroe, A. E., & Malle, B. F. (2010). From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will. *Review of Philosophy and Psychology, 1*(2), 211-224.

Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and Cognition, 27*(0), 100-108. doi:http://dx.doi.org/10.1016/j.concog.2014.04.011

Morewedge, C. K. (2009). Negativity bias in attribution of external agency. *Journal of Experimental Psychology: General, 138*(4), 535.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology, 18*(5), 561-584.

Nichols, S. (2004). The folk psychology of free will: Fits and starts. *Mind & Language, 19*(5), 473-502.

Nichols, S. (2006). Folk intuitions on free will. *Journal of Cognition and Culture, 6*(1), 57-86.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous, 41*(4), 663-685.

Paulhus, D. L., & Carey, J. M. (2011). The FAD–Plus: Measuring lay beliefs regarding free will and related constructs. *Journal of Personality Assessment, 93*(1), 96-104.

Phillips, J., & Knobe, J. (2009). Moral judgments and intuitions about freedom. *Psychological Inquiry, 20*(1), 30-36. doi:10.1080/10478400902744279

Pronin, E., & Kugler, M. B. (2010). People believe they have more free will than others. *Proceedings of the National Academy of Sciences, 107*(52), 22469-22474.

Rapoport, A., & Chammah, A. M. (1965). Sex differences in factors contributing to the level of cooperation in the prisoner's dilemma game. *Journal of Personality and Social Psychology, 2*(6), 831.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin, 121*(1), 133.

Roese, N. J., & Olson, J. M. (1997). Counterfactual thinking: The intersection of affect and function. (pp. 1-59). San Diego, CA, US: Academic Press. doi:10.1016/S0065-2601(08)60015-5

Roskies, A. (2006). Neuroscientific challenges to free will and responsibility. *Trends in Cognitive Sciences, 10*(9), 419-423.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296-320.

Seligman, M. E. P., Railton, P., Baumeister, R. F., & Sripada, C. (2013). Navigating into the future or driven by the past. *Perspectives on Psychological Science, 8*(2), 119-141. doi:10.1177/1745691612474317

Shariff, A. F., Greene, J. D., Karremans, J. C., Luguri, J. B., Clark, C. J., Schooler, J. W., . . . Vohs, K. D. (2014). Free will and punishment: A mechanistic view of human nature reduces retribution. *Psychological Science*. doi:10.1177/0956797614534693

Stillman, T. F., & Baumeister, R. F. (2010). Guilty, free, and wise: Determinism and psychopathy diminish learning from negative emotions. *Journal of Experimental Social Psychology, 46*(6), 951-960.

Stillman, T. F., Baumeister, R. F., & Mele, A. R. (2011). Free will in everyday life: Autobiographical accounts of free and unfree actions. *Philosophical Psychology, 24*(3), 381-394.

Stillman, T. F., Baumeister, R. F., Vohs, K. D., Lambert, N. M., Fincham, F. D., & Brewer, L. E. (2010). Personal philosophy and personnel achievement: Belief in free will predicts better job performance. *Social Psychological and Personality Science, 1*(1), 43-50.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458.

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will encouraging a belief in determinism increases cheating. *Psychological Science, 19*(1), 49-54.

Weisel, O., & Shalvi, S. (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences, 112*(34), 10651-10656.

Wong, K. F. E., & Cheng, C. (2013) Predictable or not? Individuals' risk decisions do not necessarily predict their next ones. *PLoS ONE 8(2),* e56811.

Zuckerman, M. (1979). Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. *Journal of Personality, 47*(2), 245-287.

