Memory & Cognition (2022) 50:1103–1123
https://doi.org/10.3758/s13421-022-01315-0

Cognitive processes in imaginative moral shifts: How judgments of morally unacceptable actions change

Beyza Tepe[1] · **Ruth M. J. Byrne**[2]

Accepted: 15 April 2022 / Published online: 9 May 2022

Abstract
How do people come to consider a morally unacceptable action, such as "a passenger in an airplane does not want to sit next to a Muslim passenger and so he tells the stewardess the passenger must be moved to another seat", to be less unacceptable? We propose they tend to imagine counterfactual alternatives about how things could have been different that transform the unacceptable action to be less unacceptable. Five experiments identify the cognitive processes underlying this imaginative moral shift: an action is judged less unacceptable when people imagine circumstances in which it *would have been* moral. The effect occurs for immediate counterfactuals and reflective ones, but is greater when participants create an immediate counterfactual first, and diminished when they create a reflective one first. The effect also occurs for unreasonable actions. We discuss the implications for alternative theories of the mental representations and cognitive processes underlying moral judgments.

Keywords Imagination · Counterfactuals · Morality · Judgments

Introduction

How do people change their minds about moral matters? People generally tend to be conservative about changing their moral decisions (e.g., Haidt, 2001; Stanley et al., 2018). In five experiments, we examined how quickly people change their minds about a situation they initially judged as morally unacceptable, when they were prompted to think about alternative possibilities. Suppose a man in an airplane about to take his seat in front of you calls the stewardess and says he does not want to sit next to a Muslim passenger seated in the row, and he tells the stewardess the passenger must be moved to another seat. To what extent is his behavior morally acceptable, in your opinion? Suppose you decide it is not at all morally acceptable, but you try to imagine very quickly whether there could be some circumstances in which his behavior would be morally acceptable. You might wonder whether the Muslim passenger had been rude to him so the reason he wanted him to be moved wasn't to do with the man's religion. Given this imagined alternative circumstance, to what extent do you now think the man's behavior is morally acceptable?

Our first question is whether immoral actions become less morally unacceptable when people imagine how they *would have been* moral, even in the absence of further facts. We aim to examine whether alternative possibilities in which an action could be considered morally justified rather than immoral, are rapidly accessible or whether the moral imagination requires effortful reflection. Intuition and deliberation may interact in various ways (e.g., Bialek & De Neys, 2017; Evans & Stanovich, 2013; Greene et al., 2001; Gürcay & Baron, 2017; Kahneman, 2011), and intriguingly, people seem to assume that immoral events are impossible at first sight (e.g., Phillips et al., 2015; Phillips et al., 2019; Phillips & Cushman, 2017). The more possible an event is, the more moral it is judged to be (e.g., Shtulman & Tong, 2013). Hence, we test the hypothesis that people can rapidly access counterfactual alternatives that go against their initial judgment that an action is immoral.

✉ Beyza Tepe
beyza.tepe@eas.bau.edu.tr

Ruth M. J. Byrne
rmbyrne@tcd.ie

[1] Department of Psychology, Bahçeşehir University, Ciragan Cd., Osmanpasa Mektebi Sk. 4-6, 34349, Besiktas, Istanbul, Turkey

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

Our second question is about unreasonable actions. Suppose the man had instead told the stewardess he did not want to sit next to any passenger and told her to move everyone in the row to somewhere else. To what extent is this behavior reasonable? Can you think of alternative circumstances in which it would have been rational? Our question is, granted that immoral actions can come to be considered less morally unacceptable, can unreasonable actions seem less irrational? Given that both immoral and irrational behaviors violate prescriptive norms, and irrational actions are also sometimes considered impossible at first sight (e.g., Phillips & Cushman, 2017), we test the hypothesis that people can also rapidly access counterfactual alternatives that go against their initial judgment that an action is irrational.

Of course, people can update their moral judgments as they learn new information, since the context of the action can change the valence of their evaluation (e.g., Andrejević et al., 2020; Simpson et al., 2016; Tepe & Aydinli-Karakulak, 2019). Our interest is whether they can do so even in the absence of further information, merely as a result of imagining alternatives. The current theories that guide our hypothesis are drawn from three separate lines of inquiry: moral judgment updating, everyday non-monotonic reasoning, and counterfactual imagination, which we consider in turn.

A key issue for theories of moral judgment has been to explain how people can tolerate exceptions to their belief in a moral principle, such as doing no harm to others, when a dilemmic conflict exists, such as needing to protect others (e.g., Goodwin, 2017; Gray & Keeney, 2015; Haidt, 2012; Mikhail, 2013; Schein & Gray, 2018). People's moral judgments are affected by beliefs about human agency (e.g., Alicke, 2000), intentionality (e.g., Cushman, 2008; Malle & Holbrook, 2012), outcome severity (e.g., Mazzocco et al., 2004), social relationships (e.g., Tepe & Aydinli-Karakulak, 2019), and by emotions (e.g., Greene et al., 2009; Greene & Haidt, 2002) and justifications (e.g., Haidt, 2001; Malle et al., 2014). Such studies clarify many of the effects of contextual information on moral judgments, but leave unanswered the question of whether people have rapid access to alternative possibilities that go against their initial moral judgments. In many circumstances, reasoning may justify, rather than revise, immediate moral judgments that have arisen from effortless, affective processes (e.g., Haidt, 2001, 2007; Luo et al., 2006; Royzman et al., 2015; Sanfey et al., 2003; Tepe et al., 2016; Ugazio et al., 2012; Valdesolo & DeSteno, 2006; Wheatley & Haidt, 2005). In some circumstances, reasoning may itself contribute to immediate moral judgments (e.g., Bloom, 2010; Haidt, 2012; Horne et al., 2015; Maki & Raimi, 2017; Paxton et al., 2012; Paxton & Greene, 2010; Wiech et al., 2013). Nonetheless, people tend not to change their initial moral judgments when presented with opposing reasons (e.g., Stanley et al., 2018), and they tend to resist negotiation on moral issues (e.g., Skitka, 2010; Skitka et al., 2005; Turiel, 2002). Updating initial moral judgments in the absence of sufficient reasons seems counterintuitive and may require some effort (e.g., Haidt, 2001, 2012). Hence, our question is, granted that immoral actions can become less morally unacceptable when people imagine how they *would have been* moral, are such counterfactual alternatives to initial moral judgments immediately accessible?

The first tenet of our theory draws on the idea that inferences are non-monotonic (i.e., defeasible), in that conclusions can be withdrawn on receipt of further information, through cognitive processes that combine premise information with background knowledge (e.g., Cariani & Rips, 2017; Espino & Byrne, 2020; Oaksford & Chater, 2018; Stenning & Van Lambalgen, 2012), so that belief revision maintains epistemically entrenched beliefs (Elio & Pelletier, 1997; Gärdenfors, 1988; Pollock, 1987). We suggest that people interpret exceptions to moral principles as counterexamples, incorporated into a cohesive model by constructing arguments to reconcile the premise that justified a conclusion (the passenger's behavior towards the Muslim man was discriminatory, therefore morally unacceptable), with additional knowledge that refines assumptions (the passenger may have been reacting to the man's rudeness), to ensure the conclusion is no longer warranted (the behavior was not discriminatory, so it is less morally unacceptable). The revision of belief in the original conclusion maintains the entrenched moral principle, by identifying a different set of facts within which to interpret the behavior, or a competing moral principle that trumps it (given that people may be resistant to imagining changes to the norm itself, e.g., Gendler, 2000).

The second tenet is that the counterfactual imagination provides one of the missing mechanisms contributing to the non-monotonicity of moral inferences. People often think about how things could have been different, especially after bad outcomes and unexpected events (e.g., Byrne, 2016; Kahneman & Tversky, 1982; Markman et al., 1993; Roese, 1997). They create models to mentally simulate actions and their outcomes (e.g., Byrne & Johnson-Laird, 2020; Byrne & Timmons, 2018; Cushman, 2013; Kahneman & Tversky, 1982; Markman et al., 2008). What they select to change in their models of reality to create a counterfactual alternative depends on the availability of alternatives, guided by norms about what is usual – physically, socially, morally, and intrapersonally – including descriptive norms based on statistical averages and prescriptive ones based on moral ideals (e.g., Bear & Knobe, 2017; Halpern & Hitchcock, 2015; Henne et al., 2019; McCloy & Byrne, 2000; Phillips et al., 2015; Roese, 1997). Abnormal events recruit their normal counterparts from memory and the retrieved default possibilities may be sampled for those that are morally good (e.g., Kahneman & Miller, 1986; Khemlani et al., 2018; Phillips et al.,

2019; Phillips & Cushman, 2017). Although it is established that counterfactuals can amplify moral judgments so that a morally bad event is considered to be even worse (e.g., Alicke, Buckinghman, Zell, & Davis, 2008; Lench et al., 2014; Malle et al., 2014; Migliore et al., 2014; Parkinson & Byrne, 2017, 2018; Timmons & Byrne, 2018), a gap in current theories is whether counterfactuals can *reverse* a moral judgment, so that a bad event is judged to be less bad. People can update moral judgments when they are explicitly provided with additional information, for example, about known reasons for an action, but whether they can do so on the basis of *imagined* counterfactual circumstances is untested (e.g., Cone & Ferguson, 2015; Mann & Ferguson, 2015; Monroe & Malle, 2019; Sabo & Giner-Sorolla, 2017; Stanley et al., 2018). People can imagine mitigating circumstances but their tendency to do so is affected by the emotions elicited by a moral transgression (e.g., Piazza et al., 2013). Yet people tend to imagine how things could be better rather than worse (e.g., De Brigard et al., 2013a, b; Rim & Summerville, 2014), and so it is plausible that they can replace something morally bad in the actual world with something morally good in an imagined alternative. Of course, it is possible to imagine an alternative scenario that is better in some way to what actually happened, but which does not necessarily alter the moral appropriateness of the event. Nonetheless, since the default representation of possibilities can be framed by morality, the rapid imagination of the possibilities for a moral behavior may be feasible (e.g., Phillips & Cushman, 2017). We suggest that immediate counterfactuals deliver the moral counterpart of an immoral action (such as that the man was not acting in a discriminatory manner), enabling rapid access to reasons (he was reacting to rudeness instead).

We consider that the process by which people construct counterfactual alternatives may not require effortful reflection. Counterfactual explanations may be constructed by processes comprised of immediate access to default possibilities, or reflective construction of considered arguments. A gap in current theories is how intuitive counterfactuals relate to deliberative ones (e.g., Goldinger et al., 2003; Roese et al., 2005). In contrast, extensive evidence has been gathered about whether moral judgments are guided by intuition or reason (e.g., Greene et al., 2004; Greene et al., 2008; Haidt, 2012; Luo et al., 2006; Moore et al., 2008; Sanfey et al., 2003; Suter & Hertwig, 2011), and on the role of emotions in moral judgment (e.g., Haidt et al., 1993; Russell & Giner-Sorolla, 2011a, 2011b; Ugazio et al., 2012; Wheatley & Haidt, 2005). No agreement currently exists on how dual processes of intuition and reason may interact. For example, they could occur sequentially, with poorer quality, fast intuitions overridden by better quality slower reflections (e.g., Evans & Stanovich, 2013). Alternatively, they could occur independently, even in parallel, with quick responses not always wrong, and slow ones not always right, and conflict detection occurring regardless of response choice (e.g., Bialek & De Neys, 2017; Bucciarelli et al., 2008; Gubbins & Byrne, 2014; Gürcay & Baron, 2017; Shtulman & Tong, 2013; Stupple & Ball, 2008; Trippas et al., 2017). Immediate counterfactuals could deliver the moral counterpart of an immoral action (the man was not acting in a discriminatory manner), enabling rapid access to reasons (he was reacting to rudeness instead). Subsequent reflection, rather than overriding an immediate thought with a competing one, can develop it into an elaborate counterargument, that is, the dual processes could operate in sequential co-operation rather than only in sequential competition.

In five experiments we tested these proposals by asking people to judge the moral acceptability of a set of immoral actions, to try to imagine alternative ways each one would be moral, and to provide their judgments of them again. If moral judgments are rigidly anchored in values driven by automatic processes, their judgments will remain immovable before deliberative reflection; if they are open to moderation by justification, then an imaginative shift will occur effortlessly, to the action being considered less morally unacceptable; our theory predicts the latter.

