Understanding „Deontology“ and „Utilitarianism“ in Moral Dilemma Judgment

–

A Multinomial Modeling Approach

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Maximilian Hennig

aus Duisburg

Tübingen

2019

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	24.02.2020
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Mandy Hütter
2. Berichterstatter:	Prof. Dr. Klaus Fiedler
3. Berichterstatter:	Prof. Dr. Andreas Glöckner

Acknowledgments

The last three and a half years that have resulted in this dissertation thesis have been a challenging and wonderfully stimulating journey. Reaching its (preliminary) end would not have been possible (or far less enjoyable and interesting) without the presence of the many people in my life who accompanied me along the way, be it on the academic path or besides it. I have much to be thankful for, and many people to who I wish to express my sincere gratitude.

First and foremost, thank you, Mandy, for having been my intellectual guide, mentor, and a true role model, and for always having had my best in mind over the course of the last years. I am grateful for the countless hours you invested in me, for your thoughtful supervision and guidance, for being approachable and always open to discussion and exchange, for having known when to push me and when to slow me down a bit, and for the opportunities to grow that are the direct result of thinking and working with you. I am truly grateful that you have taken me on the journey that has led to this dissertation thesis.

Thank you, Hein, for about four years ago giving me the wise advice to settle for Tübingen in the first place.

Thank you, all of my colleagues and former colleagues, (in order of appearance) René, David, Fabi, Max, Kathi, Niels, Birka, Thomas, Bruno, Marco, and Zach, for filling the office with life, each in your own way. Thank you for all of the times in which you have provided me with valuable feedback, emotional support, sweets, or at times much needed distraction, were kind enough to guide me in debugging my amateurish R-scripts or PHP-code, were open to the sort of thoughtful conversation that makes academic life worth participating in to begin with, or have just been fun to be around! A special thank you to my still sort of colleague Fabi

for all the time spent in open-ended department-corridor conversations that usually didn't start out with a goal in mind, but always ended up somewhere worth going.

Thank you, Denise, Margret, and the HiWis, for your work behind the scenes, which enabled a smooth workflow.

Thank you, Andi, for solving my technical problems and for frequent chats at the coffee machine, when large parts of the building were still sound asleep.

Thank you, Michael, for the many hours we spent over lunch discussing psychological science and everything surrounding it in all its forms and permutations.

Thank you, Olivier, Adrien, Christophe, Filip, Vassilis, Mathias, Julie, Vincent, and all the people who have welcomed me at UCLouvain during my time there. Special thanks to Olivier for hosting me in the first place, for challenging discussions, being encouraging on an intellectual and personal level, for being a valuable thinking partner and for providing me with an experience that I would not want to miss. Special thanks also to my friend Adrien, for having done much to make Brussels and Louvain-La-Neuve feel like home to me for the three months of my stay, for countless hours spent in real conversations, for discussing science, uncertainty, the future, and everything else that makes life valuable (with good Belgian beer and without it), for the occasional silly banter, and for being a great (intellectual) sparring partner.

Thank you, Uwe and all the people from Shin-Do, for providing an addictive opportunity to grow outside of the office and to cultivate focus, as well as for providing much needed balance to my desk-heavy life. Although it may not be obvious why, I think the dissertation might well have failed without this.

Thank you, Timo, for tons of much needed emotional support and ice cream, for your kind and stable presence, and for having done so much to help me grow as a person.

Thank you, Ira, for all the long conversations about life, opportunity, and uncertainty, be it over a distance via phone or in person over an enjoyable meal, which always helped to put things in perspective.

Thank you, Helmut and Vera, for having provided me with the opportunity to walk this path in the first place, for your continued trust in my abilities and for being the supportive and loving presence in my life that you are.

Thank you to my intellectual and comedic heroes, who continue to teach me how to take life seriously and not too seriously, respectively, and for suggesting that being able to switch flexibly between both mindsets is an artform that may well save your sanity at times.

Also, thank you to John Cleese, who has taught me that worrying about whether your acknowledgments are too long is *just plain silly*!

Table of Contents

Abstract

Over the course of the last two decades, research on moral judgment has been heavily shaped by the application of moral dilemma research. Data obtained with this paradigm are commonly interpreted by assuming a hard split between two different kinds of processes – a “deontological” sensitivity to moral norms regardless of consequences, and a “utilitarian” sensitivity to consequences regardless of moral norms. Additionally, it is frequently assumed that these two processes arise from distinct cognitive systems, implicated in “emotional” and “rational” processing, respectively.

Over the course of this thesis, I will address several methodological and conceptual assumptions of this conventional approach to understanding dilemma judgment with the application of multinomial modeling.

Specifically, the current thesis investigates the impact of factors that are systematically confounded in the context of the conventional dilemma approach, and assess whether “deontological” response patterns are indeed insensitive to consequences, as the dominant conceptualization maintains (Chapter II). Results indicate the assumption of a hard split between “norms” and “consequences” (let alone “deontology” and “utilitarianism”) as determinants of dilemma judgment to be artificial and overly simplistic, and suggest that both response patterns may be understood in terms of expected consequences, and demonstrate the potential biasing impact of prominent confounds on individual response patterns.

Subsequently, the current thesis assesses whether the findings of two of the presented experiments are consistent with the predictions of a model that avoids reliance on dual-process assumptions, and finds largely confirmatory evidence (Chapter III).

Finally, application of the previously developed multinomial model in two additional studies (Chapter IV) further demonstrates the importance of controlling for prominent confounds such as response tendencies, as results suggest that spurious effects may otherwise arise and may be misinterpreted in the context of dual-process models.

While integrating the work presented in these chapters, I discuss parallels between moral dilemma judgment and the phenomenon of “moral dumbfounding”, and subsequently integrate the empirical findings presented in this thesis with other models of moral judgments unrelated to the dilemma literature. In doing so, I suggest that the results of this thesis converge with other models developed outside the realm of dilemma research, which suggest moral judgment to be ultimately determined by the perception of harmful consequences, such that the perception of harm and immorality tend to go hand in hand.

Thus, the current thesis argues against the view that moral dilemma judgments are best understood in terms of adherence to absolute norms versus impartial utilitarian calculations, and rejects associated dual-process assumptions. Instead, it proposes that moral dilemma judgments may be viewed through a consequentialist lens, a proposal which converges with evidence obtained outside the realm of moral dilemma research.

Zusammenfassung

Die Anwendung moralischer Dilemmas hat im Laufe der letzten zwei Jahrzehnte großen Einfluss auf die empirische Forschung in der Moralpsychologie ausgeübt. Mit Hilfe dieses Ansatzes gewonnene Befunde werden gewöhnlicher Weise interpretiert, indem eine harte Unterscheidung zwischen zwei unterschiedlichen Prozessen angenommen wird – eine „deontologische“ Sensitivität für moralische Normen unabhängig von Konsequenzen, und eine „utilitaristische“ Sensitivität für Konsequenzen unabhängig von Normen. Zusätzlich wird außerdem für gewöhnlich angenommen, dass diese beiden Prozesse das Resultat unterschiedlicher kognitiver Systeme sind, die jeweils für „emotionale“ bzw. „rationale“ Prozesse verantwortlich sind.

Im Laufe dieser Dissertation werde ich unter Zuhilfenahme multinomialer Modellierung mehrere methodologische und konzeptuelle Annahmen dieses konventionellen Dilemma-Ansatzes untersuchen.

Zunächst untersucht die vorliegende Dissertation den Einfluss von Faktoren die im Kontext des konventionellen Ansatzes oftmals Störfaktoren darstellen, und erforscht ob “deontologische” Antwortmuster tatsächlich Sensitivität für Konsequenzen vermissen lassen, wie es die vorherrschende Konzeptualisierung unterstellt (Kapitel II). Resultate legen nahe, dass die Annahme einer harten Trennung zwischen „Normen“ und „Konsequenzen“ (einschließlich „Deontologie“ und „Utilitarismus“) als Determinanten von Dilemmaentscheidungen künstlich und übervereinfacht ist, deuten an, dass beide Antwortmuster gleichermaßen durch Sensitivität für erwartbare Konsequenzen verstanden werden können und demonstrieren den verzerrenden Einfluss bekannter Störvariablen auf identifizierbare Antwortmuster.

Im Anschluss untersucht die vorliegende Dissertation ob die Befunde von zwei der präsentierten Experimente im Kontext eines Modells zu erklären sind, welches zwei-Prozess Annahmen vermeidet, und findet größtenteils Bestätigung für dieses Modell (Kapitel III).

Schließlich demonstriert die Anwendung des zuvor entwickelten multinomialen Modells in zwei weiteren Studien (Kapitel IV) die Bedeutung der Kontrolle über bekannte Störvariablen und Handlungstendenzen, da andernfalls fälschliche Effekte auftreten können, welche im Kontext eines zwei-Prozess Modells fehlinterpretiert werden können.

Bei der Integration der Arbeit, die in den vorherigen Kapiteln präsentiert wurden, diskutiere ich Parallelen zwischen moralischen Dilemma-Urteilen und dem Phänomen des „moral dumbfounding“, und integriere die empirischen Befunde dieser Dissertation mit weiteren Modellen der moralischen Entscheidungsfindung, die unabhängig von der Dilemma-Literatur entwickelt wurden. In diesem Kontext schlage ich vor, dass die Befunde dieser Dissertation mit anderen Modellen außerhalb des Forschungsfeldes des Dilemmaurteilens konvergieren, welche andeuten, dass moralisches Urteilen durch die Wahrnehmung schädlicher Konsequenzen bestimmt werden, sodass die Wahrnehmung von Schaden und moralischen Übertretungen miteinander einhergehen.

Die vorliegende Dissertation argumentiert damit gegen den Standpunkt, dass Dilemmurteile am besten in Hinsicht auf absolute Normen versus Maximierung von Konsequenzen verstanden werden sollten und lehnt damit einhergehende zwei-Prozess Annahmen ab. Stattdessen schlägt sie vor, dass moralische Dilemmaurteile durch eine konsequenzialistische Linse verstanden werden sollten, ein Standpunkt der mit Befunden außerhalb des Forschungsgebietes des Dilemmaurteilens konvergiert.

Chapter I

General Introduction

Max Hennig

Eberhard Karls Universität Tübingen

1.1. Moral judgment

Determining what thoughts, statements or behaviors one considers “right” or “wrong”, worthy of praise or condemnation, or indicative of noble or deficient character constitutes a key component of everyday functioning. This process of moral judgment has been argued to be a major contributor to social cohesion (Haidt, 2012) and has been found to play a prominent role in everyday social interaction (Hofmann, Brandt, Wisneski, Rockenbach, & Skitka, 2018; Hofmann, Wisneski, Brandt, & Skitka 2014).

As such, moral judgment is an important area of empirical inquiry that has been of great interest to psychological research. Whereas early approaches to the empirical study of moral judgment have focused on the importance of rational reasoning (e.g. Kohlberg, 1969; Piaget, 1965), more recent approaches have argued this emphasis to be misplaced. In the aftermath of the cognitive revolution, much research demonstrated the impact of emotional processes on moral judgment (e.g. Haidt & Hersh, 2001; Haidt, Koller, & Dias, 1993), and some theorists have suggested that the role of reflective reasoning in moral judgment is severely limited. According for instance to the widely cited Social Intuitionist Model (SIM), reflective reasoning is usually (though not exclusively, see the correspondence between Salzstein and Kasachkoff, 2004, and Haidt, 2004) confined to post-hoc rationalizations without causal influence on the judgment process, such that moral judgments result primarily from moral intuitions (Haidt, 2001; also see Haidt, Björklund, & Murphy, 2000).

Although this radical intuitionist view is not unequivocally shared (e.g. Guglielmo, 2018; Pizarro & Bloom, 2003; Royzman, Kim, & Leeman, 2015; Stanley, Yin, & Sinnott-Armstrong, 2019; Uhlmann & Zhu, 2014), many dominant theoretical models acknowledge that moral judgment is likely to arise from an interplay of both emotional and rational processes alike (e.g. Greene & Haidt, 2002). This view, which has been prominent in the literature on human cognition more generally (e.g. Evans & Stanovich, 2013; Sloman, 2014),

has consequently been influential in shaping the paradigms that have been developed to study the moral judgment process.

1.1.1. Moral dilemmas and the Dual-Process Model of moral judgment

One particularly influential paradigm for investigating the foundations of moral cognition constitutes the application of moral dilemmas, which has been introduced into the literature by Greene, Sommerville, Nystrom, Darley, and Cohen (2001). Inspired by philosophical discussions surrounding the moral relevance of individual rights and duties versus consequences of a decision (e.g. Foot, 1967; Thomson, 1976, 1985), Greene and colleagues developed a battery of sacrificial dilemmas.

Such sacrificial dilemmas require the reader to decide whether he would actively kill one single individual in order to save a larger number of people. A canonical example would be the famous trolley-problem, in which the reader must decide whether to flip a switch in order to change the way of an incoming trolley, which is about to run over and kill five people. If the switch is flipped, the trolley will be steered out of the way of these people, but instead directed towards another track, where one single person would be overrun as a result. As intended by Greene and colleagues (2001), their dilemmas were constructed to evoke a tension between the two normative ethical systems of deontology and utilitarianism. According to deontological ethics, related to Immanuel Kant, the morality of an action is judged by its conformity to individual rights and duties that are to be accepted as binding regardless of consequences. Thus, according to this deontological view, if it is wrong to kill then it is always wrong to kill, even if multiple people are saved as a result. In contrast, according to utilitarianism related to the philosophy of John Stuart Mill, the moral value of an action is determined entirely by its consequences, such that an action that increases overall well-being should be considered moral, while an action that decreases well-being and increases suffering is deemed immoral. Consequently, the switch should be flipped and the individual sacrificed to benefit the greater good. In accordance with these ethical positions

that inspired the paradigm, the respective response options are usually labeled “deontological” and “utilitarian” (e.g. Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004).

Based on the application of this paradigm in the context of behavioral and neuroimaging research, Greene and colleagues (Greene, 2007, 2014; Greene et al., 2001, 2004, 2008) developed the influential Dual-Process Model of moral judgment (DPM). Convergent with other dual-process models (Haidt, 2001; also see Evans & Stanovich, 2013; Sloman, 2014) this theoretical account maintains that dilemma responses result from the working of two different processing systems, with different processing properties. Importantly, these processing systems are assumed to be functionally separate, and rooted in different neural mechanisms (Greene et al., 2001, 2004; Patil et al., 2019). As such, it is supposed to be the competition between these separate mechanisms that results in the experienced conflict characteristic of a proper dilemma, an assumption termed the *central tension principle* (Greene, 2014). Specifically, “deontological” responses are supposed to result from the working of a fast, automatic and effortless System 1 that is sensitive to socio-emotional stimuli. In contrast, “utilitarian” responses supposedly result from the working of a separate System 2, which is slow, deliberative, and performs effortful and rational costs-benefit analyses when the required cognitive resources are available.[1]

As for instance Valdesolo and Steno (2006) found, mood improvement can increase “deontological” judgments, which is frequently interpreted to suggest emotional processes as the mechanism underlying such responses (also see Amit & Greene, 2012; Greene, 2007; Koenigs et al., 2007; Patil & Silani, 2014). Conversely, reducing cognitive resources by applying cognitive load (Greene et al., 2008) or time pressure (Suter & Hertwig, 2011) has

[1] As will become apparent over the course of this thesis, my use of quotation marks surrounding the terms deontological and utilitarian in the context of the dilemma approach is deliberate, as I see little reason to assume that dilemma responses are most accurately described by reference to these broad ethical systems (e.g. Hennig & Hütter, 2019; Kahane, 2012, 2015; Kahane et al., 2018).

been found to reduce measures of “utilitarian” judgment (but see Tinghög et al., 2016). This is frequently accepted as indication for controlled, rational processes underlying such responses (Greene, 2014; Paxton, Ungar, & Greene, 2012).

In short, the DPM can thus be summarized as consisting of three central premises. First, dilemma responses can be meaningfully described as “deontological” and “utilitarian”, respectively. Second, these responses are produced by two functionally separate cognitive systems, engaged in automatic, intuitive, and emotional processing (System 1) and controlled, rational, deliberative processing (System 2), respectively. Third, the latent cognitive processes resulting from the working of those systems systematically lead to “deontological” (System 1) and “utilitarian” (System 2) responses, respectively.[2] All of these premises will be critically considered over the course of the current thesis. Specifically, Chapter II will mainly address premises one and three. Likewise, Chapter III will target premises one and three as well, albeit from a different angle. Finally, Chapter IV will put a closer focus on premise two.

Evidence in favor of the DPM seems substantial, and is provided by work employing different methodologies, including neuroimaging work, lesion studies, or various experimental manipulations (e.g. Amit & Greene, 2012; Greene, 2007, 2014; Greene et al., 2001, 2004, 2008; Koenigs et al., 2007; Valdesolo & Steno, 2006; Patil & Silani, 2014; Patil et al., 2019; Paxton et al., 2012; Suter & Hertwig, 2011). However, the conventional dilemma paradigm suffers from several identifiable shortcomings that significantly impact the scope of theoretical conclusions that can be confidently derived based on its application in

[2] Note that earlier versions of the DPM also contained an additional claim that required qualification upon closer inspection. As originally proposed by Greene et al. (2001, 2004), their reaction time data indicated that System 1 provided default answers to moral dilemmas, which could subsequently be overwritten by System 2. This claim endorses a clear temporal ordering of processes and characterizes the DPM as a default-interventionist model. The effect supporting this interpretation, however, has been shown to be an artefact resulting from idiosyncracies of specific dilemmas (McGuire, Langdon, Coltheart, & Mackenzie, 2009), and is not reproducible when assessed with process measures (Koop, 2013), or mathematical modeling approaches (Bago & DeNeys, 2018; Baron, Gürçay, Moore, & Starke, 2012; Gürçay & Baron, 2017; also see Cohen & Ahn, 2016).

psychological research. This, consequently, spells out some problems for the DPM. In the following sections, I will briefly discuss these problems as well as proposed solutions.

For the moment, I merely note that other models of moral judgment exist, which either conceptualize the properties of underlying processes differently (e.g. Cushman, 2013; Graham, Haidt, & Nosek, 2009; Haidt, 2001; Rosas, 2017), make no recognizable dual-process assumptions (e.g. Gray, Schein, & Cameron, 2017; Holyoak & Powell, 2016; Schein & Gray, 2018), or explicitly reject a clean mapping of processes on observable judgments (e.g. Cohen & Ahn, 2016), many of which I will touch on in later chapters of the thesis.

1.1.2. The non-independence problem and process dissociation

As first recognized by Conway and Gawronski (2013) the dependent measure of the conventional dilemma paradigm, the decision whether or not to sacrifice the individual for the greater good, is inherently ambiguous. That is, although the dilemma paradigm aims to draw conclusions about the degree to which emotional “deontological” and rational “utilitarian” processes contribute to observable judgments, both processes are captured in the same outcome measure. Therefore, the dependent measure of the conventional approach does at best represent the dominance of one process over the other. Similar arguments have also been presented by other conceptual critics, pointing out that there is no reason to assume that rejecting a deontological rule automatically entails acceptance of utilitarianism (e.g. Kahane, 2012, 2015). Phrased differently, because independent estimations of “deontological” and “utilitarian” inclinations are not provided, one cannot confidently infer underlying processes from observable responses. For instance, even if the Dual-Process Model is correct and Systems 1 and 2 are systematically connected to “deontological” and “utilitarian” responses, respectively, moral judgment data alone is insufficient for determining whether a manipulation that increases sacrificial killing is due to increased System 2- or decreased System 1 processing.

As a solution, Conway and Gawronski (2013) proposed the application of a process dissociation (PD) approach (see Jacoby, 1991) to derive independent estimations of “deontological” and “utilitarian” inclinations. In this approach, participants are not merely presented with incongruent scenarios, in which deontological and utilitarian reasoning would motivate divergent responses (i.e. proper dilemmas), but also with congruent scenarios, in which both reasoning styles would lead to the same response. The underlying theoretical model can be expressed in the form of a processing tree depicting the expected responses if a given process dominates the judgment. According to the theoretical model of the PD approach, if “utilitarian” inclinations drive a response, sacrificial killing will be accepted in incongruent and rejected in congruent scenarios. If it does not, “deontological” inclinations may determine the response in which case the killing will always be rejected. If “deontological” inclinations also do not drive the response, it is assumed that the killing will always be accepted (see Figure 1). These theoretical assumptions can be spelled out in the form of algebraic equations and subsequently solved for, yielding estimates for the parameters *U* and *D*, representing “utilitarian” and “deontological” tendencies, respectively.[3]

By now, the PD approach has been applied in multiple studies (e.g. Armstrong, Friesdorf, & Conway, 2019; Conway, Goldstein-Greenwood, Polacek, & Greene, 2018; Friesdorf, Conway, & Gawronski, 2015; McPhetres, Conway, Hughes, & Zuckerman, 2018; Muda, Niszczota, Białek, and Conway, 2018; Reynolds & Conway, 2018), and the results of many of these are interpreted as supporting the DPM. As for instance Conway and Gawronski (2013) found, the *D*-parameter of the PD approach was increased when empathic concern was

[3] For instance the likelihood of the decision to sacrifice in a congruent scenario can be expressed as: p(“sacrifice” | congruent) = $(1 - U) \times (1 - N)$. After estimating individual likelihoods dependent on congruency condition, the strength of the *U*-parameter can be determined via the following formula: $U = p$(“no sacrifice” | congruent) – p(“no sacrifice” | incongruent). Consecutively, the *D*-parameter can be determined via the following formula: $D = p$(“no sacrifice” | incongruent) / $(1 - U)$.

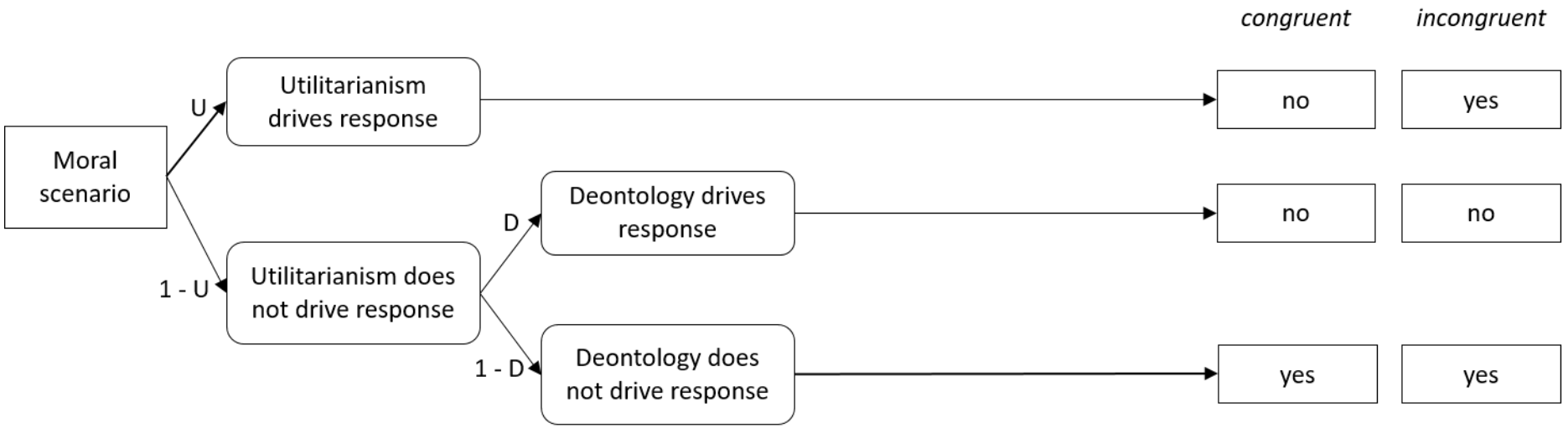

Figure 1. The PD model of Conway and Gawronski (2013), predicting sacrificial killing from utilitarian (*U*) and deontological (*D*) inclinations.

experimentally induced via the presentation of emotionally evocative pictures, while the *U*-parameter was reduced by cognitive load (see Greene et al., 2008; Patil et al., 2019; Paxton et al., 2012; Suter & Hertwig, 2011; but see Royzman, Landy, & Leeman, 2015; Tinghög et al., 2016), both of which is consistent with the central tenets of the Dual-Process Model.

1.1.3. The omission of response tendencies and multinomial processing tree modeling

Although providing a solution to the non-independence problem, the PD approach still contains another threat to the validity of its measures. Specifically, the approach rests on the strong assumption that sacrifices will always be accepted when neither “utilitarian” nor “deontological” concerns dominate the judgment process. However, as Hütter (2013) has noted, evoking equally strong deontological and utilitarian inclinations simultaneously is exactly what a proper dilemma should do *by definition*. Consequently, in such dilemmatic cases in which neither inclination dominates, a decision may be reached based upon other criteria, such that harm is not by default accepted.

Support for this assumption comes from research by van den Bos, Müller, and Damen (2011) who demonstrated that sacrificial killing judgments correlate positively with behavioral disinhibition and increase when disinhibition is experimentally induced. In a similar vein, findings by Crone and Laham (2017) suggest that, once experimentally separated, utilitarian inclinations and a general preference for action unrelated to moral considerations proper may be equally predictive of dilemma responding. These findings thus suggest general behavioral tendencies as a third influence factor underlying dilemma judgments. Consequently, the tacit assumption of the PD approach that harm will always be accepted if neither “utilitarian” nor “deontological” inclinations dominate the judgment process is likely mistaken, as in this case participants may decide based on a general preference for “doing something rather than nothing”, or vice versa, respectively.[4]

[4] This point is closely related to research on the omission bias (e.g. Baron & Ritov, 2004) and the status-quo bias (Ritov & Baron, 1992; Samuelson & Zeckhauser, 1988), as noted by Gawronski, Conway, Armstrong, Friesdorf, and Hütter (2016). In the context of moral judgment, this effect means

The problem of response tendencies has two concrete implications. First, to the extent to which dilemma judgment is determined by such general response tendencies, estimates of *U*- and *D*-parameters may be systematically distorted. Specifically, the *D*-parameter may be artificially inflated (Hütter, 2013). Second, to the extent that general response tendencies are systematically conflated with the measures of primary interest, they may contribute to spurious effects. This second concern is of high importance, because such a systematic conflation is identifiable in the scenarios used in the conventional and PD approaches alike. Specifically, the “utilitarian” resolution always requires an action and direct interference with the state of affairs described in the dilemmas, whereas the “deontological” resolution requires passivity and inertia. Thus, to the extent to which judgment is determined by these general response tendencies (Crone & Laham, 2017; van den Bos et al., 2011; also see Duke & Bègue, 2017; Patil, 2015), these may be mistaken as indication for “deontological” or “utilitarian” inclinations, respectively. Based on this recognition, Hütter (2013) proposed the application of multinomial processing tree (MPT) modeling to the study of dilemma judgment, to experimentally control for response tendencies.

Multinomial processing tree models represent a class of formal mathematical models for the analysis of categorical outcome measures, and have found wide application in psychological research (Riefer & Batchelder, 1988; for a review see Hütter & Klauer, 2016). The procedure entails several of the strengths of the approach introduced by (Jacoby, 1991). Specifically, it requires the specification of a theoretical model that makes precise qualitative predictions about the processes supposed to underlie observable responses. Compared to the PD approach, however, it also provides several important advances. First, it enables the estimation of more than two parameters, consequently allowing for the specification of more

---

that harm resulting from inaction is preferred over equivalent harm resulting from action (Cushman, Young, & Hauser, 2006). As such, judgments resulting from the omission or status quo bias may be falsely interpreted as indicative of deontological inclinations in the context of the PD paradigm (but see Baron & Goodwin, 2019).

complex and comprehensive theoretical models. Second, it applies the maximum likelihood algorithm to provide an assessment of general model fit via a chi-square test (Hu & Batchelder, 1994). Thus, unlike previous approaches, multinomial modeling allows investigating whether the underlying theoretical model provides a good explanation of the data to begin with, identifying fallacious theoretical assumptions by lack of model fit (for details see Klauer, Stahl, & Voss, 2012). As such, MPT modeling provides a powerful methodological tool, which is ideally suited to investigate the cognitive foundations of dilemma judgment.

1.1.4. The CNI model of moral judgment

Recently, a multinomial model of dilemma judgment has been developed by Gawronski et al. (2016; Gawronski, Armstrong, Conway, Friesdorf, & Hütter, 2017, 2018). As Gawronski et al. (2016, 2017) argue, in order to draw inferences about the processes underlying observable dilemma responses, relevant aspects of the supposed processes need to be manipulated independently. As such, the authors identified the relevant building blocks underlying utilitarian and deontological philosophy, respectively, to provide such manipulations (see Gawronski & Beer, 2016).

Following the logic of the PD approach (Conway & Gawronski, 2013), Gawronski et al. (2016, 2017) implemented a manipulation of congruency, by varying whether sacrificial killing would increase overall consequences (incongruent) or not (congruent). Additionally, they also presented scenarios in which commitment to moral norms would lead to the death of a single person, or not. They achieved this by manipulating whether scenarios contained a *proscriptive* norm that prohibits a decision or a *prescriptive* norm, which prescribes a decision. For instance, a proscriptive scenario would have one decide whether to torture a terrorist, in order to acquire information on where he hid several explosives (proscriptive norm: “*Do not* torture”). In contrast, a prescriptive scenario would contain judging whether

one should stop ones partner from doing so (prescriptive norm: “*Do* avoid the torturing of others”).

The resulting measurement model aims to explicitly address the problem of uncontrolled response tendencies described above by predicting moral dilemma judgments from three orthogonal processes, which are derived from the manipulations of congruency and type of norm: Sensitivity to consequences, sensitivity to norms, and inaction tendencies, represented by the parameters *C*, *N*, and *I*, respectively.

According to the CNI model, when neither sensitivity to consequences nor sensitivity to norms determine moral judgment, a preference for inaction (*I*) or a preference for action (1 – *I*) determines the response. The CNI model of moral judgment thus expands the measurement model of the PD approach by the estimation of the *I*-parameter, which represents a preference for inaction over action. As in the PD approach, the likelihood for responses dependent on experimental condition can be spelled out in the form of equations and represented in a processing tree (see Figure 2). For instance, the likelihood for norm-breaking (i.e. sacrificial killing) in a “proscriptive-congruent” scenario can be expressed as follows:

$$p(\text{“yes”}|\ \text{congruent/proscriptive}) = (1 - C) \times (1 - N) \times (1 - I)$$

The subsequent application of this model has yielded several important insights, some of which have led to a reconsideration of conclusions suggested by prior work. For instance, recent findings suggest that the relationship between testosterone-levels and moral dilemma judgment may be more complex than previously thought. Specifically, some prior research suggests that high levels of testosterone increase sacrificial killing (Carney & Mason, 2010; Chen, Decety, Huang, Chen, & Cheng, 2016; Montoya et al., 2013), an effect considered to be potentially mediated by reduced empathic concern (Hermans, Putnam, & van Honk, 2006).

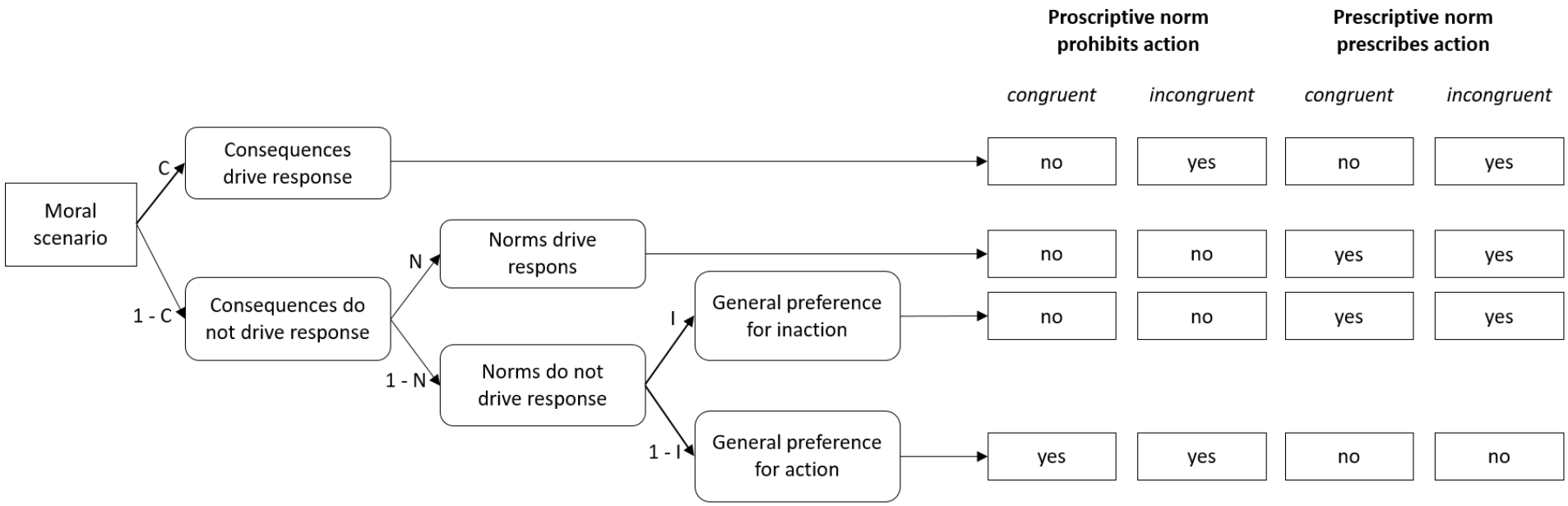

Figure 2. The measurement model underlying the CNI model of Gawronski et al. (2016, 2017), predicting norm-breaking from sensitivity to consequences (*C*), sensitivity to norms (*N*) and preference for inaction over action (*I*).

However, contrary to these findings, recent research suggests administration of exogenous testosterone to lead to *increased* norm-endorsement as represented by the CNI models *N*-parameter (Brannon, Carr, Jin, Josephs, & Gawronski, 2019). Note that this finding is the opposite of what may be expected based on the Dual-Process Model, according to which empathic concern underlies deontological norm-adherence. Specifically, if the Dual-Process Model is correct, reducing empathic concern via increasing testosterone should *decrease* norm-endorsement.[5]

Other applications of the CNI model resolve counterintuitive findings that have previously puzzled DPM-theorists. For instance, as Bartels and Pizarro (2011) have found, trait measures of psychopathy are positively related to sacrificial killing in the conventional paradigm (also see Balash & Falkenbach, 2018; Koenigs, Kruepke, Zeier, & Newman, 2012; Patil, 2015). This finding is difficult to account for based on traditional dual-process theorizing, as it does not seem reasonable to assume that high psychopathy is related to ascribing value to saving the lives of others to begin with. As Gawronski et al. (2017, Experiments 3a + 3b) showed, psychopathy is likely related to reduced endorsement of norms and consequences alike, yet also to a preference for action over inaction. As the first two effects cancel out in the context of the conventional paradigm, decreased inaction tendencies are sufficient to produce the spurious effect documented by Bartels and Pizarro (2011).

More importantly, findings by Gawronski et al. (2016, 2017) also suggest previous effects of cognitive load to be similarly spurious (e.g. Greene et al., 2008; Suter & Hertwig, 2011; Conway & Gawronski, 2013; also see Li, Xia, Wu, & Chen, 2018; Patil et al., 2019). As demonstrated in three experiments (Gawronski et al., 2016, Experiment 2; Gawronski et al., 2017; Experiments 2a + 2b), cognitive load may leave endorsement of consequences

[5] However, note also that Brannon et al. (2019) found high levels of *endogeneous* testosterone at baseline to be related to *decreased* norm-endorsement, which is consistent with the Dual-Process Model. This suggests that testosterone may influence moral judgment via more than one single mechanism.

unaffected but increase inaction tendencies instead, resulting in a spurious effect if assessed via the conventional (e.g. Greene et al., 2008; Suter & Hertwig, 2011) or PD approaches (Conway & Gawronski, 2013; Li et al., 2018; Patil et al., 2019).[6] Furthermore, when reassessing findings by Conway and Gawronski (2013) suggesting that presentation of emotionally evocative pictures increases “deontological” processing, this effect failed to replicate in two experiments (Gawronski et al., 2017, Experiments S1a + S1b).

These findings thus qualify theoretical conclusions drawn from previous research, and call cornerstone findings frequently cited as providing support for the Dual-Process Model into question. That is, evidence for the assumptions that “utilitarian” response patterns require cognitive resources, while “deontological” response patterns are selectively sensitive to socio-emotional stimuli may indeed be less powerful than commonly assumed, if proper measurement models are applied. As such, these results demonstrate the crucial importance of explicitly considering response tendencies when dilemma studies are conceptualized and interpreted, as theoretical missteps may result otherwise.

1.1.5. The proCNI model

The CNI model of moral judgment thus provides a significant improvement over previous methodological approaches, and has yielded important insights into the mechanisms underlying dilemma judgment. In addition, the model also promotes a much-needed shift towards greater precision in terminology. That is, by labeling process parameters as sensitivity to “norms” and “consequences” instead of “deontological” and “utilitarian”, respectively, it invites specific consideration of the different sorts of consequences and norms embedded in the presented scenarios. It also avoids reference to broader philosophical systems, adherence

[6] Note that recently Byrd and Conway (in press) suggested that different modes of reflection may have differential impact on moral judgment, such that *arithmetic* reflection may be related to “utilitarian” judgment, whereas *logical* reflection may be related to “utilitarian” and “deontological” judgment alike. However, though these findings are informative, the application of the PD approach renders them inconclusive. That is, as response tendencies are not controlled for these effects, similar to previous findings, may be spurious (see Gawronski et al., 2016, 2017).

to which is unlikely to be the ultimate motivator of the respective judgments. With this approach it encompasses much critique raised by conceptual critics, who have argued that the use of these philosophical descriptors is overly simplistic and ultimately misleading, and that attempts to derive normative conclusions from dilemma research may be in error (e.g. Kahane, 2012, 2015; Kamm, 2010; Kahane & Shackel, 2010; Kahane, Everett, Earp, Farias, & Savulescu, 2015; Kahane et al., 2018; but see Conway et al., 2018). However, the model also leaves room for further improvement, regarding both its employed stimulus material and experimental manipulations, as I propose.

First, and most importantly, a manipulation of norm type as prescriptive or proscriptive as employed by Gawronski et al. (2016, 2017, 2018) may lead to both statistical and conceptual problems. As results by Janoff-Bulman, Sheikh, and Hepp (2009) suggest, proscriptive norms may generally be perceived as more mandatory than prescriptive norms, whereas adherence to prescriptive norms may be perceived as more praiseworthy but less obligatory. Consequently, this suggests that sensitivity to norms may have a stronger impact on moral judgment in the “proscriptive” than the “prescriptive” versions of the CNI scenarios, which has two implications. The first implication is a conceptual one. As the *N*-parameter is averaged across norm conditions, the resulting measure may reflect a conceptually impure conglomerate of adherence to both types of norm, consequently overestimating adherence to proscriptive norms, while underestimating adherence to prescriptive norms. The second implication relates to the *assumption of invariance of processes*, which describes a critical condition that needs to be met in order to apply multinomial modeling (Hütter & Klauer, 2016). According to this assumption, processes must contribute equally to responses across different stimulus categories, otherwise parameter estimates may be distorted (Klauer, Dittrich, Scholtes, & Voss, 2015). If norms have a greater impact on dilemma judgment in the proscriptive compared to the prescriptive condition, as research by Janoff-Bulman et al. (2009) suggests, a violation of the invariance assumption seems likely.

Second, the scenarios comprised in the CNI battery vary considerably regarding their content. That is, although five of the six scenarios in the battery can be described as “sacrificial” in their incongruent version, such that the life of an individual has to be weighed against the lives of several people, these scenarios are situated in very different contexts. That is, whereas one scenario requires judging whether to commit assisted suicide, other scenarios require judging whether to approve or veto a ransom payment, or whether to stop ones colleague from performing an illegal operation. The differences in context further impede the interpretation of the *N*-parameter, as it is not obvious which norms are implemented in the experimental material. Consequently, it becomes even less clear adherence to which norms the *N*-parameter actually represents, as only some of the dilemmas entail deliberate causation of harm rather than harm as a foreseen side effect (Cushman, 2016; Moore, Clark, & Kane, 2008). Similarly, deciding whether to approve a ransom payment to save one’s fellow countrymen may be related to different moral norms (e.g. “Save your ingroup members”) than deciding whether to engage in assisted suicide (e.g. “Do not kill / respect the sanctity of human life”; see Graham et al., 2009), such that implemented norms may vary between scenarios (also see Baron & Goodwin, 2019). In addition, it could also be argued, that even within scenarios it is not always clear which response should be motivated by norm-consistency. For instance, in the assisted suicide case one could expect that norm-consistency should motivate to *abstain* from assisting the suicide (“Do not kill”), as well as to *assist* it (“Respect individual autonomy”). Thus, one could argue that in the context of the CNI model of Gawronski et al. (2016, 2017, 2018) the term “norm” remains theoretically underdeveloped and somewhat ambiguous.[7]

[7] Note that this criticism does not apply exclusively (or even specifically) to the dilemma battery of Gawronski et al. (2016, 2017). A similar critique could be made of the dilemmas in the PD-battery (Conway & Gawronski, 2013), and has been made regarding the canonical dilemmas proposed by Greene et al. (2001, 2004; see Rosas & Koenigs, 2014), which alludes to the general difficulty of constructing dilemmas that reliably assess the processes supposed to underlie observable dilemma responses (also see Bauman, McGraw, Bartels, & Warren, 2014).

We took two steps to alleviate these concerns in our research. First, we made sure to use only scenarios in which the dilemma decision entails a causation of harm that is deliberate, rather than incidental (Cushman, 2016; Moore et al., 2008). Second, we replaced the manipulation of norm-type with a manipulation of the scenarios default-state. Specifically, we experimentally manipulated whether an action leading to sacrificial killing was not initiated yet but could be started (inaction default), or was already started but could be stopped prematurely (action default). Thereby, across default-state conditions, the same norm could be adhered to by inertia and acceptance of the status-quo described in the scenario (inaction default) or by interference and change (action default), respectively. As such, the decision to accept the scenario without interference conceptually converges with a response motivated by a status-quo or default-bias (see Everett, Caviola, Kahane, Savulescu, & Faulmüller, 2015; Johnson & Goldstein, 2003; Samuelson & Zeckhauser, 1988). We thus estimated the *I*-parameter of our model from this default-state manipulation (see Figure 3).

In conjunction, these two changes allowed us to focus on implementing the same norm against deliberate killing in all of our scenarios, such that the conceptual clarity of the *N*-parameter is increased and the likelihood for violations of the invariance assumption reduced. As these adjustments were designed to achieve the implementation of the same *proscriptive* norm across scenarios, we have consequently named our model the proCNI model.

1.1.6. The current thesis

The aim of the current thesis is to assess several of the core issues introduced above, and to provide an investigation of several fundamental assumptions on which the moral dilemma approach rests. To this end, it presents three empirical chapters, addressing different aspects of relevance for moral dilemma research.

Chapter II introduces the proCNI model of moral judgment, which aims to avoid the problems identified in the original CNI model. It describes an application of this proCNI model, in which we critically investigate core assumptions of previous approaches to dilemma

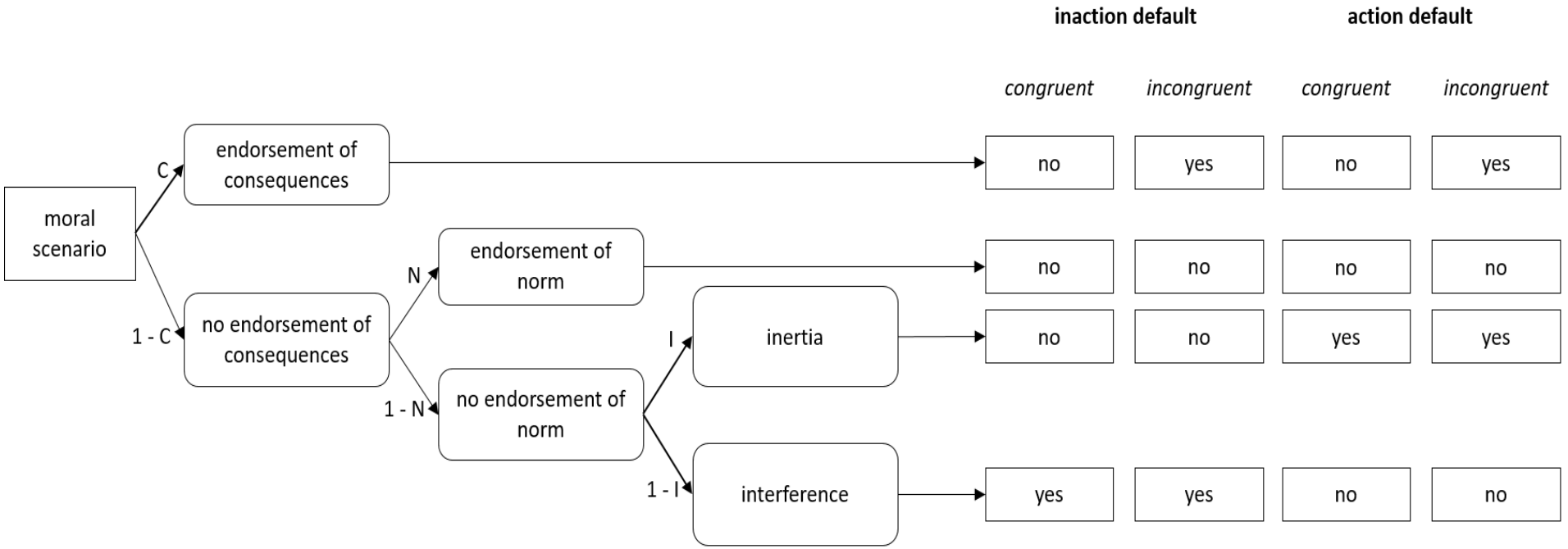

Figure 3. The proCNI model employed in the experiments presented in Chapters II and IV, predicting norm-breaking from endorsement of aggregate consequences (*C*), norm-endorsement (*N*), and inertia (*I*).

judgment and present methodological, theoretical, as well as some normative conclusions, partially based on a reassessment of previous research. Specifically, on a methodological level we investigate the impact of confounds that are identifiable in many canonically used dilemmas and discuss their potential to bias theoretical conclusions. On a theoretical level, we apply manipulations of various forms of consequences to investigate the assumption that dilemma responses are best characterized by assuming a sharp divide between “deontological” and “utilitarian” processing, as prominent dual-process accounts of dilemma judgment suggest.

Subsequently, Chapter III assesses the difficulty ratings collected in the last two experiments presented in Chapter II. In doing so, it addresses several predictions derived from Subjective Utilitarian Theory (Cohen & Ahn, 2016), an alternative account that does not operate on dual-process assumptions and aims to explain dilemma judgment parsimoniously in terms of a single process.

Chapter IV provides a critical investigation of the Dual-Process Model at the example of the foreign language effect in moral dilemma judgment (Keysar, Hayakawa, & Ahn, 2012). Here we apply the proCNI model to investigate the mechanisms underlying this effect, about which previous research using the conventional and PD approaches comes to contradictory conclusions.

Finally, Chapter V concludes the thesis by first considering parallels between moral dilemma research and research on the effect of moral dumbfounding, regarding both their canonical interpretations and some identifiable problems related to both approaches. Subsequently it integrates the presented findings with other models of moral judgment unrelated to the dilemma literature, and discusses limitations as well as opportunities for further refinement of the current paradigm.

Chapter II

Revisiting the Divide between Deontology and Utilitarianism in Moral Dilemma Judgment: A Multinomial Modeling Approach

Max Hennig & Mandy Hütter

Eberhard Karls Universität Tübingen

The following chapter contains an empirical article resulting from a cooperation between Max Hennig (lead author) and Prof. Dr. Mandy Hütter (second author). The manuscript entitled “Revisiting the Divide between Deontology and Utilitarianism in Moral Dilemma Judgment: A Multinomial Modeling Approach” is currently in press at the *Journal of Personality and Social Psychology*, and is available as an advance online publication, http://dx.doi.org/10.1037/pspa0000173. Both authors contributed equally to the research project. Specifically, both authors contributed approximately 50% to the generation of scientific ideas, data generation, analysis and interpretation, and paper writing, respectively.

2.1. Introduction

Research on moral judgment—humans’ judgments on which behaviors or attitudes they consider ”right” or ”wrong”—traditionally has been inspired by philosophical thought experiments, a well-known example of which is the trolley scenario (Foot, 1967; Thomson, 1976, 1985). In the commonly used variant of this scenario, participants are asked whether they would flip a switch in order to redirect a runaway trolley, which is heading towards five people about to be killed. By flipping the switch, their lives would be saved, but the trolley would kill one person on another track. The scenario is designed to evoke a conflict between the normative ethical systems of *deontology* and *utilitarianism*. According to deontological reasoning, related to the philosophy of Immanuel Kant, morality is based on consistency with universal moral norms relating to individual rights and duties, which are to be accepted as absolute and binding regardless of the concrete consequences of the resulting actions. If one thus considers killing to be wrong in the deontological sense, this would imply that redirecting the trolley is unacceptable even if multiple lives would be saved. In contrast, according to utilitarian reasoning, related to the philosophy of John Stuart Mill, the objective consequences of an action in the particular context in which it takes place determine its moral value. If the action produces an increase in overall well-being and decrease in suffering then it

is considered moral; if it decreases overall well-being and increases suffering, it is deemed immoral.

We consider several methodological and conceptual aspects of this approach and introduce a variant of a recently proposed multinomial processing tree (MPT) model of moral dilemma judgment (Gawronski, Armstrong, Conway, Friesdorf, & Hütter, 2017). We apply it in five experiments to investigate the impact of different types of consequences and personal involvement on the endorsement of consequences and norms. To anticipate the results of our experiments, the parameter typically thought to represent “deontological” responding shows a strong sensitivity to different types of consequences, which speaks against its conceptualization as concerned primarily with absolute moral prohibitions. Therefore, we propose that an emphasis on philosophical language (e.g., deontology, utilitarianism) is unhelpful and potentially misleading. In line with recent theories of moral judgment we conclude that the dilemma paradigm may be most parsimoniously viewed through a consequentialist lens, in which the processes underlying different response patterns represent sensitivity to different sorts of consequences.

2.1.1. Explaining moral dilemma judgment

The most influential and widely accepted theory in the domain of moral dilemma judgment is Greene and colleagues’ Dual-Process Theory (Cushman, Young, & Greene, 2010; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Paxton, Ungar, & Greene, 2012). According to this theory, “deontological” judgment reflects the output of an intuitive, fast, and effortless System 1 that is sensitive to social-emotional stimuli. By contrast, ”utilitarian” judgment is supposed to result from deliberative, slow, and effortful cost-benefit analysis performed by a functionally separate System 2.

An alternative explanatory model was proposed by Cushman (2013). According to his Dual-Systems framework that contrasts action values with outcome values, “deontological”

judgment primarily represents aversion to performing the negatively valenced *actions* that are necessary for the sacrificial killing (e.g., pushing, stabbing etc.). Crucially, there is some evidence suggesting that moral dilemma judgments are indeed sensitive to such action properties (Miller & Cushman, 2013). According to this view, it is thus not sensitivity to moral norms but sensitivity to the negativity of the necessary action, which is causally responsible for not killing the target (Miller, Hannikainen & Cushman, 2014; but see Reynolds & Conway, 2018).

Subjective Utilitarian Theory (Cohen and Ahn, 2016) proposes a much simpler single-process account. It suggests that moral dilemmas instigate a comparison process between two options (e.g., do not kill one individual versus save five individuals) in which respondents try to identify and execute the choice that is valued most by them. Contrary to Dual-Process Theory, Subjective Utilitarian Theory does not explicitly refer to the endorsement of norms and conceptualizes affective processes as only one factor that influences the overall valuation process. Thus, Subjective Utilitarian Theory would explain moral dilemma judgments by reference to the relative subjective value of the response options, without making assumptions about the emotional or rational causal antecedents of a judgment. Supporting this account, Cohen and Ahn (2016) showed that the overlap in subjective values between two options predicts both choices and response times in moral dilemmas.