Tables

Table 1

Experiment 1 - Means and standard deviations for the manipulation check and free will attributions

	Manipulation check – valence			Free will attributions – current situation			Free will attributions – future situation		
	Negative	Positive	Total	Negative	Positive	Total	Negative	Positive	Total
Self	−36.62 (68.95) [47]	78.73 (38.00) [51]	23.41 (79.71)	3.87 (.90)	3.27 (1.23)	3.56 (1.12)	4.13 (.85)	3.43 (.98)	3.76 (.98)
Other	−62.18 (58.02) [55]	75.80 (35.80) [51]	4.21 (84.50)	3.78 (1.01)	3.12 (1.24)	3.46 (1.17)	4.16 (.83)	3.33 (1.21)	3.76 (1.11)
Total	−50.40 (64.26)	64.26 (36.76)		3.82 (.96)	3.20 (1.24)		4.15 (.84)	3.38 (1.10)	

Note. Parentheses indicate standard deviation. Brackets indicate number of participants. See Figure 2 for the plots of free will attributions.

Table 2

Experiment 2 – Means and standard deviations for the manipulation check and free will attributions

	Manipulation check – valence			Free will attributions – current situation			Free will attributions – future situation		
	Negative	Positive	Total	Negative	Positive	Total	Negative	Positive	Total
Self	6.30 (64.84) [54]	47.97 (41.35) [59]	28.05 (57.55)	3.59 (1.09)	2.98 (1.12)	3.27 (1.14)	3.72 (1.12)	3.17 (1.16)	3.43 (1.17)
Other	3.24 (69.89) [51]	59.16 (34.65) [57]	32.75 (60.80)	3.45 (1.12)	2.84 (1.11)	3.12 (1.15)	3.47 (.95)	3.23 (1.17)	3.34 (1.07)
Total	4.81 (67.03)	53.47 (38.45)		3.52 (1.10)	2.91 (1.12)		3.60 (1.04)	3.20 (1.16)	

Note. Parentheses indicate standard deviation. Brackets indicate number of participants. See Figure 3 for the plots of free will attributions.

Table 3

Experiment 3 - Means and standard deviations for the free will attributions

	Free will attributions – current situation			Free will attributions – future situation			Free will attributions – predictability		
	Negative	Positive	Total	Negative	Positive	Total	Negative	Positive	Total
Self	3.42 (1.26) [19]	3.40 (.97) [47]	3.41 (1.05)	3.84 (1.17)	3.36 (1.26)	3.50 (1.24)	3.05 (1.03)	2.62 (.99)	2.74 (1.01)
Other	4.00 (.75) [26]	3.24 (1.17) [45]	3.52 (1.09)	3.73 (.87)	3.20 (1.16)	3.39 (1.09)	3.23 (.86)	2.64 (1.17)	2.86 (1.10)
Total	3.76 (1.03)	3.33 (1.07)		3.78 (1.00)	3.28 (1.21)		3.16 (.93)	2.63 (1.08)	

Note. Parentheses indicate standard deviation. Brackets indicate number of participants. See Figure 4 for the plots of free will attributions.

Table 4

Experiment 4 - Means and standard deviations for the manipulation check and free will attributions for current and future situations

	Free will attributions – recalled situation			Free will attributions – future situation		
	Negative	Positive	Total	Negative	Positive	Total
Self	3.68 (1.00) [53]	2.48 (1.09) [54]	3.07 (1.20)	3.33 (.98)	2.55 (1.04)	2.94 (1.08)
Other	3.94 (1.03) [52]	3.41 (1.05) [53]	3.68 (1.07)	3.15 (1.16)	2.79 (1.04)	2.97 (1.11)
Total	3.81 (1.02)	2.94 (1.16)		3.25 (1.07)	2.67 (1.04)	

	Free will attributions – agentic counterfactuals			Free will attributions – predictability		
	Negative	Positive	Total	Negative	Positive	Total
Self	1.36 (1.26)	0.91 (1.12)	1.13 (1.20)	3.42 (.95)	2.59 (.92)	3.00 (1.02)
Other	1.92 (2.26)	1.42 (1.47)	1.67 (1.91)	2.67 (1.04)	2.83 (.99)	2.75 (1.02)
Total	1.64 (1.84)	1.16 (1.33)		3.05 (1.06)	2.71 (.96)	

Note. Parentheses indicate standard deviation. Brackets indicate number of participants. See Figure 5 for the plots of free will attributions.