To address the second question of whether counterfactual explanations are immediately available or require reflection, we asked people to try to imagine and describe in writing alternative ways in which the behavior would be moral either very quickly or to take their time to reflect carefully (see Fig. 1). To address the question of whether counterfactuals affect immoral actions differently from irrational ones, we asked people to judge the moral acceptability of immoral actions and the rationality of unreasonable ones. If the representation of possibilities tends toward moral and rational actions, the same pattern should be found for both.

Experiment 1

The aim of the experiment was to examine whether a person's judgment of the morality of an immoral action changes after they have imagined how the action *would have been* moral. We compared an "immediate" condition in which participants imagined alternatives, under a time constraint of 20 s in which they had to read about the behavior and make their judgment, to another "reflective" condition in which they imagined alternatives under no time constraints. We included a third control condition in which participants did not imagine alternatives but instead carried out the task of providing a title that describes the action (see Fig. 2). The factual task was intended to engage participants in comparable elaborative processing by requiring them to consider a description that succinctly summarized the behavior (e.g., Bransford & Johnson, 1972). It controls for the potential of a demand characteristic, that when individuals are asked to

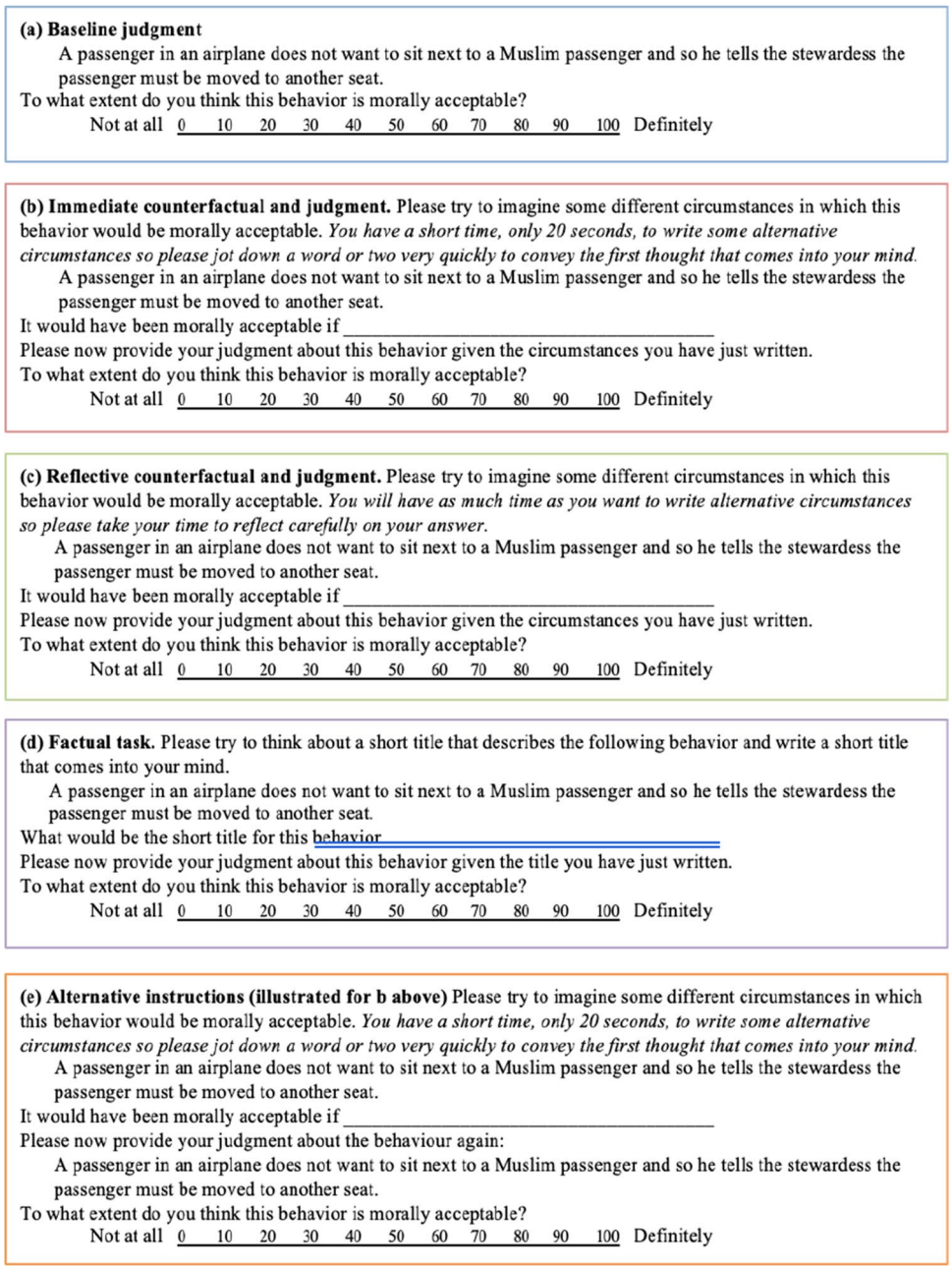

Fig. 1 Schematic representation of the experimental trials. (**a**) Example of a baseline moral judgment. In each experiment participants completed judgments in the baseline phase first. (**b**) Example of the immediate counterfactual task: Participants imagined some different circumstances and completed a counterfactual sentence stem task for each action in 20 s, and then made their judgment of it again on the next screen. (**c**) Example of a reflective counterfactual task: Participants completed the counterfactual task for each action taking as much time as they required and then made their judgment of it again. (**d**) Example of a factual task: Participants wrote a short title for each action and then made their judgment of it again. (**e**) Example of alternative instructions designed to re-focus on the described behavior rather than the counterfactual circumstances, illustrated for the immediate counterfactual task

re-evaluate their initial judgments a second time, they may believe they are expected to change their initial judgment. The experiment also examined whether a person's judgment of the rationality of an unreasonable action changes after they have imagined how the action *would have been* rational.

Method

Participants The participants were 186 US and UK volunteers recruited through the online platform Prolific who received £0.85 sterling for participation, and there were 131

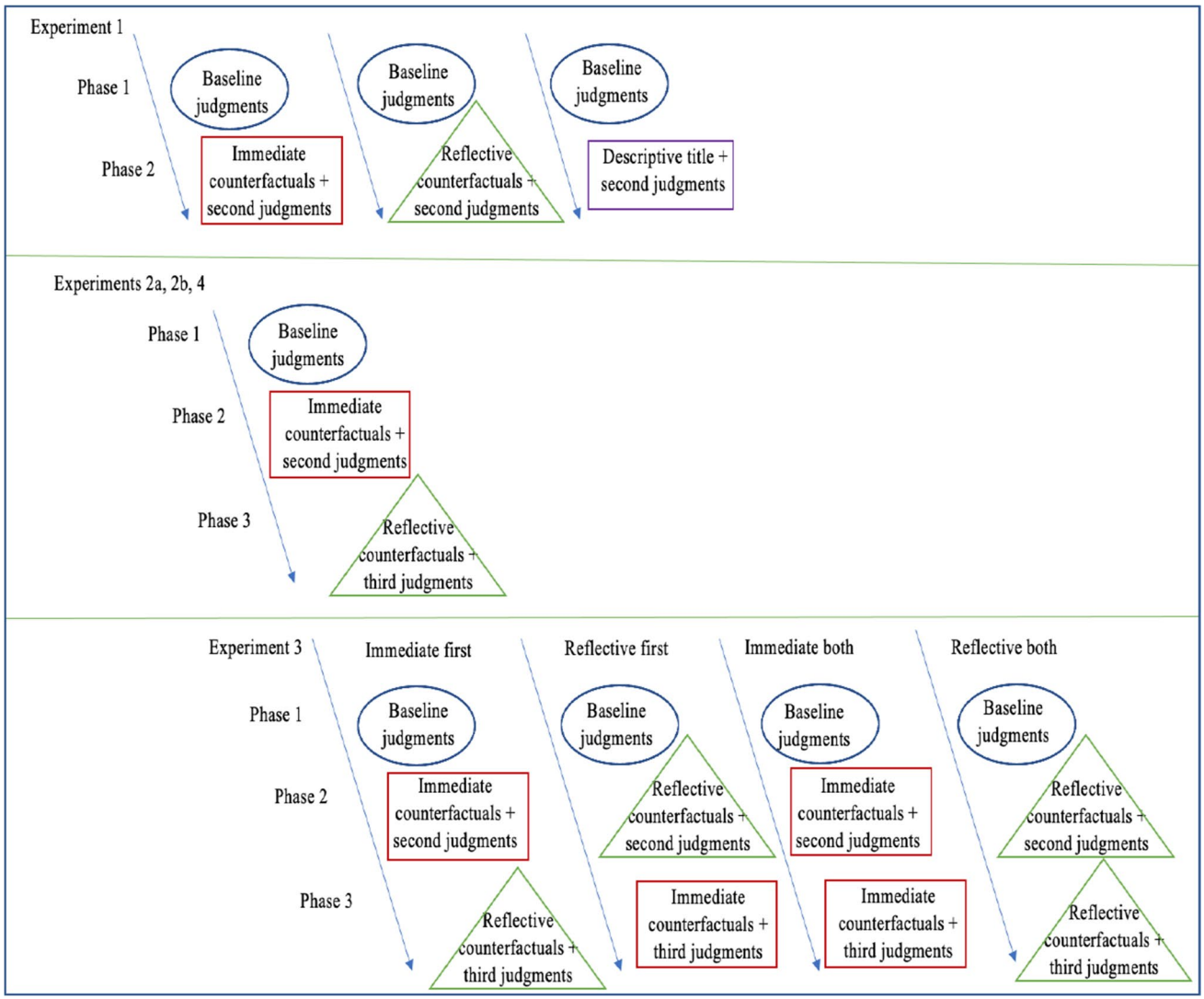

Fig. 2 Schematic representation of the experimental designs. Illustration of the sequence of events in the experiments. In all experiments, participants judged the moral acceptability of a set of immoral actions, or the rationality of a set of unreasonable actions. In the first phase, they made their baseline judgments, in the subsequent phases, they carried out a counterfactual task for each action and provided their judgment of it again. The counterfactual task required participants to complete a sentence stem 'it would have been morally acceptable if…' for the immoral actions (or 'it would have been rational if…' for the irrational actions). In the immediate counterfactual condition, they were required to do so in 20 s, in the reflective counterfactual condition they did so with no time constraints. In Experiment 1, participants completed two phases only, and the second phase was either immediate or reflective, or a factual control task, in a between-participants design. In Experiments 2a, b, and 4, participants completed three phases, the second phase was immediate and the third phase was reflective, in a within-participants design. In Experiment 3 participants completed three phases, and they corresponded to either the immediate-first sequence of the previous experiments or to a reflective-first, immediate-both or reflective-both sequence

women, 54 men and one non-binary individual, with a mean age of 34 years and an age range of 18–73 years. They were assigned at random to six groups (for distribution across groups see Table S1 in the Online Supplementary Materials (OSM)). The planned sample sizes were motivated by power analyses conducted with G*power (Faul et al., 2009). A sample size of 162 participants is required to provide at least 80% power to detect a medium-sized effect at $p < .05$, to test a main effect of immediate versus reflective counterfactuals on individuals' judgments in an analysis of variance (ANOVA) with the design of 2 (judgment phase: first vs. second judgments) × 3 (counterfactual task: immediate, reflective, factual) × 2 (judgment content: judgments of morality for immoral actions vs. judgments of rationality for irrational actions) with repeated measures on the first factor. We restricted access to Prolific participants who were native English speakers, above 18 years of age, and who answered correctly a "robot-detection" picture-matching

question (one participant was eliminated because of their incorrect answer). Prior to any data analysis we eliminated participants from the recruited 209 participants who failed to complete all the tasks (12 participants), failed the attention check question (five participants), or who had carried out a similar study (five participants). The experiments received prior approval from the School of Psychology Ethics Committees of Trinity College Dublin and Istanbul University. For all the studies, the participants gave their informed consent, and we report all of our manipulations and measures. After the experiments participants completed several demographic and personality measures (see OSM).