As our discussion of the differences between prominent theories of moral dilemma judgment already suggests, moral dilemma research has to face a noteworthy problem. Although the paradigm has been frequently employed in research since the seminal work of Greene et al. (2001) and was influential in shaping the state of the literature on moral psychology (Bauman et al., 2014; Kahane & Shackel, 2010) it is still a matter of debate how dilemma judgments should best be interpreted. In response to conceptual critics (e.g., Kahane, 2012; Kamm, 2009; see also McGuire, Langdon, Coltheart, & Mackenzie, 2009), Greene (2014) has suggested that calling a judgment “utilitarian” merely means that it is easily

justified by reference to utilitarianism, while it is more difficult to justify by reference to deontology. Defending the use of these philosophical descriptors, Conway, Goldstein-Greenwood, Polacek, and Greene (2018) have recently extended this argument by introducing a taxonomy according to which a judgment may qualify as “utilitarian” with reference to five different standards. While the mere consistency with utilitarian analysis highlighted by Greene (2014) would indicate “level-1 utilitarianism”, additional criteria would qualify the judgment to rise to higher levels (e.g., signs of *some sort of* cost-benefit analysis represents level-2 utilitarianism while *concrete concern for the greater good* in the context of the dilemma decision represents level-3 utilitarianism). While this taxonomical approach is helpful, it also highlights that conventional dilemma judgments are ambiguous, and that level-1 utilitarian judgments may arise for several reasons other than utilitarian-style cost-benefit analysis. This includes the possibility of a distinctly deontological process in which competing moral rules like “do not kill” and “save people’s lives” are weighed against one another (Kahane, 2015). Thus, while the use of the philosophical terms “deontological” and “utilitarian” has been linguistic convention for a long time, this ultimately seems misleading. Consistent with the approach recently proposed by Gawronski et al. (2017) we will therefore avoid referring to deontology and utilitarianism as far as possible, as there is little reason to believe that the processes underlying dilemma responses are best described by reference to such broad ethical systems. Instead, we refer to judgment patterns as consistent with “consequences” and “norms”, the defining features of utilitarianism and deontology. We will revisit this in the General Discussion.

The conventional dilemma approach uses scenarios like the trolley dilemma described above in which consequences and norms suggest opposite responses (e.g., Bernhard et al., 2016; Gleichgericht & Young, 2013; Koenigs et al., 2007; Laakasuo, Sundvall, & Drosinou, 2017; Sarlo et al., 2012; Shenhav & Greene, 2014; Suter & Hertwig, 2011; Youssef et al., 2012). Although this approach has sparked much research interest, critics pointed out two

problems that jeopardize the interpretability of the obtained results. First, as already noted above, conventional dilemma judgments are inherently ambiguous. For instance, flipping the switch in the trolley dilemma could imply an endorsement of consequences (the canonical interpretation) as well as a rejection of norms (Conway & Gawronski, 2013; Kahane, 2015). Consequently, an estimation of the absolute degree of endorsement of these influence factors is not provided.[8]

Second, considering “utilitarian” and “deontological” styles of information processing the only possible processes underlying moral judgment neglects reasonable alternative interpretations of the results (Gawronski et al., 2016, 2017). The decision to flip the switch and redirect the trolley may as well follow from response tendencies that are unrelated to moral considerations proper (cf. Cohen & Ahn, 2016). For instance, recent findings suggest that activating a promotion (compared to prevention) focus increases sacrificial killing, suggesting an influence of general motivational tendencies on moral dilemma judgments (Gamez-Djokic & Molden, 2016; see also Robinson, Joel, & Plaks, 2015). This finding may be subsumed under a general preference for action or interference with the status-quo that may be sufficient to facilitate an action irrespective of its consequences. This is a serious problem for the conventional approach, because it systematically confounds sacrificial killing with action and norm-adherence with inertia. Indeed, when these factors are experimentally separated, it appears that action tendencies are about as predictive of sacrificial killing as sensitivity to consequences (Crone & Laham, 2017). Further support for this assumption stems from the documented positive relationship between behavioral disinhibition and

[8] As inverse relations pose a general problem for the interpretation of the data, they do also impose boundaries on the theoretical implications that may be drawn. As Gawronski et al. (2016) pointed out, it seems inconsistent to investigate the central assumption of the Dual-Process Theory, that utilitarian and deontological responding are the result of two functionally independent processing systems (Greene et al., 2004; Greene, 2007), by using a paradigm which conceptualizes the engagement of these processes as directly related.

sacrificial killing in conventional dilemmas (van den Bos, Müller, & Damen, 2011; cf. Balash & Falkenbach, 2018; Duke & Bègue, 2015).

To address this limitation, Gawronski and colleagues (2016, 2017) introduced MPT modeling to the study of moral dilemma judgment. MPT modeling is a statistical technique that allows dissociating the processes or classes of processes underlying responses in categorical measures (Riefer & Batchelder, 1988; for a review, see Hütter & Klauer, 2016). Thus, it is particularly well suited to quantify the psychological foundations of moral dilemma judgments, which are usually derived from dichotomous responses (‘yes’ vs. ‘no’ or ‘acceptable’ vs. ‘unacceptable’). This modeling approach requires the researcher to specify a-priori the contributors to responses across different experimental conditions.

The CNI model assumes that moral dilemma judgments are driven by three orthogonal psychological bases: consequences (*C*), norms (*N*), and inaction tendencies (*I*). The model conceptualizes the endorsement of norms (*N*) as representing a commitment to abstaining from causing direct harm (adhering, for instance, to the ME-HURT-YOU criteria by Greene et al., 2004; see also Gray, Young, & Waytz, 2012; Schein & Gray, 2018), regardless of whether the resulting consequences maximize overall well-being, or whether the decision requires inaction or action. Similarly, endorsement of consequences (*C*) reflects the maximization of lives saved, such that it leads to norm-breaking only if this results in more survivors than deaths on an aggregate level (i.e., considering all individuals in a given scenario), again independent of whether inaction or action is required. Finally, inaction tendencies quantify one’s preference for inaction over action regardless of the norms or consequences involved in the decision. Applying this model, (in)action tendencies turned out to be predictive of moral dilemma judgments over and above the endorsement of consequences and norms (Gawronski et al., 2016, 2017, 2018). Moreover, removing the confound between endorsement of consequences and preference for action reversed previous findings, which suggested an influence of cognitive load on the *C*-parameter. As the application of the CNI model clarified,

cognitive load increases inaction tendencies while leaving other processes unaffected (Gawronski et al., 2016, 2017; see also Zhang, Kong, Li, Zhao, & Gao, 2018). This finding casts doubt on the claim that imposing cognitive load or time constraints interferes with utilitarian-style cost-benefit analysis, which is a central tenet of the Dual-Process Theory (e.g., Greene 2014; Greene et al. 2008; Suter & Hertwig, 2011).

2.1.2. Viewing moral norms through a consequentialist lens

The terminology proposed by Gawronski et al. (2016, 2017) increases precision and reduces potential for theoretical confusion. It thus represents an important step towards focusing on identifiable characteristics of judgment patterns while avoiding unnecessary reference to broader philosophical positions that are not well reflected in observable dilemma responses (Aktas, Yilmaz, & Bahçekapili, 2017; Conway et al., 2018; Kahane et al., 2018). We suggest that this rationale should be taken further by acknowledging that morally relevant norms and consequences are more tightly interwoven than their systematic juxtaposition in almost two decades of dilemma research may suggest (see Gray & Schein, 2012). That is, it is not at all obvious why the term “moral norm” should only apply to the principle “do not kill” but not to the principle “always maximize the greater good.”[9] Conversely, the reason why some norms are almost universally accepted (“Do not kill”) while others are not (“Do not eat shellfish”) may lie in their direct connection to tangible consequences regarding well-being and suffering. It has been suggested that a focus on moral norms may result from the overgeneralization of principles that usually produce desired consequences, and have consecutively become detached from those principles and followed for their own sake (Baron, 1994). According to this view, the reason for endorsement of norms is ultimately a concern for achieving positive consequences (Sunstein, 2005). Both of these points suggest that assuming

[9] Note that this argument does not depend on the use of the terms “norms” and “consequences”. That is, it is equally unclear in what meaningful aspects “always maximize the greater good” does not constitute a deontological rule.

a hard split between “norms” and “consequences” (let alone “deontology” and “utilitarianism”) as independent motivators of dilemma judgment may be overly simplistic.

We adopt a perspective of what Sunstein (2005) referred to as “weak consequentialism”. According to this view, the moral judgment process consists of a broadly consequentialist cost-benefit analysis, which may recognize violations of rights and duties as morally relevant consequences, and which does not have to be explicitly utilitarian. We suggest that viewing dilemma response patterns through this lens and explicitly considering the consequences, which may motivate adherence to norms, is ultimately helpful for several reasons. First, this approach is empirically sound. It converges with the findings underlying Subjective Utilitarian Theory (Cohen & Ahn, 2016), which suggest dilemma judgments to result from a single process of consequentialist weighing of costs and benefits. It also fits with accumulating evidence from areas of moral psychology unrelated to the dilemma literature, according to which morally relevant guidelines for behavior are intimately tied to the perceptions of harmful consequences (e.g., Gray & Schein, 2016; Schein, Ritter, & Gray, 2016; also see Schein & Gray, 2018).[10] Second, it increases conceptual precision by avoiding conceptualizations that falsely suggest a clear-cut difference between norms and consequences (let alone “deontology” and “utilitarianism”) as being independent from one another. Third, it invites explicit consideration of the different consequences built into the experimental material, and thereby promotes clarity regarding the conclusions that are warranted based on the application of the paradigm.

[10] Though focusing on the role of consequences as potential mediators of norm-consistent dilemma judgment, we recognize that the moral judgment process can also be conceptualized in terms of balancing the demands of different and at times conflicting moral duties (Holyoak & Powell, 2016; Kahane, 2015). The term *weak consequentialism*, which we endorse in the context of our discussion, as such has a closely related counterpart in what Holyoak and Powell (2016) termed *moderate deontology*. That is, applied to the dilemma literature, we consider focusing on the mediating role of consequences to be the more parsimonious approach, while we acknowledge that other moral questions may be better conceptualized in terms of interlocking rules.

2.1.3. The role of self-relevant consequences, death avoidability, and personal involvement

We will now consider the influence of self-relevant consequences, death avoidability, and personal involvement, on the processes underlying dilemma judgments, as we regard them informative for theoretical and methodological reasons. On a theoretical level, at least two of these factors can be considered instances of different sorts of consequences. Combined with the MPT approach, these manipulations allow us to investigate the degree to which norm-consistent response patterns show sensitivity to consequences. On a methodological level, these factors are important, because they are frequently confounded in the canonically used stimulus material introduced by Greene and colleagues (2001, 2004; Rosas & Koenigs, 2014). We assess the degree to which these confounds create problems for the interpretation of the data, as well as for the theoretical and normative conclusions that can be derived.

Self-relevant consequences. Self-relevant consequences straightforwardly concern the consequences of the moral decision for the person who judges its acceptability. As Rosas and Koenigs (2014) pointed out, many of the conventionally employed scenarios contain a strong element of self-interest (see also Kahane 2015). For instance, in several scenarios of the original set (Greene et al., 2001), sacrificial killing would lead to positive consequences only for the judge and not for others, resulting in a contrast between norm-adherence and egoistic concern for oneself, rather than impartial concern for general well-being.[11] Similarly, in several other scenarios (Greene et al., 2004) the positive consequences of sacrificial killing apply to others and to the judge alike. The *crying baby* scenario, for instance, presents one with a choice between suffocating a baby by muffling its screams or letting the baby cry. The consequence of the second option is that oneself, the baby, and several others are killed by hostile soldiers. In this case, sacrificial killing could thus be explained by an impartial concern

[11] We do not use the term “egoism” to confer a value judgment, but rather to reference to egoism as an ethical system, according to which self-interest forms the foundation of morality, and it is thus moral and/or rational to do what promotes personal interests (Shaver, 2019).

for maximizing overall well-being or a purely egoistic concern for maximizing one’s own well-being. Such self-relevant consequences have been shown to influence sacrificial killing (Christensen, Flexas, Calabrese, Gut, & Gomila, 2014; Kahane, Everett, Earp, Farias, & Savulescu, 2015; Koop, 2013; Lotto, Manfrinati, & Sarlo, 2014; Moore, Clark, & Kane, 2008; Moore, Lee, Clark, & Conway, 2011; also see Bonnefon, Shariff, & Rahwan, 2016). These results cast doubt on the interpretations of responses in line with consequences in such confounded dilemmas, and question the common assumption that such responses are determined mainly by a concern for the greater good (Balash & Falkenbach, 2018; Bartels & Pizarro, 2011; see also Conway et al., 2018). Instead, they call for a cautious interpretation of the psychological bases underlying responses that are in line with utilitarian philosophy (see Aktas et al., 2017; Kahane, 2015; Kahane et al., 2015, 2018) and underscore the importance of explicitly considering the nature of the consequences built into a dilemma when drawing theoretical and normative conclusions.

In addition, due to the limitations of the conventional approach, it is still an open question which of the underlying processes are affected by the presence of self-relevant consequences. It may be the case that self-relevant consequences enter an analysis of aggregate costs and benefits, in which they are then given higher importance than other-relevant consequences. In this case, endorsement of sacrificial killing would qualify as level-2 utilitarian according to the criteria of Conway et al. (2018), in the sense that it represents *some sort of* analysis of costs and benefits, while being ultimately based on egoism. Alternatively, it may be the case that sacrificial killing is increased because egoistic cues reduce the relevance assigned to the moral norm, such that acceptance of direct harm becomes more likely. One goal of this paper is to address this question and to provide an investigation of the effect of self-relevant consequences on the endorsement of consequences (the *C*-parameter) and norms (the *N*-parameter).

Personal involvement. A vast amount of evidence indicates that norm-endorsement as assessed with the conventional paradigm is higher when considering *personal* as opposed to *impersonal* moral violations (Cecchetto, Rumiati, & Parma, 2017; Koenigs et al., 2007; Koop, 2013; Kusev, van Schaik, Alzahrani, Lonigro, & Purser, 2016; Moore et al., 2008, 2011; Moretto, Làdvas, Mattioli, & di Pellegrino, 2010;). According to the original definition introduced by Greene et al. (2001), impersonal moral violations allow for high levels of psychological distance and low personal involvement during the act of violation, whereas personal moral violations enforce low levels of psychological distance and high levels of personal involvement. More specifically, a moral violation qualifies as personal if it is an action that causes harm and does not result from the deflection of an existing threat onto a different party, such that the deciding agent directly generates the harm herself (Greene et al., 2004; Moore et al., 2008). Personal involvement is thus closely tied to the concept of action aversion (Miller et al., 2014), which may underlie the rejection of sacrificial killing in the dilemma context (Miller et al., 2014; Reynolds & Conway, 2018).

By disentangling personal and impersonal dilemmas in our MPT model we can test a central postulate of the Dual-Process Theory (Greene et al., 2004), as well as a prediction of the Dual-Systems framework (Cushman, 2013; Miller et al. 2014). According to the Dual-Process Theory, personal violations lead to increased emotional engagement, which in turn increases sensitivity to norms while leaving sensitivity to consequences unaffected. This predicts an increase of the *N*-parameter, but no change of the *C*-parameter, when personal involvement is high as opposed to low. This effect would also be expected based on the Dual-Systems framework, because high-involvement actions possess more aversive properties than low-involvement actions (e.g., shooting someone is more aversive than delivering information that lead to that person being shot).

Death avoidability. Death avoidability represents another manipulation of morally relevant consequences. That is, when death of the single target is inevitable (as compared to

avoidable) the consequence of either decision (accept the killing of the target or not) are identical for the target (i.e., the target will die). As previous studies suggest, killing is generally judged more acceptable if it would merely hasten an inevitable death than if it would end the life of someone who would otherwise survive (Christensen et al., 2014; Koop, 2013; Moore et al., 2008, 2011; Suessenbach & Moore, 2015). On a methodological level, a high proportion of death-inevitable scenarios within a stimulus set (e.g., the crying baby scenario described above) would likely lead to an artificial inflation of sacrificial killing. Death avoidability should thus be considered in creating scenarios and interpreting results.

On a theoretical level, this factor is intriguing because it represents an interesting manipulation of consequences irrespective of action properties and norms (Cushman, 2013). That is, the action of stabbing someone in the heart is equally repellent, irrespective of whether or not the person would have survived otherwise. Likewise, an absolutist proscription of killing is also still in place, even when a person is destined to die. If it is indeed the case that the *N*-parameter is best described in terms of action aversion or emotional norm-adherence, it should remain unaffected by a manipulation of death avoidability. However, viewing dilemma response patterns through a consequentialist lens suggests a different hypothesis. If, as we maintain, adherence to “do not kill” norms is ultimately followed because it is generally efficient in achieving the consequence of avoiding unnecessary death (Baron, 1994; Sunstein, 2005; see Schein & Gray, 2018), this would predict an effect on the *N*-parameter. Thus, in the case of inevitable death the *N*-parameter should be reduced, because the desired consequences of saving the life of the individual cannot be achieved.

2.1.4. Overview of the Present Research

We conducted five experiments that implement a novel variant of the CNI model. Our proCNI model operates on proscriptive norms only.[12] To create four scenario conditions that

[12] We are aware that the term “norm” can be used in a descriptive way, such as describing general rules of conduct usually followed by members of a society or culture. However, we do not

allow for the estimation of three parameters, we manipulated the default state orthogonal to congruency. Specifically, we designed scenarios in which the same proscriptive norm may be adhered to by continuing (inaction-default) or changing (action-default) an ongoing behavior. Correspondingly, we also changed the name of the response tendency captured by the *I*-parameter, which results from this manipulation. While the CNI model conceptualizes the measured process as general inaction tendencies, we refer to it as *inertia*, as we consider this conceptualization to reduce ambiguity. That is, the term "inaction tendencies" is ambiguous about whether it refers to maintenance of a current state of affairs regardless of its content, or to endorsing the passive behavioral option in the context of the dilemma. In the CNI model both of those bases overlap, because maintenance of a status quo never requires an action. Although this also seems to be the case regarding most decisions in real life, there are some exceptions (e.g., when both administering a standard and an alternative treatment for a disease require action; e.g., Baron & Ritov, 2004; Samuelson & Zeckhauser, 1988). In accordance with previous work, we consider our model's *C*-parameter to represent the consistency of participants' responses with consequences, that is, the sum of all lives saved versus lost as the result of a decision. We refer to the *N*-parameter as representing the consistency with moral norms that forbid the causation of direct harm.

The parameters of an MPT model represent latent processes that are estimated via their relation to observable responses across a set of experimental conditions.[13] Figure 3

---

refer to norms in this descriptive sense. Instead, we use the term norm as capturing a commitment to certain moral principles centered aroung the repudiation of harm causation (for a discussion see Schein & Gray, 2018).

[13] Compared to running an ANOVA over the four scenario types, MPT modeling has important advantages. First, MPT modeling is a data reduction technique. Specifically, the proCNI model reduces the information from four conditions to three parameters, which facilitates the interpretation of results. Second, while an ANOVA can only determine *whether* conditions differ in the frequency of responses, the proCNI model can also determine *why* response frequencies differ. That is, applying MPT models allows testing hypotheses about the latent cognitive processes underlying observable responses. With regard to the proCNI model, the parameters estimate the endorsement of three psychological bases underlying moral dilemma judgments. Third, MPT modeling allows testing effects of experimental manipulations on measures of these (classes of) cognitive processes, rather than on response frequencies. While a change in response frequencies may be ambiguous (i.e., a

presents the processing tree model of the proCNI model with the responses each parameter predicts in the four experimental conditions. The parameter *C* thus expresses the likelihood that a response is consistent with consequences. In the case that a response is not in line with consequences (with the probability $1 - C$), it is norm-consistent with the probability *N*. In the case that a response is neither consistent with consequences nor with norms ($1 - C \times 1 - N$), a general preference for inertia over interference drives the response with the probability (*I*).

As a consequence, the likelihood of norm-adherence versus norm breaking in a given condition can be expressed in the form of a simple equation. These equations underlie the maximum-likelihood estimation algorithm (Hu & Batchelder, 1994) and need to be formulated for all experimental conditions. For instance, the willingness to engage in norm breaking (“yes”) in the congruent/inaction-default condition is the result of the combination of processes in the lowest branch of the model:

$$p(\text{“yes”}|\ \text{congruent/inaction-default}) = (1 - C) \times (1 - N) \times (1 - I)$$

In the incongruent/action-default condition, the combination of processes in the first and the third branch of the model lead to the breaking of the norm:

$$p(\text{“yes”}|\ \text{incongruent/action-default}) = C + (1 - C) \times (1 - N) \times (I))$$

In Experiment 1, we assessed the viability of our approach by testing the fit of the proCNI model to empirical data. Experiment Series 2 assessed the potential impact of self-relevant consequences on both the *C*- (Experiment 2a) and *N*-parameters (Experiment 2b). We did so by implementing self-relevant consequences parallel to consequences (Experiment 2a) and the proscriptive norm (Experiment 2b), respectively. In Experiment Series 3, we applied a

---

decrease in sacrificial killing may be due to increased norm-endorsement, decreased endorsement of consequences, an increase in inertia, or a combination of these), the proCNI model allows determining the source of changes in response frequencies. Note that the *I*-parameter also fulfills a more general quality control function in the sense that it contains much of the variance that does not correspond to the predictions of the preceding parameters.

revised manipulation of self-relevant consequences to investigate their impact independent of their experimental implementation. Additionally, we investigated the effects of personal involvement and death avoidability. Testing the model parameters' sensitivity to each of these factors enables a test of functional hypotheses derived from Dual-Process Theory, the Dual-Systems framework, and a weak consequentialist perspective.[14]

2.2. Experiment 1

Experiment 1 served the purpose of validating the proCNI model by assessing its fit to empirical data. We used a preliminary set of dilemmas based on some of the stimuli used by Conway and Gawronski (2013), that did not contain self-relevance, personal involvement, and death avoidability as identifiable confounds (see Appendix A).

2.2.1. Method

Participants. To estimate the required sample size we applied a widely-used rule of thumb in MPT research, according to which estimation of reliable model parameters requires that not more than 10% of the expected category frequencies are below five (Klauer, Stahl, & Voss, 2012). Assuming the parameter estimates reported by Gawronski et al. (2016), we

[14] Note that, while the proCNI model allows testing psychological theories on moral dilemma judgment, it represents a *measurement model*, not a *psychological theory*. That is, the model represents our expectations regarding the distribution of responses as a function of the experimental conditions and parameters. We make no assumptions regarding the cognitive processes (i.e., their operating principles and operating conditions) underlying the endorsement of consequences, the endorsement of norms, and inertia. Furthermore, the ordering of the parameters in the model does not imply a certain temporal sequence of processing or the assumption that one judgment basis is more influential in determining responses than another. Nevertheless, wherever theoretical considerations do suggest a certain ordering of parameters (e.g., Jacoby, 1991; Klauer & Wegener, 1998; Unckelbach & Stahl, 2009; see also Klauer, Dittrich, Scholtes, & Voss, 2015), it should be reflected by the measurement model. In the strictly successive ordering of the parameters that characterizes our measurement model, the ordering of the parameters has virtually no influence on the results. This is because the proportion of responses explained by each parameter is identical across the different model specifications, which in turn can be explained by the multiplicative relationship between parameters reflected by the equations above (e.g., the proportion of responses explained by the *C*-parameter is *C* in the proCNI model versus (1 - *N*) × C in a proNCI model). Nevertheless, we also provide the results of a pro*NC*I model, in which the *N*-parameter constitutes the dominant model parameter and the *C*-parameter is estimated conditionally on the absence of *N* in the supplemental materials. With the exception of Experiment 2b, the analyses produced identical results regarding model fit and identical conclusions regarding hypothesis tests. In the supplemental material, we also provide the results of more traditional analyses of variance, which generally converge with the results we report here.

simulated expected cell frequencies using MultiTree (Moshagen, 2010). The simulation resulted in a minimum sample size of 45 participants. In line with considerations for subsequent experiments, in which we implemented an experimental factor that targeted one of the three parameters, we doubled this number for reasons of comparability. Optimal counterbalancing of scenarios required us to collect multiples of 16. Therefore, we collected data of 96 university students of diverse majors, two of which were excluded from analysis because of missing data (16 male; M_{age} = 22.64, SD_{age} = 3.77). The experiment was conducted in a laboratory setting and participants were compensated with 8.00€ and a bar of chocolate for one hour of their time. The experiment was conducted as the last of three studies, the other two unrelated to moral judgment.

Design. This experiment implemented a 2 (congruency: congruent vs. incongruent) × 2 (default state: action-default vs. inaction-default) within-participants design. Assignment of the specific versions of scenarios was counterbalanced across participants (see below).

Materials and procedure. Participants were presented with modified versions of the scenarios *torture, hard times, vaccine policy,* and *border crossing* (see Appendix A) adapted from Conway and Gawronski (2013). Each scenario was used in four versions, differing in (1) whether endorsement of consequences would suggest sacrificial killing (incongruent) or no sacrificial killing (congruent), and (2) whether an action resulting in sacrificial killing could be committed but was not initiated yet (inaction-default), or was already initiated and could be stopped prematurely (action-default). Participants had to indicate whether they would perform the described action, thus initiating or aborting the sacrificial killing, using the response options “yes” and “no.” The proCNI model was applied to the responses on this item. Additionally, participants indicated how difficult they found reaching a judgment on a scale from 1 (very easy) to 5 (very difficult). Each participant saw each content only once, with the assignment of the different versions to the scenarios in a counterbalanced manner. Their order was fully randomized (see Schwitzgebel & Cushman, 2012).

2.2.2. Results

Analyses of the proCNI model were conducted with MultiTree (Moshagen, 2010) for all experiments. The general model fit the data well, indicated by the nonsignificant deviation of the predicted from the observed response frequencies, $G^2(1) = 1.18$, $p = .277$, $w = 0.056$. The estimate of the *C*-parameter was $C = .42$ (95% confidence interval (*CI*) [.34, .51]). The *N*-parameter's estimate was $N = .53$ (95% *CI* [.39, .67]). The *I*-parameter ($I = .56$, 95% *CI* [.41, .71]) did not differ from 0.5, $\Delta G^2(1) = 0.63$, $p = .427$, $w = 0.041$, indicating no preference for inertia or interference.[15]

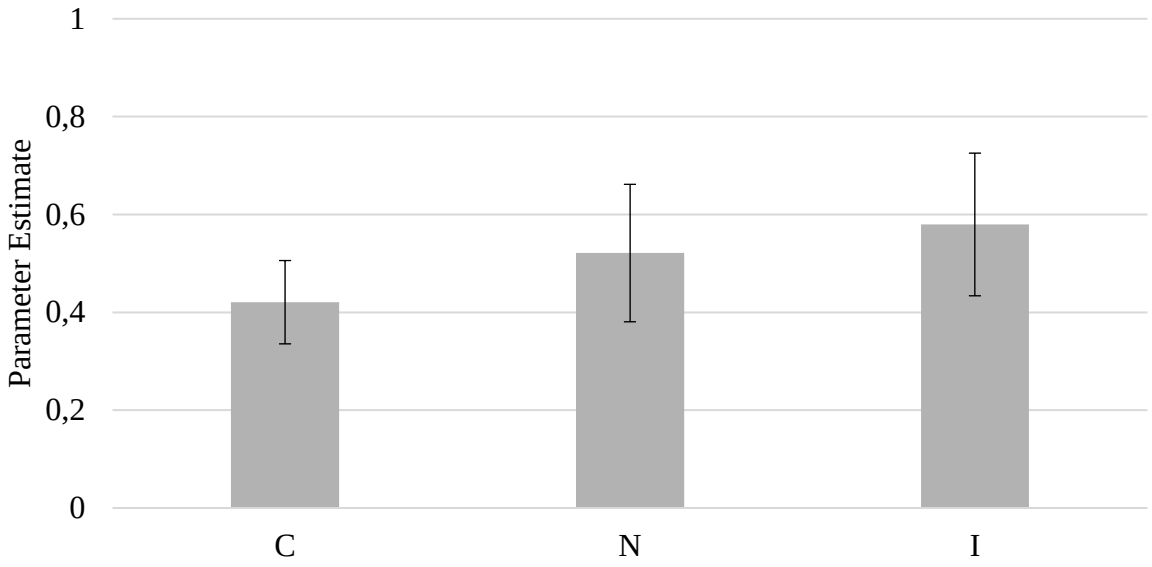

Figure 4. Parameter estimates obtained in Experiment 1. Error bars represent 95% confidence intervals.

2.2.3. Discussion

Experiment 1 provided a first test of the viability of the proCNI model. The results demonstrate that the model with the revised implementation of response tendencies via the

[15] Considering the rather small number of scenarios used in this study, we applied additional analyses to investigate whether the specified models reliably generalize across scenarios. We therefore removed the data points generated with each individual scenarios from the total data set and observed its impact on model fit. This procedure showed that goodness of fit was not affected in all four additional analyses, indicating good generalizability of our results across scenarios (all *p*'s > .30).

default state manipulation offers a good fit to the data, such that it could thus be employed in further studies.

2.3. Experiment Series 2

In Experiment Series 2, we used the proCNI model to investigate the influence of self-relevant consequences on dilemma judgment. In particular, we investigated the effect of self-relevant consequences on the *C*-parameter (Experiment 2a) and the *N*-parameter (Experiment 2b) to assess whether both the endorsement of consequences and of norms could be influenced by this factor.

2.3.1. Experiment 2a

As among others Kahane et al. (2018) have noted, consequences come in different flavors. The label “utilitarian” responding is intimately tied to the principle of act utilitarianism (choose the action that achieves the greatest well-being for the greatest number), which stresses that the number of lives saved versus lost should be assessed in an *impartial* manner (Kahane et al., 2018). However, it is conceivable that self-relevant consequences are given preferential attention in cost-benefit analyses. In Experiment 2a, we thus systematically juxtaposed self-relevant and other-relevant consequences. Here we implemented self-relevant consequences such that they always ran counter to impartial cost-benefit analysis based on bodycount. That is, to the extent to which egoistic concern for one’s own well-being is prioritized over impartial concern for aggregate well-being, the *C*-parameter will be reduced. As we expected such concern to influence dilemma judgment, we predicted self-relevant consequences to reduce the *C*-parameter. As the proCNI model allows investigating the three bases of dilemma judgment independent of one another, we also assessed the influence on the *N*- and *I*-parameters, although we saw no a-priori reasons to expect effects on these parameters.

2.3.1.1. Method

Participants. We implemented the same sample size as in Experiment 1. We collected data of 96 university students of diverse majors (31 male; M_{age} = 23.07, SD_{age} = 6.14) in a laboratory setting in return for a financial compensation of 8.00€ and a bar of chocolate for one hour of their time. The experiment was conducted as part of a two-study session, whereas the second study was unrelated to moral judgment.

Design. This experiment implemented a 2 (congruency: congruent vs. incongruent) × 2 (default state: action-default vs. inaction-default) × 2 (self-relevant consequences: present vs. absent) mixed design with repeated measures on the first two factors.

Materials and procedure. Materials and procedure are identical to Experiment 1, except that participants were randomly assigned to one of the two self-relevance conditions. For participants in the “present” condition, positive consequences that applied only to themselves could be achieved by neglecting the consequences of their decision in terms of lives saved or lost. Specifically, in congruent scenarios (in which endorsement of consequences and endorsement of norms result in the same decision) self-relevant consequences could be maximized by deciding such that norms and aggregate consequences are neglected. In the incongruent condition, self-relevant consequences could be maximized by ignoring aggregate consequences. In the case of the torture scenario (see Appendix A), for instance, participants in the “self-relevance present” *congruent* condition were informed that their supervisor assigns the solution of the case high priority, and that a positive evaluation of their work by the supervisor will improve their chances for a promotion, which provides a reason for engaging in torture. Participants in the “self-relevance present” *incongruent* condition were likewise informed that a positive evaluation by their supervisor would likely lead to a promotion, but that the supervisor assigned the protection of individual rights a high priority, which provides a reason against engaging in torture. Assignment of the different versions to the scenarios was counterbalanced within the self-relevance conditions.

2.3.1.2. Results

The model fit the data well, $G^2(2) = 1.21$, $p = .545$, $w = 0.057$. The influence of self-relevance on the strength of the *C*-parameter was significant, resulting in a lower *C*-parameter in the self-relevance present ($C_{present} = .28$, 95% CI [.16, .40]) compared to self-relevance absent condition ($C_{absent} = .51$, 95% *CI* [.40, .62]), $\Delta G^2(1) = 8.50$, $p = .037$, $w = 0.138$. The *N*-parameter ($N_{present} = .53$, 95% *CI* [.36, .70], $N_{absent} = .70$, 95% *CI* [.50, .89]) was not affected by self-relevant consequences, $\Delta G^2(1) = 1.50$, $p = .221$, $w = 0.063$. The *I*-parameter ($I_{present} = .66$, 95% *CI* [.48, .84], $I_{absent} = .49$, 95% *CI* [.17, .81]) also did not differ between conditions, $\Delta G^2(1) = 0.83$, $p = .363$, $w = 0.046$, and was thus set equal across conditions. This overall *I*-parameter ($I = .62$, 95% *CI* [.47, .77]) did not differ from 0.5, $\Delta G^2(1) = 3.08$, $p = .215$, $w = 0.090$, indicating no influence of inertia. Parameter estimates are depicted in Figure 5.

2.3.1.3.Discussion

The results of Experiment 2a indicate that the presence of self-relevant consequences can affect whether dilemma response patterns are consistent with aggregate consequences as represented by the *C*-parameter. Specifically, they show that the presence of minor self-relevant consequences (e.g., receiving a promotion) may override the decision recommended by an impartial cost-benefit analysis based on bodycount. This finding suggests that the endorsement of sacrificial killing in the canonical stimulus set reviewed by Rosas and Koenigs (2014) may reflect self-interest rather than utilitarianism. Our results also provide a plausible alternative explanation for findings that appear counterintuitive at first sight (e.g., Bartels & Pizarro, 2011). For instance, in the majority of Bartels and Pizarro’s (2011) dilemmas, sacrificial killing offered the opportunity to maximize self-relevant consequences, which may have contributed to the positive relationship between “utilitarian” judgments and antisocial tendencies they reported (see also Balash & Falkenbach, 2018; Karandikar, Kapoor, Fernandes, & Jonason, 2019).

2.3.2. Experiment 2b

In Experiment 2b, we manipulated self-relevant consequences such that they always incentivized sacrificial killing. This extension is relevant for two reasons. First, it provides a first investigation of the conditions under which the process underlying the *N*-parameter is sensitive to consequences, which has potential conceptual and terminological implications. Second, this approach directly models the confound of self-relevant consequences with endorsement of sacrificial killing present in several of the frequently employed dilemmas introduced by Greene et al. (2001, 2004; Rosas & Koenigs, 2014), which allows assessing the degree to which responses to these dilemmas may be biased by this systematic confound. That is, we directly investigate whether endorsement of sacrificial killing in such dilemmas may reflect egoism rather than utilitarianism. That is, to the extent that self-relevant consequences influence response patterns in our data, the assumption that sacrificial killing in those dilemmas reflects concern for maximizing aggregate well-being (or “level-3 utilitarianism” according to Conway et al., 2018) is called into question (Kahane, 2015; Kahane et al., 2015; also see Aktas et al., 2017). Investigating the impact of self-relevant consequences on the independent bases of moral dilemma judgments, we expect self-relevant consequences to reduce norm-endorsement (*N*), while leaving endorsement of consequences (*C*) and inertia (*I*) unaffected.

2.3.2.1. Method

Participants. As in the previous experiments, 96 university students of diverse majors (27 male; M_{age} = 25.56, SD_{age} = 6.19) participated in a laboratory study in return for a financial compensation of 8.00€ and a bar of chocolate for one hour of their time. The experiment was presented as the last one in a block of four studies, with the other three unrelated to moral judgment.

Design. This experiment implemented a 2 (congruency: congruent vs. incongruent) × 2 (default state: action-default vs. inaction-default) × 2 (self-relevance: present vs. absent) mixed design with repeated measures on the first two factors.

Materials and procedure. Materials and procedure are identical to Experiment 2a, with the exception that sacrificial killing now always produced self-relevant consequences. In the case of the torture scenario, participants in the “self-relevance present” condition now always read that their supervisor assigned the solution of the case high priority, regardless of aggregate consequences of the decision. The order of scenarios was randomized. Assignment of the different versions to the scenarios was fully random.[16]

2.3.2.2. Results

The model fit the data well, $G^2(2) = 0.57$, $p = .751$, $w = 0.039$. The self-relevance manipulation exerted an effect on the estimate of the *N*-parameter, which was smaller in the self-relevance present ($N_{present} = .10$, 95% *CI* [.00, .31]) than in the self-relevance absent ($N_{absent} = .54$, 95% *CI* [.34, .75]) condition, $\Delta G^2(1) = 8.16$, $p = .004$, $w = 0.147$. Furthermore, setting the *N*-parameter in the self-relevance present condition equal to 0 revealed no significant reduction in model fit, indicating that participants’ responses were not characterized by norm-endorsement in this condition, $\Delta G^2(1) = 0.93$, $p = .335$, $w = 0.049$. The *C*-parameter ($C_{present} = .38$, 95% *CI* [.25, .51], $C_{absent} = .46$, 95% *CI* [.34, .58]) did not differ between self-relevance conditions, $\Delta G^2(1) = 0.87$, $p = .351$, $w = 0.048$. The *I*-parameters of the two conditions ($I_{present} = .58$, 95% *CI* [.46, .69], $I_{absent} = .53$, 95% *CI* [.31, .75]) could be set equal without a loss in model fit, $\Delta G^2(1) = 0.11$, $p = .735$, $w = 0.017$. The *I*-parameter (*I*

[16] Because Experiment 2b was conducted prior to Experiments 1 and 2a, the scenarios used in Experiment 2b contain slight differences due to later adjustment. Because we consider those adjustments to be minor, Appendix A only contains the scenarios used in Experiments 1 and 2a. The scenarios used in Experiment 2b will be provided by the first author upon request.

= .57, 95% *CI* [.46, .67]) did not differ from 0.5, indicating no influence of inertia, $\Delta G^2(1) = 1.69$, $p = .429$, $w = 0.066$. The parameter estimates are depicted in Figure 5.[17]

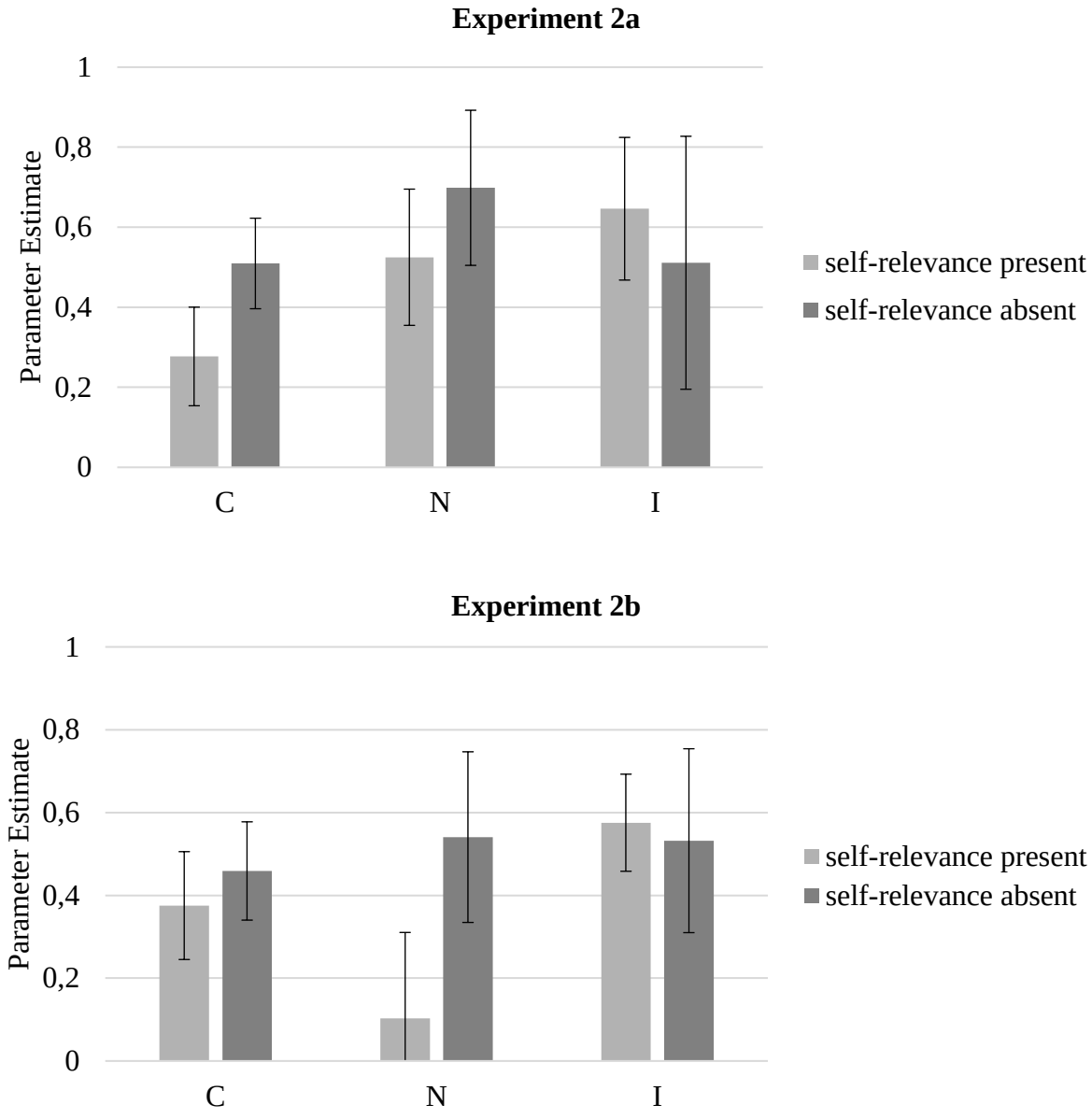

Figure 5. Parameter estimates obtained in Experiments 2a and 2b, separated by self-relevance conditions. Error bars represent 95% confidence intervals.

[17] We conducted additional analyses for Experiment Series 2 in the same manner as for Experiment 1. Again, a stepwise removal of scenarios did not indicate lack of model fit in any of the eight analyses (four scenarios × two experiments), indicating good generalizability of the respective model specification across scenarios (all *p*’s > .50).

2.3.2.3. Discussion

As the results of Experiment 2b suggest, self-relevant consequences that systematically incentivize sacrificial killing can reduce norm-endorsement as measured by the *N*-parameter. In fact, when self-relevant consequences were present there was no indication that dilemma judgments reflected endorsement of a proscriptive norm against killing. These results are informative in at least two regards. First, they demonstrate that the process underlying the *N*-parameter is not insensitive to consequences. This provides an extension of earlier work which, due to limitations inherent to the conventional dilemma design, was unable to assess the impact of self-relevant consequences on the endorsement of aggregate consequences and norms independent of another (Christensen et al., 2014; Kahane et al., 2015; Koop 2013; Lotto et al., 2014; Moore et al., 2008, 2011). Experiment Series 2 shows that both processes may be affected. Thus, in a conventional incongruent dilemma that confounds the maximization of aggregate consequences with norm-rejection, egoistic incentives may have a double biasing effect on the acceptance of sacrificial killing, which is mediated by both increased endorsement of aggregate consequences and decreased norm-endorsement. Thus, norm-breaking in response to such confounded stimuli that are frequently used in the conventional paradigm should not be uncritically accepted as indicating concern for aggregate well-being, because a desire to promote self-interest alone can explain such decisions equally well (Kahane et al., 2015; Rosas & Koenigs, 2014).

Thus, our results highlight the fact that consequences can come in different flavors, other-beneficial as well as self-beneficial, not all of which are weighed in a manner that is commensurable with act utilitarianism. Note that sacrificial killing for selfish reasons would still qualify as “level-2 utilitarian” according to the criteria of Conway et al. (2018), as long as self-relevant consequences formed part of a broadly consequentialist cost-benefit analysis. However, they would not ascend to level 3, which would require that a concern for the greater good is prioritized over selfish desires. As our analysis demonstrates, self-relevant

consequences can have a profound impact on dilemma judgments. This finding speaks against the claim that conventional “utilitarian” responses to dilemmas containing this confound can be confidently interpreted as reflecting genuine concern for the greater good (Conway et al., 2018). A practical implication of the present findings thereby concerns the construction of materials for research into moral dilemma judgment.

2.4. Experiment Series 3

Building on the results of the previous experiments, Series 3 extended our investigation of process parameters’ sensitivity to different manipulations of consequences and personal involvement. Regarding the influence of self-relevance on moral dilemma judgment, Experiment Series 3 moves beyond the approach of Experiments 2a and 2b by manipulating self-relevant consequences in a manner that effects on the endorsement of aggregate consequences (*C*-parameter) and norms (*N*-parameter) are equally likely based on the implementation. Specifically, the self-relevant consequences were identical to the consequences for the group. Thereby, the severeness of personal consequences was matched with the severeness of consequences for others. Thus, in Experiment Series 3 the implementation of self-relevant consequences did not systematically favor effects on any parameter. As a result, effects on process parameters can be ascribed more validly to the presence of self-relevant consequences irrespective of their experimental implementation.

The present experimental series also extends our consequentialist approach to moral judgment by building on results by Moore et al. (2008, 2011; see also Koop 2013; Lotto et al., 2014; Suessenbach & Moore, 2015). These authors applied the conventional dilemma paradigm to investigate the influence of self-relevant consequences, personal involvement, and death avoidability on endorsement of sacrificial killing. As their results suggest, sacrificial killing is more frequent (1) when this decision is incentivized by self-relevant consequences, (2) when the act of killing allows for low levels of personal involvement, and

(3) when the death of a potential target is inevitable regardless of one’s decision. The limitations of the conventional paradigm have hindered a direct investigation of these factors’ effects on the psychological bases of moral dilemma judgment. One purpose of Experiment Series 3 was thus to assess the functional mechanisms underlying these previously demonstrated findings.

We expected to replicate the general pattern of results demonstrated by Moore and colleagues (2008, 2011) with self-relevant consequences increasing sacrificial killing, but were agnostic about whether this effect would be mediated by an influence on norm-endorsement (*N*) or the endorsement of consequences (*C*). Although Moore et al. (2008) proposed that their effect results from changes in the endorsement of consequences, both mediators are conceivable based on the results of Experiment Series 2.

Second, personal involvement is a prominent factor in conventional dilemma research that has substantially contributed to the development of Dual-Process Theory. As famously posited by Greene et al. (2001, 2004), low psychological distance (i.e., high personal involvement) during the act of killing may activate emotional processing which in turn causes increased norm-sensitivity, resulting in a desire to avoid harming the single target. In a similar vein, Cushman (2013) proposed that norm-endorsement may be best explained by an aversion towards performing harmful actions, irrespective of their concrete consequences (cf. Miller & Cushman, 2013). As high-involvement scenarios involve causation of harm in a more direct, unmediated manner than low-involvement scenarios, the Dual-Systems framework would likewise predict an effect on norm-endorsement (Miller et al. 2014, Experiment 4; also see Waldmann & Dieterich, 2007).

Finally, death avoidability provides an intriguing opportunity to test different theoretical predictions against each other. Moore et al. (2008, 2011) found more sacrificial killing in the case of inevitable as opposed to avoidable deaths. They proposed that distinguishing between saving the life of someone whose death is avoidable (e.g., killing a

healthy person, so that his organs can be used to save the lives of several others) versus inevitable (e.g., killing a seriously wounded soldier who would otherwise be captured and tortured to death by hostile forces) requires deliberative processing. As Dual-Process Theory maintains that deliberative processing underlies utilitarian responding, they argued that effects are likely attributable to changes in cost-benefit analysis. In the proCNI model, this would be reflected in the increased endorsement of consequences (*C*) for inevitable scenarios.

A second prediction can be derived from Dual-Systems Theory (Cushman, 2013; Miller & Cushman, 2013; Miller et al., 2014; also see Reynolds & Conway, 2018), which suggests that the process underlying a rejection of sacrificial killing is best described as an aversion to performing harmful *actions* regardless of their concrete consequences. From this perspective, one should expect no effect on the *N*-parameter, as the actions in inevitable and avoidable scenarios do not show notable differences regarding their harmful properties (e.g., stabbing vs. smothering someone). If, in contrast, as we suggest, dilemma response patterns are ultimately best explained by their expected consequences (see Baron, 1994), one should expect an effect on the *N*-parameter. That is, moral dilemma judgment should be characterized by the norm-endorsement only to the extent that norm-adherence has positive consequences. Consequently, the *N*-parameter should be notably reduced when the individual one could kill is doomed to die to begin with.

To test these hypotheses Experiment Series 3 introduces a revised dilemma set based on the work by Moore et al. (2008), in which the three factors of interest are implemented orthogonally. Furthermore, all scenarios require decisions regarding sacrificial killing. This characteristic addresses a potential shortcoming of our stimulus set used in Experiment Series 2, in which one scenario represented a decision unrelated to sacrificial killing (the *hard times* scenario). All scenarios can be found in Appendix A. The structured approach we followed when constructing the scenario set for Experiment Series 3 is outlined in a manual in Appendix B.

Given recent concerns about the prevalence of false-positive findings in psychological research (e.g., Murayama, Pekrun, & Fiedler, 2014; Open Science Collaboration, 2015), Experiment 3b was a direct replication of Experiment 3a with a different sample. We interpreted only those effects that replicated across both experiments.

2.4.1. Experiment 3a

2.4.1.1. Method

Participants. We conducted an a-priori power analysis to determine the sample size necessary for achieving a power of .80 for obtaining a parameter difference of 0.1 exemplary for the personal involvement manipulation, resulting in a required sample size of 639 participants. The study was advertised via the SoSciSurvey Online Panel (Leiner, 2014) in return for the chance of winning one of ten 20.00 € gift vouchers. A total of 964 German-speaking participants completed this study. After the exclusion of participants who failed an instructional manipulation check (28.10%), the final sample consisted of 693 participants (258 male; M_{age} = 40.08, SD_{age} = 14.20).

Design. This experiment implemented a 2 (congruency: congruent vs. incongruent) × 2 (default state: action-default vs. inaction-default) × 2 (self-relevance: present vs. absent) × 2 (avoidability: death avoidable vs. death inevitable) × 2 (personal involvement: high-involvement vs. low-involvement) repeated-measures design. Killing the target was instrumental to achieving the saving of multiple lives (incongruent condition) or a minor positive outcome (congruent condition). The self-relevant consequences offered in the self-relevance present condition would always incentivize the killing. The personal involvement factor was nested within the self-relevance and avoidability factors, such that within each of the present/avoidable, present/inevitable, absent/avoidable, and absent/inevitable conditions there was a high-involvement and a low-involvement version.

Materials and procedure. Scenarios (see Appendix A) were based on those used by Moore et al. (2008). The complete stimulus set consisted of eight different scenarios in eight versions each, resulting in a set of 64 scenarios. We created eight lists representing a counterbalanced assignment of the different versions to the specific scenarios. One list was selected per participant and the respective scenarios were presented in random order. Importantly, in the present experiment self-relevant consequences were implemented in a manner that they were identical to the consequences for the group. Therefore, self-relevant consequences always incentivized norm-breaking and sensitivity to consequences alike.

Participants also rated the difficulty of the presented scenarios and completed an instructional manipulation check after the last scenario. Superficially, the manipulation check read like one of the preceding scenarios, except that after about half of the text it contained a short explanation of its purpose and the instruction not to provide an answer as in the preceding scenarios, but to click on a specific word in the body of the text instead.

2.4.1.2. Results

To investigate the hypotheses of interest we tested four different models. We first estimated an overall model, assessing the contribution of joint *C*-, *N*-, and *I*-parameters. Only then we separately assessed the influence of the self-relevance, personal involvement, and avoidability manipulations in three additional models. We followed this approach as it corresponds to the three separate hypotheses we set out to test. Furthermore, it keeps the likelihood of false positives to a minimum (e.g., Murayama et al., 2014).

Overall model. The overall model containing one *C*-parameter ($C = .17$, 95% *CI* [.14, .19]), one *N*-parameter ($N = .34$, 95% *CI* [.31, .37]), and one *I*-parameter provided a good fit to the data, $G^2(1) = 0.01$, $p = .932$, $w = 0.001$. The *I*-parameter indicated a general preference for inertia over interference, $I = .53$, 95% *CI* [.51, .56], $\Delta G^2(1) = 8.08$, $p = .005$, $w = 0.038$.

Self-relevant consequences. The model estimating *C*-, *N*-, and *I*-parameters separately for present and absent conditions fit the data well, $G^2(2) = 3{,}89$, $p = .143$, $w = 0.027$. Equating the *C*-parameters across self-relevance conditions caused a significant decrease in model fit, $\Delta G^2(1) = 89.33$, $p < .001$, $w = 0.127$, indicating stronger endorsement of consequences in the self-relevance present ($C_{present} = .29$, 95% *CI* [.25, .32]) than in the self-relevance absent condition ($C_{absent} = .05$, 95% *CI* [.01, .08]). The *N*-parameters also differed between conditions, $\Delta G^2(1) = 14.42$, $p < .001$, $w = 0.051$, indicating stronger norm-endorsement in the self-relevance absent ($N_{absent} = .39$, 95% *CI* [.35, .43]) than in the self-relevance present condition ($N_{present} = .27$, 95% *CI* [.22, .32]). The estimates of the *I*-parameter ($I_{present} = .52$, 95% *CI* [.49, .55], $I_{absent} = .54$, 95% *CI* [.51, .57]) did not differ between conditions, $\Delta G^2(1) = 0.74$, $p = .389$, $w = 0.012$. The parameter estimates are depicted in Figure 6.