Table 5

Experiment 5 - Correlations table

	M	SD	Free will	Intent	Action
Free will attributions	73.27	32.36	-		
Intent valence	−9.11	75.06	−.10	-	
Action valence	−7.60	81.95	−.27***	.64***	-
Outcome valence	−.33	94.48	−.17**	.17**	.58***

Note: * $p < .05$; ** $p < .01$; *** $p < .001$.

Table 6

Summary of findings

Ex	N	Sample	IV1	IV2	DV	Behavior	Details	Contributions
1	204	MTurk	Outcome valence	Self / Other	FW attributions past + future	High versus low risk options	Asian Disease scenario, *Tversky and Kahneman (1981)*, high risk option	Baseline effect: Negative -> higher FW attributions
2	221	MTurk	Outcome Framing valence	Self / Other	FW attributions past + future	High versus low risk options	Asian Disease scenario, *Tversky and Kahneman (1981)*, low risk option	1. Framing effect 2. Addressing # of deaths confound
3	137	MTurk	Action valence	Self / Other	FW attributions past + future + predictability	Cooperation versus defection	Prisoner's dilemma: *Rapoport and Chammah (1965)*	Action focus: fixed situation, fixed actions + predictability measure
4	212	Students	Action valence	Self / Other	FW attributions: past + future + predictability + ACF	Ordinary everyday life behaviors	Recalled real-life interactions	Real life situations + agentic counterfactuals measure
5	301	MTurk	Intent, Action, & Outcome	-	FW attributions	Moral behavior	Hospital scenario, *Phillips and Knobe (2009)*	Contrasting intent, action and outcome with FW attributions

Note. FW = free will; ACF = agentic counterfactuals.

Figures

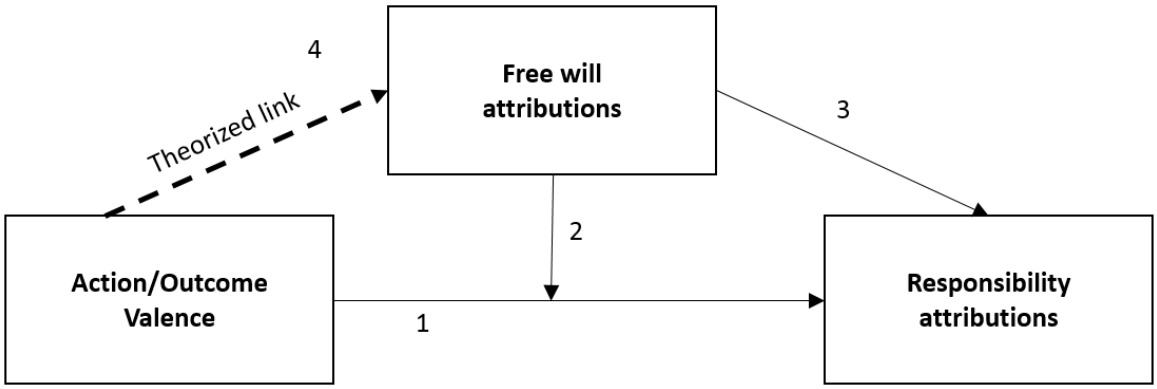

Figure 1. Free will attribution model: Attributions of responsibility for actions/outcomes depend on the perceptions of free will (#1 and #2). Theorized link: The conceptual link between free will and responsibility (#3) leads to a cognitive association between valence and free will attributions (#4) - negative actions and outcomes trigger higher attributions of free will.