Materials and design A set of immoral and irrational scenarios were used, and each consisted of a single sentence that contained a scene-setting clause and an action (see OSM). The materials were three immoral actions, and three matched unreasonable actions, adapted from the previous literature (Phillips & Cushman, 2017; Tepe & Aydinli-Karakulak, 2019). The materials were presented to the participants in their native language of English.

The key measures were answers to two questions: "To what extent do you think this behavior is morally acceptable?" and "To what extent do you think this behavior is rational?" Participants provided their judgments on a 0–100 slider scale with 0 labelled "not at all" and 100 labelled "definitely." As a control to ensure that all participants were exposed to the same sorts of judgments for every action, they judged not only the morality of immoral actions but also their rationality, and not only the rationality of irrational actions but also their morality (and their judgments to these additional measures, which were highly correlated, are provided in Table S1 in the OSM).

The experiment included *judgment content* as a between-participants factor: participants either received the set of immoral actions, or the set of irrational actions. The *type of task* participants carried out – immediate-counterfactual, reflective-counterfactual, or factual title – was the second between-participants factor. Every participant first provided their judgments about to what extent the actions were moral or to what extent the actions were rational for the set of actions, each presented on a separate screen, in the "baseline" phase (see Fig. 1a). To examine the effects of counterfactual thoughts they then received instructions on a separate screen, for either the immediate-counterfactual condition (see Fig. 1b), the reflective-counterfactual condition (see Fig. 1c), or the factual control condition (see Fig. 1d). In the "immediate" condition, a 20-s counter counted down on screen and the program moved on to the next screen automatically after 20 s. The timer started as soon as the participant moved to the screen and so the 20 s allowed includes the time taken to read the instructions and the scenario (see Fig. 1b). Twenty seconds is thus a very short time indeed to try to imagine alternatives and to jot them down, given that even to read the instructions and the scenario takes that long for the average reader.[1]

In the "reflective" condition, there was no time restriction. In the control "factual" condition, participants wrote a short title for each action and no time restriction was applied. Every participant produced judgments first in phase 1 (the baseline judgments) and second in phase 2 (after their task – immediate, reflective, or factual), and thus *judgment phase* was a within-participants variable. Hence the design included the between-participants factors of judgment content (immoral, irrational) and type of task (immediate-counterfactual, reflective-counterfactual, or factual task), and the within-participants factor of judgment phase (first vs. second judgments).

Participants completed two judgments (moral acceptability and rationality) for three actions (either immoral or irrational) in the baseline phase and the same again in the second phase, i.e., 12 judgments in total. The order of judgments was randomized in all experiments. To control for order effects, the materials were presented in two different randomized orders in each experiment, and no order effects were found in any experiment (see the section on the full statistical tests in the OSM).

Procedure The materials were presented online using Qualtrics in each experiment. Participants received instructions for each study that it aimed to examine how people think about various events and that it was a study of everyday judgment, in which the aim was to examine the sorts of answers that most people provide. They were asked to take part only if they were willing to consider the tasks seriously, and instructed that they should do the study in a quiet place where they would be uninterrupted for its duration. Each task was presented on a separate screen and participants could not return to an earlier screen once they had provided their judgment.

Results and discussion

The datasets for this experiment and the subsequent experiments are available via the Open Science Framework at: https://osf.io/mw94z/.

We compared the judgments of the morality of immoral actions and the rationality of irrational actions in an ANOVA with the design of 2 (*judgment phase*: first vs. second judgments) × 3 (*counterfactual task*: factual, immediate, reflective) × 2 (*judgment content*: judgments of morality for immoral actions vs. judgments of rationality for irrational actions) with repeated measures on the first factor. In this experiment and subsequent ones when assumptions of homogeneity of variance were violated, we corrected degrees of freedom using the Greenhouse-Geiser

[1] The time was chosen based on the duration taken by five volunteers to try to read the instruction and a scenario and jot down an imagined alternative.

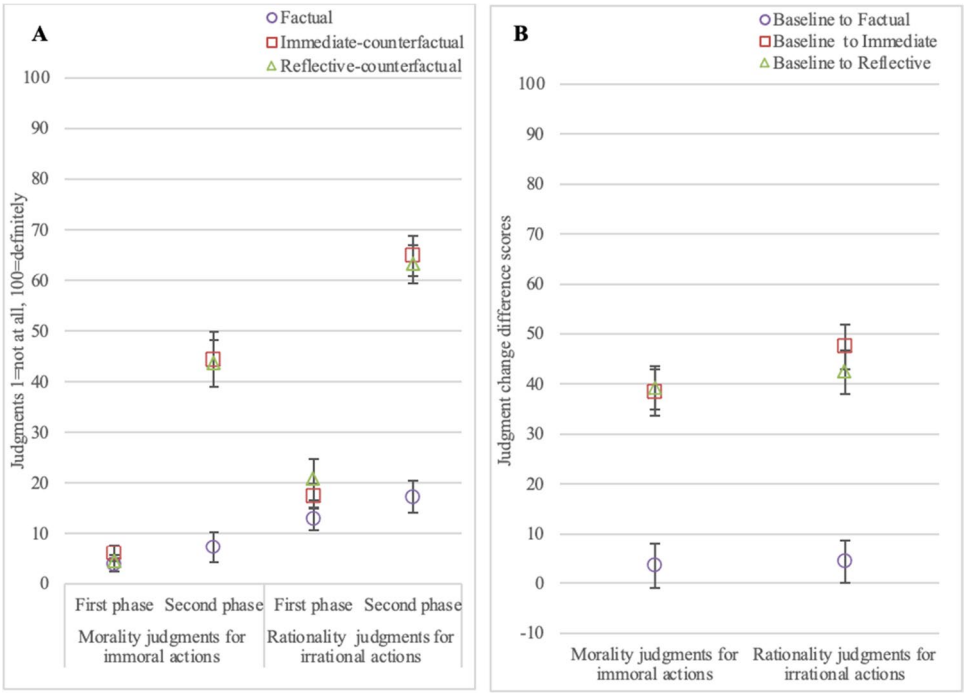

Fig. 3 In Experiment 1 participants created either immediate or reflective counterfactuals, or else constructed a factual title for the action. In (**A**) their mean judgments for the first and second phase are presented for the moral acceptability of immoral actions and the rationality of irrational actions. In (**B**) the difference scores for the judgment change from the first phase to the second are presented. Plots of data in Experiment 1 are based on 186 UK and US participants. Error bars are standard error of the mean

and Welch-Satterthwaite corrections as appropriate. The results showed that immoral actions were judged to be less unacceptable when people imagined how they could have been moral. Participants' judgments increased in the second phase compared to the first, as shown by a main effect of judgment phase, F (1,180) = 259.59, $p < .001$, $\eta p^2 = .591$, 90% CI (0.516, 0.647), and they increased when they created an immediate or reflective counterfactual compared to a factual title, as shown by a main effect of task, F (2, 180) = 57.404, $p < .001$, $\eta p^2 = .389$, 90% CI (0.300, 0.463). There was also a main effect of judgment content, F (1,180) = 51.652, $p < .001$, $\eta p^2 = .223$, 90% CI (0.139, 0.305), as judgments of the morality of immoral actions were lower than judgments of the rationality of irrational actions. Nonetheless the same pattern was observed for both sorts of content, which did not interact with phase, F (1,180) =1.416, $p = .236$, $\eta p^2 = .008$, 90% CI (0.000, 0.042) or task, F (2, 180) = 1.569, $p = .211$, $\eta p^2 = .017$, 90% CI (0.000, 0.053), and there was no interaction of all three variables, F (2,180) = 0.452, $p = .637$, $\eta p^2 = .005$, 90% CI (0.000, 0.026), see Fig. 3a.

Judgment phase and counterfactual task interacted, F (2,180) = 48.973, $p < .001$, $\eta p^2 = .352$, 90% CI (0.258, 0.428) as Fig. 3a shows. We decomposed the interaction with a Bonferroni-corrected alpha of $p < .0056$ for nine comparisons. The comparisons showed that the increase in participants' judgments in the second phase compared to the first occurred only when participants created immediate counterfactuals, t (58) = 12.957, $p < .001$, $d = 1.687$, 95% CI (1.285, 2.082), and reflective ones, t (63) = 10.424, $p < .001$, d =1.303, 95% CI (0.966, 1.634), but not when they thought of a title, t (62)=2.305, $p = .025$, $d = 0.29$, 95% CI (0.037, 0.541), on the corrected alpha of $p < .0056$. This result shows that the increase in judgments cannot be attributed to extraneous factors such as repetition, practice, or task demands.

The comparisons to decompose the interaction of phase and task also showed that participants' judgments increased in the second phase when they created an immediate counterfactual compared to a title, t (99.818) = 10.117, $p < .001$, d = 1.833; 95% CI (1.393, 2.266), and when they created a reflective counterfactual compared to a title, t (114.163) = 10.501, $p < .001$, d = 1.864, 95% CI (1.437, 2.284), but there was no difference between creating an immediate or reflective counterfactual, t (121) = 0.369, $p = .713$, d = 0.067, 95% CI (−0.288, 0.420), see Fig. 3a. Jotting down a few words quickly to convey the first thought that comes to mind was as effective as taking time to reflect carefully. Finally, as an important baseline, we confirmed that there were no differences in first-phase judgments in the three conditions: factual versus immediate, t (120) = 1.636, p = .105, d = 0.29, 95% CI (−0.060, 0.639); factual versus reflective, t (103.557) = 1.513, $p = .133$, d = 0.269, 95% CI (−0.082, 0.618); and immediate versus reflective, t (121) =

Table 1 Examples of different ways participants imagined an immoral action would have been moral, or an irrational action would have been rational, illustrated for one of the actions

Immoral action: *"A person does not want to sit beside a Muslim passenger on a plane and so he tells the stewardess the passenger must be moved to another seat"*.
1. FACTS: Action is *not* immoral (or not irrational)– other facts explain it
e.g., "if the Muslim passenger had been rude". (Action not in response to religion)
2. DILEMMA: Action *is* immoral, but is in response to a dilemmic conflict with another moral action
e.g., "if the Muslim passenger had been acting threateningly". (Action justified to protect others)
3. ALTERNATIVE NORMS: Action is moral in another possible world which has different norms
e.g., "It would have been acceptable if this action was morally good in some society".
4. OPPOSITE: It would have been moral if the action had not been taken
e.g., "it would have been acceptable if he had sat beside the Muslim passenger".
5. RESIST: Refusal to engage in imagination.
e.g., "it is never right to do this".
Irrational action: *"A person does not want to sit beside any passenger on a plane and so he tells the stewardess the passengers must be moved to other seats"*.
1. FACTS: Action is *not* unreasonable– other facts explain it
e.g., "if the person had been on a private plane".
2. DILEMMA: Action *is* unreasonable, but is in response to a dilemma with a competing action.
e.g., "if the person had a contagious disease".
3. ALTERNATIVE NORMS: Action is reasonable in another possible world which has different norms
e.g., "It would have been rational if this action was reasonable in some society".
4. OPPOSITE: It would have been reasonable if the action had not been taken
e.g., "it would have been rational if he had sat beside the passengers".
5. RESIST: Refusal to engage in imagination.
e.g., "it is never reasonable to do this".

0.198, $p = .843$, d = 0.036, 95% CI (−0.318, 0.389). (Welch-Satterthwaite df corrections were applied for judgments in the baseline and reflective phases that violated Levene's test for equality of variance, $p < .001$ in both cases.)

Judgment change scores To probe further, we constructed judgment-change scores (difference scores based on subtracting the mean judgment scores in the first phase from the second phase) for the three groups (see the OSM). We carried out a 3 (judgment change: baseline-to-factual, baseline-to-immediate, baseline-to-reflective) × 2 (judgment content: judgments of morality for immoral actions vs. judgments of rationality for irrational actions) ANOVA with repeated measures on the first factor, on the judgment change difference scores. Participants' judgments changed from the first phase to the second when they constructed an immediate counterfactual, or a reflective one, more than when they thought of a title, as shown by a main effect of judgment change, F (2, 180) = 48.971, $p < .001$, $\eta p^2 = .352$, 90% CI (0.258, 0.428), see Fig. 3b. Immoral actions became less morally unacceptable when participants imagined how they could have been moral, and the same was so for irrational actions. There was no main effect of judgment content, F (1, 180) = 1.416, $p = .236$, $\eta p^2 = .008$, 90% CI (0.000, 0.042), and no interaction, F (2, 180) = 0.452, $p = .637$, $\eta p^2 = .005$, 90% CI (0.000, 0.026).