Personal involvement. The model considering *C*-, *N*-, and *I*-parameters separately per personal involvement condition fit the data well, $G^2(2) = 0.62$, $p = .734$, $w = 0.011$. Setting the parameters equal across conditions revealed a significant effect on the *N*-parameter. Indeed, the proportion of norm-consistent judgments was higher when the killing required high levels of personal involvement ($N_{high} = .48$, 95% *CI* [.44, .52]), than when it allowed for low levels of personal involvement ($N_{low} = .20$, 95% *CI* [.16, .25]), $\Delta G^2(1) = 79.41$ $p < .001$, $w = 0.120$. There was no significant effect on the *C*-parameter ($C_{high} = .16$, 95% *CI* [.13, .20], $C_{low} = 0.17$, 95% *CI* [.14, .21]), $\Delta G^2(1) = 0.07$, $p = .790$, $w = 0.004$. Likewise, the *I*-parameter was the same across conditions ($I_{high} = .54$, 95% *CI* [.50, .58], $I_{low} = .53$, 95% *CI* [.50, .56]), $\Delta G^2(1) = 0.19$, $p = .660$, $w = 0.077$. The parameter estimates are depicted in Figure 7.

Avoidability. The model estimating separate *C*-, *N*-, and *I*-parameters for the death-avoidable and death-inevitable conditions fit the data well, $G^2(2) = 0.13$, $p = .935$, $w = 0.005$. Avoidability affected the estimate of the *C*-parameter such that endorsement of consequences was stronger when the death of the victim was avoidable ($C_{avoidable} = .20$, 95% *CI* [.17, .23]) rather than inevitable ($C_{inevitable} = .13$, 95% *CI* [.10, .17]), $\Delta G^2(1) = 7.77$, $p = .005$, $w = 0.037$.

However, avoidability also had an effect on the *N*-parameter, $\Delta G^2(1) = 526.53$, $p < .001$, $w = 0.308$, such that responses reflected stronger norm-endorsement when the death of the person to be sacrificed was avoidable ($N_{avoidable} = .71$, 95% *CI* [.67, .74]) rather than inevitable ($N_{inevitable} = .00$, 95% *CI* [.00, .05]). This effect was much larger than the effect on the *C*-parameter. Furthermore, setting the *N*-parameter in the death-inevitable condition equal to zero did not reduce model fit, indicating that responses did not reflect norm-endorsement in this condition, $\Delta G^2(1) = 0.02$, $p = .880$, $w = 0.002$. The estimates of the *I*-parameter were equal across conditions ($I_{avoidable} = .53$, 95% *CI* [.47, .59], $I_{inevitable} = .54$, 95% *CI* [.51, .56]), $\Delta G^2(1) = 0.07$, $p = .797$, $w = 0.004$). The parameter estimates are depicted in Figure 8.

2.4.1.3. Discussion

Experiment 3a provided a first look into the processes underlying the effects of three factors that have been shown to influence moral dilemma judgment (Christensen et al., 2014; Koop, 2013; Lotto et al., 2014; Moore et al., 2008, 2011), and that constitute manipulations of different consequences and personal involvement. We will provide a thorough discussion of the effects obtained after presenting Experiment 3b, which constituted a direct replication with a different participant sample aimed at evaluating the robustness of our findings.

2.4.2. Experiment 3b

2.4.2.1. Method

Participants. We aimed at recruiting a sample comparable in size to that of Experiment 3a via the university’s mailing list. A total of 753 university students and employees responded to our recruiting e-mail and participated in exchange for a chance of winning one of ten 20.00€ gift vouchers. After the exclusion of participants who failed the instructional manipulation check (21.80%), the final sample consisted of 577 participants (227 male; $M_{age} = 26.52$, $SD_{age} = 9.62$).

Design, materials, and procedure. Design, materials, and procedure were identical to those in Experiment 3a, except for some final adjustments to the scenarios that removed lingering imprecisions (see Appendix A).[18]

2.4.2.2. Results

We followed the analytic strategy employed in Experiment 3a, again testing four different models.

Overall model. The proCNI model with one *C*-parameter (*C* = .20, 95% *CI* [.17, .23]), one *N*-parameter (*N* = .26, 95% *CI* [.23, .30]), and one *I*-parameter (*I* = .54, 95% *CI* [.51, .56]) provided a good fit to the data, $G^2(1) = 1.03$, $p = .309$, $w = 0.015$. The *I*-parameter indicated a general preference for inertia over interference, $\Delta G^2(1) = 9.56$, $p = .002$, $w = 0.046$.

Self-relevant consequences. The model estimating *C*-, *N*-, and *I*-parameters separately for the present and absent conditions fit the data well, $G^2(2) = 1.98$ $p = .417$, $w = 0.020$. Equating the *C*-parameters across self-relevance conditions caused a significant decrease in model fit, $\Delta G^2(1) = 97.39$, $p < .001$, $w = 0.145$, indicating that responses reflected more endorsement of aggregate consequences in the self-relevance present ($C_{present}$ = .34, 95% CI [.30, .38]) than the self-relevance absent condition (C_{absent} = .06, 95% CI [.03, .10]). While estimates of the *N*-parameter ($N_{present}$ = .29, 95% *CI* [.24, .35], N_{absent} = .24, 95% *CI* [.20, .28]) did not differ between self-relevance conditions, $\Delta G^2(1) = 2.08$, $p = .149$, $w = 0.021$, there was a marginal effect on the *I*-parameter, suggesting that responses reflected more inertia when self-relevance was present ($I_{present}$ = .56, 95% *CI* [.53, .60]) rather than absent (I_{absent} = .52, 95% *CI* [.49, .54]), $\Delta G^2(1) = 3.78$, $p = .052$, $w = 0.029$. The parameter estimates are depicted in Figure 6.

[18] After data collection was finished, we recognized that out of the 64 scenarios, one still contained an error (Orphanage: self-relevance present, death-avoidable, impersonal congruent, action-default). Removing this item from analysis did not change any of the effects and interpretations. Therefore, we report the results of the analyses including all scenarios. Appendix A contains the corrected version of the scenario.

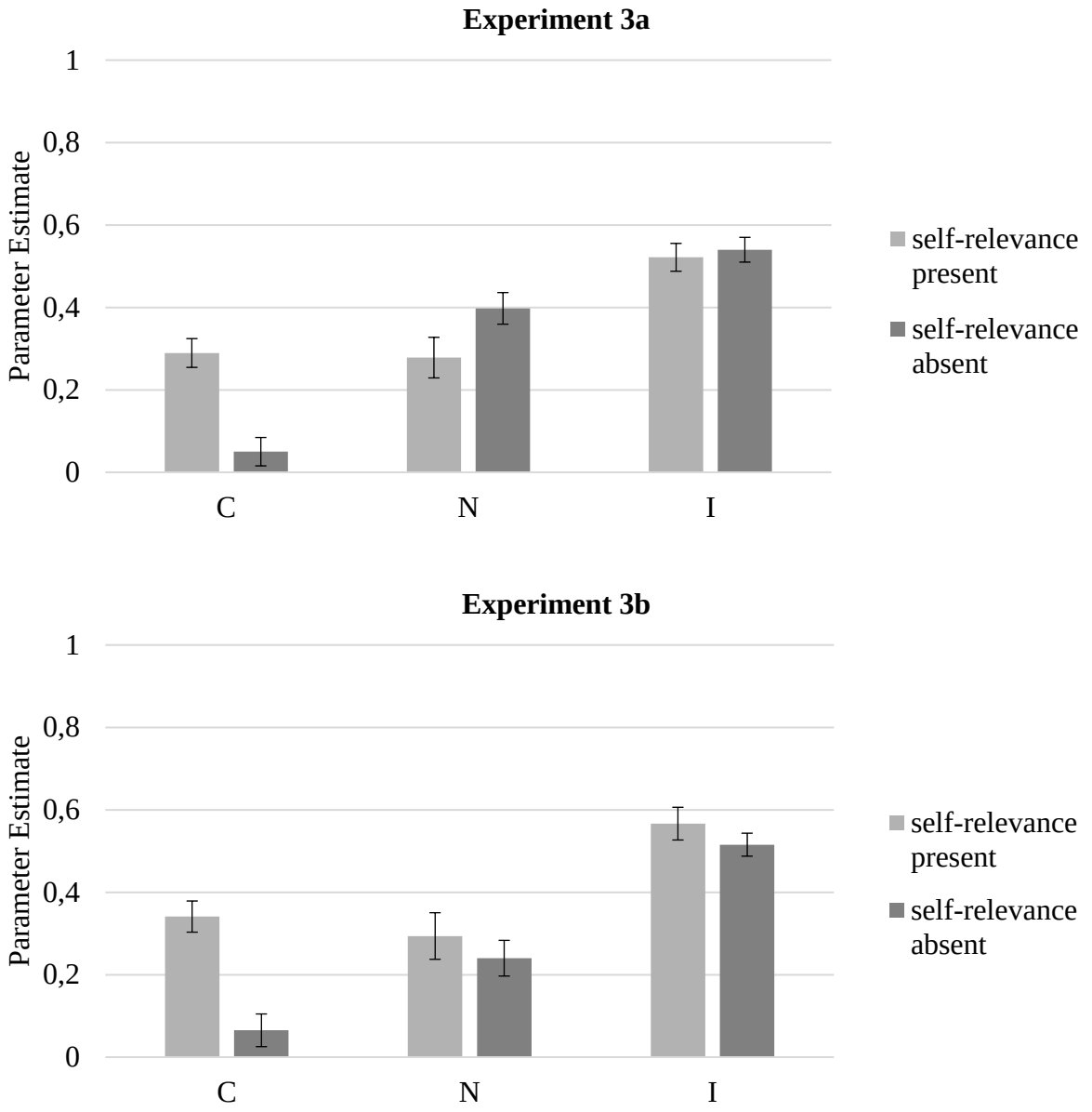

Figure 6. Parameter estimates obtained in Experiments 3a and 3b, separated by self-relevance conditions. Error bars represent 95% confidence intervals.

Personal involvement. The model estimating *C*-, *N*-, and *I*-parameters separately per personal involvement conditions fit the data well, $G^2(2) = 1.03$, $p = .597$, $w = 0.011$. Setting the parameters equal across conditions revealed a significant effect on the *N*-parameter, indicating more norm-endorsement in high-involvement ($N_{high} = .35$, 95% *CI* [.30, .40]) compared to low-involvement scenarios ($N_{low} = .17$, 95% *CI* [.12, .22]), $\Delta G^2(1) = 24.41$, $p < .001$, $w = 0.073$. There were no effects of personal involvement on the *C*-parameter ($C_{high} = .19$, 95% CI [.15, .23], $C_{low} = 0.21$, 95% *CI* [.17, .25]), $\Delta G^2(1) = 0.47$, $p = .491$, $w = 0.010$,

or the *I*-parameter (I_{high} = .56, 95% *CI* [.52, .59], I_{low} = .52, 95% *CI* [.49, .55]), $\Delta G^2(1) = 2.44$, $p = .118$, $w = 0.023$. The parameter estimates are depicted in Figure 7.

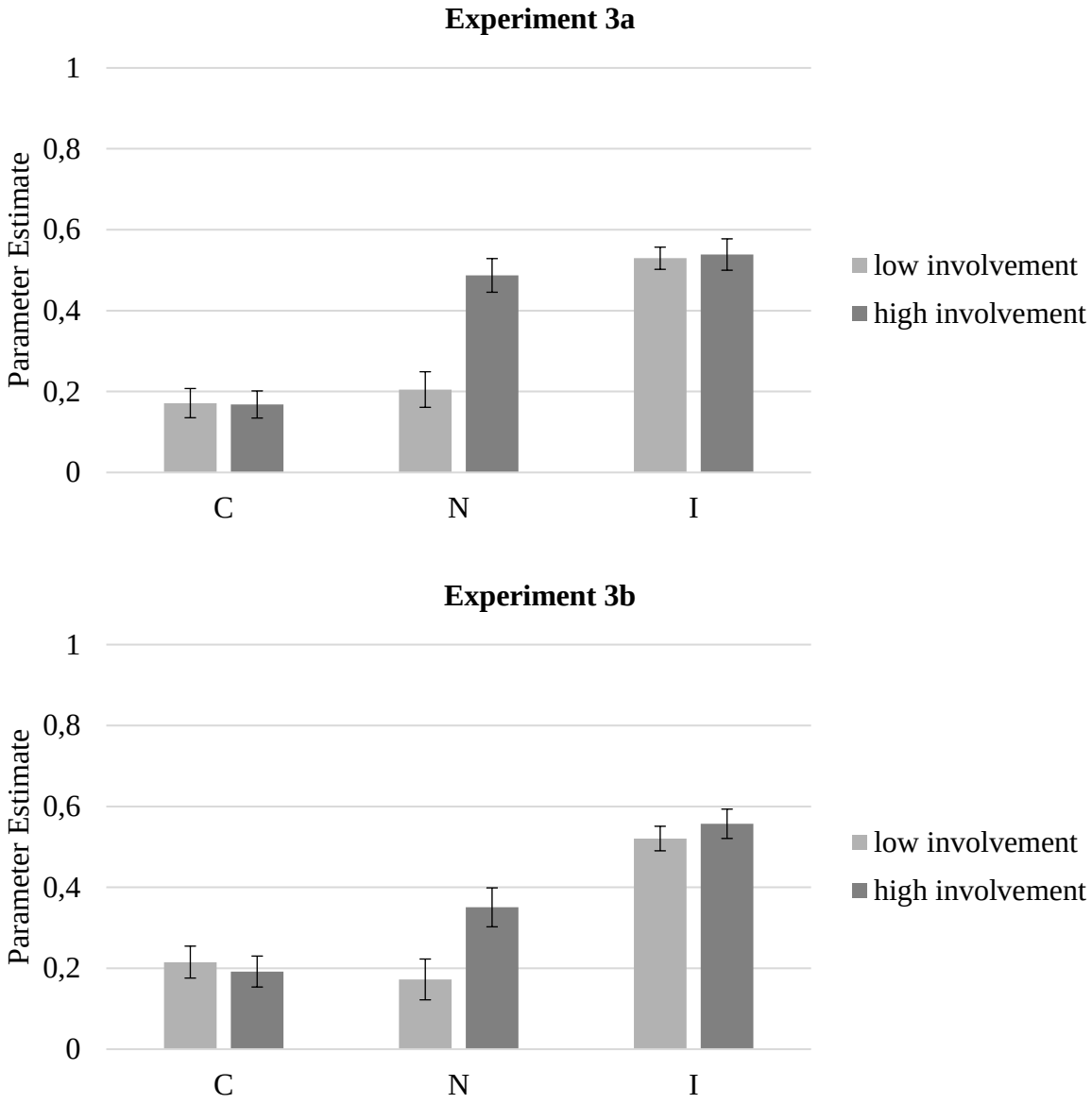

Figure 7. Parameter estimates obtained in Experiments 3a and 3b, separated by personal involvement conditions. Error bars represent 95% confidence intervals.

Avoidability. The model estimating separate *C*-, *N*-, and *I*-parameters for the death-avoidable and death-inevitable conditions did not show sufficient fit to the data, $G^2(2) = 17.40$, $p < .001$, $w = 0.061$. In this model, the avoidability manipulation did exert a significant

effect on the *C*-parameter, reflecting stronger endorsement of consequences in the avoidable ($C_{avoidable}$ = .23, 95% *CI* [.20, .26]) compared to the inevitable ($C_{inevitable}$ = .17, 95% *CI* [.13, .21]) condition, $\Delta G^2(1) = 4.38$, $p = .037$, $w = 0.031$. Avoidability also had an effect on the *N*-parameter, $\Delta G^2(1) = 436.04$, $p < .001$, $w = 0.308$, such that norm-endorsement was higher when the death of the person to be sacrificed was avoidable ($N_{avoidable}$ = .65, 95% *CI* [.60, .69]) rather than inevitable ($N_{inevitable}$ = .00, 95% *CI* [.00, .05]). Furthermore, setting the *N*-parameter in the inevitable condition to zero did not reduce model fit, indicating that responses were not characterized by norm-endorsement in this condition, $\Delta G^2(1) = 0.00$, $p = 1$, $w = 0.000$. Avoidability exerted no effect on the *I*-parameter ($I_{avoidable}$ = .52, 95% *CI* [.46, .58], $I_{inevitable}$ = .54, 95% *CI* [.52, .57]), $\Delta G^2(1) = 0.39$, $p = .530$, $w = 0.009$.

We additionally explored the source of model misfit. Given that the *N*-parameter in the inevitable condition was equal to the lower bound of the parameter space, we reversed the *N*-parameters coding in that condition to investigate the possibility of reversed effects. Note that this specification only influences estimates of the *N*-parameter and the according change in model fit, while leaving the parameter estimates of the *C*- and *I*-parameters unaffected. With this modification, the model indeed provided a good fit to the data, $G^2(1) = 1.38$, $p = .507$, $w = 0.017$. The *N*-parameter in this condition was estimated at $N_{inevitable}$ = .10, 95% *CI* [.05, .15]), indicating a small reversed effect of *N*, $\Delta G^2(1) = 16.04$, $p < .001$, $w = 0.059$. This effect suggests that in the inevitable condition, participants showed a general *acceptance* rather than rejection of sacrificial killing.[19] Parameter estimates are depicted in Figure 8.

[19] We conducted additional analyses for Experiment Series 3 in the same manner as for Experiment 1 and Series 2, this time for the general model as well as for all the models that tested the influence of the moderators. In the model investigating the influence of avoidability in Experiment 3b the fitting model with the reversed *N*-parameter was used. Out of the 64 analyses (eight scenarios × four models × two experiments) four showed a significant deviation from model fit, one of which remained significant after adjustment for multiple comparisons (all other *p*'s > .008). As the suitability of the specified model for describing the data depended on a particular scenario in only 1.56% of the cases, this suggests good generalizability of the specified models across scenarios.

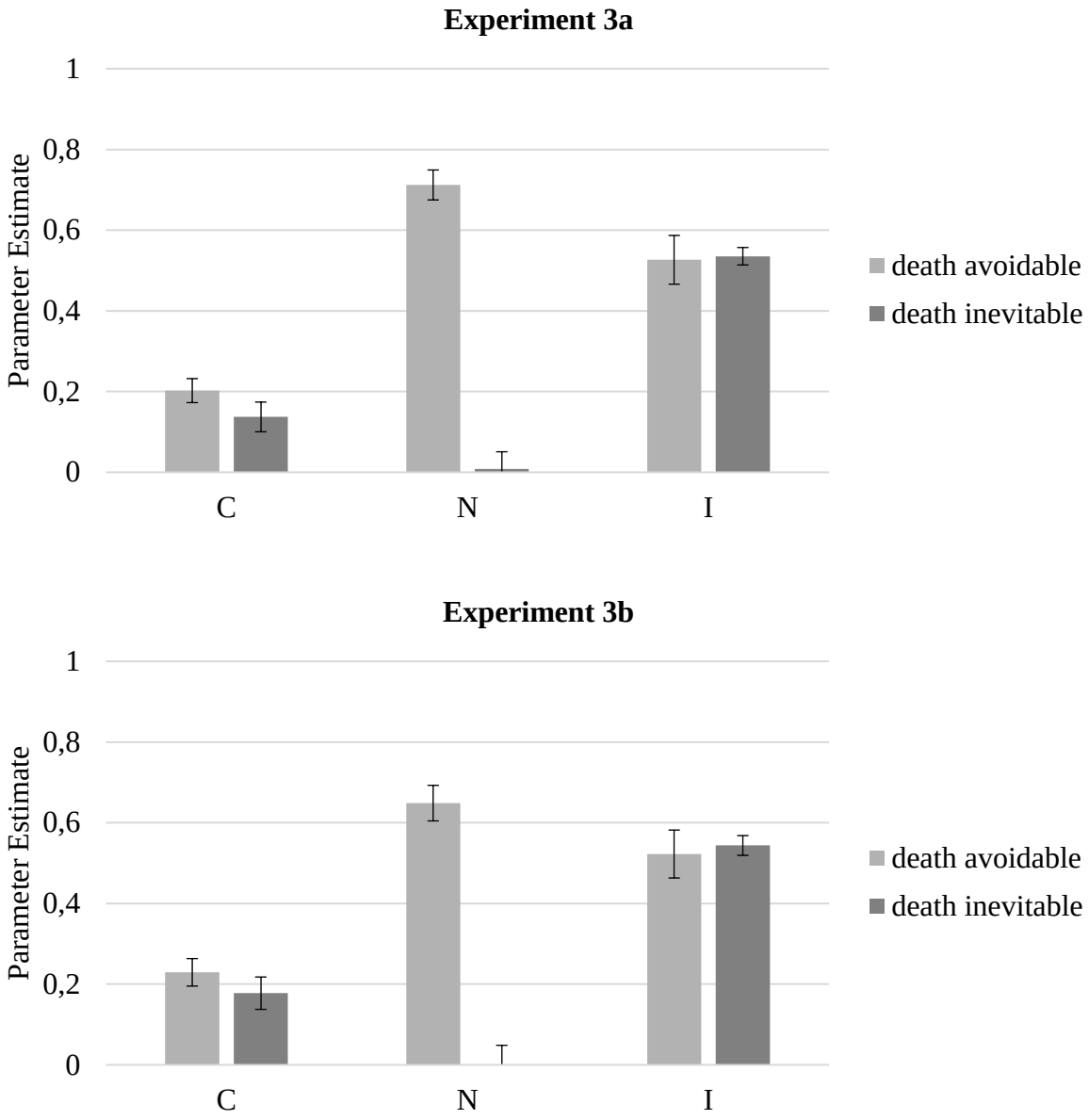

Figure 8. Parameter estimates obtained in Experiments 3a and 3b, separated by death avoidability conditions. Error bars represent 95% confidence intervals.

2.4.2.3. Discussion

Results of Experiment 3b largely replicate those of Experiment 3a. As in Experiment 3a, the self-relevance manipulation was found to increase the endorsement of consequences, personal involvement increased the norm-endorsement, and the impact of consequences was stronger when death was avoidable rather than inevitable. We also found support for our hypothesis that death avoidability affects norm-endorsement. We found evidence for

judgment-patterns that were norm-consistent only if the death of the target was avoidable in the first place. This suggests perception of harmful consequences as a candidate for the process underlying not only the endorsement of aggregate consequences, but also of norms as assessed by dilemma research (cf. Gray & Schein, 2016; Schein & Gray, 2018; Schein et al., 2016).

In line with Experiment Series 2, the results demonstrate the biasing impact of self-relevant consequences on measures of judgment processes. The improved implementation of self-relevant consequences in Series 3 furthermore suggests that introducing self-relevant consequences may have a stronger effect on the weighing of lives saved versus lost captured by the *C*-parameter than on norm-endorsement as represented by the *N*-parameter. This adds to previous findings on the effect of self-relevant consequences on dilemma judgment (Moore et al. 2008, 2011; Christensen et al., 2014; Kahane et al., 2015; Koop 2013; Lotto et al., 2014; Suessenbach & Moore, 2015) and suggests an impact on cost-benefit analysis as a likely underlying mechanism.

The last two experiments are also informative about the role of the death of the single target being avoidable or inevitable. Moore et al. (2008) proposed that, when death is inevitable, individuals place greater importance on the aggregate consequences of a decision in terms of lives saved versus lost. The proCNI model allows a direct investigation of this hypothesis and found disconfirming evidence. First, when death was inevitable, there was a small but consistent *decrease* in the size of the *C*-parameter, indicating less endorsement of consequences. This decrease suggests that the differences between major and minor consequences plays less of a role when the target’s life will be lost irrespective of participants’ decision.

More importantly, however, in both experiments, inevitability of the target’s death decreased the *N*-parameter to zero. That is, when the death of the single person was inevitable to begin with, participants showed no desire to spare the targets’ life anymore. To the best of

our knowledge, this effect on the *N*-parameter constitutes—together with our findings regarding effects of self-relevance—the first empirical evidence that the process underlying the rejection of sacrificial killing is sensitive to consequences. This result suggests that a characterization in terms of adherence to absolute or “deontological” norms regardless of consequences is overly simplistic. Thus, these findings are informative regarding the interpretation of the *N*-parameter. In the present paradigm, norm-endorsement is not well explained by action properties, or emotional factors related to the sacrificial killing, because both factors do not vary systematically between the avoidable and inevitable scenarios. This finding is thus inconsistent with both the Dual-Process Theory (Greene, 2014) and the Dual-Systems framework (Cushman, 2013).

The effect of death avoidability on the *N*-parameter also influences our interpretation of the effect of personal involvement. At first sight, the fact that sacrificial killing is decreased when participants would have to administer harm in a direct and unmediated manner is in line with both the Dual-Process Theory (Greene, 2007, 2014) and the Dual-Systems framework (Cushman, 2013). However, both theories explain this effect by reference to factors whose influence in the present paradigm is rendered unlikely based on the effect of death avoidability on the *N*-parameter (see above). We suggest that the effect can instead be incorporated into a weak consequentialist framework under the assumption that participants consider *fuzzy* personal consequences arising from their decisions. That is, if personal involvement and psychological proximity are high, individuals may consider consequences such as rumination, regret, or guilt, which may be less pressing concerns in the case of low involvement (e.g., Ghorbani, Liao, Çayköylü, & Chand, 2012). Social and self-perception concerns constitute further examples of such fuzzy consequences. That is, people may be concerned to be perceived as cold when endorsing sacrificial killing (Everett, Faber, Savulescu, & Crockett, 2018; Everett, Pizarro, & Crockett, 2016; Hughes, 2017; Robinson, Page-Gould, & Plaks, 2017; Rom & Conway 2018; Rom, Weiss, & Conway, 2017), and in

turn, strategically adjust their responses (Lee, Sul, & Kim, 2018; McDonald, Defever, & Navarette, 2017; Robinson et al., 2015; Rom & Conway, 2018; see also Uhlmann, Pizarro, & Diermeier, 2015).[20] We term these consequences *fuzzy* in the sense that they are less concrete and more difficult to quantify than the outcome of an analysis that relies on bodycount as only relevant metric.

2.5. General Discussion

Research into moral dilemma judgments has been characterized by the assumption of a strong opposition between consequences and norms as determining responses. The present research investigated the influence of a multitude of consequences on moral dilemma judgment under a weak consequentialist perspective (see Sunstein, 2005).

2.5.1. Theoretical and conceptual implications

Our investigation of the impact of self-relevant consequences, personal involvement, and death avoidability in the context of the proCNI model allowed us to investigate several predictions based on the Dual-Process Theory (Greene et al., 2001, 2004) and the Dual-Systems framework (Cushman, 2013; Cushman & Miller, 2013; Miller et al., 2014). We also assessed the extent to which norm-endorsement and endorsement of consequences should be viewed as operating independent of or intertwined with one another, and consider resulting implications for how to conceptualize the processes underlying dilemma judgment. To anticipate the upcoming discussion, the present work mainly affects our interpretation of the *N*-parameter and indices of norm-consistent responding in other methodological approaches to moral dilemma judgments. As the process underlying the *C*-parameter has been considered as consequence-sensitive since the inception of the dilemma approach, its conceptualization

[20] Note, however, that some of this work (Rom & Conway, 2018; Rom et al., 2017) relied greatly on the crying baby dilemma, which incentivizes sacrificial killing and contains a victim whose death is inevitable. Both of these aspects may be relevant for character inferences, as for example someone who refuses killing even under such extreme circumstances may be perceived as especially warm and morally principled, while someone who rejects such killing may be perceived as especially incompetent.

remains unaffected by our work. Similarly to the original CNI model, we thus consider the *C*-parameter to represent a sensitivity to aggregate consequences commensurable with the goal to maximize the number of lives saved, such that it overlaps with the instrumental harm aspect of genuine utilitarian philosophy (Kahane et al., 2018).

Our work is the first to investigate the impact of personal involvement as canonically implemented in the context of multinomial modeling, which provided us with the opportunity to directly assess the functional mechanism underlying the influence of this canonical factor. Across two experiments, we found that high involvement, thus physical proximity and low amounts of psychological distance, increases norm-endorsement in moral dilemma judgments. Although this is consistent with previous research (Cecchetto et al., 2017; Christensen et al., 2014; Koenigs et al., 2007; Koop, 2013; Kusev et al., 2016; Moore et al., 2008, 2011; Moretto et al., 2010), to our knowledge we are the first to demonstrate that this finding is attributable to a pattern of norm-consistent responding. This is predicted by the Dual-Process Theory (Greene et al., 2001, 2004; Greene, 2014), which explains this effect by changes in emotional processing (also see McDonald et al., 2017). It is equally predicted by the Dual-Systems framework (Cushman, 2013; Miller & Cushman, 2013), according to which this response pattern is attributable to sensitivity to the harmful nature of the action to be performed, regardless of its tangible consequences. That is, in contrast to the Dual-Process Theory, the Dual-Systems framework explains this response pattern as representing negative emotional reactions that result from imagining a repellent *action*, rather than from imagining the negative *outcomes* for the target of that action (Miller et al., 2014).

However, the findings of the death avoidability manipulation, which replicate the general pattern of results found by Moore and colleagues (2008), suggest that norm-consistent responding in the dilemma context is neither best understood in terms of emotional harm aversion (Greene, 2014), nor in terms of action aversion (Cushman, 2013). This is, to the best of our knowledge, the first experimental evidence demonstrating that “deontological”

response patterns in dilemma judgment show sensitivity to tangible consequences. Even more remarkably, the *N*-parameters reduction to zero even suggests a complete mediation by expected consequences. We propose that these effects tentatively suggest that norm-endorsement represents sensitivity to *consequences* that are *causally proximal*. That is, those that are produced in an unmediated manner as a result of the dilemma decision, such that the judge’s responsibility for the proximal consequence is unambiguous and marked by a high degree of causal directness (Pizarro, Uhlmann, & Bloom, 2003; Waldmann & Dieterich, 2007; also see Greene et al., 2009).

The present results are thus fully in line with our suggestion that responses that are traditionally considered “utilitarian” and “deontological” alike should be viewed through a consequentialist lens. According to the weak consequentialist view we endorse, even if people may de facto base their decisions on moral norms that they may apply heuristically (Sunstein, 2005; also see Gigerenzer, 2004), they do so because these tend to produce desired consequences (Baron, 1994). As such, these norms are valuable because they may constitute an important means for ensuring positive consequences in the long run for the self and the community alike (Axelrod, 1986; Axelrod & Dion, 1988). That is, in many real-life situations the positive consequences of potential norm-violations are likely based on a probabilistic judgment that may well be inaccurate. As such, an absolutist prohibition against killing may be generally conducive to the greater good, because it avoids making decisions based on such inaccurate assessments. For instance, a doctor concluding that harvesting the organs of one patient *may* (in principle) save the lives of five does not *know* for sure whether he would be successful in achieving the desired consequences, whereas the single patient would die with certainty.

Furthermore, once a narrow focus on bodycount as the only relevant metric for a consequentialist analysis is abandoned, a relationship between norm-adherence and consequences becomes even more apparent. For instance, a doctor’s decision to kill one

patient in order to save five others may maximize well-being in the sense of saving four lives net. However, living in a society that judges such killing as generally acceptable (i.e., abandoning a norm against killing under such circumstances) would entail living in constant fear about being sacrificed for the greater good, once one enters a doctor’s waiting room. As has been noted before, such fuzzy consequences are usually not considered explicitly in utilitarian analyses that keep a strict and unimaginative focus on bodycount. Yet, such fuzzy consequences are of clear moral relevance, given one accepts maximization of well-being and minimization of harm to lie at the heart of moral questions (Harris, 2010). From this perspective, the presence of norms forbidding killing, lying, or similar actions are thus conducive to positive consequences on a societal level. This analysis illustrates how concerns for rights and duties can be incorporated into a cost-benefit analysis under a weak consequentialist perspective (Sunstein, 2005).

This weak consequentialist view also fits well with the theoretical approach taken by Subjective Utilitarian Theory (Cohen & Ahn, 2016), which likewise conceptualizes dilemma decisions to be best explained by a single process that consists in an analysis of subjectively perceived costs and benefits. Indeed, our finding that the effect of “norms” on dilemma judgment may be mediated by tangible consequences converges well with Cohen and Ahn’s (2016) proposition that an effect of emotion on dilemma judgment is ultimately explained by changes in the perceived value of the respective response options, such that positing two separate processing systems is unnecessarily unparsimonious.

On a broader level, our results converge with process models of moral judgment, which propose that subjective perceptions of harm lie at the heart of moral analyses. As for instance the Theory of Dyadic Morality maintains, the causation of perceived harm is a necessary ingredient for a situation to be judged morally relevant, and moral norms are accepted to the degree to which they are perceived to reduce the causation of harm (Gray, Schein, & Cameron, 2017; Jackson & Gray, 2019; Schein & Gray, 2018). This proposal is

supported, among others, by results suggesting that perceptions of harm mediate the repudiation of various ostensible violations of purity, like gay marriage, uttering blasphemous statements or consuming gene-modified food (Gray & Schein, 2016; Schein et al., 2016). As such, we view our conceptual proposal as firmly embedded within an innovative and growing literature on the processes underlying moral judgment, which surpasses the narrow boundaries of dilemma research.

Moreover, evidence from within and outside the moral dilemma literature suggests a reframing of response processes, thereby moving away from a terminology that involves an artificial split between norms and consequences, let alone a systematic attachment to broader philosophical positions. Instead, processes underlying responses may be better understood in consequentialist terms (c.f. Cohen & Ahn, 2016; Schein & Gray, 2018; also see Baron, 1994; Sunstein, 2005). This avoids the use of misleading philosophical terminology that carries unintended conceptual baggage (e.g., Conway et al., 2018; Kahane, 2012). It also promotes explicitly considering the different forms of consequences that may enter the process of dilemma decision making, which reduces the danger of drawing false conclusions. For instance, an explicit consideration of the impact of self-relevant consequences decreases the likelihood that responses will be interpreted in terms of impartial cost-benefit analysis, and that theoretical or normative conclusions are drawn based on this assumption (Kahane et al., 2015; Rosas & Koenigs, 2014).

We explicitly note that our weak consequentialist conceptualization of norm-adherence in moral dilemmas is tentative and certainly open to revision, clarification or rebuttal by future research. Likewise, we want to stress that our proposal to consider dilemma response patterns through a consequentialist lens does not represent a dismissal of deontological approaches to normative ethics or to moral judgment research. Although we propose that dilemma responses may be most parsimoniously and most helpfully understood by adopting a consequentialist perspective, we do neither suggest that people do not make use

of rules in moral judgment, nor that rule-focused approaches to understanding moral judgment cannot be helpful. As Holyoak and Powell (2016) remark, moral rules may well be understood from a standpoint of *moderate deontology*, according to which they represent general guidelines which may be broken if situational demands are strong enough, or if they conflict with other moral rules that are deemed more important (also see Kahane, 2015). This, in turn, is fully commensurable with a weak consequentialist view (Sunstein, 2005), according to which concern for rules, rights and duties can form part of a broadly consequentialist cost-benefit analysis. Thus, although we propose that focusing on consequences rather than norms may be a more illuminating approach to moral analyses in most cases (Gray & Schein, 2016; Schein et al., 2016), our proposal is motivated by the recognition that—at a fundamental level—norms and consequences are intimately linked to one another. Therefore, basing theoretical conclusions on the assumption of a sharp divide between these two carries a definite risk of oversimplification (Gray & Schein, 2012).

2.5.2. Methodological implications

Our findings question a number of assumptions on which many of the studies using the conventional dilemma approach have relied, and which are also fundamental to the Dual-Process Theory of moral judgment (Greene et al. 2001, 2004, 2008; Greene 2014). Specifically, these are assumptions regarding the suitability of the employed stimuli for assessing theoretical predictions. As Rosas and Koenigs (2014) noted, several of the scenarios in the canonical set introduced by Greene et al. (2001, 2004) systematically confound the sacrificial killing choice with self-relevant consequences, which renders responses ambiguous. Experiment Series 2, which was devoted to assessing the effect of self-relevant consequences on judgment processes, suggests this concern to be valid. Results indicate that presence of self-relevant consequences may affect endorsement of aggregate consequences (Experiment 2a) and norm-endorsement (Experiment 2b) alike, depending on their implementation. Whenever self-relevant consequences are present in canonical dilemmas,

they incentivize killing. As this is consistent with endorsing aggregate consequences and norm-rejection alike, it is possible that this confound thus exerts a double biasing effect, mediated by increased cost-benefit weighing as well as decreased norm-endorsement. In a narrow sense, this finding provides potential alternative explanations for some previous findings. For instance, the positive relationship between "utilitarian" responding and psychopathy (e.g., Bartels & Pizarro, 2011, Karandikar et al., 2019) may in part result from high-psychopathic individuals being more sensitive to self-relevant consequences.

On a broader note, our experimental evidence casts doubt on the assertion that sacrificial killing responses to dilemmas in the canonical battery can be confidently interpreted as reflecting "level-3 utilitarianism," driven by concern for the greater good (Conway et al., 2018).[21] This has implications for some propositions based on this assumption. For instance, Greene (2014) proposed that *act utilitarianism* should be normatively favored over other ethical theories, because it is "not chasing intuitions" (p. 725). Specifically, this argument appears to deliberately contrast the impartial analysis of costs and benefits (e.g., lives lost versus saved) inherent in act utilitarianism with other forms of consequentialism, which also incorporate self-interest into their analysis (Greene, 2014, p.724). This normative claim, however, makes the strong assumption that sacrificial killing responses are based on genuinely utilitarian concerns for the greater good within the context of the dilemma decision, untainted by egoism (level-3 utilitarianism" according to Conway et al., 2018), which seems doubtful in light of our experimental findings (also see Moore et al. 2008, 2011; Kahane et al., 2015, 2018; Koop 2013; Lotto et al., 2014; Suessenbach & Moore, 2015). Indeed, the results of Conway et al. (2018), who obtained a relationship between

[21] Also note that Conway et al.'s (2018) analysis made use of the battery of process dissociation dilemmas introduced by Conway and Gawronski (2013), not the battery of Greene et al. (2001, 2004). This limits the extent to which their results speak to the processes underlying responses to the confounded scenarios in the classical set addressed by us as well as in other research (e.g., Kahane et al., 2015; Rosas & Koenigs, 2014).

psychopathy and dilemma judgments, already suggests problems for this normative argument (also see Gawronski et al., 2017; Reynolds & Conway, 2018).

2.5.3. Limitations and future directions

Though informative in several ways, our research is certainly not without limitations. First, we have noted that our findings regarding the role of personal involvement are consistent with the Dual-Process Theory and the Dual-Systems framework alike. Here, it should be kept in mind that the main purpose of investigating the influence of this factor was to provide a functional test of previously demonstrated findings. Therefore, we adhered to the implementation of personal involvement as anchored to the action necessary to affect the single target, because this is the implementation that has been consistently used in the history of dilemma research. As such, it is unclear in how far the effect of personal involvement on sacrificial killing reflects a genuine influence of the concept of personal involvement, or an influence of its experimental implementation. This, however, is a concern that applies primarily to Dual-Process interpretations of the data, according to which the effects of personal involvement are an indication for changes in System 1 processing (Greene et al., 2001, 2004). The proCNI model does not attempt to draw conclusions about different cognitive systems underlying the endorsement of consequences and norms. Moreover, the proCNI model is ideally equipped to investigate such theoretical tenets. That is, if the Dual-Process Theory is correct in its mechanistic assumptions and emotional processes selectively favor “deontological” responding, one should expect an influence of personal involvement on the *N*-parameter even when the manipulation is consistently aligned with the endorsement of consequences instead of norms. Future research may make use of the opportunities provided by the proCNI model to test this hypothesis directly.[22]

[22] The neat one-to-one mapping suggested by the Dual-Process Theory has also been scrutinized in other research. Some results suggests for instance that sensitivity to consequences is related to emotional aversion to imagining negative outcomes (Miller et al., 2014; Reynolds & Conway, 2018), or that prevention-focused responding may increase the tendency to explain one’s dilemma decisions via deontological reasoning (Gamez-Djokic & Molden, 2016, study 7). Note,

Among all of our experiments, participants showed a tendency for inertia only in the sacrificial killing dilemmas in Experiment Series 3. As Gawronski and colleagues (2016, 2017) demonstrated, their *Inaction* parameter contributed to a number of effects observed in the conventional dilemma approach (also see Zhang et al., 2018). These moderators were not assessed in the present research, so that it remains to be investigated whether the same factors that influence the *I*-parameter in the original CNI model may also influence the more clearly and strictly defined *I*-parameter of the proCNI model.

Research that analyses participants' responses to hypothetical scenarios is generally characterized by a number of limitations (Bauman et al., 2014). First, the approach relies on participants accepting the small-world assumptions of the dilemmas as they are presented to them (Shou & Song, 2017; Royzman, Kim, & Leeman, 2015). However, violations of this assumption are not problematic for the proCNI model and the original CNI model as long as they do not co-vary with one of the parameters. Unsystematic variation will be captured by the *I*-parameter, which also serves a quality-control function in the sense that it helps estimating the preceding parameters more validly and reliably. Moreover, we applied a structured approach in designing our stimulus material to ensure that our scenarios were as credible as possible and internally coherent.

Second, the generalizability of the findings depends on the ecological validity and breadth of the employed scenarios. Thus, it would be of value to reassess our findings with a new set of scenarios in order to warrant that the effects we obtained are not specific to the dilemmas we used, which is a pervasive concern in research employing vignettes (see e.g. Koop, 2013; McGuire et al., 2009). While the present experiments employed various scenarios, the set of scenarios was still small. Future work should attempt to increase the number of experimental stimuli to address this potential concern. Moreover, note that we

---

however, that in the latter case participants reasons were assessed after dilemmas were judged, such that this finding is suggestive, but does not establish a causal connection between deontological reasoning and "deontological" dilemma judgment.

manipulated several of our central factors under investigation via different contents (i.e., self-relevance and death avoidability), preventing hierarchical MPT analysis (e.g., Heck, Arnold, & Arnold, 2018) that account for variability across different scenario contents. Extending the scenario set in a way that these factors are manipulated within contents would address this limitation. We believe that the template provided in Appendix B is particularly helpful in serving these goals.

2.5.4. Considerations on power and composition of samples

In designing our experiments, we considered statistical power in a number of ways. First, we determined sample sizes a-priori based on the results of previous experiments and power-analyses to warrant that our experiments achieved satisfactory power. Second, we implemented all factors within-participants. Third, scenario construction removed relevant confounds and ensured that they followed a unitary structure to reduce measurement error (Appendix B).

Regarding the generalizability of our findings across the population, we note that our experiments made use of three different samples. Experiment 1 and Experiment Series 2 were conducted with students of diverse majors. Experiment 3a recruited a diverse sample of participants via the SoSciSurvey Panel (Leiner, 2014) and Experiment 3b included both students and employees recruited via the university mailing list. The viability of our approach was demonstrated in all of our experiments.

Some of the more specific findings, however, may be related to characteristics of our samples. First, participants indicating female gender approached an overall prevalence of approximately 64%. Because recent studies suggest that women show a stronger endorsement of norms as well as a greater preference for inaction than men (Gawronski et al., 2016, 2017; Friesdorf, Conway, & Gawronski, 2015), the specific values of the *N*- and *I*-parameters may be higher than in more balanced or in male-dominated samples. A second factor concerns the prevalence of religious belief. As previous work suggests, certain aspects of religious belief

may be related to norm-focused or absolutist moral thinking (Piazza, 2012; Piazza & Landy, 2013; Piazza & Sousa, 2014; Simpson, Piazza, & Rios, 2016; Szekely, Opre, & Miu, 2015). Because religious belief is assigned comparatively low importance in Germany (Pew Research Center, 2018) the size of the *N*-parameter may be affected by this boundary condition. However, in none of our experiments we were interested in the absolute values of parameter estimates, but only interpreted effects of experimental manipulations.

2.5.5. Conclusion

Our findings suggest that moral dilemma responses should ultimately be viewed through a consequentialist lens, according to which different response patterns represent sensitivity to different sorts of consequences, namely to consequences commensurable with an analysis of number of lives saved or lost (*C*), and to proximal consequences that follow as a direct result of a behavioral decision (*N*). We propose that methodological and theoretical advancement can be promoted by applying a lens of weak consequentialism, which adds to a growing body of literature that suggests that perceived consequences may lie at the heart of moral judgment.

Chapter III

Does Manipulating Psychological Value influence Dilemma Difficulty? – Testing Predictions derived from Subjective Utilitarian Theory

Max Hennig

Eberhard Karls Universität Tübingen

The following chapter contains a manuscript resulting from a cooperation between Max Hennig (lead author) and Prof. Dr. Mandy Hütter (second author) entitled “Does Manipulating Psychological Value influence Dilemma Difficulty? – Testing Predictions derived from Subjective Utilitarian Theory”. Max Hennig and Mandy Hütter contributed 60% and 40%, respectively. Specifically, both authors contributed approximately 50% to the generation of scientific ideas, data generation, and analysis and interpretation, whereas Max Hennig wrote the manuscript.

3.1. Introduction

Over the course of the last two decades, a noticeable amount of research on the processes underlying moral judgment has been inspired by philosophical thought experiments. The most prominent of those are variations of ethical dilemmas, in which it must be decided, whether one person should be killed in order to save the lives of several others. Such sacrificial dilemmas were originally designed as interesting philosophical puzzles, which explore the differences between moral considerations stressing rights and duties, and those that focus on the tangible consequences of morally relevant decisions (Foot, 1967; Thomson, 1976, 1985). Since their introduction into the psychological literature in seminal work by Greene et al. (2001), much theorizing about the mechanisms of moral judgment has resulted from the application of such dilemmas. Specifically, Greene et al. (2001) developed the Dual-Process Model of moral judgment, which has been particularly influential in shaping the literature on dilemma judgment in particular, and moral judgment in general (e.g. Greene et al., 2004).

Briefly, the theory asserts two main propositions. First, it maintains that the decision not to kill can be meaningfully described as “deontological”, because it is “naturally justified” by reference to deontology, a normative ethical system that focuses on the importance of moral rights and duties. Likewise, the decision to kill can be meaningfully described as

“utilitarian” because it is “naturally justified” by reference to utilitarianism, an ethical system that stresses the impartial maximization of consequences (Greene, 2014).[23] Second, it maintains that these judgments are the result of two functionally separate cognitive systems. Whereas a fast, intuitive and effortless System 1 sensitive to socio-emotional stimuli is the supposed causal antecedent of “deontological” judgments, a slow, deliberate and effortful System 2 is assumed to cause “utilitarian” judgments (Greene et al., 2008; see also Corey et al., 2017; Suter & Hertwig, 2011; Valdesolo & Steno, 2006). Thus, according to the Dual-Process Model it is this competition between emotional and rational processes, which causes the experience of conflict going along with the consideration of a proper dilemma.

However, several studies employing advanced methodological approaches have been conducted in recent years, the results of which have posed challenges to core tenets of the Dual-Process Model (e.g. Bago & De Neys, 2018; Gawronski et al. 2016, 2017, 2018; Hennig & Hütter, 2019; Koop, 2013; Baron & Gürçay, 2017; see also McGuire et al., 2009).

3.1.1. Subjective Utilitarian Theory

In a similar vein, Cohen & Ahn (2016) recently proposed an alternative account for explaining moral dilemma judgments. According to their Subjective Utilitarian Theory (SUT) all moral dilemma judgments can be explained by a single process, in which dilemma judges aim at identifying and subsequently executing the choice, which is of greater personal value to them. This process is supposed to be driven by the perceived value of the two respective dilemma targets. Thus, they characterize this process as *utilitarian*, because it focuses on comparing the perceived utility of both dilemma response options, saving the individual versus saving the group, respectively. Moreover, contrary to the Dual-Process Model, Cohen and Ahn (2016) stress that the perceived value of each target is the result of a *subjective*

[23] Although these philosophical descriptors are common parlance in the dilemma literature, their use has been questioned for empirical as well as conceptual reasons (e.g. Gawronski et al., 2016, 2017, 2018; Hennig & Hütter, 2019; Kahane, 2012, 2015; Kahane et al., 2015, 2018).

assessment process, the outcome of which does not need to adhere to “objective utilitarian” values as determined by a utilitarian calculus, which impartially counts the number of lives saved and lost by each decision. In this subjective process, the personal value of each target is mentally represented in the form of a normal distribution, from which the observer samples in order to determine its true value, represented by the distribution mean. As the psychological values of both targets are represented as distributions, and the decision process entails an attempt to successfully determine the true psychological value for each item, the judgment process increases in difficulty and perceived conflict to the degree that targets are similar in personal value, and to which their associated value distributions overlap. Conversely, the more targets differ in personal value and the lower the distributional overlap, the lower the perceived conflict and the easier the dilemma decision is supposed to be.

Subjective Utilitarian Theory thus contradicts both main propositions of the Dual-Process Model sketched above. First, it does not embrace the labeling of responses as “deontological” and “utilitarian”, but instead characterizes the whole decision process as fundamentally utilitarian in nature. Second, it does not rely on the assumption of two distinct processing systems and associated processing characteristics, which are supposedly systematically linked to different observable responses (e.g. Greene, 2014). Instead, it views emotional factors as fully expressed in the personal value of each target. Supporting their theoretical account, Cohen and Ahn (2016) present four experiments implementing a random walk model approach (e.g. Ratcliff, 2002). As their results indicate, overlap of psychological value distributions, which were determined in a pretest, consistently accounted for more than 90% of variance in participants’ judgment errors, as well as for 60-90% of response time data.

Thus, these findings support two central postulates of the SUT, which are directly relevant for the present analysis. First, they suggest that dilemma judgments may be comprehensively explained in terms of similarity (or divergence) of the perceived psychological values of dilemma targets. Second, they suggest that this similarity in perceived

psychological value is what leads to the experience of conflict characteristic of a proper dilemma. As Subjective Utilitarian Theory is highly specific in this regard, it can also be used to generate precise predictions. Specifically, if both postulates of Subjective Utilitarian Theory are correct, it follows that systematic manipulation of the value of dilemma targets will influence perceptions of dilemma conflict and difficulty. For instance, under the assumption that a group of people is generally assigned higher psychological value than an individual (Cohen & Ahn, Experiment 1), any manipulation that decreases the value of the individual or increases the value of the group should decrease perceived conflict and difficulty of the dilemma decision, because overlap of value distributions is reduced. Conversely, a manipulation that increases the psychological value of the individual or decreases the value of the group should increase perceived difficulty of the decision, to the extent that this increases overlap of the value distributions.[24]

3.1.2. Manipulating value of dilemma targets

Hennig and Hütter (2019) conducted several experiments that allow a direct test of these predictions. Specifically, applying a multinomial modeling approach, these authors implemented various manipulations of consequences and action properties and investigated their influence on moral dilemma judgment (Experiments 3a + 3b). As we suggest, at least three of these manipulations can be easily translated into changes of personal value of either the individual or the group implicated in the dilemma. As a consequence, precise predictions for perceived conflict can be derived based on the premises of Subjective Utilitarian Theory. As an approximation of perceived conflict, we will investigate participants’ ratings of dilemma difficulty, which were collected but not previously analyzed by the authors. A prerequisite for this approach is, however, that the psychological values of the dilemma targets is correctly assumed. That is, if Subjective Utilitarian Theory is correct, it may be

[24] However, note that particularly strong changes in psychological value may yield null effects, if the value of a target shifts but the overlap in value distributions remains identical.

erroneous to assume that the group is assigned higher value than the individual by definition. Instead of relying on this assumption, we will therefore infer psychological values from proportions of sacrificial killing found by Hennig and Hütter (2019; see supplemental analyses in Appendix C). That is, in accordance with Subjective Utilitarian Theory, we will assume proportions of sacrificial killing that exceed 50% to indicate higher psychological values for the group than for the individual, and vice versa for proportions of sacrificial killing below 50% (see Table 1).

Table 1.

Proportions of endorsement of sacrificial killing per condition, aggregated across Experiments 3a and 3b of Hennig and Hütter (2019).

	Condition	Endorsement of sacrificial killing
Personal involvement*	High	32.79%
	Low	42.42%
Self-relevance*	Present	40.50%
	Absent	34.72%
Congruency*	Incongruent	46.57%
	Congruent	28.64%
Death-avoidability*	Inevitable	51.73%
	Avoidable	23.35%

Note: * = Percentages differ significantly between levels of factor with $p < .01$.