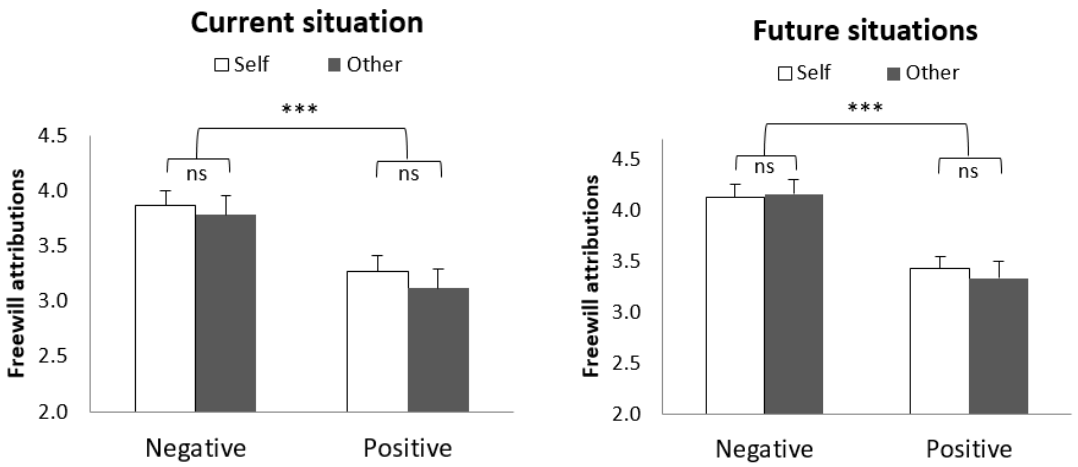

Figure 2. Experiment 1 - attributions plot. Error bar indicates standard error. *** $p < .001$, ns $p > .05$. The top comparison is for the main effect contrasting positive and negative for self–other combined. The bottom comparison is between the self–other conditions within the positive and negative conditions.

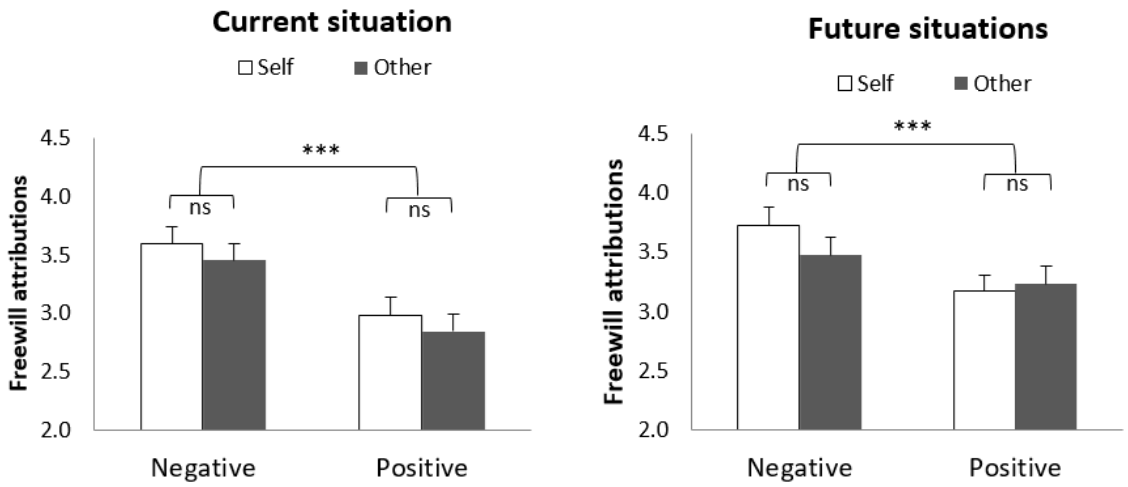

Figure 3. Experiment 2 attributions plot. Error bar indicates standard error. *** $p < .001$, ns $p > .05$. The top comparison is for the main effect contrasting positive and negative for self–other combined. The bottom comparison is between the self–other conditions within the positive and negative conditions.