Counterfactuals We categorized the counterfactuals participants created (see Table 1). Participants tended to create facts-based counterfactuals, for example, "the action would have been morally acceptable if the Muslim passenger had been rude," that indicate the action is not immoral because other facts explain it, or they created dilemma-based counterfactuals, for example, "the action would have been morally acceptable if the Muslim passenger had been acting threateningly," that indicate that the action *is* immoral, but it is in response to a dilemmic conflict with another moral action that justifies it, for example, to protect others. For immoral actions, participants produced more facts-based counterfactuals than dilemma-based ones, whereas for irrational actions they did the opposite. Counterfactual analyses are presented in the OSM.

Overall, the results show that people's judgment of the morality of an immoral action changes after they have imagined how the action *would have been* moral. An immoral action was considered less immoral when participants imagined alternatives for only 20 s, or when they deliberated in their imagination of alternatives with no time constraints. This

judgment shift did not occur when participants did not imagine alternatives but instead described the action. The same imaginative shift occurred for judgments of the rationality of an unreasonable action. The next experiments were designed to find out whether an additional moral shift occurs when participants reflect carefully, *after* the first thought that comes to mind.

Experiments 2a and 2b

The aim of the experiments was to examine whether there is an additional imaginative shift in judgments of the morality of an immoral action when participants first imagine alternatives for only 20 s, and then subsequently imagine alternatives under no time constraints (see Fig. 2). We also extend our material set to a larger set of immoral and irrational actions in Experiment 2a.

We extend the materials even further in Experiment 2b to examine not only possible actions but also impossible ones, for example, "A passenger in an airplane does not want to sit next to a Muslim passenger and so he tells the stewardess the passenger must be moved to the moon." In the previous experiment, we tested counterfactual possibilities about immoral or irrational behaviors that are physically possible, that is, they can happen in real life, and the results showed that people can rapidly imagine counterfactual possibilities that turn an immoral event into a less immoral one. Are counterfactual possibilities accessible even for situations that are physically impossible, that is, they cannot happen in real life? In both experiments, we again examine judgments of the rationality of unreasonable actions as well as judgments of the morality of immoral actions.

Method

Participants The participants in Experiment 2a were 164 students from the University of Istanbul who volunteered in return for course credits. There were 135 women and 29 men, with a mean age of 22 years and an age range of 18–59 years. They were assigned at random to two groups (see Table S3, OSM). We tested as many students as volunteered from the undergraduate module who were invited to participate. A sample size of 105 participants is required to provide at least 80% power to detect a medium-sized effect at $p < .05$ for the main effect of immediate versus reflective counterfactuals in the 3 (*judgment phase*: baseline, immediate-counterfactual, reflective-counterfactual) × 2 (*judgment content*: judgments of morality for immoral actions vs. judgments of rationality for irrational actions) design with repeated measures on the first factor. In Experiment 2b, 79 UK participants recruited through Prolific received £0.85 for participation, and there were 59 women and 20 men, with a mean age of 18 years and an age range of 18–33 years. They were assigned at random to four groups (see Table S5, OSM). In Experiment 2b a post hoc power test indicated the sample size provides 80% power to detect a large-sized effect at $p < .05$ for a main effect of possible versus impossible actions in the 3 (*judgment phase*: baseline, immediate-counterfactual, reflective-counterfactual) × 2 (*possibility*: possible actions vs. impossible actions) × 2 (*judgment content*: judgments of morality for immoral actions vs. judgments of rationality for irrational actions) design with repeated measures on the first factor (but only approximates 60% power to detect a medium sized effect), and so we consider this experiment an exploratory test and interpret its results with caution. The materials were presented to the participants in their native language of Turkish (Experiment 2a) or English (Experiment 2b).

None of the Turkish students had taken part in a similar study previously, and Prolific participants were excluded if they reported having done so (ten participants were removed from Experiment 2b). We restricted access to Prolific participants who were native English speakers, above 18 years of age, and who answered correctly a "robot-detection" picture-matching question (one participant was eliminated because they did not do so correctly). Also prior to any data analysis we eliminated participants who failed to complete all the tasks (one participant in Experiment 2a and seven in 2b), and those who failed the attention check question (five participants in 2a and none in 2b), resulting in 164 participants in Experiment 2a, and 79 participants in Experiment 2b.

Materials, design, and procedure The materials were similar to the previous experiment. Experiment 2a used a larger set of eight scenarios, which varied the content for immoral and unreasonable actions, to test further whether counterfactuals affect immoral actions differently from irrational ones. Experiment 2b used the same materials as Experiment 1, but we examined not only actions that are possible, but also matched ones that are impossible, at least in an everyday situation (see the OSM for details). The materials were chosen from a larger set tested in a pilot study (see Table S10, OSM).

The measures were also similar to the previous experiment except that in Experiment 2a participants were also asked: "To what extent is it possible to think of this behavior as morally acceptable/rational?" Their judgments to these additional measures are provided in Table S3 in the OSM, since whether the questions were phrased with certainty or in terms of possibility had no effect. Accordingly, in Experiment 2a participants completed four judgments (moral acceptability, rationality, moral possibility, rational possibility) for four actions (either immoral or irrational) in the three phases of baseline, immediate, and reflective, that is, 48 judgments in total. In Experiment 2b participants completed two judgments

(moral acceptability and rationality) for three actions (either immoral or irrational) in the three phases of baseline, immediate, and reflective phase, that is, 18 judgments.

The design of each experiment was similar to Experiment 1, with one main exception: Participants made judgments in a baseline phase, then they thought about some alternative circumstances for just 20 s and completed the judgments a second time; then they thought about alternative circumstances with no time constraints and completed the judgments a third time (see Fig. 2). In Experiment 2a the between-participant factor *judgment content* again had two levels (immoral, irrational) and the within-participant factor of *judgment phase* had three levels (baseline, immediate-counterfactual, reflective counterfactual), as participants carried out both the immediate task, and then the reflective task (see Fig. 2). In Experiment 2b there was an additional between-participants factor of *possibility*, to compare possible actions to impossible actions. The procedure in each experiment was the same as in Experiment 1.

Results and discussion

In Experiment 2a the ANOVA was a 3 (*judgment phase*: baseline, immediate counterfactual, reflective counterfactual) × 2 (*judgment content*: judgments of morality for immoral actions vs. judgments of rationality for irrational actions) design with repeated measures on the first factor. In Experiment 2b the ANOVA was a 3 (*judgment phase*: baseline, immediate-counterfactual, reflective-counterfactual) × 2 (*possibility*: possible actions vs. impossible actions) × 2 (*judgment content*: judgments of morality for immoral actions vs. judgments of rationality for irrational actions) design with repeated measures on the first factor.

Once again immoral actions were judged less unacceptable when people imagined how they could have been moral. Participants' judgments shifted as they progressed through the three phases, as shown by main effects of judgment phase, in Experiment 2a, F (1.606, 260.10) =194.526, $p < .001$, $\eta p^2 = .546$, 90% CI (0.480, 0.597), see Fig. 4a; and Experiment 2b, F (1.609, 120.69) =230.132, $p < .001$, $\eta p^2 = .75$, 90% CI (0.690, 0.794), see Fig. 4b. Judgment content also showed a main effect, in Experiment 2a, F (1, 162) =91.472, $p < .001$, $\eta p^2 = .361$, 90% CI (0.265, 0.443), and Experiment 2b, F (1, 75) =34.246, $p < .001$, $\eta p^2 = .313$, 90% CI (0.173, 0.432), as participants' judgments of the morality of immoral actions were lower than their judgments of the rationality of irrational actions. There was no main effect of the possibility or impossibility of the actions in Experiment 2b, F (1, 75) =0.770, $p = .383$, $\eta p^2 = .010$, 90% CI (0.000, 0.076), see OSM, Tables S3 and S5.

There was an interaction of judgment content with judgment phase, in Experiment 2a, F (1.606, 260.10) = 3.217, $p = .052$, $\eta p^2 = .019$, 95% CI (0.000, 0.053) (Greenhouse-Geiser corrections were applied to the degrees of freedom because of the violation of sphericity), and in Experiment 2b, F (1.609, 120.69) = 3.630, $p = .039$, $\eta p^2 = .046$, 90% CI (0.001, 0.114), see Fig. 4a and b. None of the other interactions in Experiment 2b were significant (see OSM).

We decomposed the interactions of judgment content with judgment phase using Bonferroni-corrected alphas of $p < .0056$ for nine comparisons in each experiment. Participants' judgments continued to shift when they created reflective counterfactuals in the third phase compared to immediate ones in the second phase, in Experiment 2a, for immoral actions, t (82) = 5.978, $p < .001$, $d = 0.66$, 95% CI (0.417, 0.892), and irrational actions, t (80) = 6.137, $p < .001$, $d = 0.682$, 95% CI (0.438, 0.922), and in Experiment 2b, for immoral actions, t (37) = 5.538, $p < .001$, $d = 0.873$, 95% CI (0.444, 1.243), although for irrational actions the difference was not significant on the corrected alpha of $p < .0056$, t (40) = 2.447, $p = .019$, $d = 0.382$, 95% CI (0.063, 0.697).

They also judged the immoral actions to be more immoral in the baseline condition compared to the immediate one, Experiment 2a: t (82) = 8.54, $p < .001$, d = 0.937, 95% CI (0.677, 1.194), Experiment 2b, t (37) = 8.72, $p < .001$, d = 1.415, 95% CI (0.958, 1.862), and compared to the reflective one, Experiment 2a t (82) = 11.341, $p < .001$, d = 1.245, 95% CI (0.956, 1.53), Experiment 2b, t (37) = 12.111, $p < .001$, d = 1.965, 95% CI (1.412, 2.507). Likewise, they judged the irrational actions to be more irrational in the baseline condition compared to the immediate one, Experiment 2a, t (80) = 9.431, $p < .001$, d = 1.048, 95% CI (0.774, 1.317), Experiment 2b, t (40) = 12.299, $p < .001$, d = 1.921, 95% CI (1.397, 2.436), and compared to the reflective one, Experiment 2a, t (80) = 11.784, $p < .001$, d = 1.309, 95% CI (1.010, 1.605), Experiment 2b, t (40) = 13.364, $p < .001$, d = 2.087, 95% CI (1.044, 3.131), see OSM for further comparisons.

Judgment change difference scores Judgment change difference scores from the immediate to the reflective phase were less than from the baseline to immediate phase, or from the baseline to the reflective phase, as shown by main effects of judgment change, in Experiment 2a, F (1.408, 228.14) = 75.330, $p < .001$, $\eta p^2 = .317$, 90% CI (0.236, 0.388), see Fig. 4c; and Experiment 2b, F (1.299, 97.445) = 112.013, $p < .001$, $\eta p^2 = .599$, 90% CI (0.493, 0.669), see Fig. 4d; and its interaction with content, in Experiment 2a, F (1.408, 228.14) = 5.135, $p = .014$, $\eta p^2 = .031$, 90% CI (0.003, 0.074), and Experiment 2b, F (1.299, 97.445) = 6.984, $p = .005$, $\eta p^2 = .085$, 90% CI (0.015, 0.178). For full details see the OSM.