Personal involvement. Personal involvement is a frequently manipulated factor in moral dilemma research, which has been fundamental to the development of the Dual-Process Model (Greene et al., 2001, 2004). As usually implemented, a dilemma is considered to evoke "high involvement" if it requires the causation of serious harm in a direct and unmediated manner, instead of merely deflecting preexisting harm (e.g. Greene et al., 2004; Moore et al., 2008). Thus, whereas redirecting a runaway trolley onto a track one person is standing on would not fulfill these criteria, pushing someone in front of the trolley would. Manipulation of personal involvement has been routinely found to influence moral judgment, resulting in less

sacrificial killing under circumstances of high involvement (e.g. Greene et al., 2001, 2004, 2008; Hennig & Hütter, 2019; Koop, 2013; Moore et al., 2008, 2011; Suessenbach & Moore, 2015; But see Cohen & Ahn, 2016, Experiments 2-5). From the perspective of Subjective Utilitarian Theory, such an effect is likely to result from changes in the value assigned to the individual, rather than the group. That is, as psychological proximity to the individual increases with personal involvement, the personal value of the individual is likely to increase. As the data of Hennig and Hütter (2019; see supplemental analyses in Appendix C) indicate, sacrificial killing in low-involvement scenarios was endorsed in 42.42% of the cases, suggesting higher psychological value assigned to the individual than to the group (see Table 1). Consequently, increasing psychological value of the individual due to high involvement should result in decreased overlap of value distributions. Under the assumption that dilemma difficulty represents a good approximation of perceived conflict, increasing personal involvement should lead to *decreased* difficulty.[25]

Self-relevance. Following the approach taken in previous research (e.g. Moore et al., 2008, 2011), Hennig and Hütter (2019) systematically varied whether sacrificial killing would have consequences only for others or also for oneself. This was achieved, by making the judge part of the group that would suffer negative consequences if sacrificial killing was avoided, in half of the presented dilemmas, thereby making consequences self-relevant. As previous research shows (e.g. Hennig & Hütter, 2019; Kahane et al., 2015) the presence of self-relevance reliably increases sacrificial killing. From the perspective of Subjective Utilitarian Theory, making consequences self-relevant in this manner should increase the value assigned to the group. As the judgment data of Hennig and Hütter (2019; see supplemental analyses in Appendix C) show, when self-relevant consequences were absent

[25] Note that Cohen and Ahn (2016) did not find an effect of personal involvement on reaction time and error rate in the context of their random walk model. However, given the large number of studies demonstrating an influence on sacrificial killing, as well as the effect on dilemma decision in the actual dataset, we consider our prediction on balance well justified by the available evidence.

sacrificial killing was endorsed in 34.72% of the cases, suggesting a general preference for the individual over the group (see Table 1). Consequently, increasing the value of the group by making consequences self-relevant should reduce the overlap in value distributions, resulting in *increased* difficulty.

Congruency. Third, the authors manipulated the severity of negative consequences for the group. Whereas incongruent scenarios represented “proper dilemmas” in which sacrificial killing could save several lives, congruent scenarios contained only minor negative consequences for the group (i.e. broken arms or stolen possessions).[26] This manipulation is thus likely to reduce the psychological value assigned to the group. This interpretation is in accordance with the judgment data of Hennig and Hütter (2019; see supplemental analyses in Appendix C), which indicate that reducing severity of consequences for the group decreased sacrificial killing. When considering incongruent scenarios, sacrificial killing was endorsed in 46.57% of the cases, suggesting a slight preference for the individual over the group (see Table 1). Consequently, reducing negative consequences for the group should further reduce overlap of value distributions, thereby reducing experienced conflict. Thus, Subjective Utilitarian Theory would predict that making scenarios congruent should *decrease* the difficulty of dilemma decisions.

Death-avoidability. Finally, a further manipulation of consequences was implemented, by varying whether the individual target would survive the described situation if no sacrificial killing took place, or whether her death was inevitable in the first place. As previous research suggests, under circumstances of inevitable death sacrificial killing is increased (Hennig & Hütter, 2019; Moore et al., 2008). From a Subjective Utilitarian perspective, this indicates that

[26] In the original studies, a multinomial model was applied to estimate the strength of endorsement of aggregate consequences, norm-endorsement and response tendencies as independent predictors of dilemma judgment. The manipulation of congruency was applied in order to dissociate endorsement of aggregate consequences and norm-endorsement from one another, by systematically varying whether these processes would motivate different or identical decisions. Hence the term “congruent”, which denotes the condition in which both processes should lead to identical responses.

inevitability of death decreases the personal value assigned to the individual target. Note that the amount of reduction determines the impact on experienced conflict. That is, it is possible that a strong reduction in psychological value assigned to the individual can still lead to no change in experienced conflict, when the *overlap* in value distributions remains the same. Therefore, we again took participants’ dilemma judgments into account to derive our predictions. According to the judgment data of Hennig and Hütter (2019; see supplemental analyses in Appendix C), when death was avoidable sacrificial killing was endorsed in only 23.35%, suggesting a clear preference for the individual. However, when death was inevitable, sacrificial killing was endorsed in 51.73% of the cases, which suggests a strong similarity of psychological values in this condition (see Table 1). Therefore, we suggest that Subjective Utilitarian Theory makes a clear prediction in this case as well, namely that experienced conflict should be increased when death of the single target is inevitable. Consequently, making death inevitable should *increase* perceived difficulty compared to scenarios in which death is avoidable.

3.2. The present analysis

In the present analysis of the difficulty ratings collected by Hennig and Hütter (2019) we thus investigate the four hypotheses described above. As we suggest, all of these follow directly from the premises of Subjective Utilitarian Theory and the empirically observed dilemma judgments.

3.2.1. Method

We conducted an analysis of data collected by Hennig and Hütter (2019; Experiments 3a + 3b). None of the effects described here have been previously published, as this constitutes the first analysis of the difficulty ratings collected by the authors. As the methodology used in these two experiments was identical, we conducted our analyses on the

combined datasets of both experiments. All analyses were conducted in R (R Core Team, 2018) with the packages "ez" (Lawrence, 2011) and "MASS" (Venables & Ripley, 2002).

Participants. We examined the data of 1716 German-speaking participants, who participated in one of the two studies in return for a chance of winning a 20€ gift-voucher. Participants of Experiment 3a were recruited online via the SoSciSurvey Panel (Leiner, 2014), participants of Experiment 3b responded to a message sent out via the universities mailing list. Both experiments were programmed via SoSciSurvey (Leiner, 2014) and conducted online. After the exclusion of participants who failed an instructional manipulation check (26.03%), the final sample consisted of 1269 participants (485 male; M_{age} = 33.91, SD_{age} = 14.07).

Design. Both experiments implemented a 2 (congruency: congruent vs. incongruent) × 2 (default state: action default vs. inaction default) × 2 (self-relevance: present vs. absent) × 2 (avoidability: death avoidable vs. death inevitable) × 2 (personal involvement: high involvement vs. low involvement) within-participants design.[27] The personal involvement factor was nested within the self-relevance and avoidability factors, such that within each of the present/avoidable, present/inevitable, absent/avoidable, and absent/inevitable conditions there was a high personal involvement and a low personal involvement version. Perceived difficulty constituted the dependent variable.

Materials and procedure. Eight different sacrificial scenarios were used in eight different versions each. Scenarios always required participants to choose whether to kill an individual target in order to achieve beneficial outcomes for a larger group. Scenarios were ordered in eight lists representing a counterbalanced assignment of the different factor-level combinations to individual scenarios. After selection of one list per participant, the respective

[27] Because these experiments implemented a multinomial modeling approach, default-state was implemented as additional factor, which was necessary to control for general response tendencies in a separate model parameter. As Subjective Utilitarian Theory does not provide a concrete prediction regarding this manipulations, we have not incorporated this factor into the current analysis (see General Discussion).

scenarios were presented in random order. Participants rated the difficulty of each scenario directly after indicating whether they would endorse the sacrificial killing. Specifically, participants were asked to reflect on how conflicted they felt about their decision and to what degree they felt torn between the two options. They indicated this by answering how difficult they found reaching their conclusion on a five-point likert scale item anchored at *very easy* (1) and *very difficult* (5). After the last scenario, an instructional manipulation check followed, which was designed to identify and remove participants who paid insufficient attention to actual scenario content.

3.2.2. Results

In order to investigate the effects of manipulated dilemma factors on perceived difficulty, we conducted five seperate repeated-measures ANOVAs. These analyses revealed a marginal influence of personal involvement, suggesting that judgments were considered marginally more difficult for low-involvement (M_{low} = 2.94, SD_{low} = 0.79) compared to high-involvement (M_{high} = 2.90, SD_{high} = 0.81) scenarios. However, this effect was of negligible size, $F(1, 1268) = 3.24$, $p = .072$, $\eta_p^2 < .001$. The influence of self-relevance was significant, such that judgments were considered more difficult when self-relevance was present ($M_{present}$ = 3.09, $SD_{present}$ = 0.80) rather than absent (M_{absent} = 2.76, SD_{absent} = 0.77), $F(1, 1268) = 266.02$, $p < .001$, $\eta_p^2 = .041$. Congruency had a significant influence, indicating that decisions were judged more difficult for incongruent ($M_{incongruent}$ = 3.12, $SD_{incongruent}$ = 0.84), as compared to congruent ($M_{congruent}$ = 2.73, $SD_{congruent}$ = 0.78) scenarios, $F(1, 1268) = 285.86$, $p < .001$, $\eta_p^2 = .055$. Likewise, the influence of death-avoidability was significant, indicating perceived difficulty to be higher when death of the single target was inevitable ($M_{inevitable}$ = 3.31, $SD_{inevitable}$ = 0.85) rather than-avoidable ($M_{avoidable}$ = 2.54, $SD_{avoidable}$ = 0.75), $F(1, 1268) = 1186.48$, $p < .001$, $\eta_p^2 = .187$.[28]

[28] With one exception, conducting the analyses separately per experiment led to identical results. The only divergence concerns the influence of personal involvement, in the sense that low-

3.3. Discussion

The present work constitutes an investigation of four predictions derived from Subjective Utilitarian Theory (Cohen & Ahn, 2016). According to the central postulates of this framework, 1) dilemma judgments can be comprehensively understood in terms of similarity (or difference) in the perceived psychological value of the two dilemma targets, and 2) similarity in perceived value translates into the experience of conflict, whereas dilemmas with targets of dissimilar values are perceived to be of low conflict. Based on the premises of Subjective Utilitarian Theory we avoided the assumption that participants apply “objective” utilitarianism to determine the value of dilemma targets. Instead, we derived the assumed psychological values of dilemma targets, which formed the foundation for our hypotheses, empirically by assessing observable dilemma responses. Importantly, this approach was necessary, as indicated by the fact that there was a general preference for the individual over the group, in contrast to what an assumption of “objective utilitarianism” would imply. Based on this finding, we proposed that the premises of Subjective Utilitarian Theory provide specific predictions regarding the influence of four factors manipulated in earlier work (Hennig & Hütter, 2019). Out of those four predictions, three were clearly confirmed.

Regarding the well investigated factor of personal involvement, we suggested that Subjective Utilitarian Theory would predict that experienced conflict decreases as personal involvement increases. This prediction was, in part, based on the common finding that as personal involvement increases, sacrificial killing decreases (e.g. Christensen et al., 2014, Greene et al., 2001, 2004; Koop, 2013; Moore et al., 2008, 2011; but see Cohen & Ahn, 2016, Experiments 2-5), an effect that was also found in the dataset of Hennig & Hütter (2019; see supplemental analyses in Appendix C), which provided the conflict ratings for the current

---

involvement scenarios (M_{low} = 2.91, SD_{low} = 0.82) were judged marginally more difficult than high-involvement scenarios (M_{high} = 2.86, SD_{high} = 0.84) in Experiment 3a, $F(1, 691) = 3.50$, $p = .062$, $\eta_p^2 = .005$. However, this pattern did not replicate in Experiment 3b, where difficulty of low-involvement scenarios (M_{low} = 2.97, SD_{low} = 0.75) and high-involvement scenarios (M_{high} = 2.94, SD_{high} = 0.78) did not differ, $F(1, 576) = 0.41$, $p = .520$, $\eta_p^2 < .001$.

analysis. Crucially, inspecting dilemma judgment suggested that, when sacrificial killing allowed low involvement, the individual was assigned higher value than the group (see Table 1). Consequently, based on Subjective Utilitarian Theory, we expected experienced conflict to decrease as personal involvement increases, because the increased psychological value of the individual would further reduce overlap of value distributions. However, no meaningful difference in difficulty ratings was found between high- and low-involvement scenarios.[29] This divergence between dilemma judgment and experienced conflict is arguably not fully consistent with Subjective Utilitarian Theory in its current form, as it suggests that experienced conflict and observable dilemma judgment do not necessarily go hand in hand. Specifically, it suggests that a factor may influence observable dilemma judgments and decrease sacrificial killing (in this case suggesting *decreased* similarity in psychological values) while leaving experienced conflict unaffected (suggesting *no* systematic influence on psychological values). This finding provides a potential challenge for Subjective Utilitarian Theories central assumption that observable dilemma judgment and experienced conflict alike are driven by the same psychological mechanism, namely divergence in psychological value.

The remaining effects, however, are easily incorporated into a Subjective Utilitarian framework. First, there were conceptual as well as empirical reasons to expect that self-relevance increases experienced conflict. Crucially, this prediction was based on our inspection of observable dilemma judgment, according to which the individual was generally preferred over the group when self-relevance was absent (see Table 1). Conceptually, making the dilemma judge part of the group should increase the psychological value assigned to the group. Empirically, previous research found this implementation of self-relevant consequences to increase sacrificial killing (e.g. Kahane et al., 2015; Moore et al., 2008,

[29] There was a trend towards decisions in low-involvement scenarios being judged somewhat easier than in high-involvement scenarios. However, this difference did not reach conventional levels of significance. Given the large number of datapoints covered in the analysis as well as the negligible effect size, we do not consider this effect to be meaningful and will refrain from discussing and interpreting.

2011), an effect that was also evident in the data of Hennig and Hütter (2019, see supplemental analyses in Appendix C). Consequently, we expected the presence of self-relevant consequences to make decisions more difficult, because increasing the psychological value of the group would increase overlap in value distributions. This prediction was confirmed by the results of our analyses. Thus, this finding is consistent with Subjective Utilitarian Theories assumption that dilemma judgment and experienced conflict alike are driven by the similarity of psychological values of dilemma targets.

Second, based on the judgment data suggesting that in incongruent scenarios there was a slight preference for the individual over the group (see Table 1), we expected that reducing the negative consequences for the group should reduce overlap of value distributions, resulting in less experienced conflict. As with self-relevant consequences, this prediction was confirmed, as participants judged reaching a decision in congruent scenarios easier than in incongruent scenarios.

Finally, we assessed the influence of making the death of the single target inevitable. This manipulation has been found to increase proportions of sacrificial killing in previous research (Moore et al., 2008), as well as in the experiments of Hennig and Hütter (2019; see supplemental analyses in Appendix C). As evident in the moral judgment data, when death was avoidable participants chose to abstain from sacrificial killing most of the time, suggesting a preference for the individual over the group in this condition (see Table 1). As making the death of the individual inevitable should reduce the psychological value assigned to this individual, we expected inevitability of death to increase overlap of value distributions and consequently increase experienced conflict. This prediction was confirmed, as decisions for death inevitable scenarios were judged more difficult than for death avoidable scenarios.

As described above, the results of our analyses are largely supportive of the predictions we derived from Subjective Utilitarian Theory, providing support for three out of four hypotheses. As such, they suggest that, while relying on parsimonious theoretical

assumptions, Subjective Utilitarian Theory nevertheless provides specific predictions. Thus, the present findings are largely consistent with Cohen & Ahn’s (2016) central proposition that dilemma judgments are the result of a single utilitarian process in which the subjective values of both dilemma targets are weighed against one another. Specifically, results largely converge with the assumption that differences in psychological value of dilemma targets determine both dilemma judgment and experienced conflict alike. Notably, it is not clear how the same predictions may have been derived from more traditional theories of dilemma judgment, which rely on more complex theoretical assumptions regarding the emotional or reflective nature of underlying processes (e.g. Greene, 2014; Greene et al., 2001, 2004).

Though informative, the current work is not without limitations. First, a more direct test of Subjective Utilitarian Theory may have been conducted by evaluating the influence of the described dilemma factors on participants’ response time as done in the original work underlying the theory (Cohen & Ahn, 2016), instead of approximating experienced conflict via reported difficulty, as done in the present analysis. As response time constituted the measure of conflict implemented in the original work of Cohen and Ahn (2016), this would have provided a more direct comparison with the findings that lay the foundation of the theory. Second, it should be noted that the present analysis cannot address whether difficulty or experienced conflict can explain the *response patterns* formalized in the proCNI model (Hennig & Hütter, 2019). Specifically, whereas the present analysis dealt with analyzing the difficulty assigned to judging *single* scenarios, the proCNI model estimates the *patterns of responding across different experimental conditions*. As such, the current analysis does provide some indication regarding the role of experienced conflict in norm- and consequence-consistent responding, but does not connect ideally to the response patterns estimated via the multinomial method. Future analyses may make use of hierarchical Bayesian modeling, which allows investigating the relationship between continuous predictors and model parameters, to investigate this question directly (Heck et al., 2018). Thus, whether difficulty

or experienced conflict is systematically related to meaningful response patterns remains an open question at the moment.

Finally, as Subjective Utilitarian Theory claims that dilemma responses can be comprehensively understood by assessing value distributions of dilemma targets, this also means it has to be able to explain all effects on judgment via changes in those perceived value distributions. As such, it is questionable whether all factors of influence on observable responses can indeed be reasonably integrated into the theory. For instance, as Hennig and Hütter (2019, see supplemental analyses in Appendix C) show, manipulating a dilemma default-state (thus whether a sacrificial action has already begun or not) influences observable responses, leading to more sacrifices if action rather than inaction is the default. The influence of this factor on dilemma difficulty was not investigated in the present analysis, as it is not obvious why such a manipulation would influence the value attached to dilemma targets. Conceptualized more broadly, one may argue that this manipulation should affect the value of *choices* (i.e. continue vs. stop action) rather than *targets* (i.e. one vs. five; see Johnson & Goldstein, 2003; Samuelson & Zeckhauser, 1988; Baron & Ritov, 2004). Thus, on a strict reading of Subjective Utilitarian Theory in its current form, as concerned with value of targets rather than choices, the influence of default-state does not seem easily explainable. Indeed, a pedantic reading of the framework may suggest that demonstration of any such influence on observable responses that does not translate into experienced conflict and value of dilemma targets alike may be sufficient to falsify the framework, or at least to demonstrate a severe limitation to its fundamental claims. Future work may further flesh out Subjective Utilitarian Theory by clarifying the distinction between value of targets and choices, thereby also further clarifying the conditions under which the framework may be considered falsified.

3.3.1. Conclusion

As the results of the current analysis illustrate, although some points for theoretical clarification remain., Subjective Utilitarian Theory can be applied to obtain specific and

testable predictions, while being more parsimonious in its fundamental assumptions than traditional dual-process models of dilemma judgment. Those predictions were largely confirmed outside the random walk modeling approach within which it originated. This suggests Subjective Utilitarian Theory to be a useful framework for advancing theoretical development in the field of moral dilemma judgment research.

Chapter IV

Consequences, Norms, or Willingness to Interfere – What Drives the Foreign Language Effect in Moral Dilemma Judgment?

Max Hennig

Eberhard Karls Universität Tübingen

The following chapter contains a manuscript resulting from a cooperation between Max Hennig (lead author) and Prof. Dr. Mandy Hütter (second author) entitled “Consequences, Norms, or Willingness to Interfere – What Drives the Foreign Language Effect in Moral Dilemma Judgment?”. Max Hennig and Mandy Hütter contributed 60% and 40%, respectively. Specifically, both authors contributed approximately 50% to the generation of scientific ideas, data generation, and analysis and interpretation, whereas Max Hennig wrote the manuscript.

4.1. Introduction

The ability to make judgments about what principles or actions one considers ‘right’ or ‘wrong’, ‘acceptable’ or ‘unacceptable’ is central to everyday human functioning and fulfills a crucial role in organizing social interactions and guiding behavior (Haidt, 2012). In the spirit of dual-process models of human cognition (Evans, 2008; Evans & Stanovich, 2013; Sloman, 2014) it is frequently assumed that moral judgment is determined by two different cognitive systems, one supposed to be fast, effortless and automatic (System 1), the other to be slow, effortful and deliberative (System 2). As influential models of moral cognition maintain, System 1 processes produce emotional or intuitive reactions to perceived moral violations and determine the outcome of the judgment, unless they are overridden and the judgment corrected by deliberative cost-benefit analysis or effortful, rational reasoning, which is supposed to rely on System 2 processing (Greene, 2007; Greene et al., 2004, 2008; Haidt, 2001, 2007). Much of the evidence in favor of this view has been gained by the application of hypothetical dilemmas, in which both of these systems are expected to motivate divergent responses. As conceptualized by some work, the output of System 1 may cause among others intuitive disgust responses, which leads to the moral condemnation of arguably harmless actions because they violate norms of purity (Haidt et al., 1993). Moreover, it has been reported that such actions are frequently repudiated without the ability to provide adequate

reasons for doing so (Haidt, Bjorklund, & Murphy, 2000; Haidt & Hersh, 2001). These findings have led some to conclude that intuitive processes are the main determinants of moral judgment, with reasoned consideration of tangible harm barely contributing to the evaluative process (Haidt, 2001; but see Guglielmo, 2018; Gray & Keeney, 2015b; Royzman, Kim, & Leeman, 2015; Stanley et al., 2019).

Similarly, the Dual-Process Model of moral judgment (DPM) likewise proposes that automatic System 1 processing provides a default response to situations of moral relevance, unless deliberative cost-benefit analysis supported by System 2 is engaged in and leads to a revision of this initial judgment. Specifically, data obtained with the application of sacrificial dilemmas seem to indicate that System 1 processing systematically motivates the avoidance of harmful actions, in which one person is sacrificed to save the lives of several others. System 2 processing, in contrast, is supposed to contribute to the endorsement of such sacrifices, because the lives of multiple others could be saved as a result of this decision (Greene, 2007; Greene et al., 2001, 2004). In the context of sacrificial dilemmas, the Dual-Process Model supposes that Systems 1 and 2 are functionally independent from one another and motivate responding focused on the adherence to moral norms and responding focused on the maximization of positive consequences, respectively.

It has also been argued that the differential contribution of both systems is represented by differences in responses to personal and impersonal dilemmas (Greene et al., 2001, 2004). While personal dilemmas enforce high levels of personal involvement during the act of killing (e.g. the *footbridge dilemma* in which avoiding that a trolley runs over a group of people necessitates pushing a heavy man in front of it with one’s own hands), impersonal dilemmas allow for low levels of personal involvement (e.g. the *trolley dilemma*, in which avoiding that a trolley runs over a group of people requires flipping a switch and steering it onto another track where only one person will be killed). Although the ratio of costs and benefits does not differ between those types of scenarios, it is usually found that endorsement of sacrificial

killing increases as personal involvement decreases (Koop, 2013; Moore et al., 2008, 2011). This is commonly interpreted as an indication for personal dilemmas invoking higher levels of emotional conflict than impersonal dilemmas, which then results in increased adherence to moral norms (Greene et al., 2004; Koenigs et al., 2007). Evidence consistent with the Dual-Process Model framework appears substantial, and is also frequently taken to demonstrate the existence of functionally independent cognitive systems sensitive to “hot” socio-emotional information and “cold” quantifiable indicators of costs and benefits, respectively (e.g. Bernhard et al., 2016; Lane & Sulikowski, 2017; Suter & Hertwig, 2011).

A relatively recent example of such evidence is the foreign language effect (FLE). As originally demonstrated by Keysar et al. (2012), encountering a reasoning problem in a foreign as opposed to the native language may reduce the occurrence of several cognitive biases, such as the framing effect and loss aversion. Specifically, when considering the Asian disease problem (Kahneman & Tversky, 1979), participants doing so in their native language were more likely to accept the certain death of several people for a chance of saving a larger number, when this opportunity was framed in terms of potential gains (“If you choose Medicine B there is a 33.33% chance that 200 out of 600 people will *be saved*”), than when it was framed in terms of potential losses (“If you choose Medicine B there is a 33.33% chance that 400 out of 600 people will *die*”). This risk aversion effect was not found among participants considering the problem in a foreign language, which Keysar et al. (2012) suggested to result from an increase in rational System 2 processing (also see Costa, Foucart, Arnon, Aparici & Apesteguia, 2014).

Subsequent research has carried the FLE into the realm of moral cognition, investigating the influence of language processing on judgment in the context of sacrificial dilemmas. As Costa et al. (2014) were the first to demonstrate, a similar effect can be found among participants pondering the footbridge problem – participants were more willing to push the heavy man in front of the trolley when considering the problem in a foreign than in

their native language. However, this effect was not found when participants were judging the trolley problem instead. This pattern of results has been repeatedly demonstrated in the context of the conventional dilemma paradigm (Chan, Gu, Ng, & Tse, 2016; Cipolletti, McFarlane, & Weissglass, 2016; Corey et al., 2017; Geipel, Hadjichristidis, & Surian, 2015b; Shin & Kim, 2017), and is generally interpreted as offering support for the Dual-Process Model. Specifically, many researchers have taken the FLE in moral judgment to provide evidence for what Geipel, Hadjichristidis, & Surian (2016) coined the *increased deliberation account* (Costa et al., 2014), or the *reduced intuition account* (Cipolletti et al., 2016; Corey et al., 2017; Shin & Kim, 2017), respectively (for reviews see Costa, Vives, & Corey, 2017; Hayakawa, Costa, Foucart, & Vives, 2016). According to the *increased deliberation account*, the presentation of information in a foreign as compared to the native language triggers cognitive System 2 processing and thereby increases rational deliberation, which results in more engagement in utilitarian reasoning. In contrast, the *reduced intuition account* suggests that presenting morally relevant information in a foreign language reduces System 1 processing and dials down the aversive responses triggered by the description of harmful actions, such that judgments are less guided by adherence to moral norms (e.g. “Do not kill”). Thus, both accounts would explain the same increase in sacrificial killing by reference to different underlying mechanisms. Note also that these canonical interpretations of the FLE rest on assuming the truth of the second foundational premise of the Dual-Process Model, as outlined in the general introduction in Chapter I. That is, they assume that each observable dilemma response is systematically produced by processing Systems 1 or 2, respectively.

Although past research on the FLE in dilemma judgment is informative and theoretically intriguing, we propose that there are some limitations to this work, which impact theoretical conclusions that can be confidently derived. Specifically, these limitations concern 1) the range of commonly used stimuli, 2) the suitability of the conventional dilemma approach for investigating the mechanisms underlying observable judgment, and 3) a lack of

control over general response tendencies, which may cause spurious effects. In the following sections, we will briefly discuss each of those points and describe how we aimed to address these problems in the current work.

4.1.1. Previous work relied on small samples of stimuli

It is sometimes assumed that the occurrence of an FLE in the footbridge- but not in the trolley-problem results from a difference in personal involvement between those two scenarios. That is, the footbridge scenario is supposed to be more emotionally evocative than the trolley-scenario (e.g. Greene et al., 2001, 2004; Valdesolo & Steno, 2006).[30] Based on this it has been proposed that foreign language does not affect responses in low-involvement trolley-type scenarios, because there is less of a prepotent emotional response that could be attenuated (e.g. Corey et al., 2017; Costa et al., 2014). Consequently, finding a foreign language effect only in the footbridge scenario may suggest emotional processes as underlying mechanism.

However, one reason why this conclusion may not be warranted is that studies investigating the FLE have only employed small samples of experimental stimuli. That is, most research has relied on trolley- and footbridge scenarios either heavily (e.g. Corey et al., 2017; Geipel et al. 2016, Shin & Kim, 2017) or exclusively (Cipolletti et al., 2016; Costa et al., 2014). As such, it is not clear whether effects are due to differences in personal involvement, or idiosyncratic properties of these specific dilemmas. Moreover, in the rare cases where additional scenarios were used, this sometimes led to contradictory results. For instance, whereas Shin and Kim (2017) found a foreign language effect for the high-involvement “crying baby” dilemma, Geipel et al. (2016) did not. This problem is exacerbated

[30] However, note that the evidence regarding this point is mixed. That is, whereas some research has found support for this assumption (e.g. Greene et al., 2001, 2004; Koenigs et al., 2007), some has not (e.g. Horne & Powell, 2013; Lotto et al., 2014; Nakamura, 2013). Others still have found some systematic differences in emotional evocativeness, yet conclude that these may be less impactful than commonly assumed (Horne & Powell, 2016).

when considering research, which investigated the foreign language effect with a large and more generalizable sample of high-involvement stimuli. Using the entire battery of 39 dilemmas proposed by Greene et al. (2004), Chan et al. (2016) found no evidence for a foreign language effect, neither when analyzing all scenarios, nor when conducting the analyses only for the 22 high-involvement dilemmas. Importantly, though, when analyzing only the footbridge dilemma, an FLE was found. Thus, when considering all the research employing the conventional dilemma approach, results robustly indicate only an effect in the footbridge dilemma and an absence of effect in the trolley dilemma. Thus, the FLE in moral dilemma judgment may in fact be more heavily stimulus driven than commonly considered. Specifically, attributing the effect to differences in evoked personal involvement does not seem warranted, given the limited range of commonly used stimuli. As such, it is not clear whether differential effects occur because of systematic differences in evoked emotional processing, or because of idiosyncrasies of the specific scenario(s) under investigation.

4.1.2. Previous work found no support for mechanistic assumptions

A further problem concerns the suitability of the conventional dilemma approach for assessing the supposed underlying psychological mechanisms. As has been pointed out before (Conway & Gawronski, 2013; Gawronski et al., 2016, 2017), by using only scenarios in which utilitarian and deontological styles of processing lead to divergent responses, the adherence to both principles is conflated in the same outcome measure. Due to this problem of inverse relation, independent estimations of ‘utilitarian’ and ‘deontological’ inclinations are not provided and the dependent measure of the conventional approach may, at best, represent the dominance of one process over the other. This is of relevance for the investigation of the FLE because, as a consequence, the mechanisms supposedly underlying responses cannot be investigated by assessing responses to the conventional paradigm alone. In order to provide stringent evidence in favor of the *increased deliberation* or *reduced intuition* accounts, respectively, the mechanisms supposed to underlie dilemma judgment have to be investigated

as well. One method for doing so is conducting mediation analyses. Implementing this approach, Geipel et al. (2015b) found that subjective distress did not mediate the effect of language on endorsement of sacrifices in the trolley and footbridge dilemmas, which contradicts the *reduced intuition account* (for similar results also see Chan et al., 2016; Geipel et al., 2015a).[31]

An alternative approach to testing the supposed mechanisms underlying the FLE rests on the application of methods that are able to solve the conflation of ‘deontological’ and ‘utilitarian’ inclinations in one outcome measure. As proposed by Conway & Gawronski (2013), this can be achieved by manipulating the consequences of a dilemma decision. Thus, one can present not only incongruent scenarios in which a sensitivity to beneficial consequences and a sensitivity to norms should lead to divergent responses (e.g. the classical trolley dilemma), but also with scenarios in which a sensitivity to beneficial consequences and a sensitivity to norms should lead to the same response (e.g. redirecting the trolley would only cause minor injuries instead of death, such that a sensitivity to consequences would not motivate sacrificial killing). Consecutively, parameters can be estimated, which represent the probability for endorsement of norms and consequences independent of one another, such that the problem of inverse relation can be avoided.

Several recent studies have taken this process dissociation (PD) approach and appear to converge on evidence in favor of the reduced intuition account. According to Muda, et al. (2018), effects of foreign language are demonstrable on deontological and utilitarian inclinations alike, once estimated independently, in the sense that *both* are reduced in the case

[31] On a related note, Hayakawa and Keysar (2018) have recently argued that the FLE in moral judgment may result from changes in mental imagery, specifically from imagining the single target with decreased vividness in the case of foreign language. However, the amount of variance in dilemma response explained by vividness of imagination demonstrated in their work was small (7%). Moreover, their use of a continuous rating scale allows for alternative interpretations, such as increased centrality bias in the foreign language condition (Montero-Melis, Isaksson, van Paridon & Ostarek, 2019). Thus, whereas the demonstrated mediation is suggestive and consistent with prior findings (Amit & Greene, 2012), more research is needed to draw solid conclusions.

of foreign language. Note, however, that in the context of the conventional paradigm these effects would cancel each other out, such that no foreign language effect would be observed, which makes these results inconsistent with earlier findings. Six experiments conducted by Hayakawa, Tannenbaum, Costa, Corey, and Keysar (2017) come to the same conclusion with regard to deontological inclinations as assessed by the PD procedure, finding a reduction by foreign language in five of their studies. However, regarding the effect on the PD procedure’s measure of utilitarian inclinations the results are inconsistent, demonstrating an effect of foreign language in only three experiments. Notably, inconsistent with what the increased deliberation account would suggest, in these three experiments foreign language *decreased* utilitarian inclinations. Again, this is only partially consistent with previous research, as the effects on *U*- and *D*-parameters would compensate each other and lead to a null-effect in the conventional paradigm.[32]

On balance, results of research employing the PD approach thus appear to support the reduced intuition account, while contradicting the increased deliberation account. However, in order to constitute robust evidence for this account or for the Dual-Process Model, the role of System 1 processing as underlying mechanism needs to be investigated as well. As the Dual-Process Model maintains, System 1 and System 2, the supposed causal antecedents of deontological and utilitarian judgments, are functionally independent (e.g. Greene et al., 2004). From this, it follows that an effect of foreign language on deontological inclinations should be related to measures of System 1 processing only, while remaining unrelated to measures of System 2 processing. Mediation analyses Hayakawa et al. (2017) conducted on the results of two of their studies provide only mixed support for this prediction. That is, in both cases interpersonal reactivity, a measure of general empathic concern towards others (Davis, 1983), emerged as a significant mediator of the effect of language on the PD’s

[32] Note that integration with the results of previous research is also complicated by the fact that the PD approach does not distinguish between low- and high-involvement scenarios.

measure of deontological inclinations. However, the effect was also mediated by need for cognition (Cacioppo & Petty, 1982) and scores on a cognitive reflection test (Baron, Scott, Fincher, & Metz, 2015), both indicators of System 2 processing. In fact, in both studies need for cognition and cognitive reflection combined explained a larger proportion of the FLE than the employed measure of empathic concern, which is inconsistent with the proposition of functional independence endorsed by the Dual-Process Model (e.g. Greene 2014).

4.1.3. Most previous work did not control for general response tendencies

Although the PD approach dissociates deontological and utilitarian inclinations from one another and thereby solves the problem of functional dependence, the procedure lacks control over general response tendencies. That is, responses that are motivated by neither deontological nor utilitarian inclinations are not captured by the measurement model underlying the PD approach and may end up contaminating the measures of interest. One relevant example of such confounds are general action or inaction tendencies. These are systematically confounded with ‘deontological’ and ‘utilitarian’ responses in the conventional dilemma approach and the PD approach alike, in the sense that the ‘utilitarian’ response always requires interfering and changing the described situation (e.g. acting and *pushing* the heavy man in front of the trolley) while the ‘deontological’ response requires passivity and inertia (e.g. remaining inactive and *not pushing* the heavy man in front of the trolley). This is relevant, because such action tendencies have been shown to influence dilemma responding. For instance, results by van den Bos et al. (2011) indicate that experimentally induced as well as individually assessed levels of behavioral disinhibition are related to interference in trolley and footbridge dilemmas, such that more ‘utilitarian’ judgments were observed the lower participants scored on a measure of behavioral inhibition, or when the concept was made experimentally salient. Similarly, results by Crone and Laham (2017) suggest that utilitarian inclinations may be no more predictive of responses to conventional dilemmas than action inclinations, once those two are experimentally separated. That is, if endorsement of

utilitarianism and preference for action are systematically confounded, dilemma responses may be motivated by a general preference for interference over inertia, and subsequently misinterpreted as indicative of cost-benefit analysis (see also Duke & Bègue, 2015). An analytic method that dissociates ‘utilitarian’ and ‘deontological’ inclinations from one another while also addressing this confound and controlling for general response tendencies would thus lead to less ambiguous results.

One such approach consists of the application of multinomial processing tree (MPT) modeling as proposed by Gawronski and colleagues (2016, 2017). In this approach norms and consequences, the defining features of deontology and utilitarianism, are manipulated orthogonally to one another, similar to the implementation in the PD approach. Additionally it can also be systematically varied whether a preference for interference and action leads to norm-breaking or to norm-adherence, in order to solve the conflation with (in)action tendencies. An additional advantage is that this procedure provides goodness-of-fit tests that indicate whether the specified theoretical model provides a good account of the data to begin with. Because MPT modelling also allows for the estimation of more than two independent parameters, the validity of a model can then be investigated, which predicts dilemma responses from endorsement of consequences (*C*), endorsement of norms (*N*), and a general preference for inaction over action (*I*). Additionally to controlling for (in)action tendencies the estimated *I*-parameter also serves a more general ‘quality control’ function. That is, variance in responses that cannot be explained by the *C*- and *N*-parameters is directed into the *I*-parameter, such that measurement error in the parameters of primary interest is substantially reduced and the validity of the measured concepts is increased. By now, this approach has been used in more than a dozen studies, all of which indicate that general response tendencies have to be considered for the theoretical model to provide a suitable account of the data (Białek, Paruzel-Czachura and Gawronski, 2019; Brannon et al., 2019; Gawronski et al., 2016, 2017, 2018; Hennig & Hütter, 2019; Zhang et al., 2018). Moreover, some of those

experiments demonstrate how insufficient control over response tendencies may lead to spurious results, like an apparent reduction in consequence-sensitive “utilitarian” judgment under cognitive load (e.g. Gawronski et al., 2016, 2017; see also Zhang et al., 2018).

Recently, Białek et al. (2019) have applied the CNI model to the investigation of the FLE. The results of their integrative analysis of four different polish bilingual samples were largely consistent with those obtained with a PD approach. That is, Białek et al. (2019) found foreign language to reduce sensitivity to aggregate consequences (Hayakawa et al., 2017) and norms (Hayakawa et al., 2017; Muda et al., 2018) alike, while inaction tendencies remained unaffected. Put differently, because in the case of foreign language a lower proportion of participants’ answers was guided by aggregate consequences and norms, as captured by *C*- and *N*-parameters, this means that a higher proportion of answers was guided by action tendencies and guessing, as captured by the *I*-parameter. Based on these results the authors thus concluded that foreign language, rather than having a selective influence on emotional or reflective processing, may indeed reduce moral concern as a whole.[33]

4.1.4. The present research

Given the methodological concerns discussed above, we suggest that there is still a large number of unknowns surrounding the influence of foreign language on dilemma judgment. Specifically, though providing multiple demonstrations of a FLE in dilemma judgment, previous research does not offer conclusive evidence on its underlying

[33] Note that, as in the work that applied a PD approach (Hayakawa et al., 2017; Muda et al., 2018), the reduced concern for norms and aggregate consequences in the foreign language condition cancelled out in conventional analyses, leading to a null-effect. Thus, I explicitly recognize that talking about “the foreign language effect” as a unitary phenomenon is likely in error, as the effects demonstrated in PD- and CNI-studies represent a different pattern than the effect originally demonstratd by Costa et al. (2014). Likewise, comparison with the original effect is again impeded by the fact that the battery of dilemmas used by Białek et al. (2019) does not distinguish between low- and high-involvement dilemmas. Given these complications I will attempt to keep the discussion streamlined by reserving the label “foreign language effect” to refer to the pattern demonstrated by Costa et al. (2014), as this also was the effect that motivated our studies to begin with, and touching upon the effect demonstrated in PD- and CNI-work wherever helpful. Thus, unless otherwise stated, my use of the term “foreign language effect” refers to the original effect (Costa et al., 2014).

mechanisms. Also it is currently not clear in how far the FLE is attributable to different levels of personal involvement evoked by different types of dilemmas, or rather driven by the properties of the specific stimuli usually employed (trolley vs. footbridge). Consequently, our study was designed to address both of these concerns. Using multinomial processing tree modeling, we applied the proCNI model of moral judgment (Hennig & Hütter, 2019) to estimate participants’ endorsement of aggregate consequences (*C*), norm-endorsement (*N*), and inertia (*I*) as orthogonal model parameters (see Figure 3). In contrast to the original CNI model (Gawronski et al., 2016, 2017, 2018), the proCNI model focuses specifically on scenarios implementing proscriptive norms against killing. Thus, we implemented only the norm ‘Do not kill’, in order to avoid the conflation of different and potentially non-comparable norms into one *N*-parameter (see Janoff-Bulman et al., 2009).

In order to investigate in how far the FLE results from evoked levels of personal involvement or properties of specific stimuli, we employed several personal and impersonal scenarios alike. For exploratory purposes we also collected measures of participants’ preference for intuition and preference for deliberation as indicators of System 1 and System 2 processing, respectively (Batch, 2004). With this approach, we aim to conduct a direct test of the decreased intuition and increased deliberation accounts, and in this context also provide an evaluation of the functional mechanisms underlying dilemma judgments as proposed by the Dual-Process Model.

Specifically, the increased deliberation account would predict foreign language to increase endorsement of aggregate consequences, while the decreased intuition account would predict decreased norm-endorsement (e.g. Geipel et al., 2016). However, it is also conceivable that the FLE is an artefact resulting from uncontrolled response tendencies. It may for instance be the case that foreign language does reduce inertia and thereby increases a general willingness to interfere and change a described status-quo, while leaving systematic endorsement of aggregate consequences and systematic norm-endorsement unaffected. This

would be undetectable in the context of the conventional and PD-paradigms, but instead manifest in the form of increased sacrificial killing (conventional approach) or spurious effects on *U* or *D*-parameters (PD approach; see Gawronski et al., 2016, Experiment 2; Gawronski et al., 2017, Experiments 2a + 2b; also see Zhang et al., 2018). Finally, as Białek et al. (2019) propose, foreign language may reduce moral concern as a whole while having no impact on response tendencies. This should be expressed in reduced *C* and *N*-parameters, while the *I*-parameter remains unaffected (also see Hayakawa et al., 2017).

Given the current state of the literature we remain agnostic about whether a FLE may be demonstrable in responses to all scenarios (Białek et al., 2019; Hayakawa et al., 2017; Muda et al., 2018), restricted to high-involvement scenarios (Shin & Kim, 2017), or restricted to the footbridge scenario (Chan et al., 2016; Cipolletti et al., 2016; Corey et al., 2017; Costa et al., 2014; Geipel et al., 2015b). Moreover, irrespective of language condition, the Dual-Process Model would predict norm-endorsement to be related to preference for intuition and endorsement of aggregate consequences to preference for deliberation (Greene et al., 2004, 2008), which constitute measures of preference for System 1 and System 2 processing, respectively (Betsch, 2004).

4.2. Experiment Series 4

4.2.1. Experiment 4a

Experiment 4a assessed participants’ endorsement of aggregate consequences, norm-endorsement, and inertia in the context of the proCNI model (Hennig & Hütter, 2019). Specifically, we assessed the impact of language on model parameters, and investigated differential effects of language depending on personal involvement. All presented scenarios represented instances of sacrificial killing. After completion of the scenarios, measures of preference for intuition and preference for deliberation were collected as proxies for System 1 and System 2 processing, respectively (Betsch, 2004).

4.2.1.1. Method

Participants. A total of 378 participants were recruited via the universities mailing list in exchange for winning one of ten 20.00€ gift vouchers. Following the criteria used in prior work (Muda et al., 2018; Hayakawa et al., 2017; Costa et al., 2014; Keysar et al., 2012) we excluded participants who were English speaking bilinguals or did not have German as their native language (N = 24), spent more than 12 months in an English speaking foreign country (N = 17), self-rated their understanding of the dilemmas as less than 4 on a 7-point likert scale (N = 2) or failed an instructional manipulation check (N = 88), resulting in a final sample of 247 participants (87 male; M_{age} = 23.97, SD_{age} = 5.56).

Design, materials and procedure. We implemented a 2 (congruency: congruent vs. incongruent) x 2 (default state: action default vs. inaction default) x 2 (personal involvement: high vs. low) x 2 (self-relevant consequences: present vs. absent) x 2 (language: native vs. foreign) mixed design with repeated measures on the first four factors.[34] The self-relevant consequences offered in the self-relevance present condition would always incentivize the killing. The personal involvement factor was nested within the self-relevance factor, such that within each of the self-relevance present and self-relevance absent scenarios there was a high-involvement and a low-involvement version. Language was manipulated between participants, such that participants were randomly assigned to reading all scenarios either in their native (German) or the foreign language (English).

Materials and procedure. Our set of stimuli consisted of nine different scenarios, four of which were taken from earlier work (Hennig & Hütter, 2019) and five of which were newly created or adapted for this study (See Appendix D).[35] Eight lists were created, representing a

[34] We did not have specific predictions regarding differential effects of language based on presence of self-relevant consequences. However, as self-relevance nevertheless represents a manipulated factor in our set of scenarios, we will report main effects of this factor in the results section.

[35] In four of the eight scenarios used in our prior work (see Chapter II), the death of the single individual was inevitable regardless of dilemma decision. As results of Hennig and Hütter (2019; see supplemental analyses in Appendix C) indicate, the process assumptions of the Dual-Process Model

counterbalanced assignment of the different versions to the specific scenarios. For each of the scenarios participants indicated their preferred course of action, provided difficulty ratings and completed an instructional manipulation check. This manipulation check read like one of the preceding scenarios except that it explicitly stated its purpose and gave the instruction to click into the body of the text instead of answering as in the preceding scenarios. Afterwards, participants in the foreign language condition indicated their understanding of the scenarios on a 7-point likert scale, before all of the participants filled out the preference for intuition and preference for deliberation scales (Betsch, 2004) in German. Finally, participants provided demographics and information about their mother language, the age and duration at which they started learning English, the context in which they did so, and the number of consecutive months they spent in an English speaking country.

4.2.1.2. Results

Following the analytic strategy of earlier work (Hennig & Hütter, 2019) we tested several different models relating to the hypotheses we wanted to investigate. A general model with joint *C*, *N*, and *I*-parameters assessed the viability of our general model to explain the data. Only after that did we compute separate models assessing the effect of language, self-relevant consequences, personal involvement, and the separate effects of language depending on personal involvement. Finally, we conducted recursive partitioning analyses to investigate a relationship between preference for intuition and norm-endorsement, and preference for deliberation and endorsement of aggregate consequences. Analyses of the proCNI models were conducted with MultiTree (Moshagen, 2010), recursive partitioning analyses with the “Trees” package (Wickelmaier & Zeileis, 2018) implemented in R (R Core Team, 2018).

---

are likely violated in case of inevitable death. This is indicated by lack of model fit and no evidence for norm-endorsement in participants’ responses to such scenarios. Consequently, death-inevitable scenarios were replaced by four novel death-avoidable scenarios, in order to enable an assessment of Dual-Process Model assumptions about the mechanisms underlying the FLE.

Overall model. The overall model with one *C*-parameter (*C* = .27, 95% *CI* [.23, .30]), one *N*-parameter (*N* = .68, 95% *CI* [.63, .72]), and one *I*-parameter provided a good fit to the data, $G^2(1) = 0.14$, $p = .713$, $w = 0.008$. The *I*-parameter did differ marginally from its neutral reference point at .5 (*I* = .56, 95% *CI* [.49, .62]), suggesting an influence of inertia on dilemma judgment $\Delta G^2(1) = 3.03$, $p = .082$, $w = 0.037$. To investigate a relationship between model parameters and preference for intuition and preference for deliberation, we additionally applied recursive partitioning analyses (Wickelmaier & Zeileis, 2018).[36] Specifically, we included PID-I and PID-D as covariates in the recursive partitioning model to investigate a relationship between preference for intuition and deliberation, and endorsement aggregate consequences and norm-endorsement, respectively. Results found no parameter heterogeneity based on PID-I ($S = 10.99$, $p = .551$) or PID-D ($S = 10.45$, $p = .626$), indicating no relationship between these covariates and our model parameters.

Language. The model estimating *C*-, *N*-, and *I*-parameters separately for foreign and native language conditions fit the data well, $G^2(2) = 0.56$, $p = .760$, $w = 0.016$. Setting parameters equal across language conditions indicated no effect of language on the *C*-parameter ($C_{native} = .25$, 95% *CI* [.21, .30], $C_{foreign} = .30$, 95% *CI* [.24, .35]), $\Delta G^2(1) = 1.42$, $p = .234$, $w = 0.025$. However, the language manipulation did exert an effect on the *N*-parameter, $\Delta G^2(1) = 9.13$, $p = .003$, $w = 0.064$, resulting in a lower *N*-parameter in the foreign ($N_{foreign} = .59$, 95% *CI* [.51, .66]) than in the native language condition ($N_{native} = .73$, 95% *CI* [.68, .79]). The *I*-parameter remained unaffected by the manipulation ($I_{native} = .59$, 95% *CI* [.49, .68], $I_{foreign} = .54$, 95% *CI* [.45, .63]), $\Delta G^2(1) = 0.47$, $p = .492$, $w = 0.015$. Parameter estimates are depicted in Figure 9.

[36] This method aims at identifying parameter heterogeneity resulting from differences between participants, as captured by individual difference measures, which are specified as predictors in the partitioning model. In the case of parameter heterogeneity the procedure specifies a value of the predictor as cutoff-point above and below which parameters differ significantly, such that data should consequently be analyzed separately in those two participant groups.

Personal involvement. The model estimating *C*-, *N*-, and *I*-parameters separately based on personal involvement conditions provided a good fit to the data, $G^2(2) = 1.48$, $p = .477$, $w = 0.026$. Setting parameters equal across conditions indicated different *C*-parameters, such that the *C*-parameter was higher for low ($C_{low} = .31$, 95% *CI* [.26, .36]) than for high-involvement scenarios ($C_{high} = .23$, 95% *CI* [.19, .28]), $\Delta G^2(1) = 4.76$, $p = .029$, $w = 0.046$. The *N*-parameter also differed, resulting in a higher *N*-parameter in the high ($N_{high} = .74$, 95% *CI* [.69, .80]) than low-involvement condition ($N_{low} = .59$, 95% *CI* [.52, .67]), $\Delta G^2(1) = 10.44$, $p = .001$, $w = 0.069$. The *I*-parameters did not differ between personal involvement conditions ($I_{low} = .52$, 95% *CI* [.44, .61], $I_{high} = .62$, 95% *CI* [.51, .72]), $\Delta G^2(1) = 1.83$, $p = .176$, $w = 0.029$. Parameter estimates are depicted in Figure 10.

Self-relevant consequences. The model estimating *C*-, *N*-, and *I*-parameters separately for present and absent conditions fit the data well, $G^2(2) = 2.25$, $p = .325$, $w = 0.032$. Setting parameters equal across conditions revealed an effect on the *C*-parameter, such that the *C*-parameter was higher when self-relevance was present ($C_{present} = .44$, 95% *CI* [.39, .49]) as compared to absent ($C_{absent} = .14$, 95% *CI* [.10, .18]), $\Delta G^2(1) = 70.84$, $p < .001$, $w = 0.179$. The *N*-parameter was also affected, resulting in a lower *N*-parameter when self-relevance was present ($N_{present} = .55$, 95% *CI* [.46, .64]) as compared to absent ($N_{absent} = .74$, 95% *CI* [.69, .79]), $\Delta G^2(1) = 15.31$, $p < .001$, $w = 0.083$. *I*-parameters did not differ between self-relevance conditions ($I_{present} = .51$, 95% *CI* [.42, .61], $I_{absent} = .60$, 95% *CI* [.51, .69]), $\Delta G^2(1) = 1.80$, $p = .179$, $w = 0.029$. Parameter estimates are depicted in Figure 11.

Differential effects of language dependent on personal involvement. To assess differential influences of language conditional on personal involvement we applied two additional models, estimating the influence of language for low- and high-involvement scenarios separately.

Language in low-involvement scenarios. The model estimating parameters separately for foreign and native language conditions fit the data well, $G^2(2) = 0.19$, $p = .91$, $w = 0.013$.

Setting parameters equal across language conditions indicated no effect of language on the *C*-parameter ($C_{impersonal\text{-}native}$ = .28, 95% *CI* [.22, .35], $C_{impersonal\text{-}foreign}$ = .32, 95% *CI* [.23, .40]), $\Delta G^2(1) = 0.42$, $p = .519$, $w = 0.020$. The same was true for the *I*-parameter ($I_{impersonal\text{-}native}$ = .52, 95% *CI* [.41, .64], $I_{impersonal\text{-}foreign}$ = .55, 95% *CI* [.43, .66]), $\Delta G^2(1) = 0.07$, $p = .792$, $w = 0.008$. In contrast, language did have an effect on the *N*-parameter, resulting in a lower *N*-parameter in the foreign ($N_{impersonal\text{-}foreign}$ = .47, 95% *CI* [.35, .59]) than in the native condition ($N_{impersonal\text{-}native}$ = .64, 95% *CI* [.56, .73]), $\Delta G^2(1) = 5.49$, $p = .019$, $w = 0.071$. Parameter estimates are depicted in Figure 12.