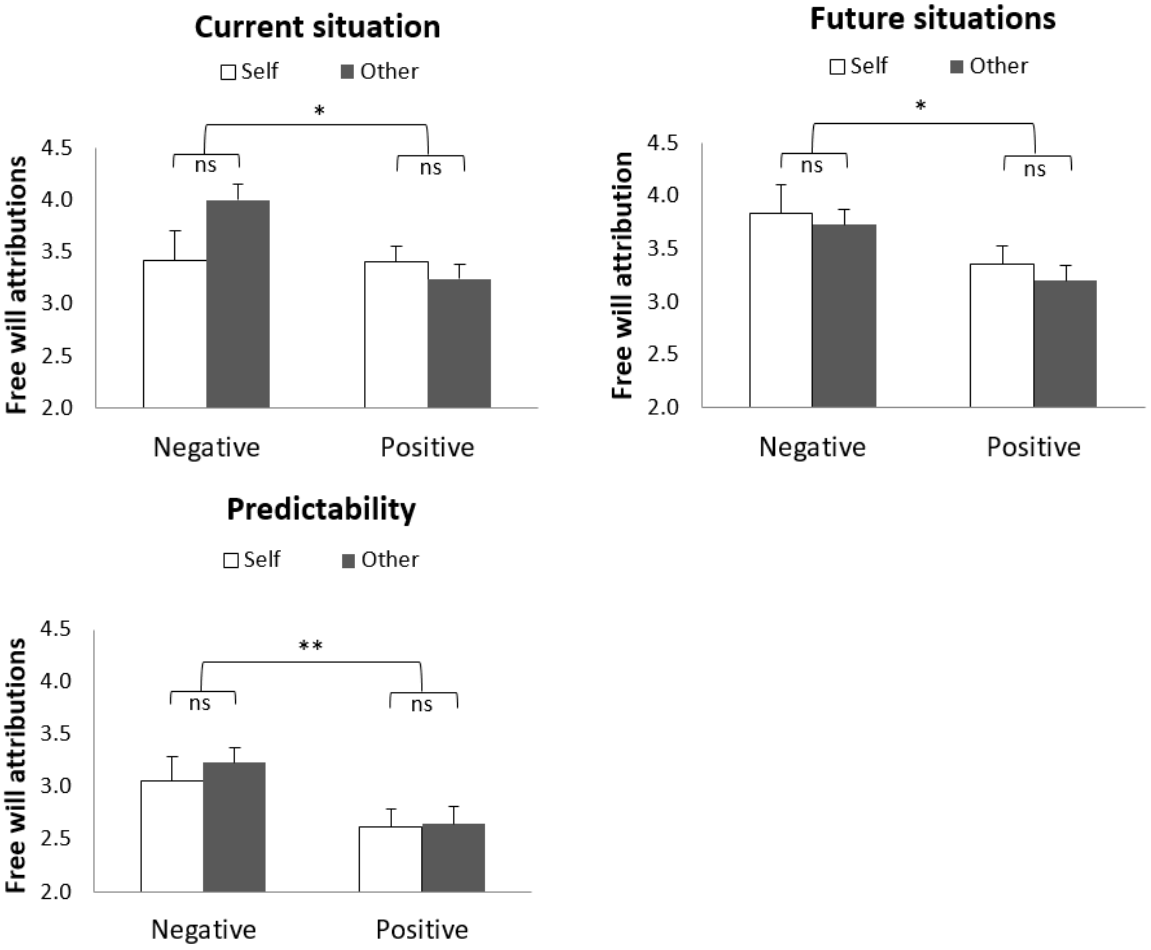

Figure 4. Experiment 3 - attributions plot. Error bar indicates standard error. *** $p < .001$, ** $p < .01$, * $p < .05$, ns $p > .05$. The top comparison is for the main effect contrasting positive and negative for self–other combined. The bottom comparison is between the self–other conditions within the positive and negative conditions.