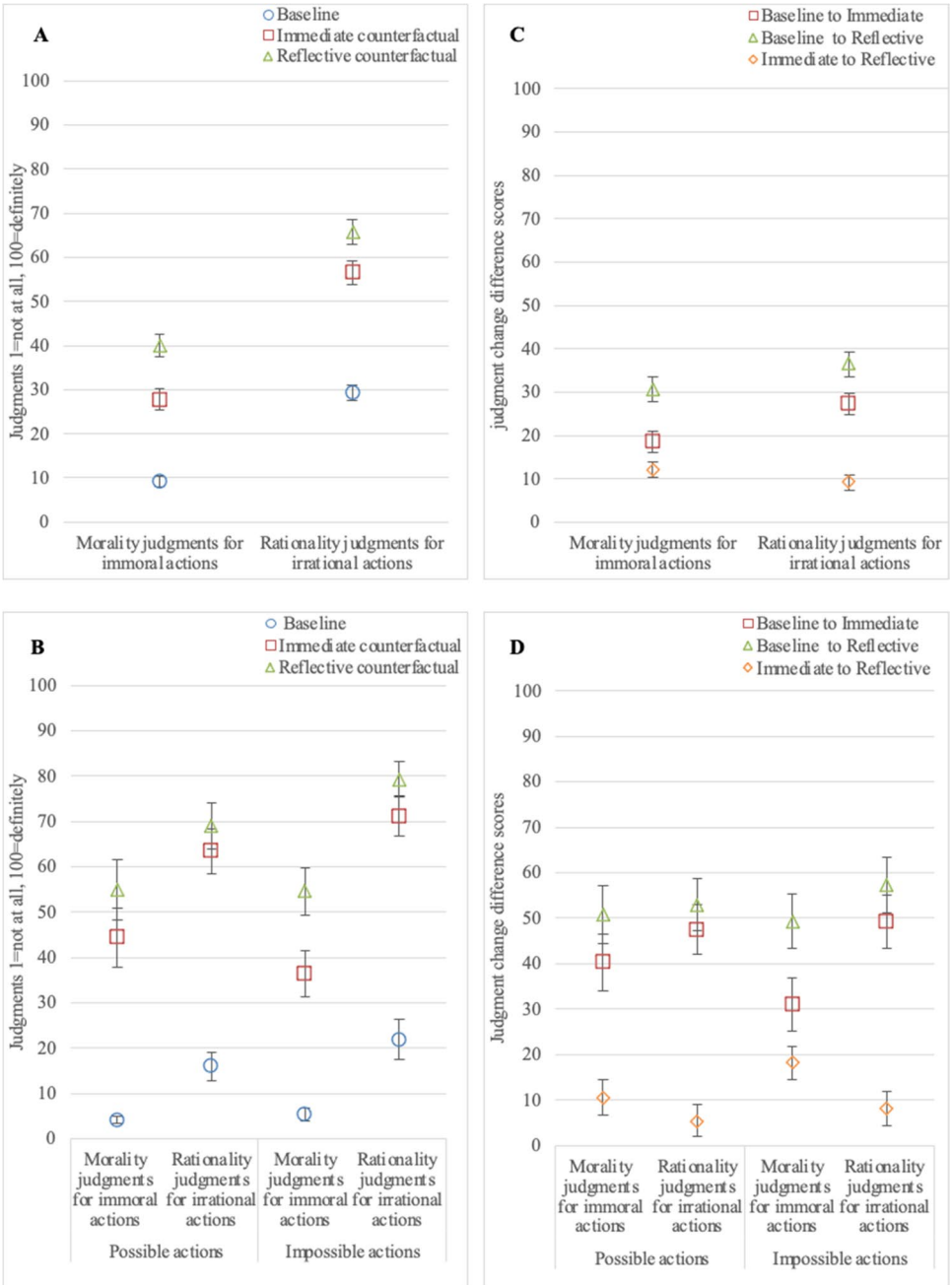

Fig. 4 In Experiments 2a and b participants constructed immediate counterfactuals and then reflective ones. Their mean judgments for the first, second, and third phase are presented for the moral acceptability of immoral actions and the rationality of irrational actions in (**A**) for Experiment 2a, and in (**B**) for Experiment 2b. The difference scores for the judgment change from one phase to another are presented in (**C**) for Experiment 2a, and in (**D**) for Experiment 2b. Plots of data in Experiment 2a are based on 164 students from the University of Istanbul, Turkey, and in Experiment 2b on 79 UK participants. Error bars are standard error of the mean

Counterfactuals Participants created reflective counterfactuals in which the action was not immoral because of other *facts*, more than ones in which it was immoral but in a *dilemma*, both for immoral and irrational actions in Experiment 2a; in Experiment 2b, they created as many facts-based as dilemma-based counterfactuals both for immoral and irrational actions. We also compared the counterfactuals participants created in the reflective phase to those they created in the immediate phase. Participants tended to focus on the *same* alternative circumstance in the reflective counterfactual as they had in the immediate counterfactual, rather than switch to a *different* alternative circumstance, in Experiment 2a and Experiment 2b, see the OSM for further details, including Tables S4 and S6.

The results show that there is an additional imaginative shift in judgments of the morality of an immoral action when participants first imagine alternatives for only 20 s, and then subsequently imagine alternatives under no time constraints. The results replicate the first experiment in showing a large shift in judgments following even just 20 s to imagine alternatives; nonetheless the results also show that there is an additional shift when participants subsequently deliberate with no time limits. The effect occurs for immoral actions and irrational ones, and not only for possible actions but also for impossible ones. Hence people tend to think about moral possibilities effortlessly, even when the moral possibilities go against their initial judgment, and even when the moral possibilities are physically impossible (see also Phillips & Cushman, 2017). Importantly, participants tended to elaborate further upon the same counterfactuals that they had created in the immediate condition, when they had no time limits in the reflective condition, rather than consider a different alternative.

Both experiments show that an additional imaginative shift occurs after reflecting carefully on alternatives. However, a potential concern is that the additional shift could arise given the opportunity to create any sort of second counterfactual, immediate or reflective. The next experiment addresses this issue.

Experiment 3

The aim of the experiment was to examine whether an additional imaginative shift in judgments of the morality of an immoral action occurs only when participants first imagine alternatives for 20 s, and then subsequently imagine alternatives under no time constraints, as in the previous experiments, or whether it also occurs when they first reflect with no time constraints, and then subsequently create counterfactuals in 20 s. Accordingly, we compared judgments made by participants who created immediate counterfactuals followed by reflective ones to those made by participants who created reflective counterfactuals followed by immediate ones. We also included two controls, one in which participants created immediate counterfactuals followed by immediate ones, and one in which they created reflective counterfactuals followed by reflective ones.

Method

Participants The participants were 355 students from Bahçeşehir University, Turkey, who volunteered in return for course credits. The participants were 264 women, 87 men, one gender-neutral individual, and three who did not provide information, with a mean age of 21 years and an age range of 17–51 years. They were randomly assigned to one of eight groups (see OSM, Table S7). The sample size had 98% power to detect a medium sized effect at $p < .05$, and we doubled the sample size for which we obtained an effect in Experiment 2a to enable us to test the predicted interaction (see Giner-Sorolla, 2018). None of the Turkish students had taken part in a similar study previously. Prior to any data analysis we eliminated participants who failed to complete all the tasks (41 participants) and those who failed the attention check question (13 participants), resulting in 355 participants.

Materials, design, and procedure We compared judgments made by participants who created immediate counterfactuals in a second phase and reflective ones in a third phase (as in Experiments 2a and 2b), to those made by participants who created counterfactuals in the opposite order, i.e., reflective counterfactuals in the second phase and immediate ones in the third phase. We included two controls, a sequence in which participants created immediate counterfactuals in both phases, and one in which they created reflective counterfactuals in both phases (see Fig. 2). The design contained the factors of *judgment content* and *judgment phase*, and in addition examined four types of *counterfactual sequence*: immediate-first, reflective-first, immediate-both, and reflective-both. The materials were the same as those in Experiment 1 and the procedure was also the same.

Results and discussion

The ANOVA was a 3 (*judgment phase*: first, second, and third judgments) × 4 (*counterfactual sequence*: immediate-first, reflective-first, immediate-both, reflective-both) × 2 (*judgment content*: judgments of morality for immoral actions vs. judgments of rationality for irrational actions) design with repeated measures on the first factor.

Immoral actions were more acceptable when people imagined how they could have been moral. Participants' judgments shifted as they progressed through the three judgment phases, as a main effect of phase showed, $F(1.767,613.04) = 624.325$, $p < .001$, $\eta p^2 = .643$, 90% CI (0.608, 0.671), see Fig. 5a. However, there was also a main effect of sequence, $F(3,347) = 2.784$, $p = .041$, $\eta p^2 = .024$, 90% CI (0.001, 0.049), as judgments were highest when participants constructed two reflective counterfactuals and lowest when they constructed two immediate counterfactuals; and judgment phase interacted with sequence, $F(5.3, 613.04) = 4.949$, $p < .001$, $\eta p^2 = .041$, 90% CI (0.013, 0.062).

We decomposed the interaction between judgment phase and counterfactual sequence with a Bonferroni-corrected alpha of $p < .0036$ for 14 key comparisons. Consistent with the previous experiments, for the immediate-first sequence,

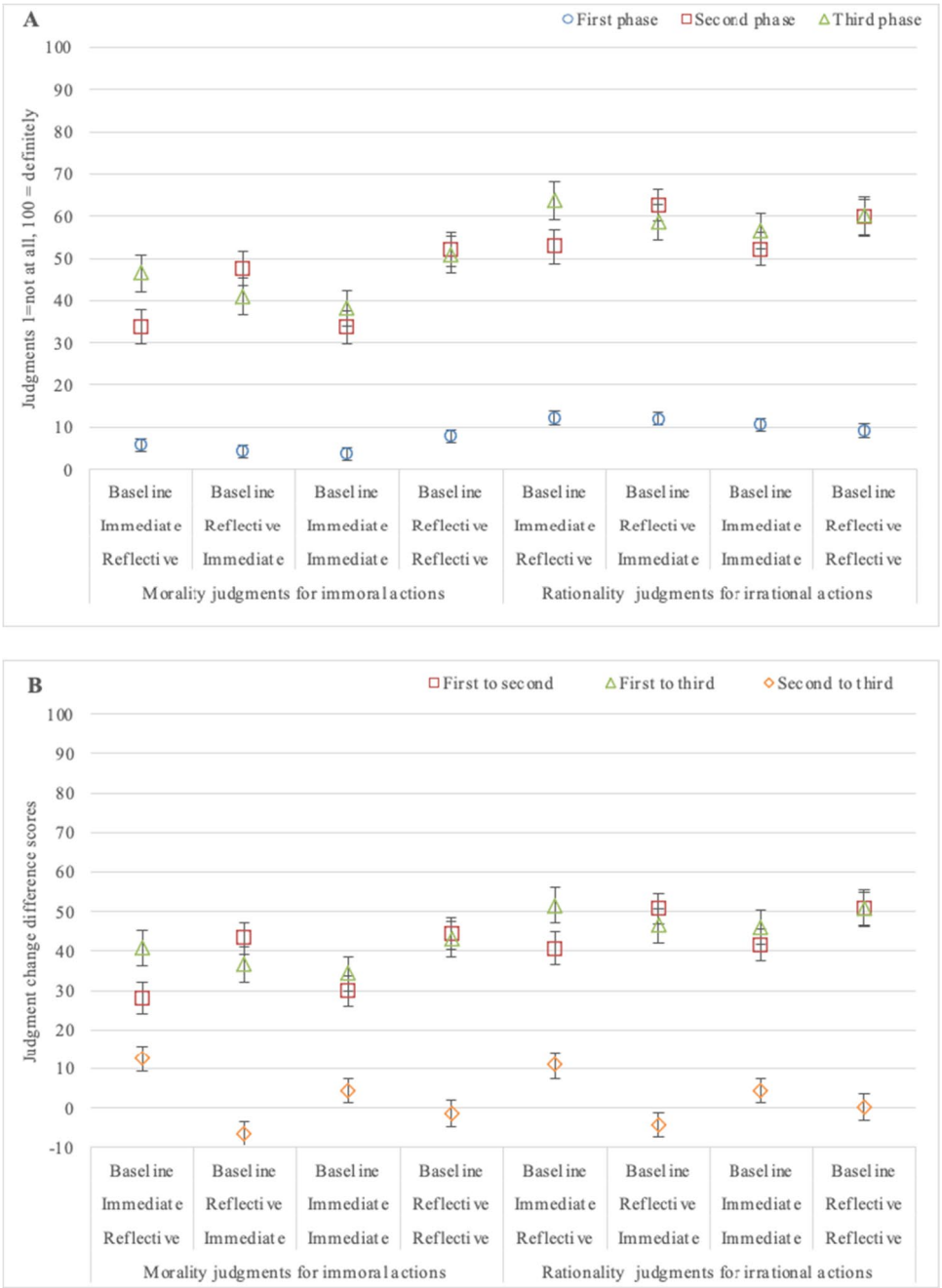

Fig. 5 In Experiment 3, participants provided judgments in a baseline phase, second phase, and third phase in one of four different sequences of immediate or reflective counterfactuals. In (**a**) their mean judgments for the moral acceptability of immoral actions and the rationality of irrational actions are presented. In (**b**) the difference scores for the judgment change from one phase to another is presented. Plots of data for Experiment 3 are based on 355 students from Bahçeşehir University, Turkey. Error bars are standard error of the mean

judgments in the third phase increased compared to the second, t (85) = 4.579, $p < .001$, $d = 0.494$, 95% CI (0.269, 0.716). For the two controls, there were no differences between the second and third phases, immediate-both, t (92) =1.99, $p = .050$, $d = 0.206$, 95% CI (0.000, 0.411), which is not significant on the corrected alpha; and reflective-both, t (83) =0.243, $p = .808$, $d = 0.027$, 95% CI (−0.188, 0.240). The result indicates it is not simply the opportunity to create a second counterfactual that leads to an increase in the third phase for the immediate-first sequence. The difference between the second and third phase was the *opposite* for the reflective-first sequence: judgments in the third phase

decreased compared to the second, t (91) = 3.018, p = .003, d = 0.315, 95% CI (0.104, 0.523), see OSM, Table S7. The result indicates that subsequent reflective elaboration on an immediate counterfactual is required to shift moral acceptability further.