Language in High-Involvement Scenarios. The model estimating parameters separately for foreign and native language conditions fit the data well, $G^2(2) = 2.47$, $p = .290$, $w = 0.047$. Setting parameters equal across language conditions indicated no effect of language on the *C*-parameter ($C_{personal\text{-}native}$ = .22, 95% *CI* [.17, .28], $C_{personal\text{-}foreign}$ = .22, 95% *CI* [.14, .30]), $\Delta G^2(1) < 0.01$, $p = .996$, $w < 0.001$. In contrast, language exerted an influence on the *N*-parameter resulting in a lower *N*-parameter in the foreign ($N_{personal\text{-}foreign}$ = .61, 95% *CI* [.50, .71]) than in the native condition ($N_{personal\text{-}native}$ = .81, 95% *CI* [.75, .88]), $\Delta G^2(1) = 11.29$, $p < .001$, $w = 0.101$. The same effect was observed for the *I*-parameter ($I_{personal\text{-}foreign}$ = .52, 95% *CI* [.40, .64], $I_{personal\text{-}native}$ = .73, 95% *CI* [.57, .89]), $\Delta G^2(1) = 4.12$, $p = .042$, $w = 0.061$. In addition, in the native language condition, the *I*-parameter differed significantly from its neutral reference point at .5, $\Delta G^2(1) = 7.56$, $p = .006$, $w = 0.058$, while no such difference was observed in the foreign language condition, $\Delta G^2(1) = 0.15$, $p = .696$, $w = 0.008$. Parameter estimates are depicted in Figure 13.

4.2.1.3. Discussion

Experiment 4a provided a first application of the proCNI model to the investigation of the foreign language effect in moral judgment. Whereas our findings are only partially consistent with those of recent studies (Białek et al., 2019; Hayakawa et al., 2017; Muda et al., 2018), they do converge with much of the results obtained in the context of the

conventional paradigm (Cipolletti et al., 2016; Corey et al., 2017; Geipel et al., 2015b; Shin & Kim, 2017; but see Chan et al., 2016). Specifically, as our results indicate reduced norm-endorsement in the foreign language condition, they provide some support for the *decreased intuition* account. This effect was demonstrable in low-involvement and high-involvement scenarios alike.

However, as our recursive partitioning analyses suggest, there is no support for the assumption that norm-endorsement is systematically related to System 1 processing, as assessed by the PID-D. Consequently, our data provides no basis for the assumption that the FLE is attributable to changes in System 1 processing. In addition, our findings also suggest that the differential effects of language on moral judgment based on levels of personal involvement (Corey et al., 2017; Costa et al., 2014; Shin & Kim, 2017) may be partially attributable to general response tendencies unrelated to moral judgment proper. That is, in low-involvement scenarios inertia did not influence participants’ judgment and was not affected by language. Among high-involvement scenarios, however, participants showed a preference for inertia when those were presented in their native language, as indicated by the increased *I*-parameter. When presented in a foreign language, this influence disappeared. This finding sheds some light on the differential effects of language on responses to low- and high-involvement dilemmas found in studies using the conventional approach (Corey et al., 2017; Costa et al., 2014; Geipel et al., 2015b; Shin & Kim, 2017; but see Chan et al., 2016), in which inertia always leads to the decision not to kill (Gawronski et al., 2016, 2017, 2018; Hennig & Hütter, 2019). In the context of the conventional paradigm, therefore, a decrease in inertia due to language in the high-involvement condition may be mistaken as indication for increased cost-benefit analysis or decreased commitment to deontological norms. Similarly, increased inertia might express itself in the form of a spurious decrease of the *D*-parameter in the context of the PD approach, which is consistent with previous findings (Hayakawa et al., 2017; Muda et al., 2018; but see Białek et al., 2019).

As such, the results of Experiment 4a suggest that the foreign language effect in moral dilemma judgment results only in part from reduced norm-endorsement in the foreign language condition. Thus, previously found effects may be partially attributable to changes in inertia unrelated to moral considerations proper.

4.2.2. Experiment 4b

To minimize the likelihood of false positive findings we followed our previously applied approach of conducting a direct replication of our initial experiment, and restricting firm conclusions to those findings that replicate across studies (Hennig & Hütter, 2019; see Gawronski et al., 2017, 2018).

4.2.2.1. Method

Participants. A total of 894 participants completed the experiment via the SoSciSurvey Online Panel (Leiner, 2014) in exchange for winning one of ten 20.00€ gift vouchers. Following the same criteria as in Experiment 4a we excluded participants who were English speaking bilinguals or did not have German as their native language (N = 32), spent more than 12 consecutive months in an English speaking foreign country (N = 64), self-rated their understanding of the dilemmas as less than 4 on a 7-point likert scale (N = 10) or failed an instructional manipulation check (N = 214), resulting in a final sample of 574 participants (215 male; M_{age} = 40.52, SD_{age} = 14.16).

Design, materials and procedure. Design, materials and procedure are identical to Experiment 4a.

4.2.2.2. Results

Overall model. The overall model with one C-parameter (C = .18, 95% CI [.16, .21]), one N-parameter (N = .70, 95% CI [.67, .73]), and one I-parameter deviated slightly from optimal model fit, $G^2(1) = 2.90$, p = .089. However, the deviation was marginal and of low strength, w = 0.024, and is therefore likely to reflect trivial effects reaching conventional

levels of significance due to large sample size (Cohen, 1988; also see Klauer, 2015). The *I*-parameter did differ from its neutral reference point at .5 (I = .56, 95% *CI* [.52, .60]), suggesting an influence of inertia on dilemma judgment $\Delta G^2(1) = 8.60$, $p = .003$, $w = 0.041$. Again, we included PID-I and PID-D as covariates in a recursive partitioning model to investigate a relationship between preference for intuition and deliberation, and endorsement of aggregate consequences and norm-endorsement, respectively. Results show no parameter heterogeneity based on PID-I ($S = 4.43$, $p = .999$) or PID-D ($S = 6.25$, $p = .992$), indicating no relationship between these predictors and the proCNI model parameters.

Language. The model estimating *C*-, *N*-, and *I*-parameters separately for foreign and native language conditions fit the data well, $G^2(2) = 3.04$, $p = .219$, $w = 0.024$. Setting parameters equal across language conditions indicated no effect of language on the *C*-parameter (C_{native} = .19, 95% *CI* [.16, .22], $C_{foreign}$ = .18, 95% *CI* [.14, .22]), $\Delta G^2(1) = 0.10$, $p = .757$, $w = 0.004$. However, the language manipulation did exert an effect on the *N*-parameter, resulting in a lower *N*-parameter in the foreign ($N_{foreign}$ = .66, 95% *CI* [.61, .70]) than in the native language condition (N_{native} = .72, 95% *CI* [.69, .75]), $\Delta G^2(1) = 4.59$, $p = .032$, $w = 0.030$. The *I*-parameter remained unaffected by the manipulation, (I_{native} = .57, 95% *CI* [.51, .62], $I_{foreign}$ = .56, 95% *CI* [.49, .62]), $\Delta G^2(1) = 0.05$, $p = .832$, $w = 0.003$. Parameter estimates are depicted in Figure 9.

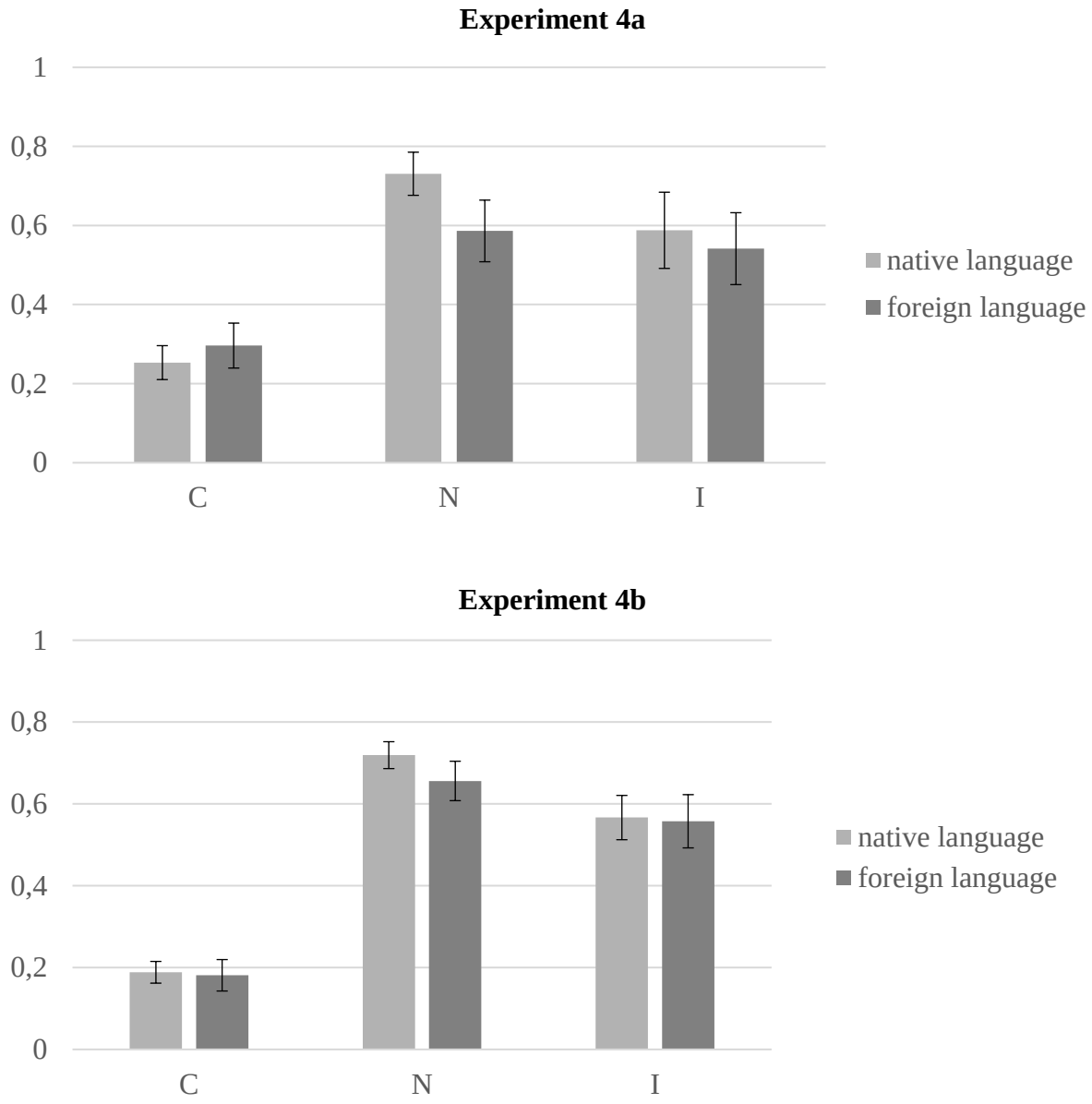

Figure 9. Parameter estimates representing endorsement of aggregate consequences (*C*), norm-endorsement (*N*) and inertia (*I*) in Experiments 4a and 4b, separated by language conditions. Error bars represent 95% confidence intervals.

Personal involvement. The model estimating *C*-, *N*-, and *I*-parameters separately for low- and high-involvement conditions provided a good fit to the data, $G^2(2) = 3.01$, $p = .222$, $w = 0.024$. Setting parameters equal across conditions indicated different *C*-parameters, such that the *C*-parameter was higher for low- ($C_{low} = .22$, 95% *CI* [.19, .25]) than for high-involvement scenarios ($C_{high} = .15$, 95% *CI* [.13, .18]), $\Delta G^2(1) = 8.60$, $p = .003$, $w = 0.041$. The *N*-parameter was also affected, resulting in a higher *N*-parameter in the high- ($N_{high} = .78$,

95% *CI* [.75, .81]) than low-involvement condition (N_{low} = .61, 95% *CI* [.56, .65]), $\Delta G^2(1)$ = 39.07, $p < .001$, w = 0.087. *I*-parameters did not differ between personal involvement conditions (I_{low} = .53, 95% *CI* [.48, .59], I_{high} = .61, 95% *CI* [.54, .68]), $\Delta G^2(1) = 2.66$, p = .103, w = 0.023. Parameter estimates are depicted in Figure 10.

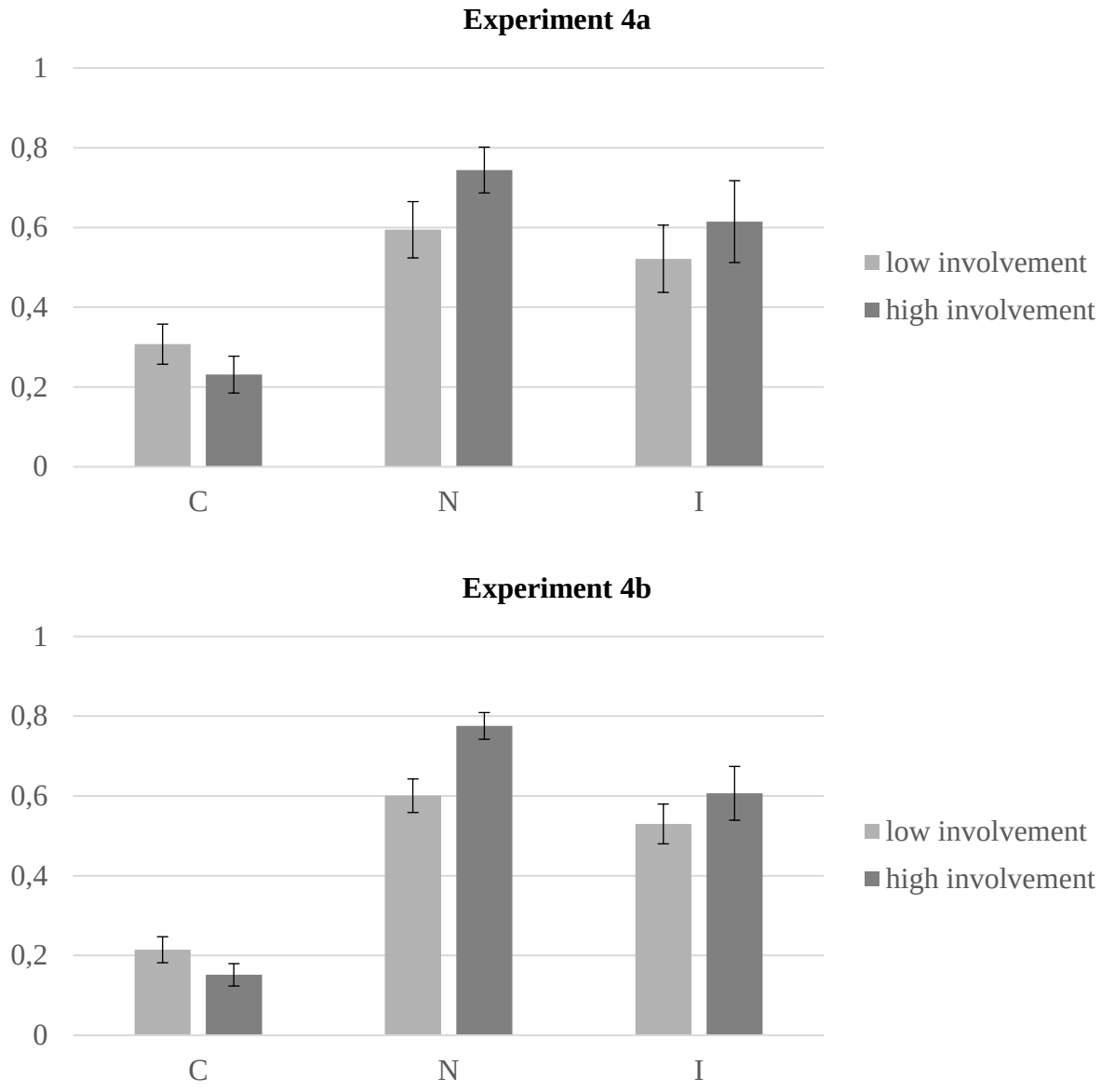

Figure 10. Parameter estimates representing endorsement of aggregate consequences (*C*), norm-endorsement (*N*) and inertia (*I*) in Experiments 4a and 4b, separated by personal involvement conditions. Error bars represent 95% confidence intervals.

Self-relevant consequences. The model estimating *C*-, *N*-, and *I*-parameters separately for present and absent conditions fit the data well, $G^2(2) = 3.62$, $p = .164$, $w = 0.027$. Setting parameters equal across conditions revealed an effect on the *C*-parameter, such that the *C*-parameter was higher when self-relevance was present ($C_{present} = .27$, 95% *CI* [.24, .30]) as compared to absent ($C_{absent} = .10$, 95% *CI* [.07, .13]), $\Delta G^2(1) = 56.43$, $p < .001$, $w = 0.105$. The *N*-parameter was also affected, resulting in a lower *N*-parameter when self-relevance was present ($N_{present} = .61$, 95% *CI* [.56, .65]) as compared to absent ($N_{absent} = .78$, 95% *CI* [.75, .81]), $\Delta G^2(1) = 38.25$, $p < .001$, $w = 0.086$. *I*-parameters did not differ between self-relevance conditions ($I_{present} = .57$, 95% *CI* [.52, .63], $I_{absent} = .55$, 95% *CI* [.49, .62]), $\Delta G^2(1) = 0.24$, $p = .621$, $w = 0.007$. Parameter estimates are depicted in Figure 11.

Language in low-involvement scenarios. The model estimating parameters separately for foreign and native language conditions fit the data well, $G^2(2) = 1.19$, $p = .551$, $w = 0.015$. Setting parameters equal across language conditions indicated no effect of language on the *C*-parameter ($C_{low\text{-}native} = .22$, 95% *CI* [.18, .26], $C_{low\text{-}foreign} = .22$, 95% *CI* [.17, .28]), $\Delta G^2(1) = 0.02$, $p = .890$, $w = 0.002$, or the *N*-parameter ($N_{low\text{-}native} = .62$, 95% *CI* [.57, .67], $N_{low\text{-}foreign} = .58$, 95% *CI* [.51, .65]), $\Delta G^2(1) = 0.79$, $p = .374$, $w = 0.012$, while the *I*-parameters differed marginally ($I_{low\text{-}native} = .50$, 95% *CI* [.44, .57], $I_{low\text{-}foreign} = .59$, 95% *CI* [.51, .68]), $\Delta G^2(1) = 2.94$, $p = .086$, $w = 0.024$. Parameter estimates are depicted in Figure 12.

Language in high-involvement scenarios. The model estimating parameters separately for foreign and native language conditions fit the data well, $G^2(2) = 2.04$, $p = .360$, $w = 0.020$. Setting parameters equal across language conditions indicated no effect of language on the *C*-parameter ($C_{high\text{-}native} = .16$, 95% *CI* [.13, .19], $C_{high\text{-}foreign} = .14$, 95% *CI* [.09, .19]), $\Delta G^2(1) = 0.27$, $p = .605$, $w = 0.007$. In contrast, language exerted an influence on the *N*-parameter resulting in a lower *N*-parameter in the foreign ($N_{high\text{-}foreign} = .73$, 95% *CI* [.66, .79]) than in the native condition ($N_{high\text{-}native} = .81$, 95% *CI* [.77, .85]), $\Delta G^2(1) = 5.03$, $p = .025$, $w = 0.031$. The same effect was observed for the *I*-parameter, such that inertia was

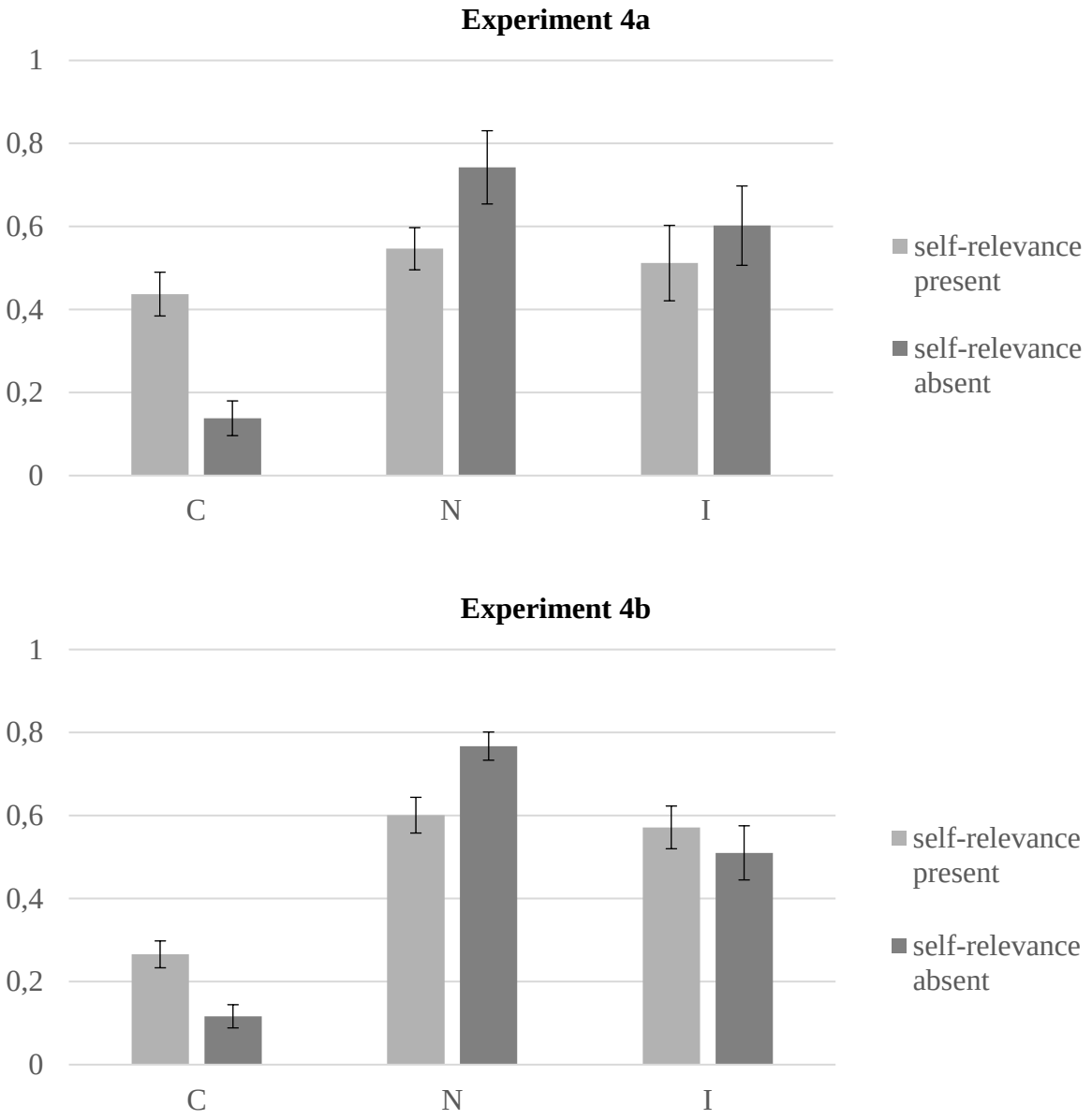

Figure 11. Parameter estimates representing endorsement of aggregate consequences (*C*), norm-endorsement (*N*) and inertia (*I*) in Experiments 4a and 4b, separated by self-relevance conditions. Error bars represent 95% confidence intervals.

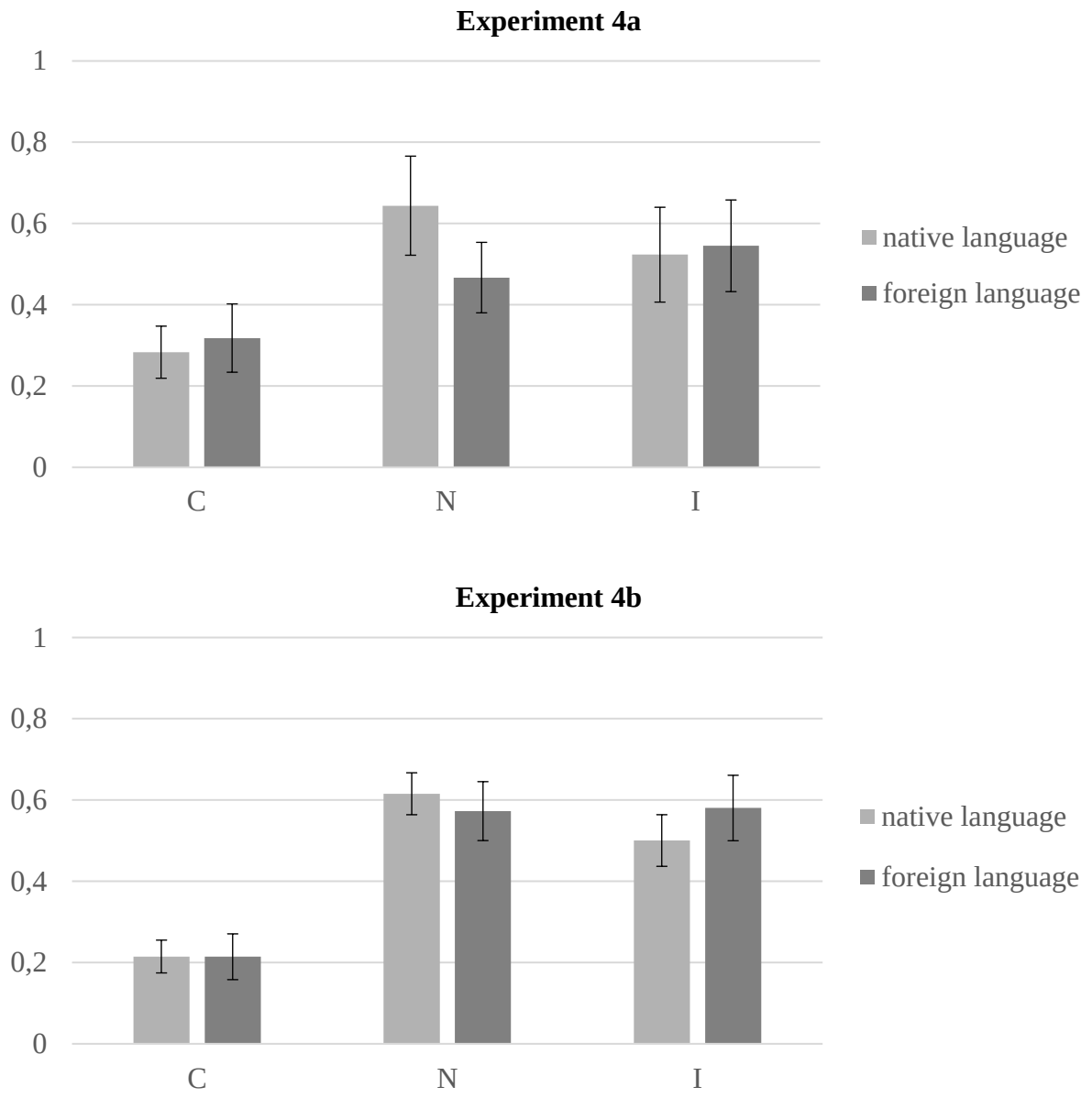

Figure 12. Parameter estimates representing endorsement of aggregate consequences (*C*), norm-endorsement (*N*) and inertia (*I*) in low-involvement scenarios in Experiments 4a and 4b, separated by language conditions. Error bars represent 95% confidence intervals.

lower in the foreign ($I_{high\text{-}foreign}$ = .51, 95% *CI* [.40, .61]) as compared to native language condition ($I_{high\text{-}native}$ = .69, 95% *CI* [.59, .78]), $\Delta G^2(1) = 6.49$, $p = .011$, $w = 0.036$. In addition, in the native language condition, the *I*-parameter differed significantly from its neutral reference point at .5, $\Delta G^2(1) = 5.75$, $p = .017$, $w = 0.033$, while only a marginal difference was observed in the foreign language condition, $\Delta G^2(1) = 2.67$, $p = .085$, $w = 0.024$. Parameter estimates are depicted in Figure 13.

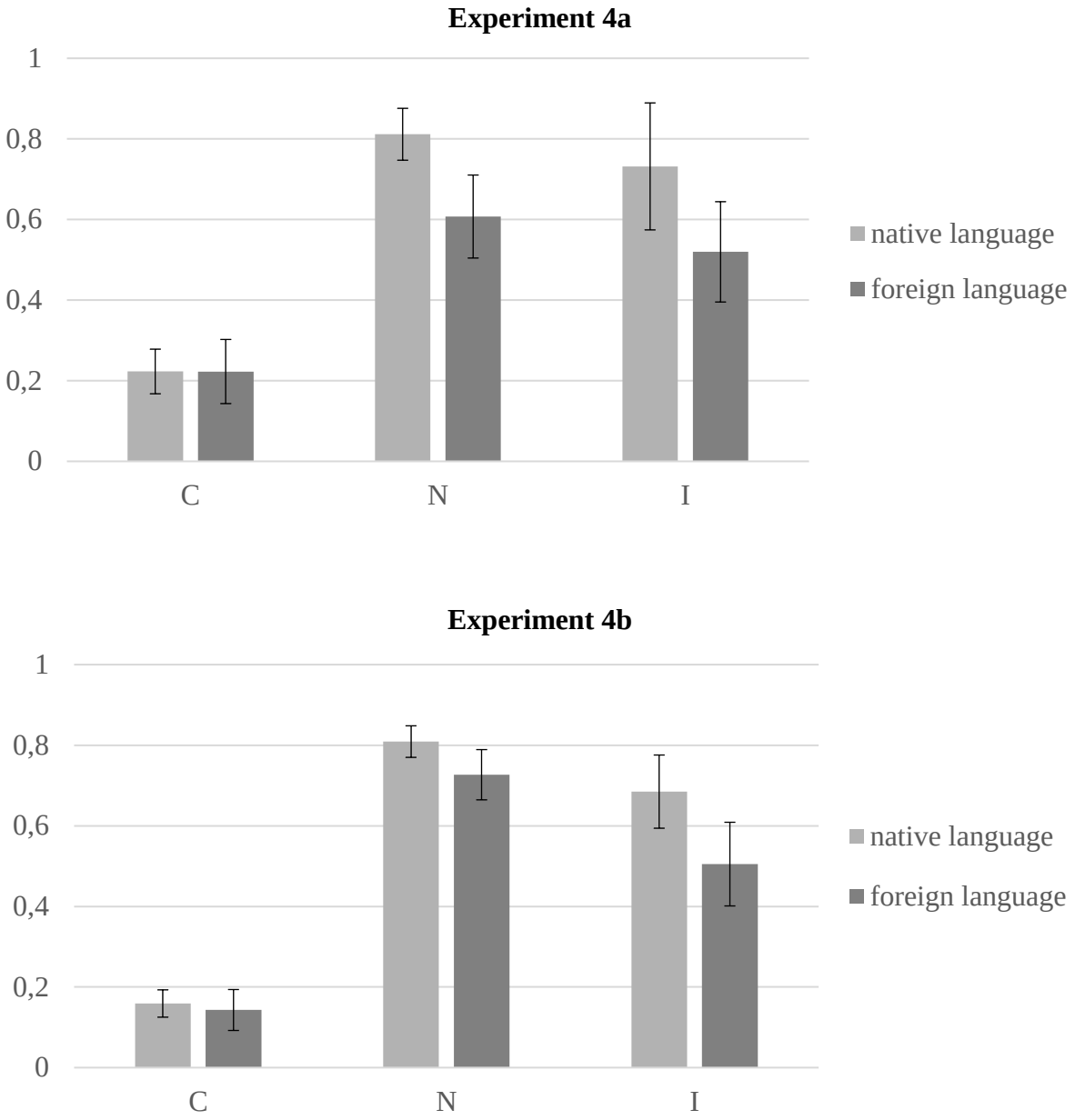

Figure 13. Parameter estimates representing endorsement of aggregate consequences (*C*), norm-endorsement (*N*) and inertia (*I*) in high-involvement scenarios in Experiments 4a and 4b, separated by language conditions. Error bars represent 95% confidence intervals.

4.3. General Discussion

In two experiments we investigated the effect of language on moral dilemma judgment with the proCNI model. Our results confirm some basic findings of earlier research, while also providing insights into underlying mechanisms. Specifically, results of both studies

indicate that reading scenarios in a foreign language did not influence participants’ endorsement of aggregate consequences. This provides no support for the *increased deliberation hypothesis*, as originally offered by Keysar et al. (2012; also see Cipolletti et al., 2016; Costa et al., 2014). Also, our results do not support the findings of some PD-studies and recent applications of the CNI model suggesting *de*creased sensitivity to consequences in the foreign language condition (Białek et al., 2019; Hayakawa et al., 2017; Muda et al., 2018). The reasons for these divergent findings are not clear at the moment, although they may be partially attributable to differences in the employed stimulus material.

In contrast, both our experiments indicate that foreign language may reduce norm-endorsement, which replicates results of recent CNI- and PD-studies (Białek et al., 2019; Muda et al., 2018; Hayakawa et al., 2017). As several studies suggest, this effect may be solely driven by responses to emotionally engaging high-involvement scenarios, and absent in low-involvement scenarios. Previous research, however, either did not assess effects of language conditional on personal involvement (Białek et al., 2019; Muda et al., 2018; Hayakawa et al., 2017), or relied almost exclusively on trolley and footbridge-scenarios when doing so (Cipolletti et al. 2016; Corey et al., 2017; Geipel et al., 2015b, Shin & Kim, 2017; 2016; but see Chan et al., 2016), such that the generalizability of those findings is unclear. As our material consisted of nine different scenarios with a high- and low-involvement version each, we think that his allows a more stringent assessment of the role of personal involvement in the foreign language effect than previous work did.

The results of our consecutive analysis of the effects of language conditional on personal involvement confirm previous findings. Specifically, no consistent effect of language on norm-endorsement was found in low-involvement scenarios, which provides no indication of a foreign language effect under such circumstances. During the supposedly more emotional high-involvement scenarios, however, foreign language consistently reduced norm-endorsement as indicated by a decreased *N*-parameter. This pattern replicates earlier work and

is consistent with the *decreased intuition account* (Costa et al., 2014; Corey et al., 2017, Geipel et al., 2015b). Furthermore it alleviates concerns about the effect being mainly stimulus driven or exclusively anchored to the footbridge problem, as may have been suspected based on some previous findings (Chan et al., 2016; Geipel et al., 2015b).

However, our results also suggest that the foreign language effect demonstrated in earlier work may partially result from lack of control over response tendencies. That is, foreign language may reduce a tendency towards inertia, such that spurious effects may arise when this is not controlled (e.g. Gawronski et al., 2016, 2017; see also Zhang et al., 2018). Specifically, in high-involvement dilemmas, we found a general preference for inertia over change when scenarios were presented in native language. This preference was reduced by presentation in foreign language. Thus, this finding suggests that the foreign language effect demonstrated in earlier work (e.g. Corey et al., 2017) is in part an artefact resulting from the systematic conflation of inertia with norm-adherence and interference with maximization of aggregate consequences, which is inherent in the conventional and PD approaches (Gawronski et al. 2016, 2017; Hennig & Hütter, 2019; see Gawronski et al., 2016, Experiments 2; Gawronski et al. 2017, Experiments 2a + 2b). As such, this finding illustrates the vital importance of controlling for general response tendencies in order to avoid misinterpretations of data due to spurious effects.

In addition, this finding is of particular importance because it shows that observable dilemma responses do not represent straightforward indicators of the emotional/intuitive or rational/deliberative nature of the processes underlying the observable judgment, which has broader theoretical implications. For instance, it is not clear how the Dual-Process Model (e.g. Greene et al., 2001, 2004, 2008; Greene, 2014), which is frequently used to interpret the foreign language effect, would be able to integrate current findings. First, as the model seems restricted to explaining dilemma responses as resulting solely from deontological norm-adherence and utilitarian maximization of consequences, it does not leave room for further

orthogonal processes determining responses. Second, and more crucially, our results seem irreconcilable with the Dual Process Models assumption of deterministic response processes (Cohen & Ahn, 2016; see Greene, 2007), stating that System 1 and System 2 processes reliably lead to deontological norm-adherence and utilitarian maximization of consequences, respectively, because under this assumption neither of those processing systems could conceivably contribute to inertia. Hence, no influence of inertia should be observed, as the output of both processing systems would be comprehensively captured by participants’ endorsement of aggregate consequences and norm-endorsement, respectively. As this neat one to one mapping of responses to processing systems thus seems unlikely to be correct, it also seems unwarranted to draw backwards inferences about the intuitive/emotional or deliberate/reflective nature of the processes underlying observable dilemma responses or estimated parameters. Consequently, we see no reason to assume that a shift in participants’ norm-endorsement is necessarily the result of changes in System 1 processing. This conclusion is further underscored by the results of our recursive partitioning analyses. As the PID-I and PID-D are established measures of the tendency to engage in affective and reflective processing (Betsch, 2004), respectively, the Dual-Process Model would have predicted positive relationships between PID-I and norm-endorsement, and PID-D and endorsement of aggregate consequences (e.g. Greene, 2014), none of which was observed in our experiments.

Finally, it should be noted that the comparison of our findings with previous research raises the question of whether it may be in error to talk about *the* foreign language effect to begin with, as the totality of the evidence appears to suggest the existence of two “foreign language effects” in dilemma judgment. Specifically, work by recent PD- and CNI model analyses suggests that foreign language may reduce moral concern as a whole, resulting in responses being determined to a stronger degree by random guessing and response tendencies unrelated to moral considerations proper, an effect that cancels out when data is analyzed via

the traditional approach (Białek et al., 2019; Muda et al., 2018; Hayakawa et al., 2017). Our findings do not replicate this pattern, but instead provide evidence for the existence of the original foreign language effect (e.g. Corey et al., 2017; Costa et al., 2014). The reasons for the occurrence of these different language effects are at the moment unclear and require more investigation. However, as the comparison with Białek et al. (2019) suggests, it seems unlikely that effects are anchored to the application of a specific analytical method and its experimental setup (i.e. the “traditional” FLE only in studies designed for the conventional method, the pattern demonstrated by Białek et al. (2019) in studies designed for PD- and MPT-analyses).

Naturally, the current work is not without limitations. First, we note that all our conclusions are bound to the dilemma paradigm we applied. That is, whether our findings meaningfully generalize to genuine in vivo moral judgment outside the context of hypothetical thought experiments is certainly up to discussion. For instance, genuine moral judgment likely contains elements not assessed in the dilemma paradigm, such as the evaluation of an agents moral character (Uhlman et al., 2015), which can diverge from the evaluation of the moral value of his actions (e.g. Uhlmann & Zhu, 2014; Uhlmann, Zhu, & Tannenbaum, 2013), or moral conviction (Mueller & Skitka, 2018; Mullen & Skitka, 2006; Skitka, Bauman, & Sargis, 2005), to name just a few. Thus, the influence of language on such phenomena remains unaddressed by the current work, which focused on the application of hypothetical sacrificial dilemmas. However, the primary focus of this work was to provide a nuanced analysis of a previously demonstrated effect, rather than drawing broad conclusions about moral judgment as a general category. Second, although the employed stimuli were constructed with great care, the English scenarios have been translated by the author, rather than by a native bilingual, which may be considered a limitation of the current experiments.

Future work may apply back-translation of scenarios by independent translators to ensure the validity of the current results.[37]

In sum, our results provide a nuanced replication and dissection of the foreign language effect in moral dilemma judgment as first demonstrated by Costa et al. (2014), which also takes its supposed underlying mechanisms into account. As our application of the proCNI model indicates, this previously demonstrated foreign language effect in high-involvement scenarios is likely due to foreign language reducing both norm-endorsement and inertia. Contrary to the predictions of the Dual-Process Model we found no relationship between our model parameters and PID-I and PID-D scores, which served as proxies for System 1 and System 2 processing. We consider our findings to provide only surface level support for the *decreased intuition account* (Geipel et al., 2015a), as we found no evidence that the reduction in participants’ norm-endorsement is related to changes in intuitive System 1 processing, as the account proposes. We do acknowledge, that the present investigation did not provide a full mediation analyses. However, we did assess the link between supposed mediator and dependent measure, in the form of a recursive partitioning analysis, which revealed no relationship between these two variables. That is, as we found no relationship between PID-I scores and norm-endorsement, our results suggest that a necessary condition for mediation was not fulfilled (see e.g. Fiedler, Schott, & Meiser, 2011).

More broadly speaking, our findings seem to suggest some boundary conditions for the scope in which the Dual-Process Model of moral judgment possesses interpretational power, as it offers no explanation for how three different processes contributing to observable dilemma responses could be deterministically produced by only two cognitive systems (see

[37] Note, however, that whereas some previous researchers did apply back-translation to their scenarios (Cipolletti et al., 2016; Corey et al., 2017; Costa et al., 2014; Hayakawa & Keysar, 2018; Hayakawa et al., 2017; Muda et al., 2018) others did not (Białek et al., 2019; Chan et al., 2016; Geipel et al., 2015a, 2015b; Shin & Kim, 2017), nevertheless yielding largely identical results (compare e.g. Costa et al., 2014 and Geipel et al., 2015a, 2015b; Hayakawa et al. 2017 and Białek et al., 2019), which suggests no strong threat to the validity of our findings.

Cohen & Ahn, 2016). As this problem indicates, future dilemma research may benefit from critically considering the assumption of deterministic response processes in order to avoid potentially unwarranted inferences about the intuitive/deliberative nature of the processes underlying observable dilemma response patterns.

Chapter V

General Discussion

Max Hennig

Eberhard Karls Universität Tübingen

5. General Discussion

The goal of the current thesis is to contribute to research on moral dilemma judgment by applying multinomial modeling, and to investigate several methodological as well as conceptual premises, on which current approaches to dilemma research rest.

In Chapter II, the proCNI model was introduced. We applied this methodological tool to investigate the assumption that dilemma response patterns can be meaningfully described as “deontological” and “utilitarian”, respectively. As our findings suggest, this clean-cut separation may be artificial and oversimplified. This is in line with several concerns expressed by conceptual critics, and also of relevance for dominant accounts of dilemma judgment, which propose a systematic mapping of two specific kinds of processes (e.g. System 1 and System 2) on specific response patterns (e.g. “deontological” and “utilitarian”).

After considering fundamental assumptions of the dual-process account, Chapter III investigated four hypotheses we derived from Subjective Utilitarian Theory (SUT). The predictions of this account, which aims to explain dilemma responding by reference to a single process and without reliance on dual-process assumptions, were largely confirmed. This investigation illustrates how single-process conceptualizations of dilemma judgment may possess equal or greater explanatory power than more conventional dual-process models, while remaining more parsimonious in their fundamental theoretical assumptions.

In Chapter IV we applied our multinomial model to investigate the foreign language effect in dilemma judgment, which is commonly interpreted in the context of the Dual-Process Model (DPM). In this investigation, we touched upon several of the issues raised in Chapter II, among which the importance of careful stimulus creation and selection for reaching valid and generalizable conclusions. We also provide an additional demonstration of the importance of controlling for response tendencies. Specifically, our results suggest the foreign language effect to apply only in the context of high-involvement dilemmas, and to be in part an artefact resulting from uncontrolled response tendencies. In addition, we found no

support for the processing claims of the Dual-Process Model (DPM), which predicts a relationship between System 1 processing and a norm-focused response pattern.

In the following sections, I will critically discuss the key contributions of the work presented in the previous chapters. The structure will roughly correspond to the three empirical chapters presented, such that sections 5.1., 5.2., and 5.3. put slight emphases on Chapters II, IV, and III, respectively, although overlap will be apparent. In section 5.3 I will identify similarities between the dual-process assumptions investigated by our research, and the claims made by early research on the phenomenon of *moral dumbfounding* (Haidt, 2001; Haidt et al., 2000), and consider potential problems shared by the dilemma and dumbfounding approaches, before integrating the current work with other models of moral judgment. In section 5.4. I close by discussing some limitations of our work and considering suggestions for future research.

5.1.What do dilemma judgments represent?

A fundamental premise on which most current approaches to dilemma judgment rest, is that dilemma response options (Greene, 2014; Greene et al., 2001, 2004) or response patterns (Conway & Gawronski, 2013; Conway et al., 2018; Reynolds & Conway, 2018) can be meaningfully described as “deontological” and “utilitarian”, respectively. As we explored in Chapter II, there are methodological as well as conceptual reasons to criticize this approach to understanding dilemma responses. Specifically, our results demonstrate some problems with relating dilemma responding to broad ethical systems.

5.1.1. “Deontology” and “Utilitarianism” in moral dilemma judgment

One key finding illustrated in Chapter II is that application of many of the canonical stimuli introduced by Greene et al. (2001, 2004) carries the risk of drawing unwarranted theoretical conclusions, as the decision to sacrifice is systematically incentivized by identifiable confounds. Specifically, in many scenarios purely egoistic concerns are sufficient

to motivate the choice of this “utilitarian” option, thereby casting doubt on the conceptual validity of the measure for those scenarios (Rosas & Koenigs, 2014; see Christensen et al., 2014; Moore et al., 2008, 2011). This is in line with the claims of previous critics, who have noted “utilitarian” judgment to be positively related to psychopathy and some measures of egoism, while finding no relationship to genuine concern for the greater good, as assessed by donation behavior (e.g. Kahane et al., 2015; Bartels & Pizarro, 2011). In light of such findings, one may reasonably wonder why the sacrificial option in such dilemmas should be labeled “utilitarian” rather than “egoistic”, as both descriptors relate to equally possible reasons underlying the judgment.

This finding already hints at a more general concern regarding the labeling of dilemma response patterns, which we investigated in Chapter II. As expressed there, I see no reason for assuming that responses or underlying processes are best described as “deontological” and “utilitarian” to begin with, as these labels carry conceptual baggage that is not adequately captured by the dilemma paradigm. Tentatively accepting a conceptualization in terms of endorsement of “norms” and “consequences” (Gawronski et al., 2016, 2017, 2018), we could instead show that both processes may be more closely related than commonly suggested. That is, as the results of manipulating death-avoidability suggest, sacrificial killing may only be rejected when it is perceived to lead to negative consequences, and accepted otherwise. As such, perceived consequences may lie at the heart of “deontological” dilemma judgment as well (e.g. Schein & Gray, 2018), in contrast to what dominant dual-process models would suggest (e.g. Cushman, 2013; Greene, 2014). Consequently, using labels that imply response patterns to be best understood as relating to “moral rules regardless of consequences” and vice versa seems imprecise and potentially misleading. Indeed, one could argue that this separation is artificial to begin with, as it does not seem obvious why “Do not kill” should be considered a moral rule, whereas the same should not be the case for “Always maximize well-being”. This criticism is in line with earlier remarks made by other critics, pointing out that

“utilitarian” dilemma judgments may be just as well reached by weighing against one another incompatible moral rules, which would reflect a distinctly deontological process (Kahane, 2012). Likewise, it has also been pointed out that such judgments lack the impartiality characteristic of genuine utilitarian reasoning (Kahane et al., 2018).

5.1.2. Dilemma response patterns – “characteristically utilitarian”?

As a response to such conceptual criticisms, defenders of the traditional philosophical descriptors have argued that the labeling merely denotes the consistency with what the respective ethical systems would recommend. E.g. Greene (2014, p. 699) has argued that decisions to sacrifice should be considered *characteristically* utilitarian, in the sense that they are “naturally justified by impartial cost-benefit reasoning”. As such, these descriptors should not be mistaken to suggest anything about the psychology underlying respective judgments, and therefore “imply nothing about the judge’s *reasons”* (Greene, 2014; p.699). This argument can be helpfully approached through the lens of the five-level taxonomy of “utilitarian” judgment proposed by Conway et al. (2018) discussed in Chapter II. According to this taxonomy, judgments reflect level-1-utilitarianism if they are *consistent* with what utilitarian reasoning would recommend, level-2-utilitarianism if they reflect *some sort of* cost-benefit reasoning, and level-3-utilitarianism if they reflect genuine commitment to maximizing the greater good. Translated into this taxonomy, it seems that Greene’s argument may be summarized as “As dilemma judgment implies nothing about the judge’s *reasons*, the dilemma approach assesses level-1-utilitarianism only. Consequently, whether or not to name the response option “utilitarian” is a question of mere semantics”. However, I would consider this argument to fall short for several reasons.

First, it does not actually seem to address the problem demonstrated by our and other research on a substantive level. That is, the concern that understanding dilemma response patterns by assuming a hard split between “deontological” and “utilitarian” processes may be artificial and oversimplified, remains unaddressed (Kahane 2012, 2015; Kahane et al., 2018).

Specifically, it does not seem to be the case that there are two processes underlying observable responses, of which one is selectively concerned with consequences of a decision whereas the other one is insensitive to them. This is a problem for the Dual-Process Model regardless of whether the respective response patterns are labeled “deontological” and “utilitarian”, or otherwise.

Even more problematic, the reasoning of Greene (2014) directly implies that responses to confounded dilemmas could be labeled *characteristically* “egoistic” with equal justification, as they are after all “naturally justified” by reference to egoism (or “level-1-egoism”, by applying the terms of Conway et al., 2018). Under this framing, findings with the conventional approach that demonstrated positive correlations between sacrificial killing and measures of egoism, psychopathy, or dark personality traits (e.g. Balash & Falkenbach, 2018; Bartels & Pizarro, 2011; Conway & Reynolds, 2018; Karandikar et al., 2019; Koenigs et al., 2012; Patil, 2015) could be cited as support for a conceptualization of such responses as “egoistic”, in the same way in which correlations with measures of cognitive reflection or manipulations of cognitive resources (e.g. Conway & Gawronski, 2013; Greene et al., 2008; Li et al., 2018; Patil et al., 2019; Suter & Hertwig, 2011) *are* routinely cited to bolster their conceptualization as “utilitarian”.

Second, and more importantly, the argument that dilemma responses are only “characteristically utilitarian” contradicts the theoretical conclusions that are frequently derived from dilemma response data. Sometimes such contradictions are apparent within the same piece of academic work. For instance, within the same article Greene first argues that dilemma responses should be considered “characteristically” utilitarian/deontological only (Greene, 2014; p. 699), and then uses dilemma response data to propose that “act consequentialism should get points for not chasing intuition and that some of its competitors (including some forms of consequentialism) should lose points for doing so.” (Greene, 2014; p. 725). Thus, here Greene suggests that empirical data indicate act utilitarianism should be

normatively favored over deontology. However, this conclusion can only follow logically if one assumes that dilemma response data do indeed provide measures of level-3 utilitarianism. Note that Greene’s normative claim even attempts to differentiate utilitarianism from *other forms of consequentialism*. This clearly entails the assumption that dilemma judgments do not only reflect level-2-utilitarianism (*some sort of* cost-benefit reasoning; Conway et al., 2018; p. 243), but level-3-utilitarianism (*concrete concern for the greater good*; Conway et al., 2018; p. 243), as this distinction would otherwise not be warranted. Likewise, Greene (2014, p: 725) proposes with the same underlying reasoning that deontological approaches to ethics should be disfavored, because they are based in intuition. Again, this normative argument is by necessity also based on a latent “level-2/3 deontology” assumption, as otherwise conclusions regarding this broad ethical system could not be warranted.

Thus, as this example illustrates, debates about how to conceptualize dilemma response patterns go beyond mere semantics, as the slope leading from “response label” to “normative ethical argument” seems slippery and easy to slide down indeed. Specifically, there seems to be a concrete danger to default to level-1-arguments when justifying the dilemma methodology, while sticking to level-3-assumptions when drawing theoretical conclusions. This, I suggest, is a further and very concrete reason in favor of abandoning a reference to broad philosophical labels and to focus on the factors that are actually manipulated in the context of the conducted experiments instead (see Gawronski et al., 2016, 2017, 2018). Specifically, conceptualizing responses in terms of endorsement of “norms” and “consequences” reduces the danger for theoretical confusion and fosters conceptual clarity regarding the precise nature of the norms and consequences under consideration. Furthermore, it explicitly invites considering the degree to which “norms” and “consequences” constitute related, instead of sharply divided, concepts.

Based on these substantive concerns, Chapter II proposes to view dilemma judgments through a lens of weak consequentialism (Sunstein, 2005), according to which dilemma

response patterns should be understood as endorsement of different sorts of consequences. The findings on which we base this proposal, and which suggest a strict split into underlying processes as “deontological” and “utilitarian” (e.g. Greene, 2007; 2014) to be artificial and overly simplistic, pose potential problems to the Dual-Process Model. That is, if both response patterns may be likewise explained by sensitivity to consequences, it is not clear to what extent a systematic mapping of process (e.g. System 2) on response pattern (e.g. “utilitarian”) possesses actual explanatory power, rather than creating merely an illusion of understanding.

5.1.3. Why response tendencies are important

The current thesis also contributes to the accumulating evidence, which suggests the importance of controlling for general response tendencies in dilemma judgment. First, and most fundamentally, in all of the analyses we presented, removing the *I*-parameter significantly reduced model fit. This replicates findings of Gawronski et al. (2016, 2017, 2018; also see Brannon et al., 2019; Crone & Laham, 2017; Duke & Bègue, 2015; van den Bos et al., 2011; Zhang et al., 2018) and underscores the importance of controlling for general response tendencies in order to avoid that parameters of main interest are contaminated by error variance. In addition, out of the four experiments that make use of the carefully constructed stimuli developed over the course of Chapter II (specifically, Experiments 3a + 3b in Chapter II, and Experiments 4a + 4b in Chapter IV), three found an *I*-parameter that was significantly larger than its neutral reference point at .5. Due to the systematic conflation of inertia with “deontological” responding, this preference for inertia or status-quo bias (Ritov & Baron, 1992; Samuelson & Zeckhauser, 1988) may be misinterpreted as decreased “utilitarian” or increased “deontological” inclinations in the conventional and PD-paradigms.