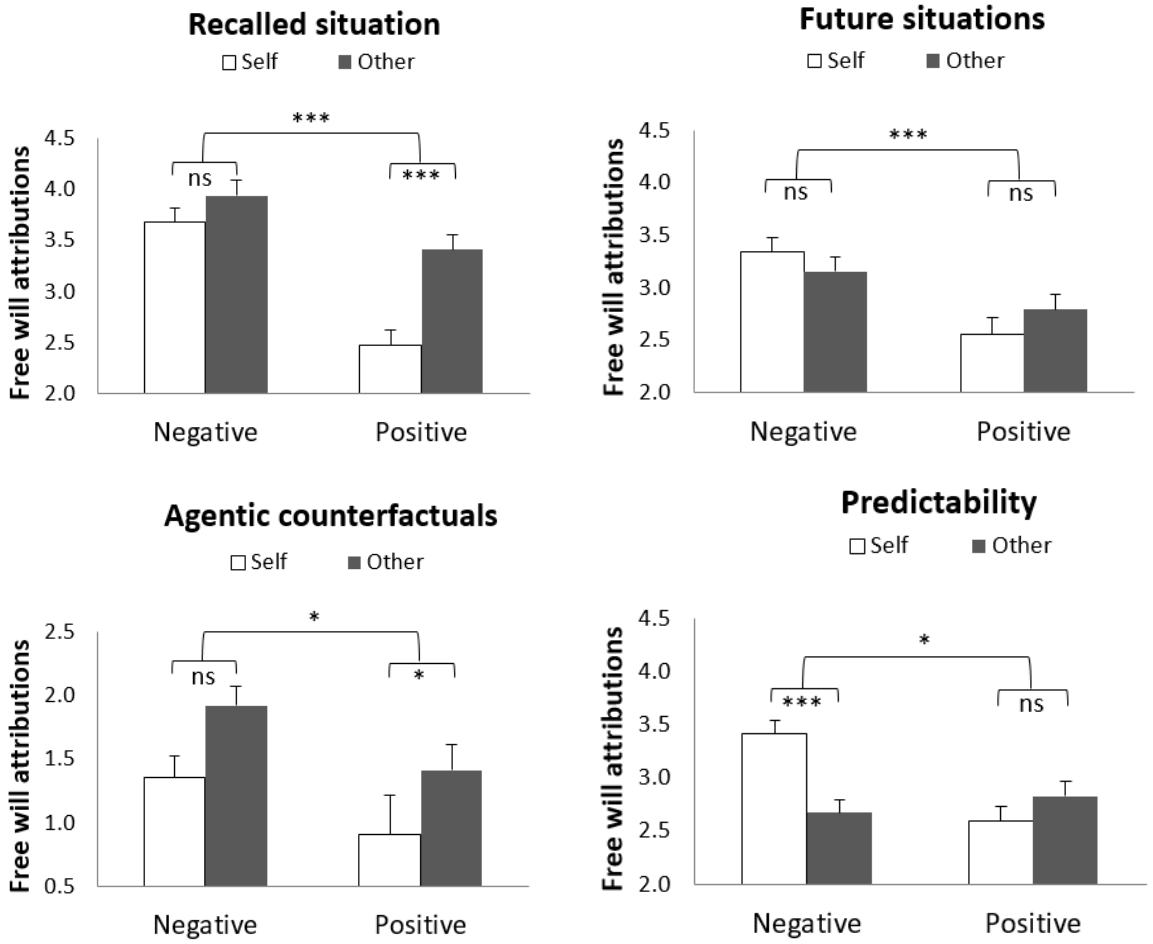

Figure 5. Experiment 4 - attributions plot. Error bar indicates standard error. *** $p < .001$, ** $p < .01$, * $p < .05$, ns $p > .05$. The top comparison is for the main effect contrasting positive and negative for self–other combined. The bottom comparison is between the self–other conditions within the positive and negative conditions.