Judgments in the second phase increased more for sequences with a reflective second phase than for those with an immediate one: reflective-first versus immediate-first, t (176) =2.982, p = .003, d = 0.447, 95% CI (0.149, 0.744) and versus immediate-both, t (183) =2.96, p = .003 d = 0.435, 95% CI (0.143, 0.726); reflective-both versus immediate-first, t (168) = 3.210, p = .002, d = 0.492, 95% CI (0.187, 0.797), and versus immediate-both, t (175) = 3.143, p = .002, d = 0.473, 95% CI (0.173, 0.772); and not when the second phase was the same sort of counterfactual, immediate-first versus immediate-both, t (177) = 0.064, p = .949; and reflective-first versus reflective-both, t (174) = 0.075, p = .941. The full set of comparisons is in the OSM.

The ANOVA also showed a main effect of judgment content, F (1,347) = 39.952, p < .001, ηp^2 = .103, 90% CI (0.056, 0.153), as judgments that the immoral actions were immoral were lower than judgments that the irrational actions were irrational; and content interacted with judgment phase, F (1.767, 613.04) = 8.288, p < .001, ηp^2 = .023, 90% CI (0.007, 0.045), see Fig. 5a.[2]

The differences occurred at each phase, baseline, t (307.566) = 5.038, p < .001, d = 0.535, 95% CI (0.322, 0.747), second phase: t (353) = 5.286, p < .001, d = 0.561, 95% CI (0.349, 0.773), and third phase: t (353) = 5.069, p < .001, d = 0.538, 95% CI (0.326, 0.750), as shown in the decomposition of the interaction of content with phase, see OSM.

Judgment change scores The judgment change difference scores are consistent with these results. The scores from the second to third phase were less than those from the baseline to second, or the baseline to third phase, as the main effect of judgment change shows, F (1.36, 472.858) = 432.296, p < .001, ηp^2 = .555, 90% CI (0.508, 0.594), see Fig. 5b. The change from baseline to second was more than from baseline to third for the immediate-first sequence, t (85) = 4.579, p < .001, d = 0.494, 95% CI (0.267, 0.716), and reflective-first sequence, t (91) = 3.018, p = .003, d = 0.315, 95% CI (0.104, 0.523); but the two control sequences showed no differences, immediate-both, t (92) =1.909, p = .050, d = 0.206, 95% CI (0.000, 0.411), which was not significant on the corrected alpha, and reflective-both, t (83) = 0.243, p = .808, d = 0.027, 95% CI (−0.188, 0.240), with a corrected alpha of p < .004 on the decomposition of the interaction of judgment change with sequence, F (4.088, 472.858) = 10.099, p < .001, ηp^2 = .080, 90% CI (0.040, 0.115). Judgment change difference scores were less for immoral actions than irrational ones, as the main effect of content showed, F (1,347) = 10.162, p = .002, ηp^2 = .028, 90% CI (0.007, 0.063), and the difference occurred from the baseline to second phase, t (353) = 3.335, p = .001, d = 0.354, 95% CI (0.144, 0.564), and baseline to third phase t (353) = 3.174, p = .002, d = 0.337, 95% CI (0.127, 0.546); there was no difference for the second to third phase, t (353) =0.153, p = .879, d = 0.016, 95% CI (−0.192, 0.224). For details see the OSM.

Counterfactuals Participants created counterfactuals in which the action was not immoral because of other *facts*, more than those in which the action was immoral but a *dilemma,* both for immoral and irrational actions, in each of the counterfactual sequences. Participants created counterfactuals that focused on the *same* alternative circumstance in the second and third counterfactuals rather than *different* ones for immoral actions, whereas for irrational actions there was no difference. They created counterfactuals that focused more on the same alternative than a different one in the immediate-first sequence and reflective-first one; there was no difference for the immediate-both sequence, and the pattern was the opposite for the reflective-both one, see the OSM, including Table S8, for further details.

The results show that an additional imaginative shift in judgments of the morality of an immoral action occurs only when participants first imagine alternatives for 20 s, and then subsequently imagine alternatives under no time constraints. The opposite occurs when they first reflect with no time constraints, and then subsequently create counterfactuals in 20 s – their judgments return to be closer to their original baseline.

A potential concern is that the instructions throughout have required participants to provide their subsequent judgments of the action "given the circumstances you have just written," which encourages participants to focus on their re-interpreted version of the action in their second or third judgments of its morality, rather than on the original action itself. Arguably, the instruction may have introduced a demand characteristic in which participants believed they were expected by the experimenter to alter their judgment in the light of new circumstances they had written about. The final experiment attempts to probe further how participants view the original immoral action itself after they have considered ways in which it could have been moral.

[2] There was no interaction between content and counterfactual sequence, F (3,347) = 1.089, p = .354, ηp^2 = .009, 90% CI (0.000, 0.025), and no interaction of the three variables, F (5.3, 613.04) = 0.173, p = .977, ηp^2 = .001, 90% CI (0.000, 0.000).

Experiment 4

Our interest has been in how participants judge the morality of an immoral action after they have considered ways in which it could have been moral, and hence we instructed them to provide their second and third judgments of the action "given the circumstances you have just written." To guard against the possibility that this instruction introduces an implicit task demand for participants to change their judgments, in the final experiment we changed the instructions to explicitly return participants' focus to the original action. We asked them to provide their subsequent judgments by requesting them to "now provide your judgment about the behavior again" and, moreover, we repeated the description of the original behavior again to bring it to the forefront of their attention again (see Fig. 1e). According to our theory, participants' imagination of alternative circumstances in which the action is moral should lead them to update their judgments of its morality, even given instructions that orient them back to focus on the original action.

Method

Participants The participants were 120 students from Bahçeşehir University who volunteered in return for course credits. There were 102 women, 17 men, and one person who did not record their gender, and they had a mean age of 21 years with an age range from 19 to 26 years. They were assigned at random to two groups (see Table S4, OSM). We tested as many students as volunteered from the undergraduate module who were invited to participate. A sample size of 116 participants is required to provide at least 90% power to detect a medium sized effect at $p < .05$ for the main effect of immediate versus reflective counterfactuals in the 3 (*judgment phase*: baseline, immediate counterfactual, reflective counterfactual) × 2 (*judgment content*: judgments of morality for immoral actions vs. judgments of rationality for irrational actions) design with repeated measures on the first factor. Prior to any data analysis we eliminated participants who failed to complete all the tasks (11 participants), or who failed to answer correctly a "robot-detection" picture-matching question (one participant), or who failed the attention check question (three participants), resulting in 120 participants.

Materials, design, and procedure The design of the experiment was similar to Experiment 2a, with one exception: after participants made judgments in a baseline phase, and thought about some alternative circumstances for 20 s, their instructions to complete the judgments a second time were as follows: "Please now provide your judgment about the behavior again: *A passenger in an airplane does not want to sit next to a Muslim passenger and so he tells the stewardess the passenger must be moved to another seat.*" Participants then thought about some alternative circumstances with no time limit, and their instructions to complete the judgments a third time were again the new instructions: "Please now provide your judgment about the behavior again," with the description of the behavior included again.

Participants were assigned to two groups (judgment content, irrational or immoral), and they completed two judgments (moral acceptability, rationality) for three actions in the three judgment phases of baseline, immediate, and reflective, i.e., 18 judgments in total. Hence, the between-participant factor *judgment content* again had two levels (immoral, irrational) and the within-participant factor of *judgment phase* had three levels (baseline, immediate-counterfactual, reflective-counterfactual), see Fig. 2. The materials and measures were based on those in Experiment 3.

Results and discussion

We carried out a 3 (*judgment phase*: baseline, immediate-counterfactual, reflective-counterfactual) × 2 (*judgment content*: judgments of morality for immoral actions vs. judgments of rationality for irrational actions) ANOVA with repeated measures on the first factor on participants' judgments. Immoral actions were more acceptable when people imagined how they could have been moral, even with the new instructions. Participants' judgments shifted as they progressed through the three phases, replicating the previous experiments, as shown by a main effect of judgment phase, F (1.328, 156.72) = 60.913, $p < .001$, $\eta p^2 = .34$, 90% CI (0.242, 0.424), see Fig. 6a. As in the previous experiments, they judged the immoral actions to be more immoral in the baseline condition compared to the immediate condition, and compared to the reflective condition, and their judgments continued to shift when they created reflective counterfactuals in the third phase compared to immediate ones in the second phase. There was a main effect of judgment content, F (1, 118) = 3.984, $p = .048$, $\eta p^2 = .03$, 90% CI (0.000, 0.099), as participants' judgments of the morality of immoral actions were lower than their judgments of the rationality of irrational actions. There was no interaction between the two variables, F (1.328, 156.72) = 1.284, $p = .270$, $\eta p^2 = .011$, 90% CI (0.000, 0.049), see OSM, Table S9.

Judgment change difference scores Judgment change difference scores from the immediate to reflective phase were less than from the baseline to the immediate phase, and change

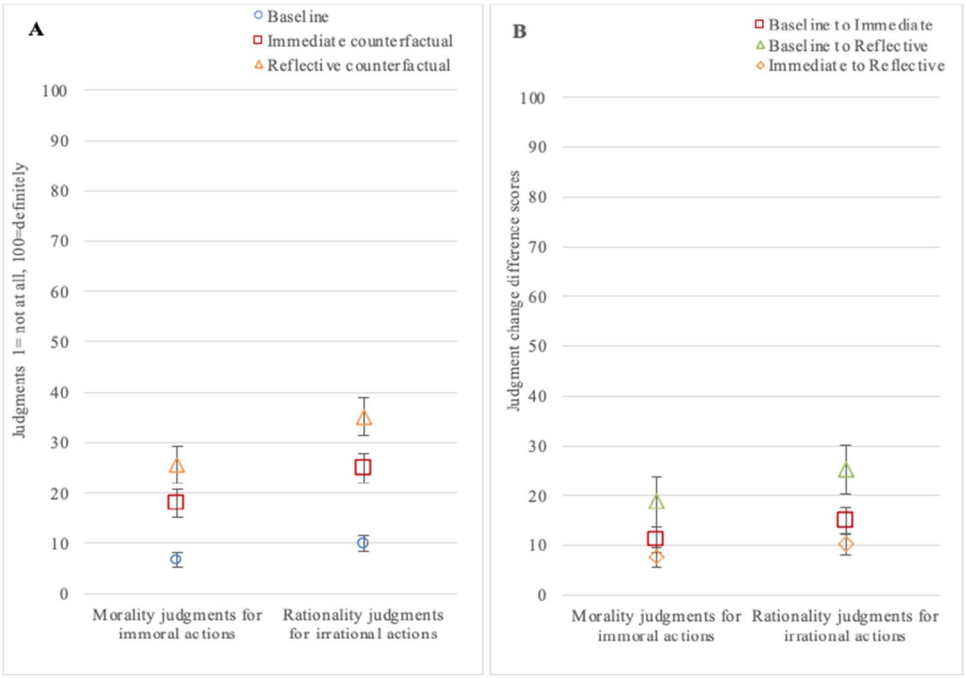

Fig. 6 In Experiment 4 participants constructed immediate counterfactuals and then reflective ones. Their mean judgments for the first, second and third phases are presented for the moral acceptability of immoral actions and the rationality of irrational actions in (**A**); the difference scores for the judgment change from one phase to another are presented in (**B**). Plots of data are based on 126 students from Bahçeşehir University. Error bars are standard error of the mean

scores from the baseline to immediate phase were less than from the baseline to the reflective phase, as shown by a main effect of judgment change, F (1.732, 204.43) = 29.104, $p <$ .001, $\eta p^2 = .198$, 90% CI (0.119, 0.272), see Fig. 6b. There was no main effect of judgment content, F (1, 118) = 1.513, $p = .221$, $\eta p^2 = .013$, 90% CI (0.000, 0.065), and no interaction of the two variables, F (1.732, 204.429) =0.613, $p =$ .520, $\eta p^2 = .005$, 90% CI (0.000, 0.028).