These findings add to those demonstrated by Gawronski et al. (2017), who showed at the example of cognitive load (Experiments 2a + 2b) and psychopathy (Experiments 4a + 4b) that spurious effects on “utilitarian” responding in the conventional and PD approaches may

arise if response tendencies are ignored, which has clear implications for the interpretation of previous work (e.g. Bartels & Pizarro, 2011; Koenigs et al., 2012) as well as broader theoretical implications regarding claims about the dual-process nature of dilemma judgment (Greene et al., 2008; Suter & Hertwig, 2011).

Notably, although the importance of general response tendencies has been consistently demonstrated in more than a dozen published studies by now, many recent publications do not consider their importance when drawing theoretical conclusions. Specifically, in much recent research relevant findings by Gawronski et al. (2016, 2017, 2018) are either briefly mentioned but not applied when theoretical conclusions are drawn (e.g. Byrd & Conway, in press; McPhetres et al., 2018; Patil et al., 2019; Reynolds & Conway, 2018; Rom & Conway, 2018) or ignored altogether (e.g. Christov-Moore, Conway, & Iacoboni, 2017; Conway et al., 2018; Fleischmann, Lammers, Conway, & Galinsky, 2019; Muda et al., 2018; Tannenbaum et al., 2017). This is particularly problematic for studies that aim at relating dilemma responses to measures of cognitive capacity or reflective reasoning (e.g. Byrd & Conway, in press; Hayakawa et al., 2017; McPhetres et al., 2018; Patil et al., 2019), as Gawronski et al. (2016, 2017) have demonstrated the danger of spurious effects resulting from changes in general response tendencies as represented by the *I*-parameter (also see Zhang et al., 2018).

Indeed, the mere presence of a third parameter already suggests problems for the Dual-Process Model. That is, the model is well equipped for explaining how two distinct kinds of processes can lead to two different kinds of responses or response patterns. However, it seems much more difficult to explain how Systems 1 and 2 may give rise to a third response pattern, which is independent of sensitivity to consequences and norms, respectively.

5.2. What does “the foreign language effect” reveal about dilemma judgment?

The previous sections have spelled out two challenges for the Dual-Process Model, which, as I suggest, follow directly from our data. First, our findings suggest that assuming a

hard split between “deontological” and “utilitarian” processing in order to describe response patterns may be in error. If this argument is correct then assuming a neat mapping of processing characteristics on response patterns (e.g. System 1 on “deontological”) seems unlikely to be theoretically helpful. Second, the impact of a third latent process (general response tendencies and preference for inertia) on observable responses is difficult to reconcile with the Dual-Process Model, and may lead to spurious findings if not controlled for. As I suggest, our investigation of the Foreign Language Effect (FLE) provides support for both of these concerns.

5.2.1. Investigating mechanistic assumptions based on the Dual-Process Model

As described in Chapter IV, the FLE, the finding that acceptance of sacrificial killing is increased when (supposedly emotionally evocative high-involvement) scenarios are presented in a foreign language (Costa et al., 2014) is usually interpreted in the context of the Dual-Process Model. Whereas several studies using the conventional dilemma approach conclude that this effect is driven by reduced emotional involvement (e.g. Cipolletti et al., 2016; Corey et al., 2017; Shin & Kim, 2017), most of this research relied on the dual-process assumptions that “deontological” responding is served by emotional processing (but see Muda et al., 2018), which I already critically considered above. The few studies actually investigating a mediating effect of emotional System 1 processing found no evidence for this assumption (Chan et al., 2016; Geipel et al., 2015b), or suggest that System 2 processing contributes to the effect as well (Hayakawa et al., 2017; supplemental analyses), both of which seems inconsistent with the processing assumptions of the Dual-Process Model (Greene, 2007, 2014; Greene et al., 2001, 2004).

The analysis presented in Chapter IV aimed at addressing both the conceptual and the mechanistic levels of the foreign language effect with the application of the proCNI model. That is, on the conceptual level, we investigated whether the FLE would be driven by decreased norm-endorsement or increased endorsement of aggregate consequences, thereby

testing the *decreased intuition* and *increased deliberation* accounts, respectively (see Geipel et al., 2016). On a mechanistic level, our application of recursive partitioning analysis (Wickelmaier & Zeileis, 2018) provided a tentative look at whether endorsement of aggregate consequences and norm-endorsement are systematically related to indicators of System 1 and System 2 processing, respectively (Betsch, 2004).

This investigation provides some clarification surrounding several aspects of the FLE. Specifically, we could provide a specification of the boundary conditions under which the effect may operate. First, we found evidence of reduced norm-endorsement when scenarios were presented in a foreign language, which is consistent with the assumptions of the *reduced intuition* account (see Cipolletti et al., 2016; Corey et al., 2017; Hayakawa & Keysar, 2018; Shin & Kim, 2017). Second, as our findings apply a more comprehensive set of stimuli than most studies with the conventional approach, which relied heavily or even exclusively on trolley- and footbridge-scenarios (Cipolletti et al., 2016; Corey et al., 2017; Costa et al., 2014; Geipel et al., 2015b; Hayakawa & Keysar, 2018; Shin & Kim, 2017; but see Chan et al., 2016), we could alleviate the concern that the FLE may be largely or even exclusively stimulus-driven. Third, our findings suggest that the FLE may be confined to high-involvement dilemmas. Fourth and finally, our results contribute to the interpretation of the FLE, by suggesting that it may be in part attributable to changes in participants’ inertia. That is, we provide a concrete demonstration of a response pattern that may express itself as a spurious effect when assessed with the conventional or PD-paradigm, similar to findings of Gawronski et al. (2016, 2017; also see Zhang et al., 2018). Thus, on the conceptual level, we find some support for the *reduced intuition* account while also demonstrating the importance of controlling for response tendencies to increase theoretical precision and avoid misconceptions.

However, it should be noted that our support for the reduced intuition account remains on the surface level, as we could not find any evidence for a relationship between System 1

processing and norm-endorsement. That is, the results of our exploratory recursive partitioning analyses indicated no relationship between PID-I and PID-D (Betsch, 2004), which served as our respective measures of preference for System 1 and System 2 processing, and any of our model parameters. I note that this finding should not be overinterpreted, as null-findings naturally do not prove the absence of a relationship, but merely provide no evidence for its presence. Furthermore, as the applied recursive partitioning technique (Wickelmaier & Zeileis, 2018) was introduced only recently, I am not aware of principles for applying power-analyses and determining required sample-sizes a priori. As a consequence, it is conceivable that even the large sample size of Experiment 4b reported in Chapter IV possessed insufficient power to detect an effect.

Having noted these limitations, the lack of relationship between PID-I/PID-D and model parameters is consistent with some work that found no mediation of the FLE by measures of System 1 or System 2 processing (Chan et al., 2016; Geipel et al., 2015a). However, it is inconsistent with findings of Hayakawa et al. (2017), who have found a mediation of the effect by measures of *both* System 1 and System 2 processing. Specifically, Hayakawa et al. (2017, supplemental analyses) found that emotional reactivity, (Davis, 1983), need for cognition (Cacioppo & Petty, 1982), and cognitive reflection (Baron et al., 2015) all mediate an effect of language on “deontological” responding, as assessed by the PD approach. As such, this finding is in fact inconsistent with the Dual-Process Model, which maintains that System 1 and System 2 operate in a manner that is functionally independent, such that activation of each system systematically favors one respective response pattern (Greene, 2014; Greene et al., 2001, 2004, 2008). Thus, although I would not claim that our recursive partitioning analysis has provided a stringent (let alone conclusive) test of mediation, I do note that, broadly speaking, our findings converge with the results of all mediation tests that have to the best of my knowledge been conducted on the FLE until now. That is, similar to these tests, we find no evidence supporting the Dual-Process Model.

5.2.2. The importance of careful stimulus design

Attempting to interpret the FLE across different measurement approaches also reveals another fact that further complicates this endeavor. That is, the findings obtained with the PD approach and the CNI model do not replicate the response pattern originally found with the conventional paradigm. Specifically, the FLE demonstrated with the conventional paradigm represents an increase in sacrificial killing in the case of foreign language for the footbridge- but not the trolley-scenario, and is usually explained in terms of differing levels of personal involvement (Costa et al., 2014; Corey et al., 2017; Geipel et al., 2015a). In contrast, this effect is not found for the data gathered with the PD- and CNI-approaches. Instead, when conducting an ANOVA on only incongruent scenarios, none of the seven experiments employing the PD approach (Hayakawa et al., 2017; Muda et al., 2018) or those conducted with the CNI model (Białek et al., 2019) reveals a significant increase of sacrificial killing in the foreign language condition.

This carries a noteworthy implication, as it suggests that the FLE demonstrated with the PD- and CNI-approaches (Białek et al., 2019; Hayakawa et al., 2017; Muda et al., 2018) represents a different effect than the FLE demonstrated in earlier studies with the conventional approach (e.g. Costa et al., 2014; Corey et al., 2017; Shin & Kim, 2017). This is also underscored by the fact that among three of the six studies they conducted, Hayakawa et al. (2017) found foreign language to *reduce* the *U*-parameter of the PD model as well, which is inconsistent with the FLE demonstrated with the conventional approach (also see Białek et al., 2019). Although the reasons for these divergent results between the conventional and PD-studies are not clear, there are three identifiable differences between these experimental setups. First, obviously, the PD-and CNI-studies also present participants with congruent scenarios, and in the case of CNI-studies also with scenarios implementing prescriptive norms. Second, unlike most of the experiments with the conventional approach, which rely on small samples of in between 2 and 4 stimuli (but see Chan et al., 2016), the PD-studies

employ a total of 20 scenarios (10 incongruent, 10 congruent), the CNI-work a total of 24 (six scenarios in four versions each). Third, unlike all of the studies with the conventional approach, the dilemma battery used in PD- and CNI-studies does not distinguish between low-involvement and high-involvement dilemmas.

Note that the first and second of these points also distinguish the work conducted with the conventional paradigm from our work. That is, the experiments reported in Chapter IV also presented congruent scenarios, and they also employed a larger set of scenarios (8 versions of 9 scenarios). However, these do deliberately distinguish between low- and high-involvement scenarios, and the pattern of effects found among high-involvement scenarios is fully compatible with the FLE demonstrated in studies with the conventional approach. This difference in stimuli may account for the different FLEs demonstrated in conventional and PD/CNI-studies. That is, the conventional FLE has been found almost exclusively in response to the high-involvement footbridge dilemma (though see Shin & Kim, 2017), and this pattern was supported by the results of the experiments reported in Chapter IV. Although this is not conclusive evidence, this cautiously suggests that the divergent effect demonstrated in work with the PD- and CNI-approaches may be related to lack of control about levels of personal involvement in the stimuli employed in these studies. Narrowly speaking, it may be more precise to distinguish between the *conventional* FLE, the pattern of which is supported by the findings described in Chapter IV, and the *PD/CNI FLE*, as both “effects” do not converge.

More broadly speaking, this illustrates a noteworthy challenge to dilemma research that goes beyond the interpretation of the foreign language effect. That is, it highlights the difficulties of employing narrative stimuli like dilemmas, which are likely to vary in many aspects besides those that are supposed to vary as a result of experimental manipulation. Among those narrative stimuli, dilemmas in particular require readers to uncritically accept their small-world assumptions and to suspend disbelief (Shou & Song, 2017). Also, a formulaic “kill one to save many” nature makes them fairly predictable, such that they may be

perceived as artificial and amusing rather than sobering and genuinely ethically challenging (Bartels et al., 2014). At the same time, reliance on small samples of stimuli always carries the risk of producing artificial effects that result from the idiosyncrasies of particular dilemmas. This is a pervasive concern for dilemma research, which aims to draw conclusions that are supposed to apply to “moral judgment” as a general category, rather than restricted to the contents of the specific scenario presented (e.g. Inbar, Pizarro, Knobe, & Bloom, 2009; Tannenbaum, Uhlmann, & Diermeier, 2011). As McGuire et al. (2009) have shown, the danger of artefacts that subsequently misinform theoretical conclusions is a very concrete one, which suggests that employed dilemmas should adhere to a clearly defined structure. Addressing this limitation was the purpose of the template provided in Appendix B.

Thus, dilemma research in general has to face the challenging task of simultaneously employing 1) a large number of stimuli 2) that adhere to a coherent structure, ideally 3) without being overly repetitive and predictable. When balancing these demands against one another trade-offs need to be accepted, as not all can be fulfilled simultaneously. In contrast to other approaches that emphasize external validity of stimuli (e.g. Brannon et al., 2019; Gawronski et al., 2017), our work adds to the literature on a methodological level, by providing a template that stresses the importance of internal coherence and comparability of stimuli with one another.

5.3.Moral judgment revisited

The previous sections of this thesis were dedicated to assessing several aspects of moral dilemma research, relating the paradigm to dual-process interpretations of dilemma findings. I have argued that those interpretations rely on conceptual and methodological assumptions that may not be warranted. As a consequence, to what extent dilemma judgment in particular and moral judgment in general may be understood without reliance on such assumptions is a question worthy of consideration. To illustrate this I will briefly sketch how

moral dumbfounding, another hugely impactful phenomenon in the field of moral psychology, has been traditionally understood by reliance on dual-process assumptions. I will then describe several findings of more recent research, which suggest that these assumptions may be in error and provide alternative and more parsimonious explanations. Finally, I will suggest that findings from both these lines of research are compatible with process models of moral judgment that avoid reliance on unparsimonious dual-process assumptions and stress the perception of tangible harm as central to moral judgment.

5.3.1. The parallels between dilemma research and moral dumbfounding

In their seminal study on moral dumbfounding Haidt et al. (2000) presented participants with a short story about Julie and Mark, a pair of siblings on a vacation trip. One night, when they stay together in a cabin at the beach, they both decide that it would be “interesting and fun if they tried making love” together. As the vignette stresses, the siblings use two kinds of birth control and both enjoy the experience, but nevertheless decide to never do it again. As a result, they keep the night as a special secret between the two of them, which makes them feel even closer.

As Haidt et al. (2000) explain, the vignette is deliberately constructed to pit *intuitive* and *rational* considerations against one another. That is, the incest scenario is likely to evoke a strong negative gut-reaction triggered by the violations of norms surrounding sexual purity, such that the behavior is intuitively perceived as repellent, disgusting, an ultimately immoral. Simultaneously, the authors argue, the vignette ensures that the siblings behavior has no tangible negative consequences, such that no rational defense of moral condemnation could be provided (the danger of incestuous offspring is eliminated, no emotional harm is done, etc.). Nevertheless, when presented with this scenario participants were found to stick to their judgment that the siblings’ behavior was immoral, even after a lack of reasonable justification was pointed out to them during a pre-structured conversation with the experimenter, who acted as a devil’s advocate, stressing that “no harm was done”. Specifically, participants

tended to take refuge in unsupported declarations (e.g. “It’s just wrong!”) or openly stated that they were dumbfounded (e.g. “I can’t explain my judgment”), a pattern that was not found in response to the “Heinz” dilemma (Kohlberg, 1969), in which “intuitive” and “rational” processing would lead to the same response. Although the incest scenario is the most prominent example of dumbfounding effects, other purity based scenarios are frequently used as well (e.g. using a chicken carcass for masturbation or eating one’s dog after it was killed by a car; Haidt et al., 1993; Schnall, Benton, & Harvey, 2008).

According to the interpretation of Haidt et al. (2000), the demonstration of dumbfounding indicates the dominance of intuitive over reflective processing. Famously, this argument was further fleshed out in the form of the Social Intuitionist Model of moral judgment (SIM; Haidt, 2001). According to this influential framework moral judgment is driven by two independent cognitive systems, outputting intuition and reasoning, respectively. Thus, on a basic level the SIM operates on dual-process assumptions similar to those embedded in the Dual-Process Model (DPM; Greene, 2014). According to the SIM, however, rational reasoning is almost exclusively post-hoc, such that moral intuitions possess ultimate causal power in the judgment process (but see the correspondence between Salzstein & Kasachkoff, 2004, and Haidt, 2004). Although much other evidence is provided to support the tenets of the SIM (Haidt, 2001), moral dumbfounding is frequently used illustrate its central principles, as it is supposed to provide an accurate representation of genuine in vivo moral judgment as characterized by the model (also see Haidt & Hersh, 2001; Haidt et al., 1993). However, the interpretation of moral dumbfounding has been subject to some criticism, closely related to those applicable to moral dilemma research.

First, as the phenomenon of dumbfounding itself already indicates, the incest scenario describes a situation that is strikingly counterintuitive. One implication is that this is the case in part because the scenario is inherently unbelievable. This is exactly what results by Royzman, Kim, and Leeman (2015) indicate. Conducting conceptual replications of the

original dumbfounding study, the authors replicated the original dumbfounding effect. However, as the analysis of additional credulity items indicates, the majority of participants rejected central harm-negating aspects of the scenario. Specifically, participants rated the believability of the claims that 1) the siblings would abstain from future sexual activity, 2) their relationship would not be damaged by their experience, and that 3) no other negative consequences would follow, to lie in between 20 and 37 percent only. This poses challenges to the dual-process interpretation of the dumbfounding effect, as it suggests that participants may have had harm-based reasons for their moral judgments after all, such that rational considerations would not necessarily be opposed to intuitive ones (also see Guglielmo, 2018; Stanley et al., 2019). Consequently, the endorsement of dumbfounding statements (“I can’t explain my judgment”) may have reflected *private disbelief* but public compliance resulting from perceived social pressure evoked by the experimental interview, rather than genuine moral confusion and puzzlement. Thus, there may be rational reasons for moral condemnation of the siblings’ behavior that were ignored, because they contradicted researcher assumptions, such that the experimental setup created a mere appearance of dumbfounding. This finding thus connects closely to similar work in the realm of dilemma judgment, showing that *perceived* (rather than *scenario-described*) outcome probability may predict dilemma judgment (Shou & Song, 2017), and which argues for the construction of ecologically valid scenarios (Bauman et al., 2014).

A similar alternative interpretation becomes apparent when considering work on act-character dissociations in moral judgment. As proposed by Uhlmann et al. (2015), it should be explicitly considered under what circumstances moral judgment actually represents *character* rather than *act* evaluations. As they argue, when acts possess certain properties, specifically statistical rarity (Ditto & Jemmott, 1989) and low attributional ambiguity (Snyder, Kleck, Strenta, & Mentza, 1979), this increases their informational value regarding the moral character of the agent, which consequently increases the likelihood for character instead of act

based moral judgments (also see e.g. Uhlmann et al., 2013). When this theoretical perspective is directly applied to the dumbfounding phenomenon, expected act- character dissociations emerge (Uhlmann & Zhu, 2014). Specifically, participants rated the *act* of committing dumbfounding-like disgusting but harmless transgressions (e.g. masturbating into a chicken carcass) as less immoral than of committing harmful transgressions (e.g. stealing a chicken carcass from a supermarket), while at the same time judging the chicken-masturbator to possess a more morally deficient *character* than the chicken-thief. Noteworthy, this finding extends to the dumbfounding effect itself. Thus, participants indicated higher endorsement of dumbfounding regarding whether using a chicken carcass for masturbation is an immoral *act* compared to whether stealing a chicken is. At the same time, participants showed lower endorsement of dumbfounding regarding whether the chicken-masturbator compared to the chicken-thief was of deficient moral *character*.

Although the authors correctly point out that their results replicate the original dumbfounding effect, their findings suggest that character evaluations may be involved in the evaluation of the dumbfounding effect due to the specific properties of the described action. Building on this finding, other research has directly investigated whether specific properties of dumbfounding-like purity violations may contribute to the moral condemnation of these ostensibly harmless actions. Specifically, as Gray and Keeney (2015b) showed, such purity violations seem to differ systematically from harm-based violations regarding two identifiable aspects. Specifically, their data suggests that purity violations are perceived as lower in *severity* than harm-based violations, and that they are simultaneously perceived to be higher in *weirdness*. Both effects are unsurprising and converge with the results of Uhlmann and Zhu (2014) discussed above. Importantly, they also provide a possible explanation for the previously described act-character dissociation. Specifically, as results by Gray and Keeney (2015b) indicate, whereas both severity and weirdness are positively related to judging an action as immoral, the impact of weirdness may depend on perceived severity. That is, for

lowly severe transgressions the relationship between weirdness and judging an action as immoral was stronger than for highly severe transgressions.

Taken together, these results suggest that moral condemnation of purity violations is mostly driven by the weird nature of these actions, not by their perceived severity. This suggests that dumbfounding may likewise be understood in terms of weirdness rather than severity, and that less weird purity violations may not produce dumbfounding effects. Consequently, the popular assumption that dumbfounding provides a valid representation of *in vivo* moral judgment (Haidt, 2001; Haidt et al., 2000) may be less sound than sometimes suggested. This, again, is in line with criticism leveled against the moral dilemma paradigm, stressing the importance of careful stimulus construction (McGuire et al., 2009) which, in order to avoid theoretical misconceptions, should possess external validity if general conclusions should be drawn (Bauman et al., 2014).

5.3.2. Are dual-process theories necessary for explaining moral (dilemma) judgment?

I want to stress that the purpose of the above discussion was not to provide a de facto rebuttal of the moral dumbfounding effect. Rather it was to illustrate that there are similar problems identifiable in the paradigms closely related to two of the dominant models of current moral psychology, dilemma and dumbfounding research related to the DPM and SIM, respectively. Specifically, the purpose was to show that there are reasonable alternative explanations for the dumbfounding effect, which are more theoretically parsimonious than the assumption of two distinct cognitive systems that systematically favor one of the two respective response options allowed in the paradigm. As such, the present (and the previous) section provides an attempt to integrate my discussion of dilemma judgment research, which forms the core of this thesis, into a broader body of literature on moral judgment, and to identify relevant similarities between influential models of moral judgment.

As the above discussion suggests, the dumbfounding effect may be understood without reliance on dual-process assumptions. Crucially, most of the discussed findings

suggest that, contrary to the original claims surrounding the dumbfounding paradigm, it may be understood in purely consequentialist terms. That is, findings surrounding act-character dissociations (Uhlmann & Zhu, 2014) and weirdness (Gray & Keeney, 2015b) suggest that the scenario may be deemed immoral, in part because incest is an uncanny behavior perceived as indicative of questionable moral character, and individuals of questionable moral character may commit harmful actions in the future. Furthermore, not accepting the harm-negating provisos implemented in the scenarios provides a direct consequentialist rationale for judging the behavior immoral (Royzman, Kim, & Leeman, 2015). If, counter to experimenter claims, the incest did cause damage to the siblings relationship or may occur again in the future, potentially without protection, these considerations can be naturally incorporated into a broadly consequentialist cost-benefit analysis (also see Guglielmo, 2018, and Stanley et al., 2019, for similar arguments).

This, I suggest, constitutes another similarity to the moral dilemma research discussed in the bulk of this thesis. Especially, as tentatively explored in Chapter III, it is possible to understand moral dilemma judgments as the result of a single process, in which subjectively construed costs and benefits are weighed against one another. As our application of Subjective Utilitarian Theory (Cohen & Ahn, 2016) shows, this single-process framework can be used to generate predictions that find confirmation outside the random walk modeling procedure within which the theory originated. This finding thus converges fully with the weak consequentialist framework proposed in Chapter II, according to which moral dilemma response patterns should be understood as representing sensitivity to different sorts of consequences. As I suggested above, when adopting this conceptualization a hard split between norms and consequences as conceptually distinct determinants of moral judgment evaporates, as people are likely to adhere to those norms which they perceive to produce desirable consequences (Gray & Schein, 2012; Harris, 2010).

This view converges with several findings obtained outside the realm of dilemma research, suggesting that all moral judgment is intimately tied to the perception of harm. For instance, the assumption that moral judgment may be dissociated from concrete harmful consequences (Haidt, 2001; Haidt et al., 2000) or explained in terms of deontological norms regardless of consequences (Greene et al., 2001, 2004, 2008) is undermined by findings suggesting that perceptions of harm and impurity correlate strongly (Gray & Keeney, 2015b). This suggests that purity norms are considered morally relevant precisely because negative consequences are expected to result from their violation. More directly, it has also been shown that such ostensibly harmless violations are perceived to have victims and induce the perception of harm and suffering alike (Gray, Schein, & Ward, 2014). Finally, perception of harm has also been demonstrated to explain some effects previously attributed to dogmatic or absolutist moral positions. For instance, whereas some work suggests that harmless acts like homosexual kissing are judged immoral because they are seen as disgusting (Inbar et al., 2009), subsequent work found this relationship may be completely mediated by anticipated negative consequences, as measured via belief in a dangerous world, irrespective of whether moral disapproval was assessed explicitly or implicitly (Schein et al., 2016). Similarly, it has been suggested that disgust alone may lead to an absolutist (i.e. insensitive to consequentialist considerations) moral opposition to gene-modified food (Scott, Inbar, & Rozin, 2016). However, reanalysis of these data showed that a relationship between disgust and endorsement of regulations drops to insignificance once perceived risk of gene-modified food is incorporated as a predictor. Even more strikingly, among those self-identifying as absolutist opponents of gene-modified food, perceived risk predicted endorsement of regulations *more strongly* than among non-absolutist opponents (Gray & Schein, 2016).

Thus, there is a substantial and growing body of evidence suggesting that perception of tangible harm forms a core component of moral judgment, arguing against conceptualizations of moral judgment as determined by absolutist norm-adherence. Recently,

this perspective has been fleshed out in the writings on the Theory of Dyadic Morality (TDM). Briefly, this framework proposes that perception of moral relevance is determined by the co-occurrence of 1) a perceived norm violation, 2) negative affect, and 3) the perception of harm. Harm, in turn, is supposed to be determined by comparison to a fuzzy cognitive template consisting of 1) a moral agent, acting upon 2) a moral patient (Schein & Gray, 2018). Thus, in contrast to what “objective consequentialist” accounts would assume, the Theory of Dyadic Morality stresses that the perception of harm emerges through a constructivist process. Similar to how constructivist accounts of emotion suggest that different emotions arise from combinations of core affect, itself consisting of valence and arousal, and conceptualization, dyadic morality embraces pluralistic conceptions of morality via combinations of its core components (Gray & Keeney, 2015a; Gray et al., 2017). Most relevant for the context of this discussion, moral judgment is thereby supposed to be subject to the processes of *dyadic comparison* (perceiving harm leads to the judgment of an action as immoral) and *dyadic completion* (deeming something immoral leads to the perception of some sort of harm). Thus, as proponents of this view have argued, moral judgment is thereby intimately tied to the perception of minds capable of suffering (Gray, Young, & Waytz, 2012; Gray, Waytz, & Young, 2012; Gray & Wegner, 2010; also see Harris, 2010).

As such, the Theory of Dyadic Morality (Schein & Gray, 2018) conceptually converges with Subjective Utilitarian Theory (Cohen & Ahn, 2016) and weak consequentialism (Hennig & Hütter, 2019; Sunstein, 2005; also see Harris, 2010) in that all frameworks assign the perception of consequences a central role in moral judgment. To some extent, the same can be said about the deontological coherence framework (Holyoak & Powell, 2016), which conceptualizes moral judgment as a process of weighing competing moral rules and duties against one another. Crucially, this framework points out likewise that moral rules are not accepted “for their own sake”, but rather that rules are understood from a perspective of *moderate deontology*, such that every rule may be broken when in conflict with

another rule (or set of rules) deemed more important. Similarly, all approaches reject a systematic one-to-one mapping of processing characteristic to observable judgment. Thus, Subjective Utilitarian Theory characterizes emotional processes as part of the overall cost-benefit analysis (Cohen & Ahn, 2016, p. 1362), while making no specific assumptions about their relative importance. Likewise, the Theory of Dyadic Morality stresses harm perception to be an intuitive process (Schein & Gray, 2018), however, without making strong claims about a general dominance of intuition over reasoning (Haidt, 2001). Deontological coherence, not yet a full-fledged theory as pointed out by the authors, remains tacit regarding this question (Holyoak & Powell, 2016). Finally, the weak consequentialist framework, likewise more a framework than a theory, deliberately avoids claims regarding systematic connections between cognitive systems and response patterns (Hennig & Hütter, 2019), as such claims do on balance not seem warranted by the evidence. As such, all of these perspectives do not endorse the assumptions of more traditional dual-process models (e.g. Haidt, 2001; Greene, 2014).

As the discussion of these models of moral judgment suggests, the field of moral psychology appears to have substantially shifted in its fundamental theoretical assumptions over the course of the last two decades, trending away from strict dual-process conceptualizations and in the direction of *process-agnosticism*. As such, the results of the current thesis contribute to integrating theoretical approaches to understanding dilemma judgment with a more extant literature on moral judgment more generally. In the spirit of process-agnosticism the presented work does neither argue for the “hardcore rationalism” of the cognitive revolution (e.g. Kohlberg, 1969), nor for fairly “radical intuitionism” (Haidt, 2001). However, it recognizes that the present results seem incompatible with several fundamental assumptions of dual-process models of dilemma judgment (Greene, 2014), and instead converge better with constructivist approaches to moral judgment that focus on the

importance of perceived harm (Cohen & Ahn., 2016; Schein & Gray, 2018; Gray et al., 2017; also see Harris, 2010).

5.4. Limitations and future directions

The current thesis applied multinomial modeling to the investigation of several aspects of research on moral dilemma judgment. In line with previous research employing the multinomial method (Białek et al., 2019; Brannon et al., 2019; Gawronski et al., 2016, 2017, 2018; Zhang et al., 2018), the present results demonstrate the value of this endeavor and, as I argued, may also provide some improvements over those previous multinomial approaches. However, despite all of this, there are some apparent limitations to the current work.

First, some methodological improvements may be applied, that are specific to the approach taken in the studies presented here. For instance, great care was taken to construct stimuli that are internally consistent and comparable in structure, resulting in a structured template for creating scenarios. However, this resulted in scenarios that are, compared to previously employed stimuli (Greene et al., 2001, 2004; Conway & Gawronski, 2013; Gawronski et al., 2017; Moore et al., 2008), long, and may be perceived as convoluted, overly repetitive or tiresome to read. Thus, future work may invest in reworking the template provided in Appendix B to yield a set of scenarios that is shorter, yet maintains its internal structure.

This may also help to address a second methodological limitation. Specifically, the modeling procedure described here did not estimate parameters individually per participant, but rather pooled observation across all participants. Therefore, in contrast to other approaches (e.g. Conway & Gawronski, 2013; Reynolds & Conway, 2018), the current method of analysis is not ideally equipped for investigating correlations between continuous predictors and model parameters. It is possible to investigate such relationships via median-split analyses (e.g. Brannon et al., 2019; Gawronski et al., 2017, Experiments 4a+4b; Zhang et al., 2018), or the application of a recursive partitioning procedure (Wickelmaier & Zeileis,

2018), as tentatively conducted in Chapter IV. A yet more promising approach would be the application of hierarchical Bayesian modeling, which allows control over participant heterogeneity as well as the estimation of correlations between continuous predictors and model parameters (e.g., Heck et al., 2018). In order to conduct these analyses, however, more datapoints per participants would be desirable which, in turn, requires the application of shorter scenarios to warrant high data quality and avoid participant fatigue and random responding.

In a next step, it may be useful to conduct a more systematic model validation, by correlating parameters of the proCNI model to other measures of moral reasoning to establish convergent validity. In this regard, the Consequentialist Thinking Style scale (Piazza & Landy, 2013; Piazza & Sousa, 2014) and the Oxford Utilitarianism Scale (Kahane et al., 2018) may provide useful orientation. Specifically, as moral dilemmas are designed to track the instrumental harm aspect of utilitarian philosophy, a positive relationship between this subscale of the Oxford Utilitarianism Scale and the *C*-parameter of the proCNI model could be expected.

Second, as I already discussed above, there are further points of criticism that apply to our work, because those are fairly generic and apply to dilemma (or even vignette) research as a whole. Specifically, the current work does not claim to be above general criticism regarding the artificial nature of sacrificial dilemmas and consequently a potential lack of external validity (Bauman et al., 2014; also see Gray & Keeney, 2015b). Closely related, it is not clear to what extent dilemma research can rely on the assumption that participants do accept and believe the fundamental premises of the described scenarios as the researchers intended (Shou & Song, 2017; also see Baron & Goodwin, 2019; Royzman, Kim,. & Leeman, 2015; Stanley et al., 2019). This, again, is a concern that becomes more pressing as employed scenarios become more counterintuitive, and applies to dilemma research as a whole. Although the development of a unitary dilemma structure was intended to minimize this concern and avoid

past missteps (e.g. McGuire et al., 2009), it is of course not clear to what extent this endeavor was ultimately successful.

Third, it seems important to acknowledge limitations of the weak consequentialist view in its current form as proposed in Chapter II. For instance, a critic of the framework may point out that weak consequentialism may be consistent with the findings of roughly two decades of dilemma research, but offers little predictive power beyond the effects it has demonstrated already. Although I want to point out that I do not see how the influence of death avoidability (Hennig & Hütter, 2019) may be parsimoniously explainable by other current models of dilemma judgment (e.g. Cushman, 2013; Greene, 2014), I would agree with the spirit of the criticism. That is, at least at the moment, I consider weak consequentialism to constitute a valuable *framework* for thinking about dilemma judgments, rather than a full-fledged *process theory* of dilemma judgment. Thus, the framework may benefit from a clearer definition of potential working mechanisms.

As Hennig and Hütter (2019) have proposed, the main difference between the consequences captured by the *C* and *N*-parameters may lie in their sensitivity to *causal proximity*. Specifically, the *C*-parameter does conceptually converge with previous models, in the sense that it is assumed to capture sensitivity to *overall* consequences relatively unbothered by causal proximity. The *N*-parameter, in contrast, may primarily capture sensitivity to consequences that are *causally proximal*, in the sense that they constitute a direct and relatively unmediated result of the behavioral decision. Put differently, part of the reason for deciding to not sacrifice the single victim may be the concrete consideration of the likelihood of the consequences of a decision. For instance, a participant pondering the footbridge problem may conclude that pushing a fat man in front of a trolley represents an action with a *proximal* consequences that will occur with high certainty due to a simple causal chain (i.e. push --> death of individual). In contrast, they may reasonably worry about whether the *distal* consequences of saving the five workers may be achieved, due to a more

complex causal chain (i.e. push --> stop trolley -- > no death of group), allowing for more things to go wrong in the process. For instance, whereas pushing the man in front of the trolley is likely to kill him, this may actually fail to stop the trolley. Thus, it is conceivable that the achievement of desirable distal consequences (i.e. saving the group) is generally assigned lower subjective outcome probability than the achievement of desirable proximal consequences (i.e. saving the single person) due to a longer and more complex causal chain (see Shou & Song, 2017; also see Stanley et al., 2019).

Under the assumption that the *N*-parameter is indeed sensitive to causal proximity, i.e. represents a preference for short and predictable causal chains and certain rather than more speculative outcomes, one may expect the priming of causality to increase the *N*-parameter while leaving the *C*-parameter unaffected, relative to a control condition. A possible way to test this may be to prime causal reasoning via simple conditional reasoning statements (i.e. Klauer, Beller, & Hütter, 2010), which contain clear indicators of causality (“If John decides to party tonight, he will likely do poorly on tomorrows exam.”), and letting participants pick the correct consequence out of a list of several options (i.e. “John gets a good grade”, “John gets an average grade”, “John gets a bad grade” etc.). Furthermore, note that this approach may offer an additional opportunity for testing predictions of weak consequentialism and the Dual-Process Model against one another. That is, from a Dual-Process Model view it could easily be argued that priming a causal reasoning mindset should induce reflective, rational, System 2 type processing, such that “utilitarian” tendencies should be increased and an effect on the *C*-parameter should be expected instead (Greene, 2014; Greene et al., 2009; Suter & Hertwig, 2011). Thus, this approach may be helpful for further fleshing out weak consequentialism and defining its potential psychological mechanism more clearly than could be done in previous work (Hennig & Hütter, 2019).

Whether these predictions will ultimately be supported, naturally, is an empirical question. However, I think that even in the absence of additional data a reasonable case can be

made that weak consequentialism may be favored over the Dual-Process Model for reasons of theoretical parsimony alone. That is, the Dual-Process Model relies on a number of assumptions, including that observable response patterns should be interpreted as related to broad philosophical positions with conceptual baggage hard to properly address in the context of sacrificial dilemmas, that those two response patterns are in turn systematically produced by two different processing systems, and that these systems operate independent of one another. As I have argued in the present thesis, the evidence in favor of each of those claims appears rather thin, at best. In contrast, weak consequentialism requires only the assumption that actual or anticipated consequences constitute the driving force underlying moral dilemma judgment. This claim is, as I have argued, empirically sound as well as more parsimonious than the multiple foundational assumptions of the Dual-Process Model.

Finally, there is the more fundamental question regarding the purpose of dilemma judgment research for moral psychology more generally. That is, the dilemma approach has arguably constituted the most influential single paradigm in the field of moral psychology since its inception about two decades ago. Likewise, the theoretical models based on this approach attempt to draw broad conclusions about moral judgment in general, which surpass the narrow boundaries of the paradigm (e.g. Greene, 2014; Greene & Haidt, 2002). Yet, at the same time some of its most influential findings (e.g. Conway & Gawronski, 2013; Greene et al., 2008; Suter & Hertwig, 2011) appear to be less robust than commonly assumed upon closer inspection (e.g. Baron, et al., 2015; Tinghög et al., 2016; Gawronski et al., 2016, 2017), frequently used stimuli suffer from systematic confounds (e.g. Hennig & Hütter, 2019; McGuire et al., 2009; Rosas & Koenigs, 2014), responses and response patterns may not be clearly divisible into sensitivity to norms and consequences regardless of the other, and their precise conceptual meaning certainly seems up to debate (Cohen & Ahn, 2016; Hennig & Hütter, 2019; Kahane, 2012, 2015), and the artificial nature of sacrificial dilemmas poses boundaries on believability and external validity (Baumann et al., 2014; Shou & Song, 2017).

At the same time, alternative approaches to studying moral judgment exist that take different perspectives, stressing the importance of character evaluations (Uhlmann et al., 2015) the constructivist nature of harm perception (Gray et al., 2017; Schein & Gray, 2018), the process of moralization (Feinberg, Kovacheff, Teper, & Inbar, 2019; Skitka, Wisneski, & Brandt, 2018; Wisneski & Skitka, 2017), the role of moral conviction in social judgment and behavioral intentions (Luttrell, Petty, Briñol, & Wagner, 2016; Mueller & Skitka, 2018; Mullen & Skitka, 2006; Skitka et al., 2005), the importance of meta-ethical beliefs in moral judgment (Piazza & Landy, 2013; Rai & Holyoak, 2013; Zijlstra, 2019), and investigation of actual in vivo moral judgment (Hofmann et al., 2014, 2018). In my view, it should be considered whether many of the questions addressed by these bodies of work may be more theoretically interesting and practically relevant for understanding actual moral judgment than the consideration of sacrificial dilemmas.

Naturally, this is not to say that dilemma research should deserve no place in the methodological toolkit of modern moral psychology. However, it is to say that, though this thesis focused on dilemma research, I explicitly recognize that a healthy and productive science of moral judgment surely benefits from applying a plurality of empirical approaches, while avoiding rigid reliance on individual paradigms.

References

Aktas, B., Yilmaz, O., & Bahçekapili, H. G. (2017). Moral pluralism on the trolley tracks – Different normative principles are used for different reasons in justifying moral judgments. *Judgment and Decision Making, 12,* 297-303.

Armstrong, J., Friesdorf, R., & Conway, P. (2019). Clarifying gender differences in moral dilemma judgments: the complementary roles of harm aversion and action aversion. *Social Psychological and Personality Science, 10,* 353-363.

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review, 80,* 1095-1111.

Axelrod, R., & Dion, D. (1988). The further evolution of cooperation. *Science, 242,* 1385-1390.

Balash, J., & Falkenbach, D. M. (2018). The ends justify the meanness: An investigation of psychopathic traits and utilitarian moral endorsement. *Personality and Individual Differences, 127,* 127-132.

Baron, J. (1994). Nonconsequentialist decisions. *Behavioral and Brain Sciences, 17,* 1-10.

Baron, J., & Goodwin, G. P. (2019). *Consequences, norms, and inaction: A comment.* Manuscript submitted for publication. Retrieved from https://www.sas.upenn.edu/~baron/

Baron, J., Gürçay, B., Moore, A. B., & Starcke, K. (2012). Use of a Rasch model to predict response times to utilitarian moral dilemmas. *Synthese, 189,* 107-117.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*, 74-85.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, *4*, 265-284.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121*, 154–161.

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass, 8*, 536–554.

Bernhard, R. M., Chaponis, J., Siburian, R., Gallagher, P., Ransohoff, K., Wikler, D., Greene, J. D. (2016). Variation in the oxytocin receptor gene (OXTR) is associated with differences in moral judgment. *Social Cognitive and Affective Neuroscience, 11*, 1872–1881.

Betsch, C. (2004). Präferenz für intuition und deliberation (PID). *Zeitschrift für Differentielle und Diagnostische Psychologie, 25*, 179-197.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*, 1573-1576.

Byrd, N., & Conway, P. (in press). Not all who ponder count costs: Arithmetic reflection predicts utilitarian tendencies, but logical reflection predicts both deontoloical and utilitarian tendencies. *Cognition*.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*, 116-131.

Cecchetto, C., Rumiati, R. I., & Parma, V. (2017). Relative contribution of odour intensity and valence to moral decisions. *Perception, 46*, 447–474.

Chan, Y. L., Gu, X., Ng, J. C. K., & Tse, C. S. (2016). Effects of dilemma type, language, and emotion arousal on utilitarian vs deontological choice to moral dilemmas in Chinese–English bilinguals. *Asian Journal of Social Psychology, 19*, 55-65.

Christensen, J. F., Flexas, A., Calabrese, M., Gut, N. K., & Gomila, A. (2014). Moral judgment reloaded: A moral dilemma validation study. *Frontiers in Psychology*, *5*, 1–18.

Christov-Moore, L., Conway, P., & Iacoboni, M. (2017). Deontological dilemma response tendencies and sensorimotor representations of harm to others. *Frontiers in Integrative Neuroscience, 11*, 34.

Cipolletti, H., McFarlane, S., & Weissglass, C. (2016). The moral foreign-language effect. *Philosophical Psychology*, *29*, 23-40.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General*, *145*, 1359–1381.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology, 104*, 216–235.

Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition, 179*, 241-265.

Corey, J. D., Hayakawa, S., Foucart, A., Aparici, M., Botella, J., Costa, A., & Keysar, B. (2017). Our moral choices are foreign to us. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1109-1128.

Costa, A., Foucart, A., Arnon, I., Aparici, M., & Apesteguia, J. (2014). “Piensa” twice: On the foreign language effect in decision making. *Cognition, 130*, 236-254.

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keysar, B. (2014). Your morals depend on language. *PloS ONE, 9*, e94842.

Costa, A., Vives, M. L., & Corey, J. D. (2017). On language processing shaping decision making. *Current Directions in Psychological Science*, *26*, 146-151.

Crone, D. L., & Laham, S. M. (2017). Utilitarian preferences or action preferences? De-confounding action and moral code in sacrificial dilemmas. *Personality and Individual Differences, 104*, 476–481.

Crone, D. L., & Laham, S. M. (2015). Multiple moral foundations predict responses to sacrificial dilemmas. *Personality and Individual Differences, 85,* 60-65.

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

Cushman, F., Young, L., & Greene, J. D. (2010). Our multi-system moral psychology: Towards a consensus view. *The Oxford handbook of moral psychology*, 47-71.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science, 17,* 1082-1089.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*, 113-126.

Ditto, P. H., & Jemmott, J. B. (1989). From rarity to evaluative extremity: Effects of prevalence information on evaluations of positive and negative characteristics. *Journal of Personality and Social Psychology, 57,* 16-26.

Duke, A., & Bègue, L. (2015). The drunk utilitarian: Blood alcohol concentration predicts utilitarian responses in moral dilemmas. *Cognition, 134*, 121–127.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255-278.

Everett, J. A., Caviola, L., Kahane, G., Savulescu, J., & Faber, N. S. (2015). Doing good by doing nothing? The role of social norms in explaining default effects in altruistic contexts. *European Journal of Social Psychology, 45*, 230-241.

Everett, J. A., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology, 79,* 200-216.

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*, 772–787.

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

Feinberg, M., Kovacheff, C., Teper, R., & Inbar, Y. (2019, March 14). Understanding the process of moralization: How eating meat becomes a moral issue. *Journal of Personality and Social Psychology*. Advance online publication. http://dx.doi.org/10.1037/pspa0000149

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology, 47,* 1231-1236.

Fleischmann, A., Lammers, J., Conway, P., & Galinsky, A. D. (2019). Paradoxical effects of power on moral thinking: why power both increases and decreases deontological and utilitarian moral decisions. *Social Psychological and Personality Science, 10*, 110-120.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5–15.

Friesdorf, R., Conway, P., & Gawronski, B. (2015). Gender differences in responses to moral dilemmas: A process dissociation analysis. *Personality and Social Psychology Bulletin, 41,* 696-713.

Gamez-Djokic, M., & Molden, D. (2016). Beyond affective influences on deontological moral judgment: The role of motivations for prevention in the moral condemnation of harm. *Personality and Social Psychology Bulletin*, *42,* 1522-1537.

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology, 113,* 343–376.

Gawronski, B., & Beer, J. S. (2016). What makes moral dilemma judgments “utilitarian” or “deontological”? *Social Neuroscience, 12,* 626-632.

Gawronski, B., Conway, P., Armstrong, J., Friesdorf, R., & Hütter, M. (2016). Understanding responses to moral dilemmas: Deontological inclinations, utilitarian inclinations, and general action tendencies. In J. P. Forgas, L. Jussim, & P. A. M. Van Lange (Eds.), *Social Psychology of Morality* (pp. 91–110). New York, NY: Psychology Press.

Gawronski, B., Conway, P., Armstrong, J., Friesdorf, R., & Hütter, M. (2018). Effects of incidental emotions on moral dilemma judgments: An analysis using the CNI model. *Emotion, 18,* 989-1008.

Geipel, J., Hadjichristidis, C., & Surian, L. (2015a). How foreign language shapes moral judgment. *Journal of Experimental Social Psychology, 59,* 8-17.

Geipel, J., Hadjichristidis, C., & Surian, L. (2015b). The foreign language effect on moral judgment: The role of emotions and norms. *PloS ONE, 10,* e0131529.

Geipel, J., Hadjichristidis, C., & Surian, L. (2016). Foreign language affects the contribution of intentions and outcomes to moral judgment. *Cognition, 154,* 34-39.

Gigerenzer, G. (2004). Fast and frugal heuristics: The tools of bounded rationality. In D. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making,* (pp. 62–88). Oxford, UK: Blackwell.

Gleichgericht, E., & Young, L. (2013). Low Levels of Empathic Concern Predict Utilitarian Moral Judgment. *PLoS ONE, 8,*e60418.

Gray, K., & Keeney, J. E. (2015). Disconfirming moral foundations theory on its own terms: Reply to Graham (2015). *Social Psychological and Personality Science, 6,* 874-877.

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,* 859-868.

Gray, K., & Schein, C. (2012). Two minds vs. two philosophies: Mind perception defines morality and dissolves the debate between deontology and utilitarianism. *Review of Philosophy and Psychology, 3,* 405-423.

Gray, K., & Schein, C. (2016). No absolutism here: Harm predicts moral judgment 30x better than disgust – Commentary on Scott, Inbar, & Rozin (2016). *Perspectives on Psychological Science, 11*, 325–329.

Gray, K., Schein, C., & Cameron, C. D. (2017). How to think about emotion and morality: circles, not arrows. *Current Opinion in Psychology, 17,* 41-46.

Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General, 143,* 1600-1615.

Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry, 23,* 206-215.

Gray, K., & Wegner, D. M. (2010). Blaming God for our pain: Human suffering and the divine mind. *Personality and Social Psychology Review, 14,* 7-16.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23,* 101-124.

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences, 11,* 322–323.

Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. *Journal of Experimental Social Psychology, 45*, 581–584.

Greene, J. D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics, 124,* 695-726.

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,* 364-371.

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

Guglielmo, S. (2018). Unfounded dumbfounding: How harm and purity undermine evidence for moral dumbfounding. *Cognition, 170,* 334-337.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review, 108,* 814-834.

Haidt, J. (2004). The emotional dog gets mistaken for a possum. *Psychological Review, 8,* 283-290.

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

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion.* New York, NY, US: Pantheon/Random House.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. *Unpublished manuscript, University of Virginia*.

Haidt, J., & Hersh, M. A. (2001). Sexual Morality: The Cultures and Emotions of Conservatives and Liberals. *Journal of Applied Social Psychology, 31,* 191-221.

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,* 613-628.

Harris, S. (2010). *The moral landscape: How science can determine human values.* Free Press.

Hayakawa, S., Costa, A., Foucart, A., & Keysar, B. (2016). Using a foreign language changes our choices. *Trends in Cognitive Sciences, 20,* 791-793.

Hayakawa, S., & Keysar, B. (2018). Using a foreign language reduces mental imagery. *Cognition, 173,* 8-15.

Hayakawa, S., Tannenbaum, D., Costa, A., Corey, J. D., & Keysar, B. (2017). Thinking more or feeling less? Explaining the foreign-language effect on moral judgment. *Psychological Science, 28,* 1387-1397.

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods, 50,* 264-284.

Hennig, M., & Hütter, M. (2019, September 2). Revisiting the divide between deontology and utilitarianism in moral dilemma judgment: A multinomial modeling approach. *Journal of Personalityand Social Psychology.* Advance online publication. http://dx.doi.org/10.1037/pspa0000173

Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral punishment in everyday life. *Personality and Social Psychology Bulletin, 44,* 1697-1711.

Hofmann, W., Wisneski, D. C., Brandt, M. J., Skitka, L. J. (2014). Morality in everyday life. *Science, 345,* 1340-1343.

Holyoak, K. J., & Powell, D. (2016). Deontological coherence: A framework for commonsense moral reasoning. *Psychological Bulletin, 142,* 1179-1203.

Horne, Z., & Powell, D. (2013). More than a feeling: When emotional reactions don't predict moral judgments. *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Horne, Z., & Powell, D. (2016). How large is the role of emotion in judgments of moral dilemmas?. *PLoS ONE, 11*, e0154780.

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika, 59*, 21-47.

Hughes, J. S. (2017). In a moral dilemma, choose the one you love: Impartial actors are seen as less moral than partial ones. *British Journal of Social Psychology, 56*, 561-577.

Hütter, M., & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology, 27*, 116–159.

Inbar, Y., Pizarro, D. A., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion, 9*, 435.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513-541.

Jackson, J. C., & Gray, K. (2019, February 7). When a good god makes bad people: Testing a theory of religion and immorality. *Journal of Personality and Social Psychology.* Advance online publication. http://dx.doi.org/10.1037/pspp0000206

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology, 96,* 521-537.

Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the Fundamentals of Financial Decision Making: Part I* (pp. 99-127).

Kamm, F. (2009). Neuroscience and moral reasoning: A note on recent research. *Philosophy & Public Affairs, 37,* 330-345.

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*, *10*, 551–560.

Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). “Utilitarian” judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, *134*, 193–209.

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology, *Psychological Review*, *125*, 164-164.

Kahane, G., & Shackel, N. (2010). Methodological issues in the scientific study of moral judgement, *Mind and Language*, *25*, 1–20.

Karandikar, S., Kapoor, H., Fernandes, S., & Jonason, P. K. (2019). Predicting moral decision-making with dark personalities and moral values. *Personality and Individual Differences, 140,* 70-75.

Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychological Science*, *23*, 661-668.

Klauer, K. C. (2015). Mathematical modeling. In B. Gawronski & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 371-389). New York, NY: Guilford Press.

Klauer, K. C., Beller, S., & Hütter, M. (2010). Conditional reasoning in context: A dual-source model of probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 298-323.

Klauer, K. C., Dittrich, K., Scholtes, C., & Voss, A. (2015). The invariance assumption in process-dissociation models: An evaluation across three domains. *Journal of Experimental Psychology: General*, *144*, 198-221.

Klauer, K. C., Stahl, C., & Voss, A. (2012). Multinomial models and diffusion models. In K.

C. Klauer, C. Stahl, & A. Voss (Eds.), *Cognitive Methods in Social Psychology*, (pp. 367-390). New York: Guilford Press.

Klauer, K. C., & Wegener, I. (1998). Unraveling social categorization in the “who said what?” paradigm. *Journal of Personality and Social Psychology, 75,* 1155-1178.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446,* 908-911.

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience, 7*, 708-714.

Koop, G. J. (2013). An assessment of the temporal dynamics of moral decisions. *Judgment and Decision Making, 8,* 527-539.

Kusev, P., Schaik, P., Alzahrani, S., Lonigro, S., & Purser, H. (2016). Judging the morality of utilitarian actions: How poor utilitarian accessibility makes judges irrational. *Psychonomic Bulletin & Review, 23,* 1–8.

Laakasuo, M., Sundvall, J., & Drosinou, M. (2017). Individual differences in moral disgust do not predict utilitarian judgments, sexual and pathogen disgust do. *Scientific Reports, 7*, 1–10.

Lane, D., & Sulikowski, D. (2017). Bleeding-heart conservatives and hard-headed liberals: The dual processes of moral judgements. *Personality and Individual Differences, 115,* 30-34.

Lawrence, MA. (2011). ez: Easy analysis and visualization of factorial experiments. R package version 3.0-0.

Lee, M., Sul, S., & Kim, H. (2018). Social observation increases deontological judgments in moral dilemmas. *Evolution and Human Behavior, 39,* 611-621.

Leiner, D. J. (2014). SoSci Survey (Version 2.6.00-i) [Computer software]. Available at https://www.soscisurvey.de

Lotto, L., Manfrinati, A., & Sarlo, M. (2014). A new set of moral dilemmas – norms for moral acceptability, decision times, and emotional salience. *The Journal of Behavioral Decision Making, 27,* 57–65.

Luttrell, A., Petty, R. E., Briñol, P., & Wagner, B. C. (2016). Making it moral: Merely labeling an attitude as moral increases its strength. *Journal of Experimental Social Psychology, 65*, 82-93.

McDonald, M. M., Defever, A. M., & Navarrete, C. D. (2017). Killing for the greater good: Action aversion and the emotional inhibition of harm in moral dilemmas. *Evolution and Human Behavior, 38,* 770-778.

McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology*, *45*, 577–580.

Miller, R., & Cushman, F. (2013). Aversive for me, wrong for you: First-person behavioral aversions underlie the moral condemnation of harm. *Social and Personality Psychology Compass, 7,* 707-718.

Miller, R., Hannikainen, I. A., & Cushman, F. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion, 14,* 573-587.

Montero-Melis, G., Isaksson, P., van Paridon, J., & Ostarek, M. (2019, September 19). Does using a foreign language reduce mental imagery? https://doi.org/10.31234/osf.io/5r28e

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill ? *Psychological Science*, *19,* 549–557.

Moore, A. B., Lee, N. L., Clark, B. A., & Conway, A. R. (2011). In defense of the personal/impersonal distinction in moral psychology research: Cross-cultural validation of the dual process model of moral judgment. *Judgment and Decision Making, 6,* 186-195.

Moretto, G., Làdavas, E., Mattioli, F., Pellegrino, G., & di Pellegrino, G. (2010). A psychophysiological investigation of moral judgment after ventromedial prefrontal damage. *Journal of Cognitive Neuroscience, 22*, 1888–1899.

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42,* 42–54.

Muda, R., Niszczota, P., Białek, M., & Conway, P. (2018). Reading dilemmas in a foreign language reduces both deontological and utilitarian response tendencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*, 321-326.

Mueller, A. B., & Skitka, L. J. (2018). Liars, damned liars, and zealots: The effect of moral mandates on transgressive advocacy acceptance. *Social Psychological and Personality Science, 9*, 711-718.

Mullen, E., & Skitka, L. J. (2006). Exploring the psychological underpinnings of the moral mandate effect: Motivated reasoning, group differentiation, or anger? *Journal of Personality and Social Psychology, 90*, 629.

Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review*, *18*, 107–118.

Nakamura, K. (2013). A closer look at moral dilemmas: Latent dimensions of morality and the difference between trolley and footbridge dilemmas. *Thinking & Reasoning, 19,* 178-204.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 1–8.

Patil., I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier,F., Calò, M., Silani, G., Cikara, M., & Cushman, F. (2019). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. Retrieved from https://psyarxiv.com/q86vx/download?format=pdf

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

Pew Research Center. (2018). The age gap in religion around the world.

Piazza, J. (2012). “If you love me keep my commandments”: religiosity increases preference for rule-based moral arguments. *International Journal for the Psychology of Religion, 22,* 285–302.

Piazza, J., & Landy, J. F. (2013). “Lean not on your own understanding”: Belief that morality is founded on divine authority and non-utilitarian moral judgments. *Judgment and Decision Making, 8,* 639–661.

Piazza, J., & Sousa, P. (2014). Religiosity, Political Orientation, and Consequentialist Moral Thinking. *Social Psychological and Personality Science, 5,* 334–342.

Piazza, J., Sousa, P., & Holbrook, C. (2013). Authority dependence and judgments of utilitarian harm. *Cognition, 128*, 261-270.

Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: A comment on Haidt (2001). *Psychological Review, 110,* 193-196.

R Core Team (2018). R: A language and environment for statistical computing. A Foundation for Statistical Computing, Vienna, Austria.

Rai, T. S., & Holyoak, K. J. (2013). Exposure to moral relativism compromises moral behavior. *Journal of Experimental Social Psychology, 49,* 995-1001.

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review, 9,* 278-291.

Reynolds, C. J., & Conway, P. (2018). Not just bad actions: Affective concern for bad outcomes contributes to moral condemnation of harm in moral dilemmas. *Emotion, 18,* 1009-1023.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95,* 318–339.

Robinson, J. S., Joel, S., & Plaks, J. E. (2015). Empathy for the group versus indifference toward the victim: Effects of anxious and avoidant attachment on moral judgment. *Journal of Experimental Social Psychology*, *56,* 139-152.

Robinson, J. S., Page-Gould, E., & Plaks, J. E. (2017). I appreciate your effort: Asymmetric effects of actors' exertion on observers' consequentialist versus deontological judgments. *Journal of Experimental Social Psychology*, *73,* 50-64.

Rom, S. C., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology*, *74,* 24-37.

Rom, S. C., Weiss, A., & Conway, P. (2017). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology*, *69,* 44-58.

Rosas, A. (2017). On the cognitive (neuro)science of moral cognition: Utilitarianism, deontology, and the “fragmentation of value”. In A. Ibáñez, L. Sedeño, & A. M. García (Eds.) *Neuroscience and Social Science* (pp. 199-215). Springer, Cham.

Rosas, A., & Koenigs, M. (2014). Beyond “utilitarianism”: Maximizing the clinical impact of moral judgment research. *Social Neuroscience, 9,* 661–667.

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,* 296-313.

Salzstein, H. D., & Kasachkoff, T. (2004). Haidt’s moral intuitionist theory – A psychological and philosophical critique. *Psychological Review, 8,* 273-282.

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1,* 7-59.

Sarlo, M., Lotto, L., Manfrinati, A., Rumiati, R., Gallicchio, G., & Palomba, D. (2012). Temporal Dynamics of Cognitive–Emotional Interplay in Moral Decision-making. *Journal of Cognitive Neuroscience, 24,* 1018–1029.

Sarlo, M., Lotto, L., Rumiati, R., & Palomba, D. (2014). If it makes you feel bad, don’t do it! Egoistic rather than altruistic empathy modulates neural and behavioral responses in moral dilemmas. *Physiology and Behavior, 130,* 127-134.

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

Schein, C., Ritter, R. S., & Gray, K. (2016). Harm mediates the disgust-immorality link. *Emotion, 16,* 862-876.

Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological Science, 19,* 1219-1222.

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgments in professional philosophers and non-philosophers. *Mind & Language, 27,* 135-153.

Scott, S. E., Inbar, Y., & Rozin, P. (2016). Evidence for absolute moral opposition to genetically modified food in the United States. *Perspectives on Psychological Science, 11,* 315-324.

Shaver, R. (2019). Egoism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Spring 2019 Edition).* Metaphysics Research Lab, Stanford University.

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. *Journal of Neuroscience, 34,* 4741-4749.

Shin, H., & Kim, J. (2017). Foreign Language Effect and Psychological Distance. *Journal of Psycholinguistic Research, 46,* 1339-1352.

Shou, Y., & Song, F. (2017). Decisions in moral dilemmas: The influence of subjective beliefs in outcome probabilities. *Judgment & Decision Making, 12,* 481-490.

Simpson, A., Piazza, J., & Rios, K. (2016). Belief in divine moral authority: Validation of a shortened scale with implications for social attitudes and moral cognition. *Personality and Individual Differences, 94,* 256–265.

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.

Skitka, L. J., Wisneski, D. C., & Brandt, M. J. (2018). Attitude moralization: Probably not intuitive or rooted in perceptions of harm. *Current Directions in Psychological Science, 27,* 9-13.

Sloman, S. (2014). Two systems of reasoning: An update. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), Dual-Process Theories of the Social Mind (pp. 69-79). New York, NY, US: Guilford Press.

Snyder, M. L., Kleck, R. E., Strenta, A., & Mentzer, S. J. (1979). Avoidance of the handicapped: an attributional ambiguity analysis. *Journal of Personality and Social Psychology, 37,* 2297-2306.

Stanley, M. L., Yin, S., & Sinnott-Armstrong, W. (2019). A reason-based explanation for moral dumbfounding. *Judgment and Decision Making, 14,* 120-129.

Suessenbach, F., & Moore, A. B. (2015). Individual differences in the explicit power motive predict “utilitarian” choices in moral dilemmas, especially when this choice is self-beneficial. *Personality and Individual Differences, 86,* 297–302.

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

Szekely, R. D., Opre, A., & Miu, A. C. (2015). Religiosity enhances emotion and deontological choice in moral dilemmas. *Personality and Individual Differences, 79,* 104–109.

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology, 47,* 1249-1254.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist, 59*, 204–217.

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal, 6*, 1395–1415.

Tinghög, G., Andersson, D., Bonn, C., Johannesson, M., Kirchler, M., Koppel, L., & Västfjäll, D. (2016). Intuition and moral decision-making – The effect of time pressure and cognitive load on moral judgment and altruistic behavior. *PLoS ONE, 11*, e0164012.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*, 72-81.

Uhlmann, E. L., & Zhu, L. (2014). Acts, persons, and intuitions: Person-centered cues and gut reactions to harmless transgressions. *Social Psychological and Personality Science, 5,* 279-285.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition, 126,* 326-334.

Unkelbach, C., & Stahl, C. (2009). A multinomial modelling approach to dissociate different components of the truth effect. *Consciousness and Cognition, 18,* 22-38.

van den Bos, K., Müller, P. A., & Damen, T. (2011). A behavioral disinhibition hypothesis of interventions in moral dilemmas. *Emotion Review*, *3*, 281-283.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S. (4th Ed.)*. New York, NY: Springer.

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb. *Psychological Science, 18,* 247-253.

Wickelmaier, F., & Zeileis, A. (2018). Using recursive partitioning to account for parameter heterogeneity in multinomial processing tree models. *Behavior Research Methods*, *50*, 1217-1233.

Wisneski, D. C., & Skitka, L. J. (2017). Moralization through moral shock: Exploring emotional antecedents to moral conviction. *Personality and Social Psychology Bulletin, 43,* 139-150.

Youssef, F. F., Dookeeram, K., Basdeo, V., Francis, E., Doman, M., Mamed, D., Maloo, S., Degannes, J., Dobo, L., Ditshotlo, P., & Legall, G. (2012). Stress alters personal moral decision making. *Psychoneuroendocrinology, 37,* 491-498.

Zijlstra, L. (2019). Folk moral objectivism and its measurement. *Journal of Experimental Social Psychology, 84,* 103807.

Appendix A (for Hennig & Hütter, 2019 – Chapter II)

Scenarios Implemented in the Present Experiments

Scenarios used in Experiments 1, 2a, and 2b

The scenarios presented in Tables A1 to A4 were adapted from Conway and Gawronski (2013).

Note that not all of the scenario versions were used in each and every experiment. Specifically, in Experiment 1, the scenarios contained no self-relevant consequences. In Experiment 2a, scenarios in the self-relevance present condition incentivized norm-breaking in the congruent condition, while they incentivized norm-adherence in the incongruent condition. Hence, self-relevant consequences consistently opposed the endorsement of aggregate consequences. In Experiment 2b, scenarios in the self-relevance present condition always incentivized norm-breaking. Hence, self-relevant consequences consistently opposed endorsement of consequences in Experiment 2a and norm-endorsement in Experiment 2b.

Table A1.
Vaccine Policy

	Incongruent	**Congruent**
	You are a doctor in a health clinic overrun by patients with an aggressive viral infection. The symptoms are painful and very inconvenient for the affected patients …	
	and if untreated, the disease leads to the patients' death in most cases.	but the virus does not cause permanent damage, and the disease is cured completely over the course of a few weeks.
	Because it is a new virus, which has not yet been intensively studied, no safe and reliable medications have been developed yet. However, you have just received a shipment of drugs that are still in the stage of testing, which can cure the viral infection but show severe side-effects. If you would administer the drugs to the patients, a small number would die from the side-effects, but the majority would live. If you would not administer the drugs to the patients, …	
	the majority would die from the disease.	they would continue to suffer the painful symptoms of the disease, for the time being.

Self-relevant reason to regard aggregate consequences / to break norm	Until now you have always carried out your work in congruence with the policies and regulations of the hospital, which is why you are on the brink of a promotion. The head physician, who makes this decision, has always supported the use of unorthodox methods, if those would spare the patients from severe suffering. You know that your promotion depends on whether you stick to the treatment policies in this case as well, and do administer the untested drugs.
Self-relevant reason to disregard aggregate consequences / to adhere to norm	Until now you have always carried out your work in congruence with the policies and regulations of the hospital, which is why you are on the brink of a promotion. The head physician, who makes this decision, has strictly forbidden the use of untested drugs due to their potential danger. You know that your promotion depends on whether you stick to the treatment policies in this case as well, and do not administer the untested drugs.
Inaction-Default	At the moment you are considering to administer the drug to the patients, and to prepare the necessary syringes. How would you behave in this situation? Would you begin the preparation and administer the drug to the patients?
Action-Default	You have already decided to administer the drug to the patients and are currently in the process of preparing the necessary syringes. How would you behave in this situation? Would you abort the preparation and deprive the patients of the drug?

Table A2.
Hard Times

	Incongruent	**Congruent**
	You are a single mother of five children in a poor country. Since the recent death of your husband you do not know how to …	
	feed your children, because you barely make any money and there is no opportunity for your children to work.	enable your children a comfortable life, because luxury goods are very expensive and you barely make enough money, to feed your family.
	A man in a suit comes to visit you and offers you the opportunity to sell one of your children. It would then by given up for adoption and grow up in a rich society. The child that you are talking about is your 11-year old daughter, a shy and timid girl that always felt very uncomfortable around strangers. You doubt that she would get along in an unfamiliar country and are afraid that she is not up to the challenge. Irrespective of what the people in your region think of his offer, the man is known far and wide as an honest businessman, who stands by his word. If you sold him one of your children, he would pay you a sum of money that is sufficient to secure …	
	the survival of the other four children.	a comfortable life for the other four children for a foreseeable future.
	However, you could never make contact with your sold child ever again.	

Self-relevant reason to regard aggregate consequences / to break norm	For the people in your village it is very important to be willing to take risks and use the few opportunities that offer themselves. If you would demonstrate that you are willing to offer your child the chance to a life outside of the poor village, the other villagers would show you considerably more respect because of it. This would considerably increase your personal reputation, which would have great medium-term influence on your social status and your daily life.
Self-relevant reason to disregard aggregate consequences / to adhere to norm	For the people in your village it is very important to stick together as a family and to get through all problems, no matter how threatening, together. If you would demonstrate that your family belongs and sticks together, the other villagers would show you considerably more respect because of it. This would not help you to feed your children, but it would considerably increase your personal reputation, which would have great medium-term influence on your social status and your daily life.
Inaction-Default	You are now considering to accept the deal and to sell the man one of your children. How would you behave in this situation? Would you accept the deal and sell one of your children to the man?
Action-Default	You have already signaled to the man that you are going to accept the deal. He has already sent you the money and is now awaiting you to hold your part of the agreement. How would you behave in this situation? Would you revoke your agreement with the man and return his money to him?

Table A3.
Border Crossing

	Incongruent	**Congruent**
	You are a soldier guarding a border checkpoint between your nation and one troubled by insurgent violence. You notice a young man in a cheap car approaching the checkpoint with a determined look on his face. You recognize the man from a mug shot to be somebody who already …	
	executed multiple violent assaults on checkpoints at border crossings, and wounded or even killed soldiers in the process.	attempted multiple times to cross the border in order to work illegally in your country.
	He is approaching so rapidly that he would certainly …	
	reach the checkpoint in no time, …	break through the border crossing, …
	unless you act quickly. If you shot the man you would have to accept that he is injured or even killed. However, making use of your firearm is the only possibility to stop him from approaching further.	
Self-relevant reason to	The government of your country assigns the protection of the border a very high priority and you, as a soldier, are instructed to show no tolerance in case of border violations. You know that, if you stick to this instruction, you would be judged to be a	

regard aggregate consequences / to break norm	capable security expert. Thereby you would have very good chances to get a promotion, which would entail more pay as well as more convenient work.
Self-relevant reason to disregard aggregate consequences / to adhere to norm	The government of your country assigns the maintenance of the assumption of innocence a very high priority and you, as a soldier, are instructed to show potentially deadly force only if you or others are threatened by force of arms. You know that, if you stick to this instruction, you would be judged to be a capable security expert. Thereby you would have very good chances to get a promotion, which would entail more pay as well as more convenient work.
Inaction-Default	Now you are trying to decide as quickly as possible, whether you want to level your rifle and shoot the man. How would you behave in this situation? Would you level your rifle and shoot the man?
Action-Default	You have already leveled your rifle in order to be able to shoot the man. How would you behave in this situation? Would you take your rifle down and let the man approach?

Table A4.
Torture

	Incongruent	**Congruent**
	You are a police officer, and have recently caught a criminal you have been hunting for some time. Circumstantial evidence clearly indicates that he has rigged a series of explosive devices all over town: Some that have already gone off and some that have yet to detonate.	
	He places the explosives outside of city cafes and sets them to go off at a time when people are drinking coffee on the patios. In this manner, he has already injured or killed many people.	He places paint cans that contain explosives outside of city cafes and sets them to go off at a time when no one is around. In this manner, he has already sprayed many cafes with paint.
	Now that the criminal is in custody, you want to know where the unexploded bombs are so you can defuse them. Although you have applied all legal interrogation techniques, he still refuses to talk. If you would not make the criminal talk, further explosives would detonate.	
Self-relevant reason to regard aggregate consequences / to break norm	Your supervising officer assigns the solution of this case highest priority, and his evaluation of your capability as an investigator will strongly depend upon how you behave in this difficult situation. You know that you would have very good chances for a promotion if you proved yourself to be a capable investigator to him.	
Self-relevant reason to disregard aggregate	Your supervising officer assigns the protection of individual rights highest priority, and his evaluation of your capability as an investigator will strongly depend upon how you behave in this difficult situation. You know that you would	

consequences / to adhere to norm	have very good chances for a promotion if you proved yourself to be a capable investigator to him.
Inaction-Default	In order to make him talk you are now considering the application of „aggressive interrogation techniques", like for example holding his head under water or beating him. In order to be able to use these techniques effectively you would have to lead him to a soundproof room, and provide a bowl of water and some other equipment. How would you behave in this situation? Would you initiate the respective preparations and apply aggressive interrogation techniques?
Action-Default	In order to make him talk you have already started the preparation for the application of „aggressive interrogation techniques", like for example holding his head under water or beating him. In order to be able to use these techniques effectively you have already led him to a soundproof room, and must now merely provide a bowl of water and some other equipment. How would you behave in this situation? Would you abort the preparation and return to the conventional interrogation techniques?

Scenarios used in Experiments 3a and 3b

For the construction of these scenarios, we adhered to a well-structured approach that is outlined in Appendix B. The scenarios presented in Tables A5 to A12 were adapted from Moore, Clark, and Kane (2008).

Self-relevance present / death avoidable

Table A5.
Orphanage

	High-involvement		**Low-involvement**	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
You work in an orphanage in a war torn country. Rebel soldiers have staged a successful takeover of the government. The soldiers have entered your town and are …				
	killing …	dispossessing …	killing …	dispossessing …
all those who did not support their regime, or whose loyalty is in question. You are furthermore aware, that the soldiers want to kill everyone of a particular ethnic group in the country, as part of a goal-directed genocide. You have one child of this ethnic group in your orphanage. It will shortly be picked up by some relatives, whereby it would escape the soldiers. However, because contrary to the relatives and the soldiers the orphanage does not own a vehicle, there is no way for you and the other children and workers in the orphanage to escape. You know that the soldiers will…				

	kill everyone they find in the orphanage.	dispossess everyone they find in the orphanage.	kill everyone they find in the orphanage.	dispossess everyone they find in the orphanage.
	You reach the conclusion that you could smother the child, which belongs to the ethnic group focused on by the soldiers, …		You reach the conclusion that you could instruct a nurse to overdose the child, which belongs to the ethnic group focused on by the soldiers, on sleeping pills, …	
	before its relatives arrive to pick it up. Since the soldiers hate the ethnic group it is a part of, they would see this as supporting their regime. This way you could make sure that the soldiers would spare you and the other children and workers instead of…			
	killing you, although this would necessitate the death of the child.	dispossessing you, although this would necessitate the death of the child.	killing you, although this would necessitate the death of the child.	dispossessing you, although this would necessitate the death of the child.
	If you do not smother the child though, the soldiers will …		If the child is not overdosed on sleeping pills, though, the soldiers will …	
	kill you and the other workers and children in the orphanage.	dispossess you and the other workers and children in the orphanage.	kill you and the other workers and children in the orphanage.	dispossess you and the other workers and children in the orphanage.
Inaction-Default	You have not yet attempted anything of this sort and are outside the room of the child. How would you behave in this situation? Would you get a pillow, enter the room of the child, and smother it with the pillow while it is sleeping?		You have not yet attempted anything of this sort and are outside the hospital ward. How would you behave in this situation? Would you enter the hospital ward, talk to the nurse, and instruct her to overdose the child on sleeping pills?	
Action-Default	You are already inside the room of the child and stand with a raised pillow next to its bed, in order to smother it. How would you behave in this situation? Would you abort your plan, take down the pillow, and leave the room of the child again?		You have already entered the hospital ward and talked to the nurse, in order to instruct her to overdose the child on sleeping pills. How would you behave in this situation? Would you enter the hospital ward, talk to the nurse, and instruct her to overdose the child on sleeping pills?	

Table A6.
Ebola

	High-involvement		**Low-involvement**	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	You are a Peace Corps health-worker who is volunteering in a rural African village. A man from a nearby village has contracted a dangerous virus that is extremely contagious and incurable. It is known to you that, apart from causing other mild symptoms, the virus regularly leads to …			
	death within a week. Miraculously, though he is showing the mild symptoms, the man has been spared of death for a month, and so he must be a rare carrier who is	several days of weakness and extreme nausea within a week. Miraculously, though he is showing the mild symptoms, the man has been spared of weakness and nausea	death within a week. Miraculously, though he is showing the mild symptoms, the man has been spared of death for a month, and so he must be a rare carrier who is	several days of weakness and extreme nausea within a week. Miraculously, though he is showing the mild symptoms, the man has been spared of weakness and nausea

	immune to the deadly effects of the virus.	for a month, and so he must be a rare carrier who is immune to those effects of the virus.	immune to the deadly effects of the virus.	for a month, and so he must be a rare carrier who is immune to those effects of the virus.
	However, he wrongly believes that your health center can cure his symptoms and approaches the village, without reacting to shouting or gesturing. Should he enter the village, he would pass the virus on to you and the other villagers so that, unlike him, all of you would …			
	die.	suffer several days of weakness and extreme nausea.	die.	suffer several days of weakness and extreme nausea.
	It comes to your mind that there is a loaded gun in the health center, which you could use to shoot him. This would stop him from entering the village. However, because due to the risk of infection nobody could treat his injuries, he would die from the bullet wound. If you do not shoot him, however, he will pass the virus on to you and the rest of the village.		It comes to your mind that you could warn the village council, which would lead to one of the members shooting the man. This would stop him from entering the village. However, because due to the risk of infection nobody could treat his injuries, he would die from the bullet wound. If you do not warn the village council, however, he will pass the virus on to you and the rest of the village.	
	In order to avoid that you yourself and the other villagers …			
	die from the virus, …	suffer several days of weakness and extreme nausea from the virus, …	die from the virus, …	suffer several days of weakness and extreme nausea from the virus, …
	you would have to shoot the man before he enters the village.		you would have to warn the village council before he enters the village.	
Inaction-Default	You are standing outside the health center and have not yet picked up the gun that is located inside. How would you behave in this situation? Would you get the gun, level it, and shoot the man?		You have not yet informed the village council, in front of whose meeting place you are situated. How would you behave in this situation? Would you enter the meeting place, talk to the members of the village council, and announce the information?	
Action-Default	You have already gotten the gun and leveled it outside, in order to shoot the man. How would you behave in this situation? Would you abort your plan, take the gun down, and bring it back to the health center?		You have already entered the meeting place of village council and started talking to the members, in order to announce the information. How would you behave in this situation? Would you abort your plan, end the conversation prematurely, and leave the meeting place again?	

Self-relevance present / death inevitable

Table A7.

Rescue 911

	High-involvement		**Low-involvement**	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	You are the sole paramedic riding on a rescue helicopter, which is on a mission to save the victims of a devastating flood. You have rescued several injured people and you have been treating them to keep them alive. Suddenly, the pilot notifies you that there has been an engine malfunction and the helicopter can no longer support the weight of you, the pilot, and all the injured people on board. If the helicopter crashes, …			

	it will kill everyone on board.	a very heavy injured person would die because of the injuries he already suffered, and additional injuries would be inflicted to everyone else on board.	it will kill everyone on board.	a very heavy injured person would die because of the injuries he already suffered, and additional injuries would be inflicted to everyone else on board.
	All medical equipment is bolted to the ground and you, yourself, are needed so the injured people are taken care of. However, you could decrease the helicopter’s weight sufficiently by …			
	throwing an injured person, who is very heavy, out of the helicopter. This would lighten the load enough to keep the helicopter in the air, but you would also cause the death of the person that you are throwing off. However, if you do not throw this injured person off, …		cutting the cable that is raising up a last injured person, who is very heavy, on board. This would lighten the load enough to keep the helicopter in the air, but you would also cause the death of the person that is hanging on the cable. However, if you do not cut the cable, …	
	the helicopter will crash. In order to avoid that because of the crash the heavy injured person, you yourself, and all the other people on board die, …	the helicopter will crash. In order to avoid that because of the crash the heavy injured person dies, and you yourself and all the other people on board suffer additional injuries, …	the helicopter will crash. In order to avoid that because of the crash the heavy injured person, you yourself, and all the other people on board die, …	the helicopter will crash. In order to avoid that because of the crash the heavy injured person dies, and you yourself and all the other people on board suffer additional injuries, …
	you would have to throw this injured person out of the helicopter.		you would have to cut the cable.	
Inaction-Default	You have not undertaken anything yet, and the heavy injured person is located some meters away from you. How would you behave in this situation? Would you approach the injured person, pull him to the door, and throw him out of the helicopter?		You have not undertaken anything yet, and the cable is located some meters away from you. How would you behave in this situation? Would you approach the cable, pull your knife, and cut it?	
Action-Default	You have already approached the heavy injured person and dragged him to the door, in order to throw him out of the helicopter. How would you behave in this situation? Would you abort your plan, pull the injured person back to his old position, and move away from him again?		You have already approached the cable and pulled your knife, in order to cut it. How would you behave in this situation? Would you abort your plan, put your knife back, and move away from the cable again?	

Table A8.
Cliffhanger

	High-involvement		**Low-involvement**	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	You are a construction worker and are on a high scaffold together with your crew, working on a skyscraper. Suddenly, the scaffolding partially collapses, which causes you and your colleagues to lose your footing and tumble. Fortunately, you and your colleagues manage to hold on to a crossbar. While dangling on the crossbar, you recognize that it can under no circumstances hold all of your weight. The worker next to you slips off the crossbar and grabs a hold of …			
	your free arm.		another part of the scaffold.	

	You realize that the entire scaffold is about to give way, whereby all of you would fall …			
	into the depth. This would result in you yourself, and all the other workers dying from the fall.	onto the rooftop of a nearby house. This would result in the co-worker who is holding on to you to be buried under the scaffold and die, while you yourself, and all the other workers would suffer injuries.	into the depth. This would result in you yourself, and all the other workers dying from the fall.	onto the rooftop of a nearby house. This would result in the co-worker who is holding on to you to be buried under the scaffold and die, while you yourself, and all the other workers would suffer injuries.
	You could kick the co-worker, who is holding on to you, in the stomach and chest, until he lets go of you.		You could loosen a latching device and thereby undock the part of the scaffold, which your co-worker is holding on to.	
	This would remove just enough weight in order to ensure, that you and the others could make it to safety before the scaffold collapses. However, your co-worker would die from the fall into the depth. If you do not …			
	kick your co-worker, though, the scaffold will collapse. In order to avoid that your co-worker, who is holding on to you …		If you do not loosen the latching device, though, the scaffold will collapse. In order to avoid that your co-worker, who is holding on to the other part of the scaffold …	
	, you yourself, and all the other workers fall to death, …	dies from the fall, and you yourself, and all the other workers suffer injuries, …	, you yourself, and all the other workers fall to death, …	dies from the fall, and you yourself, and all the other workers suffer injuries, …
	you would have to kick this co-worker in the stomach and chest, until he lets go of your hand.		you would have to loosen the latching device, so that the part of the scaffold that your co-worker is holding on to is being undocked.	
Inaction-Default	Your leg is still dangling in the air, without you having done anything. How would you behave in this situation? Would you pull back your leg, kick your colleague and continue doing so, until he lets go of your hand?		The latching device is still an arm’s length away from you, without you having done anything. How would you behave in this situation? Would you grab the latching device, shake it and continue doing so, until this part of the scaffold is undocked?	
Action-Default	You have already pulled back your leg and begun to kick your colleague, in order to make him let go of your hand. How would you behave in this situation? Would you abort your plan, stop kicking, and let your leg dangle again?		You have already grabbed the latching device and started shaking it, in order to undock this part of the scaffold. How would you behave in this situation? Would you abort your plan, stop shaking, and let go of the latching device again?	

Self-relevance absent / death avoidable

Table A9.
Transplant

	High-involvement		**Low-involvement**	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	You are a doctor in a hospital. You have five patients, each of whom is …			
	about to die due to a failing organ, …	going to suffer from irreversible physical impairments like recurrent feelings of weakness, chronic	about to die due to a failing organ, …	going to suffer from irreversible physical impairments like recurrent feelings of weakness, chronic

		pain or episodes of dizziness for the rest of their lives due to organ damage, …		pain or episodes of dizziness for the rest of their lives due to organ damage, …
	unless they have donor organs implanted within a very short period of time. A new patient is rushed into the hospital after a serious car accident. This patient is severely wounded and in need of immediate surgery. While performing a long and complicated surgery on the victim of the accident you realize that you could …			
	inconspicuously cut through one of his arteries, without anyone ever being able to prove that you …		instruct a nurse to administer the wrong medication to him, without anyone ever being able to prove that you or the nurse …	
	did anything on purpose. Hereby you would receive organ donations for the other five patients, the victim of the accident would die, though. If you do not cause the death of the accident victim and then transplant his organs, however, the other five patients will …			
	die.	live the rest of their lives with irreversible impairments.	die.	live the rest of their lives with irreversible impairments.
	In order to avoid that the other five patients die due to failing organs …	In order to avoid that the other five patients will spend the rest of their lives with recurrent feelings of weakness, chronic pain, fits of dizziness or similar impairments due to organ damage …	In order to avoid that the other five patients die due to failing organs …	In order to avoid that the other five patients will spend the rest of their lives with recurrent feelings of weakness, chronic pain, fits of dizziness or similar impairments due to organ damage …
	you would have to cut through one of the arteries of the accident victim, so that you could transplant his organs after his death.		you would have to instruct a nurse to administer the wrong medication to the accident victim, so that you could transplant his organs after his death.	
Inaction-Default	You are currently performing the surgical operation and have until now followed the prescribed procedure. How would you behave in this situation? Would you deviate from the prescribed procedure, lead your scalpel to one of the big arteries of the accident victim, and cut it inconspicuously?		You are currently performing the surgical operation and have until now followed the prescribed procedure. How would you behave in this situation? Would you deviate from the prescribed procedure, start talking to a nurse, and instruct her to administer the wrong medication to the accident victim?	
Action-Default	You are currently performing the surgical operation, have already deviated from the prescribed procedure and have lead your scalpel to one of the big arteries of the accident victim, in order to cut it inconspicuously. How would you behave in this situation? Would you abort your plan, remove your scalpel from the artery, and return to the prescribed procedure?		You are currently performing the surgical operation, have already deviated from the prescribed procedure and have started talking to a nurse, in order to instruct her to administer the wrong medication. How would you behave in this situation? Would you abort your plan, end the conversation prematurely, and return to the prescribed procedure?	

Table A10.
Nobel Prize

	High-involvement		Low-involvement	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	You and a fellow researcher are working in a small laboratory on the development of new technologies. You have recently developed a powerful energy converter, which you consider advanced and efficient. Irrespective of potential other areas of application for this converter, your colleague, who is in charge of the research project, has already decided to sell the converter to a local arms dealer and split the profit evenly between the two of you. You know that he plans to contact the buyer, who would use the energy converter for the manufacturing of a weapon, through which several people would be …			
	killed, today.	irreversibly injured, today.	killed, today.	irreversibly injured, today.
	You could poison him, by bringing him in contact with a chemical, you are in possession of, and which leads to immediate death in the case of direct skin contact. Because this chemical is normally found in laboratories like yours, everybody would think that it was just a lab accident. You merely would have to trickle a drop of the chemical on his skin in an unobserved moment. This way you would ensure that the energy converter is not used as a weapon, though you would also cause the death of your colleague. If you do not poison your colleague, however, he will sell the energy converter to the arms dealer, whereby several people will …		You could release an odorless gas in the laboratory, which would explode, when your colleague turns on the gas burner. Because this gas is normally found in laboratories like yours, everybody would think that it was just a lab accident. You merely would have to turn open the container with the gas. This way you would ensure that the energy converter is not used as a weapon, though you would also cause the death of your colleague. If you do not release the flammable gas, however, he will sell the energy converter to the arms dealer, whereby several people will …	
	lose their lives. In order to avoid that due to the sale of the converter to the arms dealer several people die, …	suffer injuries. In order to avoid that due to the sale of the converter to the arms dealer several people are irreversibly injured, …	lose their lives. In order to avoid that due to the sale of the converter to the arms dealer several people die, …	In order to avoid that due to the sale of the converter to the arms dealer several people are irreversibly injured, …
	you would have to poison your colleague.		you would have to release the flammable gas.	
Inaction-Default	You have not undertaken anything yet and are currently in the laboratory close to the container with the deadly chemical. How would you behave in this situation? Would you put on protective gloves, inconspicuously approach your colleague with a small amount of the chemical, and trickle a drop of it on his skin?		You have not undertaken anything yet and are currently alone in the laboratory close to the container with the flammable gas. How would you behave in this situation? Would you approach the container, turn open the valve, and release the gas in the laboratory?	
Action-Default	You are currently in the laboratory, have already put on protective gloves and have inconspicuously approached your colleague with a small amount of the chemical, in order to trickle a drop of it on his skin. How would you behave in this situation? Would you abort your plan, distance yourself from your colleague, and dispose of the chemical and the protective gloves again?		You are currently alone in the laboratory, have already approached the container with the flammable gas and grabbed the valve, in order to release the gas in the laboratory. How would you behave in this situation? Would you abort your plan, let go of the valve, and distance yourself from the container again?	

Self-relevance absent / death inevitable

Table A11.
Euthanasia

	High-involvement		Low-involvement	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	You are the leader of a small group of soldiers, and all of you are out of ammunition. You are on your way back from a completed mission deep in enemy territory when one of your men steps in a trap set by the enemy. The soldier is in possession of confidential information, which, if they fell into the hands of your enemies, would cause …			
	the unmasking and killing of several of your double agents.	several of your double agents to be unable to use various of their secret meeting points in the future.	the unmasking and killing of several of your double agents.	several of your double agents to be unable to use various of their secret meeting points in the future.
	His leg is badly injured and caught in the trap and you cannot free him from the trap, without killing him. However, if you leave him behind, the enemy troops will find him, torture him, and kill him. You and your sergeant agree that it would be very dangerous to just leave him behind, because he would release the confidential information under the torture sooner or later. You know that the enemy troops are closing in on your position and that it is not safe for you and your men to remain with your trapped comrade any longer. You reach the conclusion that you could …			
	stab the soldier in the heart with your …		instruct your sergeant to stab the soldier in the heart with his …	
	knife, before he is captured and tortured to death. In this case, you would avoid the loss of the confidential information but also kill the soldier, however. If you do not …			
	stab the soldier to death, …		give the command to stab the soldier to death, …	
	though, he will release the confidential information under the torture inflicted by your enemies sooner or later, before they will eventually kill him. In order to avoid that several of your double agents …			
	are unmasked and killed …	lose access to several of their secret meeting points …	are unmasked and killed …	lose access to several of their secret meeting points …
	due to the loss of the confidential information, you would have to stab the soldier to death, before he is found, tortured, and killed by your enemies.		due to the loss of the confidential information, you would have to give the command to stab the soldier to death, before he is found, tortured, and killed by your enemies.	
Inaction-Default	The soldier is still caught in the trap, without you having done anything. How would you behave in this situation? Would you approach the soldier, draw your knife, and kill him by stabbing him in the heart?		The soldier is still caught in the trap, without you having done anything. How would you behave in this situation? Would you approach your sergeant, start talking to him, and command him to kill the soldier by stabbing him in the heart?	
Action-Default	You have already approached the soldier, and have drawn your knife, in order to kill him by stabbing him in the heart. How would you behave in this situation? Would you abort your plan, put your knife away, and distance yourself from the soldier again?		You have already approached your sergeant, and have started talking to him, in order to command him to kill the soldier by stabbing him in the heart. How would you behave in this situation? Would you abort your plan, end the conversation prematurely, and distance yourself from your sergeant again?	

Table A12.
Spy

	High-involvement		Low-involvement	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	You are an officer in the military during a war. Your soldiers have located a high-ranking enemy spy, who was trapped and injured inside of a collapsed building. Shortly afterwards you have been informed over the radio that a large group of enemy soldiers is planning an assault on one of the nearby villages.			
	These soldiers would kill many civilians in the village …	These soldiers would rob many of the civilians in the village of their food …	These soldiers would kill many civilians in the village …	These soldiers would rob many of the civilians in the village of their food …
	if you and your soldiers do not stop them, but you do not know which of the villages they plan to attack. You have questioned the trapped spy who, because of this high rank, surely knows these plans. But he has refused to tell you which village is the target of the impending attack. It is unavoidable, that he will soon die from the many wounds that he has suffered in the collapse of the building.			
	You could start to beat, cut, burn and otherwise physically torture him, …		You could order an expert interrogator to beat, cut, burn and otherwise physically torture him, …	
	whereby you would definitely cause a release of this information. You and your soldiers could then protect the village from the assault, though the spy would die from the torture. If you do not torture the spy, however, your enemies will be able to execute the assault on the village unimpededly. In order to avoid that many civilians lose …			
	their lives in the impeding attack on the village, …	food in the impeding attack on the village, …	their lives in the impeding attack on the village, …	food in the impeding attack on the village, …
	you would have to torture the spy to death before he dies from his injuries, so that he tells you which village his military forces will attack.		you would have to give the command to torture the spy to death before he dies from his injuries, so that he tells you which village his military forces will attack.	
Inaction-Default	Until now, you have only interrogated the spy with the conventional interrogation techniques. How would you behave in this situation? Would you walk over to the spy, take your knife and your lighter, and start inflicting targeted cuts and burns upon him?		Until now, you have only interrogated the spy with the conventional interrogation techniques. How would you behave in this situation? Would you walk over to the radio set, call for the interrogation expert, and command him to inflict targeted cuts and burns upon the spy?	
Action-Default	You have already walked over to the spy, and have taken your knife and your lighter, in order to inflict targeted cuts and burns upon him. How would you behave in this situation? Would you abort your plan, put knife and lighter away, and return to the conventional interrogation techniques?		You have already walked over to the radio set, and have called for the interrogation expert, in order to command him to inflict targeted cuts and burns upon the spy. How would you behave in this situation? Would you abort your plan, cut the radio connection, and return to the conventional interrogation techniques?	

Appendix B (for Hennig & Hütter, 2019 – Chapter II)

Towards Comparable Scenarios in Moral Dilemma Research: A Manual for Creating Scenarios

Table B1 provides a short writing manual that outlines the procedure we followed in writing our scenarios. The underlying motivation of our approach was to reduce inter-item variation and develop stimuli that are comparable to one another with regard to their underlying elements and structure. Specifically, we attempted to construct scenarios that are internally consistent and to avoid common confounds. The manual constitutes a first attempt towards the development of a formalized approach to writing scenarios suitable for the application with the proCNI model and related approaches.

Table B1.
A general manual and template for the development of scenarios for the investigation of moral judgment.

Block	Purpose	Elements	Template	Example
1	Provide some introductory sentences describing the general setting in broad strokes. Mention the undesired result and the target’s potential contribution to avoiding it.	▪ Situation, which leads to undesired result ▪ Undesired result		You are a Peace Corps health-worker who is volunteering in a rural African village. A man from a nearby village has contracted a dangerous virus that is extremely contagious and incurable. It is known to you that, apart from causing other mild symptoms, the virus regularly leads to death within a week. Miraculously, though he is showing the mild symptoms, the man has been spared of death for a month, and so he must be a rare carrier who is immune to the deadly effects of the virus. However, he wrongly believes that your health center can cure his symptoms and approaches the village, without reacting to shouting or gesturing. Should he enter the village, he would pass the virus on to you and the other villagers so that, unlike him, all of you would die from it.

2a	Introduce the action, which leads to the targets death.	▪ Action, which causes sacrifice	You could perform the action, which causes sacrifice.	It comes to your mind that there is a loaded gun in the health center, which you could use to shoot him.
2b	Connect the action to the situation, which leads to the desired result.	▪ Desired result, emerging if action is performed	By performing the action, you would achieve the desired result.	This would stop him from entering the village.
2c	Connect the action with the death of the target. Use a signal-word indicating **conflict** (e.g., however, though).	▪ Death of the target	**However**, doing this would also kill the target.	**However**, because due to the risk of infection nobody could treat his injuries, he would die from the bullet wound.
3a	Connect **not performing the action** with the undesired result.	▪ Undesired result, emerging if action is not performed	If you **do not** perform the action, **however**, the undesired result does emerge.	If you **do not** shoot him, however, he will pass the virus on to you and the rest of the village.
3b	Establish **causal** connection between action and desired result. Formulate desired result as the **avoidance of the undesired result**.	▪ Desired result ▪ Action, which causes sacrifice	**In order to** achieve desired result, you would have to perform the action, which causes sacrifice.	**In order to avoid** that you yourself and the other villagers die from the virus, you would have to shoot the man before he enters the village.
4	Provide a quick primer of the default state, which indicates whether norm breaking will result from inertia or change of the situation.	▪ Default state	You are in the action-default (inaction-default) state and norm breaking has (not) been initiated.	You are standing outside the health center and have not yet picked up the gun that is located inside.
5	Present the decision question of the scenario, and ask for a resolution.	▪ Action to be judged by participant	How would you behave in this situation? Would you perform / abort the final action, and the steps that lead up to / away from it (depending on default state condition)?	How would you behave in this situation? Would you get the gun, level it, and shoot the man?

Note. The general structure of Blocks 4 and 5 is further unpacked below for action- and inaction-default states separately.

Blocks 4 and 5 are separated into different stages and actions to allow for a clean manipulation of the default state. The stages and respective actions are: **distal**, **proximal**, and **final**. The **distal** action leads from the **distal** stage to the **proximal** stage, the **proximal** action to the **final** stage,

the **final** action to the resolution of the scenario (completion or rejection of the sacrifice). Distal, proximal and final actions are identical for inaction and action-default conditions. However, in the inaction-default none of these actions has been performed, while in the action-default distal and proximal actions have been performed. See Table B2 for the general outline and template.

Table B2.
Implementation of the inaction- and action-default states.

Block	Inaction-Default		Action-Default	
	Template	**Example of content**	**Template**	**Example of content**
4	You are in the **distal** stage and have not yet performed the **distal** action, thereby not initiated the sacrifice.	You are **standing outside** the health center and have **not yet picked up the gun** that is located inside.	You are in the **final** stage and have already performed **distal** and **proximal** action, in order to perform **final** action.	You have already **gotten the gun** and **leveled it** outside, in order to **shoot the man**.
5	How would you behave in this situation? Would you perform **distal**, **proximal** and **final** action, thereby completing the sacrifice?	How would you behave in this situation? Would you **get the gun**, **level it**, and **shoot the man**?	How would you behave in this situation? Would you abort performance of **final** action, revert the **proximal** and **distal** action, and thereby reject the sacrifice?	How would you behave in this situation? Would you **abort your plan**, **take the gun down**, and **bring it back** to the health center?

Appendix C (for Hennig & Hütter, 2019 – Chapter II)

Supplemental Materials

This supplement contains additional analyses for the data presented in the main article. Specifically, we present the results of loglinear analyses, conventional analyses of variance (ANOVAs), and additional MPT analyses. The ANOVAs were conducted for reasons of comparability with prior dilemma research, in which this is the most widely applied analysis. However, as the small number of data points per participant per cell in Experiment 1 and Experiment Series 2 renders conventional ANOVAs a technically unsuitable approach, we also report loglinear analyses for these experiments. MPT analyses were conducted to follow up on the results of the ANOVAs and analyses reported in the main paper.

MPT analyses were conducted with MultiTree (Moshagen, 2010), ANOVAs and loglinear analysis were conducted in R (R Core Team, 2018) with the packages “ez” (Lawrence, 2011) and “MASS” (Venables & Ripley, 2002).

For all experiments, we provide a short discussion of the results of the supplemental loglinear analyses and ANOVAs, whenever they directly relate to the experimental findings reported in the main paper, thereby linking the effects of the supplemental analyses to the findings of the MPT analyses.

In all experiments, we implemented the following analytical strategy. We first tested for main effects of all manipulated factors. Second, we tested the two-way interactions of the factors congruency and default-state, which are integral to the estimation of the parameters of the MPT model, with all the other manipulated factors, namely self-relevance, personal involvement, and death avoidability. We also investigated all two-way interactions with the factor death avoidability. We did this because the results of the MPT analyses reported in the main paper suggest this factor to have a strong influence on the processes captured by the proCNI model. This was apparent in the effect of avoidability on the *N*-parameter as well as

in its effect on general model fit in Experiment 3b. We report all of these analyses in Section 1 of this supplement.

Section 2 of this supplement provides additional MPT analyses that assess the influence of self-relevant consequences and personal involvement separately for death-avoidable and death-inevitable scenarios for Experiments 3a and 3b. The results corroborate the notion that the process assumptions of conventional dilemma research apply to death-avoidable scenarios only. Consequently, findings based on the application of death-inevitable scenarios are unlikely to provide a good measurement of concern for consequences and norms, respectively.

For the sake of transparency, we also report all MPT analysis of the main manuscript conducted with an NCI instead of a proCNI model in Section 3. That is, a measurement model in which the *N*-parameter is dominant, and endorsement of consequences (*C*) is assumed to characterize dilemma responses only if endorsement of norms (*N*) does not characterize the response. We note at the outset that model fit indices do not differ depending on whether a proCNI or a proNCI model is employed. Likewise, model choice changed the interpretation of significance tests only with regard to one finding, the effect of self-relevant consequences on the *C*-parameter in Experiment 2a. We provide an explanation for this effect in the appropriate section.

Finally, we also assess the fit of different specifications of PD models, as conceptualized by Conway and Gawronski (2013). The results of these analyses are reported in Section 4.

I. Additional Analyses of dilemma responses

Experiment 1

Loglinear Analysis. A 2 (congruency: congruent vs. incongruent) × 2 (default state: inaction vs. action) loglinear analysis indicated a significant main effect of congruency, $\chi^2(1) = 49.73$, $p < .001$. The odds ratio indicated that the odds for sacrificing were 11.26 times higher in the incongruent compared to the congruent condition. The main effect of default-state and the default-state × congruency interaction were both non-significant, $\chi^2(1) = 2.13$, $p = .144$ and $\chi^2(1) = 1.74$, $p = .187$, respectively.

ANOVA. A 2 (congruency: congruent vs. incongruent) × 2 (default state: inaction vs. action) repeated-measures ANOVA revealed a significant main effect of congruency, $F(1, 93) = 98.63$, $p < .001$, $\eta p^2 = .195$, reflecting a higher proportion of sacrifices in the incongruent ($M_{incongruent} = 0.59$, $SD_{incongruent} = 0.37$) than the congruent ($M_{congruent} = 0.14$, $SD_{congruent} = 0.23$) condition. Proportion of sacrifices did not differ between default-state conditions, $F(1, 93) = 0.11$, $p = .748$, $\eta_p^2 < .001$ ($M_{inaction} = 0.34$, $SD_{inaction} = 0.32$; $M_{action} = 0.36$, $SD_{action} = 0.32$). The default-state × congruency interaction was non-significant, $F(1, 93) = 1.43$, $p = .235$, $\eta_p^2 = .003$.

Discussion. The findings of these additional analyses are in line with the results of the MPT analysis. The significant effect of congruency found in these analyses, representing more sacrificial killing in incongruent than in congruent scenarios, expresses itself in the C-parameter, which is significantly different from zero in the MPT analysis. No significant effect of default-state was found, which converges with the findings of the MPT analyses, in which the I-parameter does not differ significantly from .5. In addition, note that the MPT analysis is necessary to investigate whether endorsement of norms had an independent influence on dilemma judgment, which cannot be assessed by loglinear analysis or ANOVA as this factor is kept constant across all scenario versions (i.e., there always is a proscriptive norm present).

Experiment 2a

Loglinear Analysis. A 2 (congruency: congruent vs. incongruent) × 2 (default state: inaction vs. action) × 2 (self-relevance: present vs. absent) loglinear analysis indicated a significant main effect of congruency, $\chi^2(1) = 14.17$, $p < .001$. The odds ratio indicated that the odds for sacrificing were 6.60 times higher in the incongruent compared to the congruent condition. Both the main effect of default-state, $\chi^2(1) = 3.58$, $p = .059$, and the main effect of self-relevance were non-significant $\chi^2(1) = 0.08$, $p = .776$. The self-relevance × congruency interaction was non-significant, $\chi^2(1) = 0.74$, $p = .388$. Likewise, all other interactions remained non-significant with $\chi^2(1) \leq 1.97$, $p \geq .161$.

ANOVA. A 2 (congruency: congruent vs. incongruent) × 2 (default state: inaction vs. action) × 2 (self-relevance: present vs. absent) mixed ANOVA revealed a significant main effect of congruency, $F(1, 94) = 85.63$, $p < .001$, $\eta p^2 = .181$, reflecting a higher proportion of sacrifices in the incongruent ($M_{incongruent} = 0.52$, $SD_{incongruent} = 0.37$) than the congruent ($M_{congruent} = 0.13$, $SD_{congruent} = 0.26$) condition. The proportion of sacrifices did not differ between inaction-default ($M_{inaction} = 0.29$, $SD_{inaction} = 0.28$) and action-default ($M_{action} = 0.35$, $SD_{action} = 0.33$) conditions, $F(1, 94) = 2.09$, $p = .152$, $\eta_p^2 = .005$. While the main effect of self-relevance was non-significant, $F(1, 94) = 0.14$, $p = .707$, $\eta_p^2 < .001$, the interaction between self-relevance and congruency was significant, $F(1, 94) = 7.66$, $p = .007$, $\eta_p^2 = .019$. Follow-up analyses revealed an effect of congruency in the self-relevance present condition, $F(1, 94) = 18.73$, $p < .001$, $\eta_p^2 = .140$, resulting in a higher proportion of sacrifices in incongruent ($M_{presentIncongruent} = 0.45$, $SD_{presentIncongruent} = 0.39$) than congruent ($M_{presentCongruent} = 0.17$, $SD_{presentCongruent} = 0.30$) scenarios. In the self-relevance absent condition this effect was exacerbated, resulting in a higher proportion of sacrifices in incongruent ($M_{absentIncongruent} = 0.59$, $SD_{absentIncongruent} = 0.35$) than congruent ($M_{absentCongruent} = 0.08$, $SD_{absentCongruent} = 0.21$) scenarios, $F(1, 94) = 83.58$, $p < .001$, $\eta_p^2 = .45$ (see Figure 1). All other interactions were non-significant, $F(1, 94) \leq 1.32$, $p \geq .254$, $\eta_p^2 \leq .004$.