The results show that participants update their moral judgments when they consider some alternative ways in which an immoral action could have been moral, even when the instructions are careful to remove any implicit task demand to do so. The experiment replicates the findings of the previous experiments, with quite different instructions designed to re-focus participants on the original action. We can conclude that even though participants judge an action to be morally unacceptable initially, once they imagine alternative circumstances in which the action could have been morally acceptable, they revise their initial judgment and consider the immoral action to be less immoral. Once again, they can do so even after they consider alternative circumstances for just a very short time.

General discussion

How does an immoral act come to be considered less morally unacceptable? An important mechanism is that people can imagine ways in which it *would have been* moral, even in the absence of any further facts about the matter. People judged immoral actions to be not at all moral, but when they imagined alternative circumstances in which they would be moral, a striking shift in their judgments about the actions' moral unacceptability was observed in all five experiments. Arguably, after they have imagined alternative circumstances, people do not consider the situation to be the same. The finding implies that people possess the moral flexibility to allow circumstance to moderate their assessment of others' moral behavior, rather than being tied to their initial interpretations, and even when the circumstance is entirely imagined rather than based on further information (e.g., Cone & Ferguson, 2015; Graham & Haidt, 2012; Harman, 1975; Mann & Ferguson, 2015; Mikhail, 2013; Monroe & Malle, 2019; Piazza et al., 2013; Sabo & Giner-Sorolla, 2017; Sinnott-Armstrong, 2002; Stanley et al., 2018). Remarkably, they do so even though they receive

no external confirmation that their imagined circumstances apply validly to the situation. They deploy a repertoire of counterfactual argumentation strategies to do so, including counterfactuals that deny the action was immoral by introducing additional facts to modify the interpretation of the situation, or ones that accept the immorality of the action but introduce a dilemma with a competing moral action to justify the violation. They rarely exhibited resistance to the idea that the immoral action could be considered moral, but equally rarely engaged in any attempt to modify the norm upon which it was based (e.g., Gendler, 2000; Haidt, 2012). The same pattern was observed for thoughts about immoral actions, and those about unreasonable ones.

People can readily imagine, in a matter of seconds, alternative circumstances in which an immoral action would have been moral. Quickly jotting down a few words in just 20 s to convey the first thought that comes to mind had a significant impact on subsequent moral judgments. The short time frame of 20 s for reading the instruction, the scenario, and typing an answer leaves very little time indeed to spend on imagining an alternative. Alternative moral possibilities appear to be immediately accessible (e.g., Phillips & Cushman, 2017). Moreover, there is an added effect of reflecting carefully on alternative circumstances without any time constraints, which shifts moral judgments further in the same direction of increased moral acceptability. Participants constructed counterfactuals in the reflective phase that elaborated on the same idea as the one they first thought of in the immediate phase. The counterfactual possibility they generated briefly in the first phase may have appeared to them to warrant further elaboration. However, when a reflective counterfactual phase was followed by an immediate one, a reversal in judgments of moral acceptability was observed. Participants tended to think of the same idea again, but the 20-s limit on thinking about it reduced their judgment of the action's moral acceptability from the level attained by reflection. This unexpected result may indicate that revisiting the original counterfactual possibility so briefly somehow undermined its effectiveness; the finding merits further investigation.

Unreasonable actions can also come to be considered less irrational when people imagine how they would have been rational, just like immoral actions. Although the immoral actions were considered more morally unacceptable than the irrational actions were considered unreasonable, nonetheless the same pattern was observed for both. People may expect others to behave in ways that are moral and reasonable, and so these default possibilities may be readily available (e.g., Cushman, 2020; Phillips et al., 2015, 2019; Phillips & Cushman, 2017). The finding is consistent with the idea that moral cognition relies on the same sorts of cognitive processes that underpin reasoning about non-moral matters (Bucciarelli et al., 2008; Cushman & Young, 2011; Knobe, 2018; Rai & Holyoak, 2010; Uttich & Lombrozo, 2010; see also Haidt, 2012; Young & Saxe, 2011). The potential sorts of cognitive processes that are implicated by these discoveries are sketched in Table 2.

The experiments included participants from different cultures. Participants from the USA and UK (in Experiments 1 and 2b) and those from Turkey (in Experiments 2a, 3, and 4) judged the immoral actions to be similarly morally unacceptable at the baseline phase (generally about 5–10 on the 0–100 scale). However, the US and UK participants exhibited a greater moral shift than the Turkish participants, at the second phase (generally to about 40 on the scale vs. to about 30, respectively) and at the third phase (generally about 60 vs. about 40, respectively). In contrast, both populations judged the irrational actions to be similarly irrational at the baseline phase, and exhibited a similar shift in the second and third phases. Cultural and content effects on imaginative moral shifts are worth examining further, given that counterfactuals are pervasive, occurring regardless of linguistic or cultural convention and throughout the lifespan (e.g., Au, 1983; Beck et al., 2006; Byrne, 2005; Dudman, 1988; Harris, 2000; Walsh & Byrne, 2001).

The role of the counterfactual imagination in moral mitigation has implications for its preparatory function of supporting intentions to change (e.g., De Brigard, Addis, et al., 2013a; Rim & Summerville, 2014; Roese & Epstude, 2017; Smallman & McCulloch, 2012; Timmons et al., 2021; Van Hoeck et al., 2013) and its emotional amplification of feelings of regret or relief (Kahneman & Tversky, 1982; Sweeny & Vohs, 2012). In our experiments, participants were directed to think about whether there were alternative circumstances in which the actions would be moral, and further study of the extent to which participants engage spontaneously in such moderation is needed, as is examination of the effects of being provided with arguments based on such alternatives. Moreover, participants were asked to imagine ways in which the morally unacceptable behavior could be acceptable. Of course, they could have imagined a worse-world than the actual world, in which all such morally unacceptable actions are acceptable, for example, a world in which everyone believes racism is acceptable, but instead, most participants created a better-world or upward counterfactual, in which the specific morally unacceptable action was acceptable, because of a change to the facts, for example, the action was not in fact an instance of racism, or because of a dilemma, for example, the action was racist but carried out in service of another moral principle, for example, protecting others. Since they created upward counterfactuals about how the action could be interpreted as a better

Table 2 An illustration of the cognitive processes in counterfactual imaginative moral shifts

1. A description of an immoral action is the input, e.g., *'A passenger in an airplane does not want to sit next to a Muslim passenger and so he tells the stewardess he must be moved to another seat'.* The following processes are required:
   a. Construct a model to simulate the event (e.g., Khemlani et al., 2018).
   b. Incorporate into the model of the event additional information from background knowledge including moral norms such as the proscription of discrimination on the basis of religion, race, or ethnicity.
   c. Deduce that the action is morally unacceptable.
2. On receipt of the output of the first step, activate a set of counterfactual processes (e.g., Walsh and Byrne, 2001) that includes processes such as the following:
   a. Select an aspect of the model of the event to modify, e.g., the man does not want to sit next to<a Muslim>.
   b. Retrieve available alternatives to this aspect guided by norms.
   c. Construct a model of an alternative to the simulated event, choosing one of several available strategies, e.g.,
      i. Delete the selected aspect and replace it with the retrieved alternative, e.g., the man does not want to sit next to <a person who has been rude to him>.
      ii. Expand the selected aspect by adding something new to it, e.g., the man does not want to sit next to < a Muslim man who is engag ing in threatening behavior>.
3. The counterfactual set of processes contain immediate processes and reflective processes at each step:
   a. At the selection step immediate processes identify the more salient aspects of the event in the foreground, e.g., the Muslim man, whereas reflective processes identify more implicit features e.g., something about the actor himself.
   b. At the retrieval step, immediate processes access defaults whereas reflective processes sample possibilities more thoroughly.
   c. At the construction step, immediate processes engage in simple deletion, whereas reflective processes engage in elaborative addition.
4. The output from the counterfactual set of processes is treated as follows:
   a. The output is further information to be combined with the initial information as a counterexample <the man does not want to sit next to a Muslim passenger, but the behavior is not morally unacceptable discrimination>.
   b. The combination requires processes that combine premise information with background knowledge in a cohesive model, to ensure inferences can be withdrawn in a manner that maintains epistemically entrenched beliefs (e.g., it is unacceptable to discriminate on the basis of religion).
   c. The reconciliation of the premises with additional background knowledge ensures the conclusion is no longer warranted (the behavior was not based on discrimination so is less morally unacceptable), and the output from step 1 is withdrawn to be modified.

one, they updated their moral judgments also in an upwards direction to be more favorable towards the action. It may be fruitful in future research to examine the effects of directing participants to imagine a downward, worse-world counterfactual, that is, how a behavior could be even less morally acceptable, to examine whether they also update their moral judgments downwards to be even harsher, again based solely on imagination. A demonstration that the moral imaginative shift occurs in either direction, upwards or downwards, would provide further support for our argument that imagination alone can alter moral judgments even in the absence of further facts.

It is notable that judgments about unreasonable actions shifted from being considered to be not at all rational, to being considered rational, whereas judgments about immoral actions shifted from being considered to be not at all moral, towards the mid-point of the 0–100 scale, but rarely beyond the mid-point. Of course, the baseline judgments for irrational actions about unreasonable actions was higher than the baseline judgments for immoral actions. Nonetheless, the action's immorality was neutralized rather than transformed to be moral, and so whether mechanisms other than the counterfactual imagination can bring about greater transformations remains an open question.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13421-022-01315-0.

Acknowledgements Some of the research was carried out while the first author was a visiting PhD researcher at Trinity College Dublin for six months in 2019, supported by a TUBITAK (Scientific and Technological Research Council of Turkey) 2214, International Research Fellowship Program (53325897-115.02-166509) award to the first author. We thank Ayşe Betül İlgen, Begüm Yıldırım, Aybuke Eker, Greta Warren, Yinyue Dai and Mark Keane for their help and discussion. Some of the results were presented at the European Society for Cognitive Psychology conference in Tenerife in September 2019, and at the Vienna Forum for Analytic Philosophy conference in Vienna in June 2020 (online because of COVID-19). Data from the four studies reported in this manuscript are available through the Open Science Framework repository at: https://osf.io/mw94z/.

Author contributions Beyza Tepe and Ruth Byrne developed the study concept and designed the experiments. Beyza Tepe collected and analysed the data. Beyza Tepe and Ruth Byrne discussed the results and implications. Ruth Byrne and Beyza Tepe wrote the paper. The authors declare no conflicts of interest with respect to the authorship or the publication of this article.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*(4), 556.

Alicke, M. D., Buckingham, J., Zell, E., & Davis, T. (2008). Culpable control and counterfactual reasoning in the psychology of blame. *Personality and Social Psychology Bulletin., 34*, 1371–1381.

Andrejević, M., Feuerriegel, D., Turner, W., Laham, S., & Bode, S. (2020). Moral judgements of fairness-related actions are flexibly updated to account for contextual information. *Scientific Reports, 10*(1), 1–17.

Au, T. K.-F. (1983). Chinese and English counterfactuals: The Sapir-Whorf hypothesis revisited. *Cognition, 15*(1-3), 155–187.

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition, 167*, 25–37.

Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006). Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child Development, 77*, 413–426F.

Bialek, M., & De Neys, W. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment and Decision making, 12*(2), 148–167.

Bloom, P. (2010). How do morals change? *Nature, 464*, 490–490.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 717–726.

Bucciarelli, M., Khemlani, S., & Johnson-Laird, P. N. (2008). The psychology of moral reasoning. *Judgment and Decision Making., 3*(2), 121–139.

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

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

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

Byrne, R. M. J., & Johnson-Laird, P. N. (2020). If and or: Real and counterfactual possibilities in their truth and probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(4), 760–780.

Cariani, F., & Rips, L. J. (2017). Conditionals, context, and the suppression effect. *Cognitive Science, 41*(3), 540–589.

Chakroff, A., & Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition, 136*, 30–37.

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108*(1), 37.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*, 353–380.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292.

Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences, 43*.

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science, 35*, 1052–1075.

De Brigard, F., Addis, D. R., Ford, J. H., Schacter, D. L., & Giovanello, K. S. (2013a). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia, 51*(12), 2401–2414.

De Brigard, F., Szpunar, K. K., & Schacter, D. L. (2013b). Coming to grips with the past effect of repeated simulation on the perceived plausibility of episodic counterfactual thoughts. *Psychological Science, 24*(7), 1329–1334.

Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. *Psychology of Learning and Motivation, 50*, 307–338.

Dudman, V. H. (1988). Indicative and subjunctive. *Analysis, 48*, 113–122.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.

Espino, O., & Byrne, R. M. J. (2020). The suppression of inferences from counterfactual conditionals. *Cognitive Science, 44*(4), e12827.

Elio, R., & Pelletier, F. J. (1997). Belief change as propositional update. *Cognitive Science, 21*(4), 419–460.

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*(4), 1149–1160.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. MIT press.

Gendler, T. S. (2000). The puzzle of imaginative resistance. *The Journal of Philosophy, 97*(2), 55–81.

Giner-Sorolla, R. (2018). Powering your interaction. (Retrieved from) https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/

Graham, J., & Haidt, J. (2012). Sacred values and evil adversaries: A moral foundations approach. In M. Mikulincer & P. R. Shaver (Eds.), *Herzliya series on personality and social psychology. The social psychology of morality: Exploring the causes of good and evil* (pp. 11–31). American Psychological Association.

Gray, K., & Keeney, J. E. (2015). Impure or just weird? Scenario sampling bias raises questions about the foundation of morality. *Social Psychological and Personality Science, 6*(8), 859–868.

Goldinger, S. D., Kleider, H. M., Azuma, T., & Beike, D. R. (2003). "Blaming the victim" under memory load. *Psychological Science, 14*(1), 81–85.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364–371.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences, 6*(12), 517–523.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*, 1144–1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105–2108.

Goodwin, G. P. (2017). Is morality unified, and does this matter for moral reasoning? In J. F. Bonnefon & B. Trémolière (Eds.), *Moral inferences* (pp. 17–44). Psychology Press.

Gubbins, E., & Byrne, R. M. (2014). Dual processes of emotion and reason in judgments about moral dilemmas. *Thinking & Reasoning., 20*(2), 245–268.

Guglielmo, S. (2015). Moral judgment as information processing: An integrative review. *Frontiers in Psychology, 6*, 1637.

Gürcay, B., & Baron, B. (2017). Challenges for the sequential two-system model of moral judgement. *Thinking & Reasoning, 23*, 49–80.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814.

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*(5827), 998–1002.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Pantheon Books.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65*(4), 613.

Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *The British Journal for the Philosophy of Science, 66*(2), 413–457.

Harman, G. (1975). Moral relativism defended. *The Philosophical Review, 84*(1), 3–22.

Harris, P. L. (2000). *The work of the imagination*. Blackwell.

Henne, P., Niemi, L., Pinillos, Á., De Brigard, F., & Knobe, J. (2019). A counterfactual explanation for the action effect in causal judgment. *Cognition, 190*, 157–164.

Horne, Z., Powell, D., & Hummel, J. (2015). A single counterexample leads to moral belief revision. *Cognitive Science, 39*(8), 1950–1964.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*(2), 136–153.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). Cambridge University Press.

Khemlani, S., Byrne, R. M. J., & Johnson-Laird, P. N. (2018). Facts and possibilities: A model-based theory of sentential reasoning. *Cognitive Science., 42*(6), 1887–1924.

Knobe, J. (2018). There is no important distinction between moral and nonmoral cognition. In K. Gray & J. Graham (Eds.), *The atlas of moral psychology* (pp. 556–565). Guilford Press.

Lench, H. C., Domsky, D., Smallman, R., & Darbor, K. E. (2014). Beliefs in moral luck: When and why blame hinges on luck. *British Journal of Psychology, 106*(2), 272–287.

Luo, Q., Nakic, M., Wheatley, T., Richell, R., Martin, A., & Blair, R. J. R. (2006). The neural basis of implicit moral attitude? An IAT study using event-related fMRI. *Neuroimage, 30*, 1449–1457.

Maki, A., & Raimi, K. T. (2017). Environmental peer persuasion: How moral exporting and belief superiority relate to efforts to influence others. *Journal of Environmental Psychology, 49*, 18–29.

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

Malle, B. F., & Holbrook, J. (2012). Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology, 102*(4), 661.

Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology, 108*(6), 823.

Markman, K. D., Gavanski, I., Sherman, S. J., & McMullen, M. N. (1993). The mental simulation of better and worse possible worlds. *Journal of Experimental Social Psychology, 29*(1), 87–109.

Markman, K. D., McMullen, M. N., & Elizaga, R. A. (2008). Counterfactual thinking, persistence, and performance: A test of the reflection and evaluation model. *Journal of Experimental Social Psychology, 44*(2), 421–428.

Mazzocco, P. J., Alicke, M. D., & Davis, T. L. (2004). On the robustness of outcome bias: No constraint by prior culpability. *Basic and Applied Social Psychology, 26*(2-3), 131–146.

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

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

Mikhail, J. (2013). New perspectives on moral cognition: Reply to Zimmerman, Enoch, and Chemla, Egre, and Schlenker. *Jerusalem Review of Legal Studies, 8*(1), 66–114.

Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology, 116*(2), 215.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science, 19*, 549–557.

Oaksford, M., & Chater, N. (2018). Dynamic inference and everyday conditional reasoning in the new paradigm. *Thinking & Reasoning, 19*(3–4), 346–379.

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

Parkinson, M., & Byrne, R. M. J. (2018). Judgments of moral responsibility and wrongness for intentional and accidental harm and purity violations. *Quarterly Journal of Experimental Psychology., 71*(3), 779–789.

Paxton, J. M., & Greene, J. D. (2010). Moral reasoning: Hints and allegations. *Topics in Cognitive Science, 2*, 511–527.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*, 163–177.

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *PNAS., 114*(18), 4649–4654.

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*(12), 1026–1040.

Piazza, J., Pascale, S. R., & Sousa, P. (2013). Moral emotions and the envisaging of mitigating circumstances for wrongdoing. *Cognition & Emotion, 27*(4), 707–722.

Pollock, J. L. (1987). Defeasible reasoning. *Cognitive Science, 11*(4), 481–518.

Rai, T. S., & Holyoak, K. J. (2010). Moral principles or consumer preferences? Alternative framings of the trolley problem. *Cognitive Science, 34*, 311–321.

Rim, S., & Summerville, A. (2014). How far to the road not taken? The effect of psychological distance on counterfactual direction. *Personality and Social Psychology Bulletin, 40*(3), 391–401.

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

Roese, N. J., & Epstude, K. (2017). The functional theory of counterfactual thinking: New evidence, new challenges, new insights. In *Advances in experimental social psychology* (Vol. 56, pp. 1–79). Academic Press.

Roese, N. J., Sanna, L. J., & Galinsky, A. D. (2005). The mechanics of imagination: Automaticity and control in counterfactual thinking. In R. R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *Oxford series in social cognition and social neuroscience. The new unconscious* (pp. 138–170). Oxford University Press.

Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of Julie and Mark: Unraveling the moral dumbfounding effect. *Judgment and Decision making, 10*, 296.

Russell, P. S., & Giner-Sorolla, R. (2011a). Social justifications for moral emotions: When reasons for disgust are less elaborated than for anger. *Emotion, 11*(3), 637.

Russell, P. S., & Giner-Sorolla, R. (2011b). Moral anger is more flexible than moral disgust. *Social Psychological and Personality Science, 2*(4), 360–364.

Sabo, J. S., & Giner-Sorolla, R. (2017). Imagining wrong: Fictitious contexts mitigate condemnation of harm more than impurity. *Journal of Experimental Psychology: General, 146*(1), 134.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science, 300*, 1755–1758.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32–70.

Shtulman, A., & Tong, L. (2013). Cognitive parallels between moral judgment and modal judgment. *Psychonomic Bulletin & Review, 20*(6), 1327–1335.

Simpson, A., Laham, S. M., & Fiske, A. P. (2016). Wrongness in different relationships: Relational context effects on moral judgment. *The Journal of Social Psychology, 156,* 594–609. https://doi.org/10.1080/00224545.2016.1140118

Sinnott-Armstrong, W. (2002). Moral relativity and intuitionism. *Philosophical Issues, 12*, 305–328.

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass, 4*, 267–281.

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology, 88*, 895–917.

Smallman, R., & McCulloch, K. C. (2012). Learning from yesterday's mistakes to fix tomorrow's problems: When functional counterfactual thinking and psychological distance collide. *European Journal of Social Psychology, 42*(3), 383–390.

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition, 119*(3), 454–458.

Stanley, M. L., Dougherty, A. M., Yang, B. W., Henne, P., & De Brigard, F. (2018). Reasons probably won't change your mind: The role of reasons in revising moral decisions. *Journal of Experimental Psychology: General, 147*(7), 962.

Stenning, K., & Van Lambalgen, M. (2012). *Human reasoning and cognitive science*. MIT Press.

Stupple, E. J., & Ball, L. J. (2008). Belief–logic conflict resolution in syllogistic reasoning: Inspection-time evidence for a parallel-process model. *Thinking & Reasoning, 14*(2), 168–181.

Sweeny, K., & Vohs, K. D. (2012). On near misses and completed tasks: The nature of relief. *Psychological Science, 23*, 464–468.

Tepe, B., & Aydinli-Karakulak, A. (2019). Beyond harmfulness and impurity: Moral wrongness as a violation of relational motivations. *Journal of personality and social psychology., 117*(2), 310–337.

Tepe, B., Piyale, Z. E., Sirin, S., & Sirin, L. R. (2016). Moral decision-making among young Muslim adults on harmless taboo violations: The effects of gender, religiosity, and political affiliation. *Personality and Individual Differences, 101*, 243–248.

Timmons, S., & Byrne, R. M. (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. J. (2021). Imagined alternatives to episodic memories of morally good acts. *Journal of Positive Psychology., 16*(2), 178–197.

Trippas, D., Thompson, V. A., & Handley, S. J. (2017). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. *Memory & Cognition, 45*(4), 539–552.

Turiel, E. (2002). *The culture of morality: Social development, context, and conflict*. Cambridge University Press.

Ugazio, G., Lamm, C., & Singer, T. (2012). The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion, 12*(3), 579.

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition, 116*(1), 87–100.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17*, 476–477.

Van Hoeck, N., Ma, N., Ampe, L., Baetens, K., Vandekerckhove, M., & Van Overwalle, F. (2013). Counterfactual thinking: An fMRI study on changing the past for a better future. *Social Cognitive and Affective Neuroscience, 8*, 556–564.

Walsh, C. R., & Byrne, R. M. (2001). A computational model of counterfactual thinking: The temporal order effect. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 23.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science, 16*(10), 780–784.

Wiech, K., Kahane, G., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2013). Cold or calculating? Reduced activity in the subgenual cingulate cortex reflects decreased emotional aversion to harming in counterintuitive utilitarian judgment. *Cognition, 126*, 364–372.

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120*(2), 202–214.

Open Practices Statements The data and Online Supplemental Material including materials and additional results for all experiments are available via the Open Science Framework at: https://osf.io/mw94z/. None of the experiments were pre-registered.