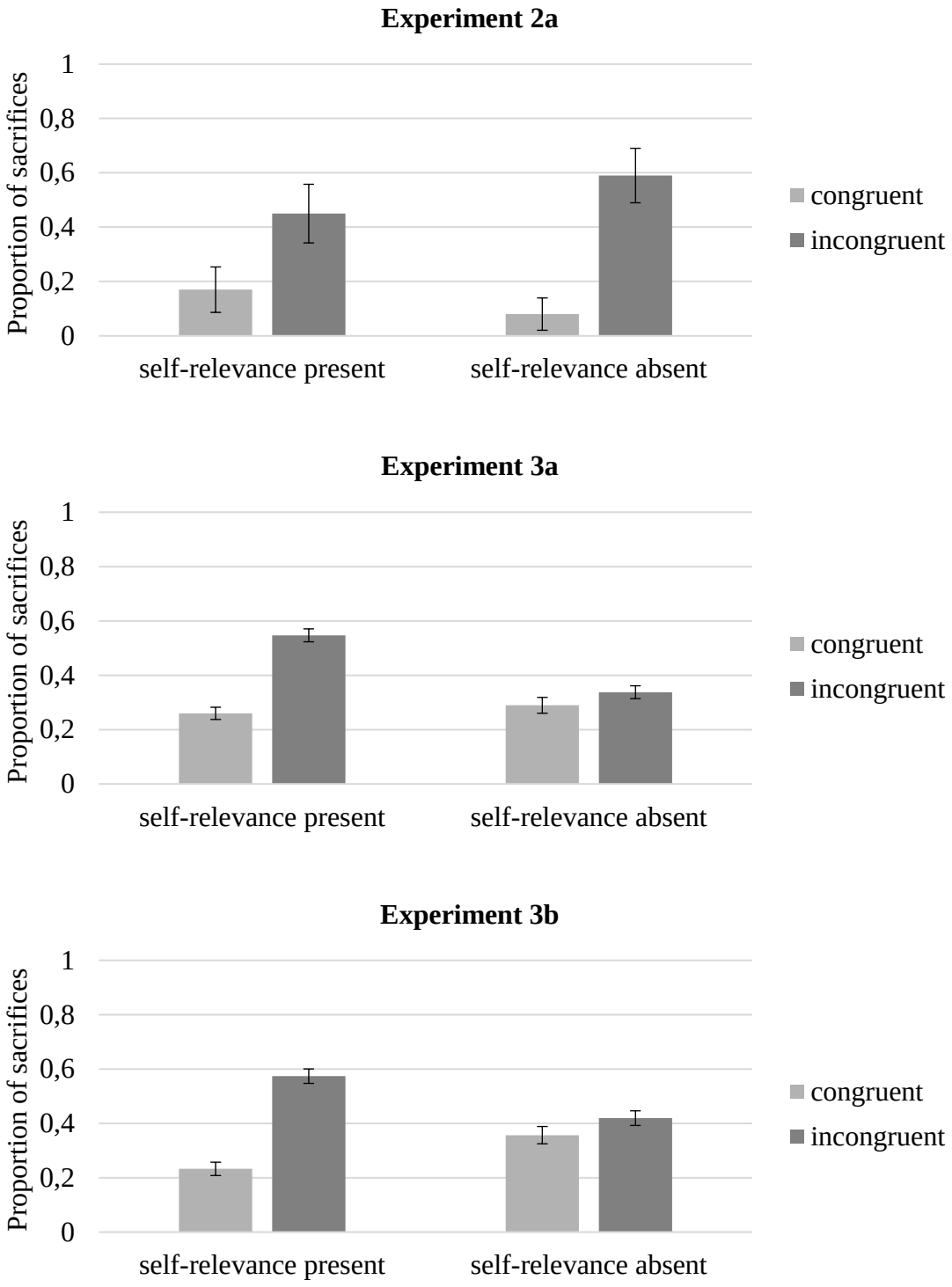

Figure 1. Proportion of sacrifices dependent on congruency and self-relevance separately for Experiments 2b, 3a, and 3b. Error bars represent 95% confidence intervals.

Discussion. Results of the additional analyses generally converge with the results of the MPT analysis. As in Experiment 1, a significant effect of congruency on sacrificial killing was found, such that endorsement was more likely in incongruent as compared to congruent scenarios. This effect is expressed in the *C*-parameter of the proCNI model. Likewise, no significant effect of default-state was found, which converges with the results of the MPT analysis according to which the *I*-parameter does not differ from .5. Based on the effect of self-relevance on the *C*-parameter of the MPT-analysis, an interaction between self-relevance and congruency was expected, such that self-relevance should increase sacrificial killing for congruent and decrease it for incongruent scenarios. This effect, however, was only found when analyzed via an ANOVA, while the loglinear analysis yielded no significant interaction effect.

Experiment 2b

Loglinear Analyses. A 2 (congruency: congruent vs. incongruent) × 2 (default state: inaction vs. action) × 2 (self-relevance: present vs. absent) loglinear analysis indicated a significant main effect of congruency, $\chi^2(1) = 15.85$, $p < .001$. The odds ratio indicated that the odds for sacrificing were 5.61 times higher in the incongruent compared to the congruent condition. The main effect of default-state was non-significant, $\chi^2(1) = 1.29$, $p = .255$, while the main effect of self-relevance was marginal $\chi^2(1) = 2.76$, $p = .097$, suggesting the odds for sacrificing to be 2.56 times higher in the self-relevance present compared to the self-relevance absent condition. None of the interaction effects were significant, $\chi^2(1) \leq 1.48$, $p \geq .230$.

ANOVA. A 2 (congruency: congruent vs. incongruent) × 2 (default state: inaction vs. action) × 2 (self-relevance: present vs. absent) mixed ANOVA revealed a significant main effect of congruency, $F(1, 94) = 91.26$, $p < .001$, $\eta p^2 = .183$, reflecting a higher proportion of sacrifices in the incongruent ($M_{incongruent} = 0.62$, $SD_{incongruent} = 0.33$) than the congruent ($M_{congruent} = 0.20$, $SD_{congruent} = 0.30$) condition. Default-state did not affect proportion of sacrifices, $F(1, 94) = 0.84$, $p = .362$, $\eta_p^2 = .002$ ($M_{inaction} = 0.39$, $SD_{inaction} = 0.31$; $M_{action} = 0.43$,

SD_{action} = 0.33). The effect of self-relevance was significant, $F(1, 94) = 6.33$, $p = .014$, $\eta_p^2 = .017$, reflecting a higher proportion of sacrifices in the present ($M_{present}$ = 0.47, SD_{presen} = 0.24) than the absent (M_{absent} = 0.35, SD_{absent} = 0.21) condition. None of the interaction effects were significant, $F(1, 94) \leq 0.92$, $p \geq .341$, $\eta_p^2 \leq .002$.

Discussion. Again, results are generally in line with the results of the MPT analyses. The main effect of congruency was significant, resulting in more sacrificial killing in incongruent than congruent scenarios as expressed in the *C*-parameter of the proCNI model. As in the previous studies, default-state did not exert a main effect, which is consistent with the results of the MPT analysis, according to which the *I*-parameter does not differ from .5. Descriptively, there was more sacrificial killing when self-relevant consequences were present as compared to absent, which is the pattern of results expected based on the effect of self-relevance on the *N*-parameter in the MPT analysis. However, this main effect was significant only when data were analyzed with an ANOVA, and just missed the significance criterion in the loglinear analysis.

Experiment 3a

In Experiment Series 3, we collected sufficient data points per participants to conduct ANOVAs. However, because the number of data points per cell per participant did not allow the estimation of one ANOVA containing all experimental factors, we report main effects as well as all the two-way interactions with the factors default-state, congruency, and avoidability.

Main effects. Results of these analyses revealed a significant main effect of congruency, reflecting a higher proportion of sacrifices in the incongruent ($M_{incongruent}$ = 0.44, $SD_{incongruent}$ = 0.30) than in the congruent ($M_{congruent}$ = 0.28, $SD_{congruent}$ = 0.25) condition, $F(1, 692) = 147.99$, $p < .001$, $\eta_p^2 = .085$.

Proportion of sacrifices was also influenced by default state, as apparent in more sacrifices in the action default (M_{action} = 0.38, SD_{action} = 0.25) than in the inaction default ($M_{inaction}$ = 0.34, $SD_{inaction}$ = 0.26) condition, $F(1, 692) = 11.06$, $p < .001$, $\eta_p^2 = .005$.

The main effect of self-relevance was significant as well, resulting from a higher proportion of sacrifices in the present ($M_{present}$ = 0.40, $SD_{present}$ = 0.26) compared to the absent (M_{absent} = 0.32, SD_{absent} = 0.24) condition, $F(1, 692) = 75.36$, $p < .001$, $\eta_p^2 = .032$.

A significant main effect of personal involvement was found, such that the proportion of sacrifices was higher in the low-involvement (M_{low} = 0.42, SD_{low} = 0.27) compared to the high-involvement (M_{high} = 0.30, SD_{high} = 0.26) condition, $F(1, 692) = 89.16$, $p < .001$, $\eta_p^2 = .047$.

The main effect of avoidability was significant, resulting in more sacrifices in the death-inevitable ($M_{inevitable}$ = 0.50, $SD_{inevitable}$ = 0.31) compared to the death-avoidable ($M_{avoidable}$ = 0.22, $SD_{avoidable}$ = 0.20) condition, $F(1, 692) = 554.19$, $p < .001$, $\eta_p^2 = .226$.

Interactions congruency. The congruency × self-relevance interaction was significant, $F(1, 692) = 138.68$, $p < .001$, $\eta_p^2 = .031$ (see Figure 1). Further investigation of the simple effects revealed that self-relevance decreased the proportion of sacrifices in the congruent ($M_{congruentPresent}$ = 0.26, $SD_{congruentPresent}$ = 0.30; $M_{congruentAbsent}$ = 0.29, $SD_{congruentAbsent}$ = 0.32) condition, while increasing the proportion of sacrifices in the incongruent ($M_{incongruentPresent}$ = 0.55, $SD_{incongruentPresent}$ = 0.39; $M_{incongruentAbsent}$ = 0.34, $SD_{incongruentAbsent}$ = 0.32) condition, $F(1, 692) = 4.63$, $p < .032$, $\eta_p^2 = .002$ and $F(1, 692) = 189.72$, $p < .001$, $\eta_p^2 = .079$, respectively. The effect of congruency was significant when self-relevance was present, $F(1, 692) = 251.30$, $p < .001$, $\eta_p^2 = .144$, as well as absent, $F(1, 692) = 4.63$, $p = .032$, $\eta_p^2 = .002$.

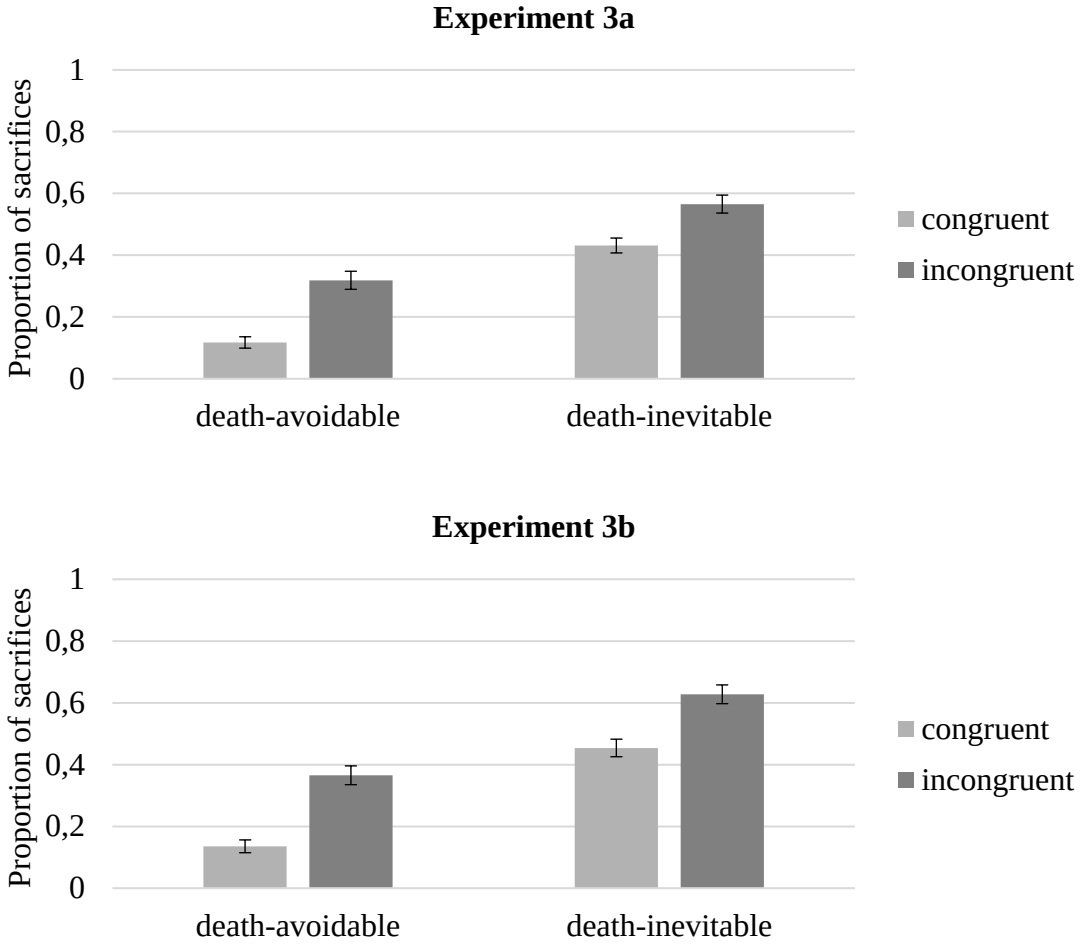

Figure 2. Proportion of sacrifices dependent on congruency and avoidability separately for Experiments 3a, and 3b. Error bars represent 95% confidence intervals.

The congruency × avoidability interaction was significant, $F(1, 692) = 10.82$, $p = .001$, $\eta_p^2 = .002$, and is presented in Figure 2. Investigation of simple effects revealed that inevitable deaths increased the proportion of sacrifices in both the congruent ($M_{congruentInevitable} = 0.43$, $SD_{congruentInevitable} = 0.39$; $M_{congruentAvoidableo} = 0.12$, $SD_{congruentAvoidable} = 0.25$) and incongruent ($M_{incongruentInevitable} = 0.57$, $SD_{incongruentInevitable} = 0.39$; $M_{incongruentAvoidable} = 0.32$, $SD_{incongruentAvoidable} = 0.33$) condition, $F(1, 692) = 387.24$, $p < .001$, $\eta_p^2 = .186$ and $F(1, 692) = 256.57$, $p < .001$, $\eta_p^2 = .106$, respectively. The effect of congruency was significant when death was inevitable, $F(1, 692) = 55.00$, $p < .001$, $\eta_p^2 = .029$, as well as avoidable, $F(1, 692) = 155.25$, $p = .032$, $\eta_p^2 = .108$.

The congruency × personal involvement interaction was non-significant, $F(1, 692) = 0.15$, $p = .698$, $\eta_p^2 < .001$.

Interactions default-state. The default-state × avoidability interaction was significant, $F(1, 692) = 4.93$, $p = .027$, $\eta_p^2 = .001$ (see Figure 3). Investigation of simple effects revealed that inevitable deaths increased the proportion of sacrifices in both the inaction-default ($M_{inactionInevitable} = 0.47$, $SD_{inactionInevitable} = 0.38$; $M_{inactionAvoidableo} = 0.21$, $SD_{inactionAvoidable} = 0.27$) and the action-default ($M_{actionInevitable} = 0.53$, $SD_{actionInevitable} = 0.37$; $M_{actionAvoidable} = 0.22$, $SD_{actionAvoidable} = 0.28$) condition, $F(1, 692) = 263.35$, $p < .001$, $\eta_p^2 = .131$ and $F(1, 692) = 324.38$, $p < .001$, $\eta_p^2 = .174$, respectively. However, the effect of default state was only significant when death was inevitable, $F(1, 692) = 14.33$, $p < .001$, $\eta_p^2 = .007$, not when death was avoidable, $F(1, 692) = 0.52$, $p = .471$, $\eta_p^2 < .001$.

The default-state × self-relevance, $F(1, 692) = 0.54$, $p = .464$, $\eta_p^2 < .001$, and default-state × personal involvement, $F(1, 515) = 0.02$, $p = .0.89$, $\eta_p^2 < .001$, interactions were both non-significant.[38]

Interactions avoidability. The self-relevance × avoidability interaction was significant, $F(1, 692) = 85.79$, $p < .001$, $\eta_p^2 = .023$ (see Figure 4). Investigation of simple effects indicated that self-relevant consequences increased the proportion of sacrifices when death was avoidable ($M_{presentAvoidable} = 0.31$, $SD_{presentAvoidable} = 0.28$; $M_{absentAvoidable} = 0.12$, $SD_{absentAvoidable} = 0.24$), but not when death was inevitable ($M_{presentInevitable} = 0.49$, $SD_{presentInevitable} = 0.38$; $M_{presentAvoidable} = 0.50$, $SD_{presentAvoidable} = 0.38$), $F(1, 692) = 210.79$, $p < .001$, $\eta_p^2 = .116$ and $F(1, 692) = 0.28$, $p = .600$, $\eta_p^2 < .001$, respectively. The effect of avoidability was significant for self-relevance present, $F(1, 692) = 126.23$, $p < .001$, $\eta_p^2 = .069$, as well as self-relevance absent scenarios, $F(1, 692) = 579.09$, $p < .001$, $\eta_p^2 = .265$.

[38] Due to design limitations, it was necessary to drop some participants for the analysis of this interaction effect. The analysis was conducted with the data points of 516 participants.

The personal involvement × avoidability interaction was non-significant, $F(1, 692) = 0.88$, $p = .347$, $\eta_p^2 < .001$.

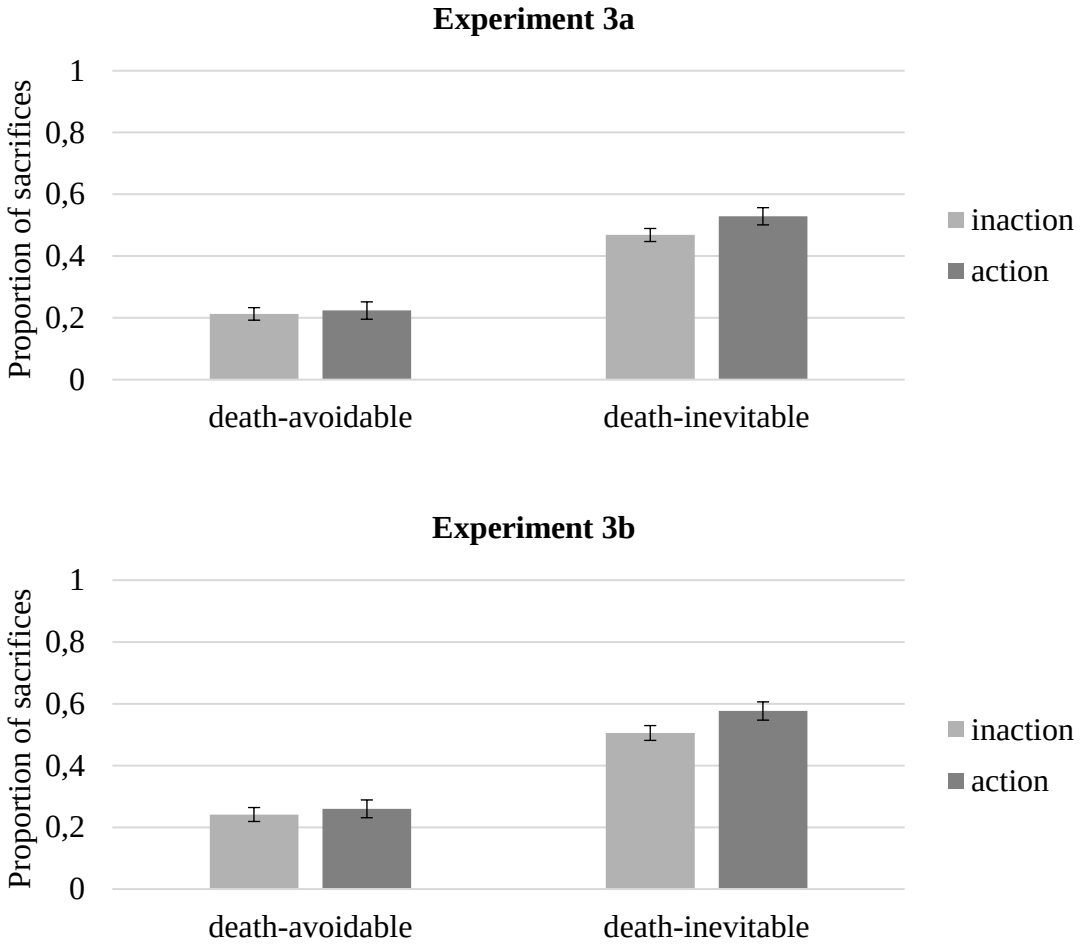

Figure 3. Proportion of sacrifices dependent on default-state and avoidability separately for Experiments 3a and 3b. Error bars represent 95% confidence intervals.

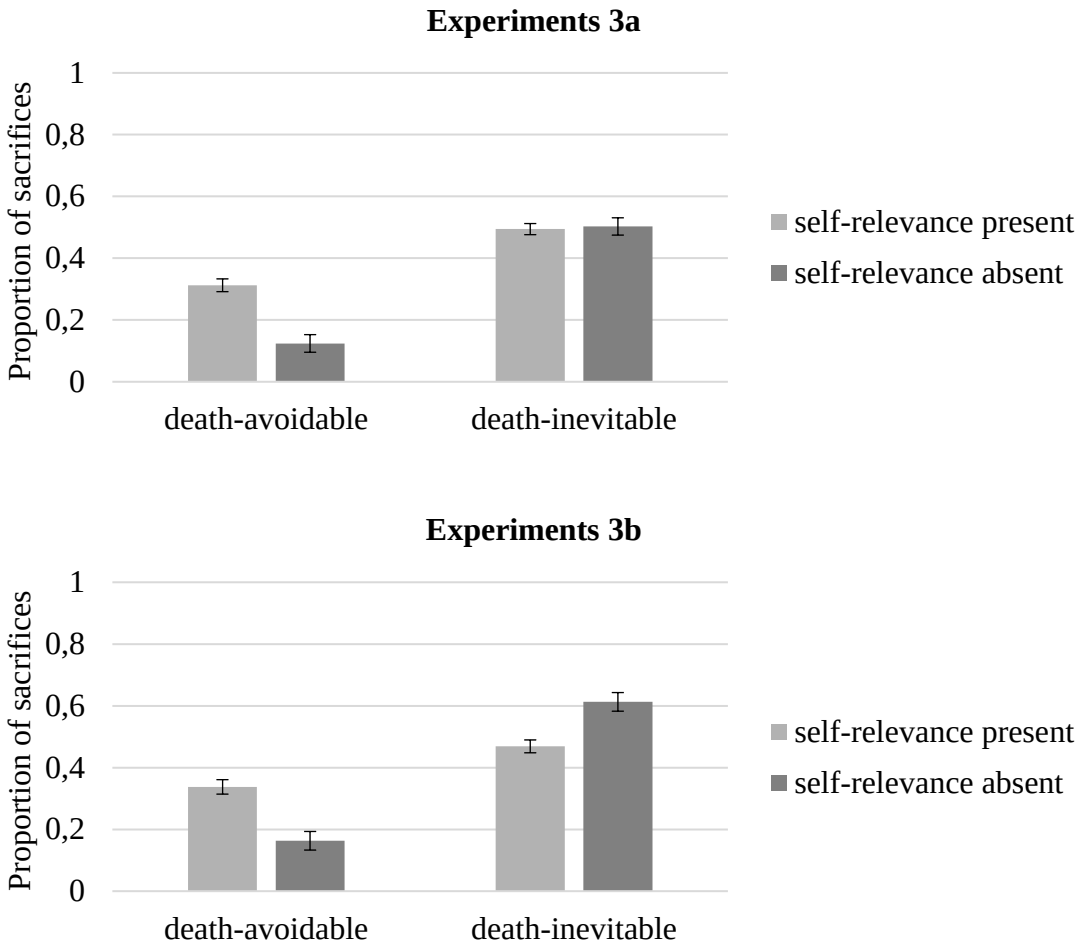

Figure 4. Proportion of sacrifices dependent on self-relevance and avoidability separately for Experiments 3a and 3b. Error bars represent 95% confidence intervals.

Discussion. The results of the ANOVAs mirror and support the findings obtained with the MPT analyses. As in the previous studies, the significant effect of congruency, resulting in more sacrificial killing in incongruent compared to congruent scenarios, is expressed in a significant *C*-parameter in the MPT analyses. This time, the effect of default-state was significant, resulting in more sacrificial killing in action- as compared to inaction-default scenarios. This converges with the results of the MPT analysis, in which we obtained a slight preference for inertia over change, as expressed by an *I*-parameter larger than .5.

The main effects of self-relevance, personal involvement, and death avoidability were significant and in the direction anticipated by the MPT analyses. More sacrificial killing was accepted when self-relevance was present as opposed to absent, which is expressed in a lower

N-parameter for present scenarios in the MPT analysis. Likewise, more sacrificial killing was endorsed in low- as opposed to high-involvement scenarios, which is expressed in a lower *N*-parameter for high-involvement scenarios in the MPT analysis. More sacrificial killing was accepted when death was inevitable as opposed to avoidable, which finds its expression in the reduced *N*-parameter for inevitable scenarios. Thus, all factors that exerted a significant effect on the *N*-parameter in the MPT analyses also exerted significant main effects on sacrificial killing when analyzed with ANOVAs.

The effects of the self-relevance and avoidability factors depended on congruency, again converging with the results of the MPT analyses. Self-relevant consequences increased sacrificial killing for incongruent but decreased sacrificial killing for congruent scenarios. This is expressed in an increased *C*-parameter for self-relevance present compared to absent scenarios in the MPT analysis. Similarly, inevitable deaths increased sacrificial killing more strongly for congruent than for incongruent scenarios. This is expressed in the decreased *C*-parameter for inevitable scenarios in the MPT analysis. Thus, all factors that exerted a significant effect on the *C*-parameter in the MPT analyses also significantly interacted with congruency when analyzed with an ANOVA. Personal involvement, which did not affect the *C*-parameter in the MPT analysis, did not interact with congruency in the ANOVA.

Experiment 3b

We followed the same analytic strategy as in Experiment 3a.

Main effects. Results of our analyses revealed a significant main effect of congruency, reflecting a higher proportion of sacrifices in the incongruent ($M_{incongruent}$ = 0.50, $SD_{incongruent}$ = 0.30) than the congruent ($M_{congruent}$ = 0.30, $SD_{congruent}$ = 0.25) condition, $F(1, 576) = 167.38$, $p < .001$, $\eta_p^2 = .121$.

Proportion of sacrifices was also influenced by default state as evident in more sacrifices in the action-default (M_{action} = 0.42, SD_{action} = 0.25) than in the inaction-default ($M_{inaction}$ = 0.37, $SD_{inaction}$ = 0.25) condition, $F(1, 576) = 13.20$, $p < .001$, $\eta_p^2 = .008$.

A significant main effect of personal involvement was found, such that the proportion of sacrifices was higher in the low-involvement (M_{low} = 0.43, SD_{low} = 0.25) compared to the high-involvement (M_{high} = 0.36, SD_{high} = 0.25) condition, $F(1, 576) = 31.19$, $p < .001$, $\eta_p^2 = .020$.

The main effect of avoidability was significant, resulting in more sacrifices in the death-inevitable ($M_{inevitable}$ = 0.54, $SD_{inevitable}$ = 0.29) compared to the death-avoidable ($M_{avoidable}$ = 0.25, $SD_{avoidable}$ = 0.20) condition, $F(1, 576) = 537.01$, $p < .001$, $\eta_p^2 = .255$.

The main effect of self-relevance was not significant. Thus, the proportion of sacrifices did not differ between self-relevance present ($M_{present}$ = 0.40, $SD_{present}$ = 0.25) and self-relevance absent (M_{absent} = 0.39, SD_{absent} = 0.24) scenarios, $F(1, 576) = 1.58$, $p = .209$, $\eta_p^2 < .001$.

Interactions congruency. A significant congruency × self-relevance interaction was obtained, $F(1, 576) = 146.50$, $p < .001$, $\eta_p^2 = .041$ (see Figure 1). Analysis of the simple effects revealed that self-relevant consequences decreased the proportion of sacrifices in the congruent ($M_{congruentPresent}$ = 0.23, $SD_{congruentPresent}$ = 0.30; $M_{congruentAbsent}$ = 0.36, $SD_{congruentAbsent}$ = 0.33) condition, $F(1, 576) = 59.78$, $p < .001$, $\eta_p^2 = .038$, while increasing the proportion of sacrifices in the incongruent ($M_{incongruentPresent}$ = 0.57, $SD_{incongruentPresent}$ = 0.39; $M_{incongruentAbsent}$ = 0.42, $SD_{incongruentAbsent}$ = 0.33) condition. The effect of congruency was significant when self-relevance was absent, $F(1, 576) = 11.37$, $p < .001$, $\eta_p^2 = .009$, as well as present, $F(1, 576) = 285.67$, $p < .001$, $\eta_p^2 = .194$.

The congruency × avoidability interaction was significant, $F(1, 576) = 6.12$, $p = .014$, $\eta_p^2 = .002$ (see Figure 2). Investigation of simple effects revealed that inevitable deaths increased the proportion of sacrifices in both the congruent ($M_{congruentInevitable}$ = 0.45, $SD_{congruentInevitable}$ = 0.37; $M_{congruentAvoidableo}$ = 0.14, $SD_{congruentAvoidable}$ = 0.25) and incongruent ($M_{incongruentInevitable}$ = 0.62, $SD_{incongruentInevitable}$ = 0.37; $M_{incongruentAvoidable}$ = 0.37, $SD_{incongruentAvoidable}$ = 0.35) condition, $F(1, 576) = 366.20$, $p < .001$, $\eta_p^2 = .201$ and $F(1, 576) =$

231.28, $p < .001$, $\eta_p^2 = .117$, respectively. The effect of congruency was significant when death was inevitable, $F(1, 576) = 79.14$, $p < .001$, $\eta_p^2 = .052$, as well as when death was avoidable, $F(1, 576) = 144.52$, $p < .001$, $\eta_p^2 = .125$.

The congruency × personal involvement interaction was not significant, $F(1, 576) = 1.09$, $p = .297$, $\eta_p^2 < .001$.

Interactions default-state. The default-state × avoidability interaction was significant, $F(1, 576) = 4.85$, $p = .028$, $\eta_p^2 = .002$ (see Figure 3). Investigation of simple effects revealed that inevitable deaths increased the proportion of sacrifices in both the inaction-default ($M_{inactionInevitable} = 0.47$, $SD_{inactionInevitable} = 0.38$; $M_{inactionAvoidableo} = 0.21$, $SD_{inactionAvoidable} = 0.27$) and the action-default ($M_{actionInevitable} = 0.53$, $SD_{actionInevitable} = 0.37$; $M_{actionAvoidable} = 0.22$, $SD_{actionAvoidable} = 0.28$) condition, $F(1, 576) = 259.47$, $p < .001$, $\eta_p^2 = .148$ and $F(1, 576) = 299.73$, $p < .001$, $\eta_p^2 = .188$, respectively. However, the effect of default-state was only significant when death was inevitable, $F(1, 576) = 16.52$, $p < .001$, $\eta_p^2 = .010$, not when death was avoidable, $F(1, 576) = 1.16$, $p = .281$, $\eta_p^2 = .001$.

The default-state × self-relevance, $F(1, 576) = 1.61$, $p = .204$, $\eta_p^2 < .001$, and default-state × personal involvement, $F(1, 446) = 1.97$, $p = .161$, $\eta_p^2 < .001$, interactions were both not significant.[39]

Interactions avoidability. The self-relevance × avoidability interaction was significant, $F(1, 576) = 168.25$, $p < .001$, $\eta_p^2 = .057$ (see Figure 4). The analysis of the simple effects indicated that self-relevant consequences increased the proportion of sacrifices when death was avoidable ($M_{presentAvoidable} = 0.34$, $SD_{presentAvoidable} = 0.29$; $M_{absentAvoidable} = 0.16$, $SD_{absentAvoidable} = 0.25$), but decreased it when death was inevitable ($M_{presentInevitable} = 0.47$, $SD_{presentInevitable} = 0.37$; $M_{presentAvoidable} = 0.61$, $SD_{presentAvoidable} = 0.37$), $F(1, 576) = 132.29$, $p < .001$, $\eta_p^2 = .094$ and $F(1, 576) = 57.05$, $p = .001$, $\eta_p^2 < .037$, respectively. The effect of

[39] Due to design limitations, it was necessary to drop some participants for the analysis of this interaction effect. The analysis was conducted with the data points of 447 participants.

avoidability was significant for self-relevance present, $F(1, 576) = 53.50$, $p < .001$, $\eta_p^2 = .038$, as well as self-relevance absent scenarios, $F(1, 576) = 686.54$, $p < .001$, $\eta_p^2 = .336$.

The personal involvement × avoidability interaction was not significant, $F(1, 576) = 0.96$, $p < .327$, $\eta_p^2 < .001$.

Discussion. Results of the analyses of Experiment 3b are almost completely identical with those of Experiment 3a, for MPT analyses and ANOVAs alike. Hence, we merely note that the discussion of results from Experiment 3a fully applies to the results of Experiment 3b, with the following exception: In Experiment 3b, the main effect of self-relevance on sacrificial killing was not significant. This, again, converges with the results of the MPT analysis, which did not indicate an influence of self-relevance on the *N*-parameter in Experiment 3b.

II. Additional MPT analyses on the combined datasets of Experiment Series 3

Baseline models for avoidable and inevitable scenarios. For scenarios in which the death of the target was avoidable, the proCNI model with one *C*-parameter (*C* = .21, 95% *CI* [.19, .24]), one *N*-parameter (*N* = .68, 95% *CI* [.65, .71]), and one *I*-parameter (*I* = .52, 95% *CI* [.48, .57]) provided a good fit to the data, $G^2(1) = 0.23$, $p = .631$, $w = 0.005$. The *I*-parameter did not differ from its neutral reference point of 0.5, $\Delta G^2(1) = 1.30$, $p = .254$, $w < 0.001$. For scenarios in which death of the target was inevitable, the model with one *C*-parameter (*C* = .15, 95% *CI* [.13, .18]), one *N*-parameter (*N* = .00, 95% *CI* [-.03, .03]), and one *I*-parameter (*I* = .54, 95% *CI* [.52, .56]) did not explain the data well, $G^2(1) = 6.82$, $p = .009$, although the size of this deviation was small, $w = 0.026$.

As the value of the *N*-parameter was once again at the lower bound of the parameter space, we once more investigated whether the meaning of the *N*-parameter changes in the case of inevitable deaths. That is, in the case of inevitable deaths the *N*-parameter may not represent rejection of proximal harm. Instead, its meaning may reverse such that it represents a general acceptance of proximal harm. We therefore again tested the fit of a model, in which the *N*-parameter would represent the tendency to always sacrifice regardless of consequences and default-state. This model fit the data well, $G^2(1) = 0.21$, $p = .649$, $w < 0.001$, and suggested a small willingness to generally accept killing regardless of consequences and default-state, as the *N*-parameter (*N* = .04, 95% *CI* [.01, .07]) differed significantly from zero, $\Delta G^2(1) = 6.62$, $p = .010$, $w < 0.026$. This result suggests that the process assumptions presupposed in dilemma research, namely that responses are determined by inclinations to maximize consequences and adhere to no-harm rules, are violated for death-inevitable scenarios. Consequently, we conducted additional analyses of the influence of self-relevance and personal involvement for death-avoidable and death-inevitable scenarios separately.

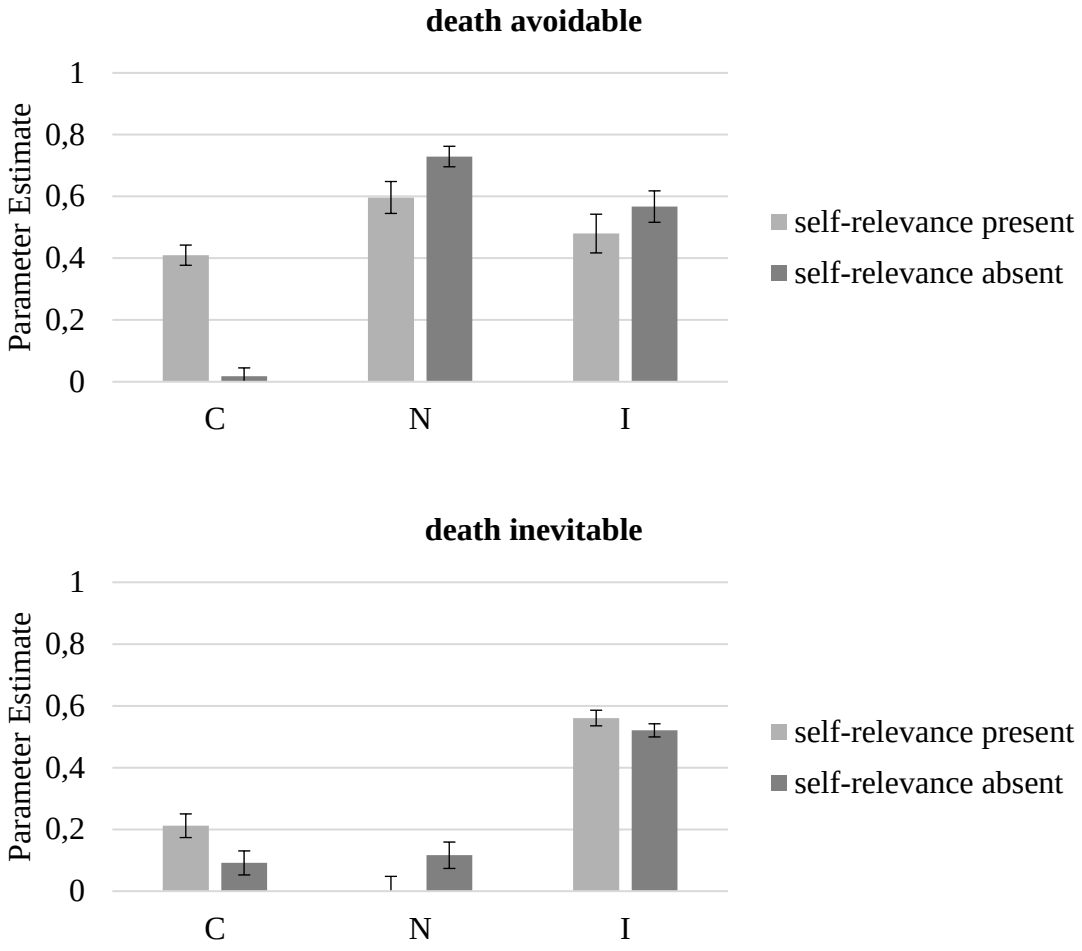

Figure 5. Parameter estimates representing endorsement of consequences (*C*), endorsement of norms (*N*) and inertia (*I*) based on the combined data of Experiments 3a and 3b, separated by self-relevance conditions. Error bars represent 95% confidence intervals.
Note. For death-inevitable scenarios, the model with the reversed *N*-parameter representing general *acceptance* of direct harm is depicted.

Avoidability × self-relevance interaction. When considering only responses to death-avoidable scenarios, the model estimating separate parameters for the self-relevance present and absent condition fit the data well, $G^2(2) = 1.29$, $p = .524$, $w = 0.011$. The parameter estimates of this model are presented in Figure 5. Equating the *C*-parameters across self-relevance conditions resulted in a significant decrease in model fit, $\Delta G^2(1) = 301.59$, $p < .001$, $w = 0.172$, indicating that *responses* adhered to aggregate more strongly in the present ($C_{present}$ = .40, 95% *CI* [.38, .44]) than the absent condition (C_{absent} = .02, 95% *CI* [-.01, .05]). The manipulation also exerted an effect on the size of the *N*-parameter indicating that responses

were less frequently consistent with norms when self-relevance was present ($N_{present}$ = .60, 95% *CI* [.55, .65]) compared to absent (N_{absent} = .73, 95% *CI* [.70, .76]), $\Delta G^2(1) = 18.83$, $p < .001$, $w = 0.043$. Finally, participants showed less inertia when self-relevance was present ($I_{present}$ = .48, 95% *CI* [.42, .54]) rather than absent, (I_{absent} = .57, 95% *CI* [.52, .62]), $\Delta G^2(1) = 4.49$, $p = .034$, $w = 0.021$. When considering only responses to death-inevitable scenarios, the model estimating separate parameters for the self-relevance present and absent condition did not fit the data well, $G^2(2) = 29.38$, $p < .001$, $w = 0.054$. Again, applying a model in which the *N*-parameter represented a general acceptance rather than rejection of proximal harm alleviated the lack of fit, $G^2(2) = 3.93$, $p = .140$, $w = 0.020$. Equating the *C*-parameters across self-relevance conditions caused a significant decrease in model fit, $\Delta G^2(1) = 19.00$, $p < .001$, $w = 0.043$ indicating that responses adhered to aggregate consequences more strongly in the present ($C_{present}$ = .21, 95% *CI* [.17, .25]) than in the absent condition (C_{absent} = .09, 95% *CI* [.05, .13]). The manipulation also exerted an effect on the size of the *N*-parameter, such that participants showed a higher tendency to cause proximal harm regardless of aggregate consequences when self-relevance was absent (N_{absent} = .12, 95% *CI* [.07, .16]) as compared to present ($N_{present}$ = .00, 95% *CI* [-.05, .05]), $\Delta G^2(1) = 20.79$, $p < .001$, $w = 0.045$. Finally, participants showed more inertia when self-relevant consequences were present ($I_{present}$ = .56, 95% *CI* [.53, .58]) rather than absent, (I_{absent} = .52, 95% *CI* [.50, .55]), $\Delta G^2(1) = 3.93$, $p = .048$, $w = 0.020$. Thus, for two of the model parameters, *N* and *I*, self-relevant consequences had opposite effects depending on whether death was avoidable or inevitable. While self-relevant consequences decreased the endorsement of norms in the case of avoidable death, they increased endorsement of this norm in the case of inevitable deaths. Similarly, while self-relevant consequences reduced inertia when death was avoidable, they increased inertia when death was inevitable.

Avoidability × personal involvement interaction. When considering only responses to death-avoidable scenarios, the model estimating separate parameters for the self-relevance

present and absent conditions fit the data well, $G^2(2) = 0.34$, $p = .846$, $w = 0.006$. Setting the parameters equal across personal involvement conditions revealed a significant effect of the manipulation on the *N*-parameter, resulting in more responses characterized by norm-endorsement when involvement was high ($N_{high} = .77$, 95% *CI* [.73, .81]) rather than low ($N_{low} = .59$, 95% *CI* [.54, .63]), $\Delta G^2(1) = 39.48$, $p < .001$, $w = 0.062$. The *I*-parameter was also affected by the manipulation, representing more inertia when personal involvement was high ($I_{high} = .59$, 95% *CI* [.51, .66]) rather than low ($I_{low} = 0.49$, 95% *CI* [.44, .54]), $\Delta G^2(1) = 4.33$, $p = .038$, $w = 0.021$. There was also a marginal effect on the *C*-parameter, suggesting more sensitivity to aggregate consequences when personal involvement was low ($C_{low} = 0.24$, 95% *CI* [.20, .27]) rather than high ($C_{high} = .19$, 95% *CI* [.17, .22]), $\Delta G^2(1) = 3.11$, $p = .078$, $w = 0.018$. When considering only responses to death-inevitable scenarios, the model estimating separate parameters for the low and high personal involvement condition did not fit the data well, $G^2(2) = 33.31$, $p < .001$, $w = 0.057$. This time the alternative model, in which the *N*-parameter represented a general acceptance rather than rejection of harm, also did not show a good fit to the data, $G^2(2) = 24.31$, $p < .001$, although the strength of the deviation was slightly smaller, $w = 0.050$. However, due to the misfit we consider parameter estimates for this model to be uninterpretable and refrain from investigating the effect of personal involvement on individual parameter estimates.

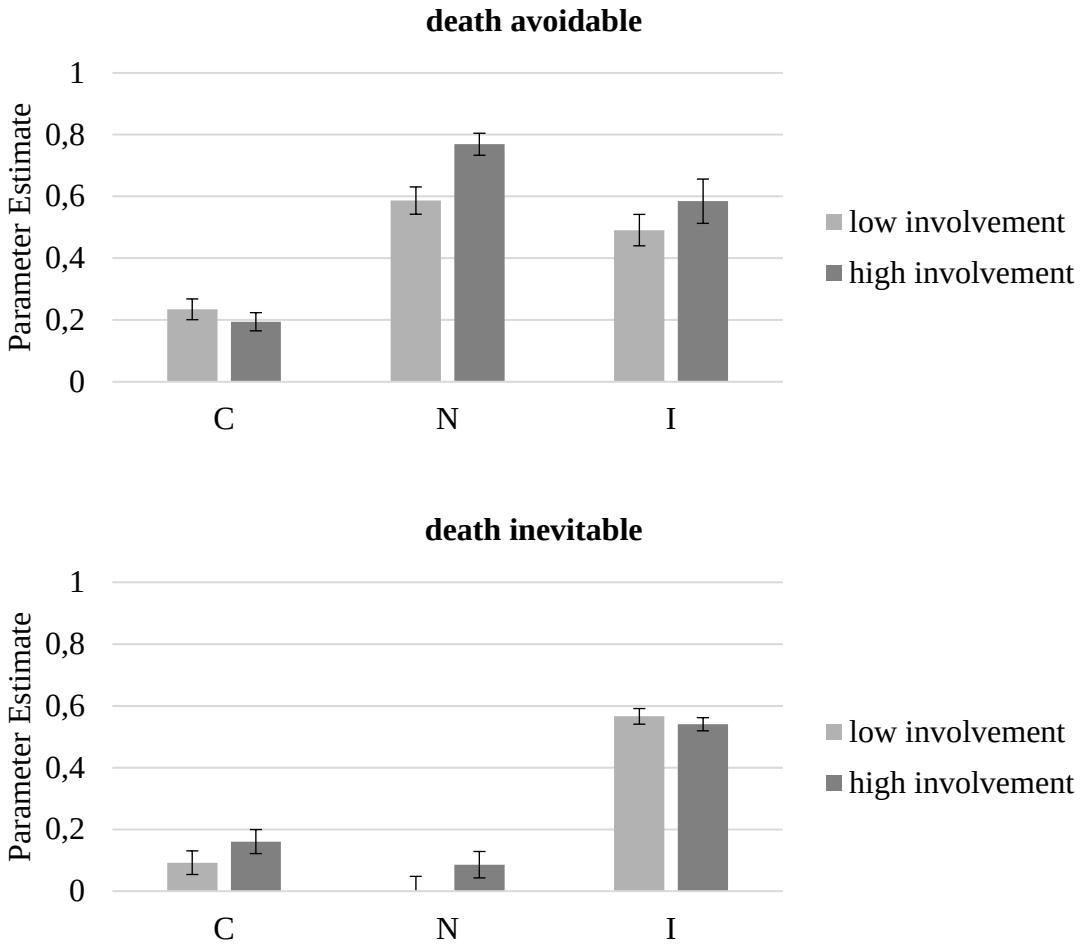

Figure 6. Parameter estimates representing endorsement of consequences (*C*), endorsement of norms (*N*) and inertia (*I*) based on the combined data of Experiments 3a and 3b, separated by personal involvement conditions. Error bars represent 95% confidence intervals.
Note. For death-inevitable scenarios, the model did not fit the data and, hence, parameter estimates should not be interpreted.

III. MPT analyses using a proNCI model

Experiment 1

Results. The general model fit the data well, indicated by the nonsignificant deviation of predicted from the observed response frequencies, $G^2(1) = 1.18$, $p = .277$, $w = 0.056$. The estimate of the *C*-parameter was $C = .61$ (95% *CI* [.49, .73]). The *N*-parameter’s estimate was $N = .31$ (95% *CI* [.22, .39]). The *I*-parameter ($I = .56$, 95% *CI* [.41, .71]) did not differ from 0.5, $\Delta G^2(1) = 0.63$, $p = .427$, $w = .041$, indicating no preference for inertia or change.

Discussion. We will use this section to illustrate some general points regarding MPT analyses. First, note that the model fit, thus the suitability of the measurement model to explain the empirical data, does not differ depending on whether a proCNI or an proNCI model is applied (see Table S1). The size of the *I*-parameter is also not affected by the ordering of *C*- and *N*-parameters as it is estimated conditional on the absence of *C* and *N* in both models. For his reason, we will only report the results regarding *C*- and *N*-parameters in the remainder of this supplement. In summary, analyzing the data of Experiment 1 with a proNCI instead of proCNI model leads to identical results.

Table S1.

Parameter estimates and the proportion of responses explained by parameters for proCNI and proNCI models for Experiment 1

Parameter	Estimate in proCNI model	Estimate in proNCI model	Proportion of responses explained
C	.42	.61	.42
N	.53	.31	.31
I	.56	.56	.15
1-I	.44	.44	.12

Experiment 2a

Results. The model fit the data well, $G^2(2) = 1.21$, $p = .545$, $w = 0.057$. The influence of self-relevance on the estimate of the *C*-parameter was significant, resulting in a lower *C*-parameter in the present ($C_{present} = .45$, 95% *CI* [.26, .65]) than the absent condition ($C_{absent} = .78$, 95% *CI* [.63, .92]), $\Delta G^2(1) = 6.54$, $p = .011$, $w = .131$. The *N*-parameter ($N_{present} = .38$, 95% *CI* [.26, .50], $N_{absent} = .34$, 95% *CI* [.23, .45]) was not affected by self-relevant consequences, $\Delta G^2(1) = 0.23$, $p = .630$, $w < .001$.

Discussion. As estimated with a proNCI model, the presence of self-relevant consequences reduced the estimate of the *C*-parameter while leaving the estimate of the *N*-parameter unaffected. Thus, for Experiment 2a, the proNCI and proCNI models lead to identical results.

Experiment 2b

Results. The model fit the data well, $G^2(2) = 0.57$, $p = .751$, $w = 039$. The self-relevance manipulation exerted an effect on the size of the *N*-parameter, which was smaller in the present *(*$N_{present} = .06$, 95% *CI* [-.07, .19]) than in the absent ($N_{absent} = .29$, 95% *CI* [.17, .41]) condition, $\Delta G^2(1) = 6.44$, $p = .011$, $w = .130$. Furthermore, testing the *N*-parameter in the present condition against 0 revealed no significant reduction in model fit, indicating that participants did not endorse norms in this condition, $\Delta G^2(1) = 1.50$, $p = .682$, $w = .063$. In contrast to the results of the analysis with the proCNI model, when analyzed with an proNCI model the *C*-parameter ($C_{present} = .40$, 95% *CI* [.25, .55], $C_{absent} = .65$, 95% *CI* [.49, .81]) was also affected by self-relevance, $\Delta G^2(1) = 4.65$, $p = .031$, $w = .110$.

Discussion. The presence of self-relevant consequences reduced the strength of the *N*-parameter. In contrast to results obtained with a proCNI model, however, the proNCI model finds a significant effect of self-relevant consequences on the *C*-parameter. However, this apparent discrepancy is attributable to the strong effect of the manipulation on the *N*-parameter. Multinomial modeling is based on a maximum-likelihood estimation procedure. As

a result, when comparing different measurement models, the parameter that is being moved from a dominant to a non-dominant position (e.g., the *C*-parameter when the model is changed from proCNI to proNCI) will increase in size, such that the underlying process still represents the same amount of variance (see Table S2). This change in size corresponds to the size of the now dominant *N*-parameter. That is, the larger the *N*-parameter the larger the increase in the *C*-parameter when models are switched from proCNI to proNCI. Conversely, the lower the *N*-parameter, the lower the increase in the *C*-parameter. As a consequence, when the dominant parameter is strongly affected by a manipulation, this may distort the effect of the manipulation on the non-dominant parameter resulting in a spurious effect, as is the case in this experiment. Thus, when applying the proNCI model, in the absent condition, where the *N*-parameter is larger than zero, the *C*-parameter is increased. In the present condition, however, where the *N*-parameter is equal to zero, the *C*-parameter remains the same. In conjunction, a significant effect on the *C*-parameter does arise as an artefact.

Table S2.

Parameter estimates and the proportion of responses explained by parameters for proCNI and proNCI models for Experiment 2b

Parameter	Estimate in proCNI model	Estimate in proNCI model	Proportion of responses explained
$C_{present}$	.38	.40	.38
$N_{present}$	.10	.06	.06
$I_{present}$	.58	.58	.32
$1\text{-}I_{present}$	.42	.42	.23
C_{absent}	.46	.65	.46
N_{absent}	.54	.29	.29
I_{absent}	.53	.53	.13
$1\text{-}I_{absent}$	.47	.47	.12

Thus, in this case the results of the proCNI model analysis is more reliable, because the effect on the *C*-parameter in the proNCI model is an identifiable artefact. Note that out of all of the hypothesis tests reported in the main article, this is the only case where model choice influences whether a non-significant effect becomes significant or vice versa. Also, the influence of self-relevant consequences on the *N*-parameter found with the proCNI model is replicated in this analysis.

Experiment 3a

Overall model. The overall model containing one *C*-parameter (C = .23, 95% *CI* [.20, .27]), one *N*-parameter (N = .28, 95% *CI* [.26, .31]), and one *I*-parameter provided a good fit to the data, $G^2(1) = 0.01$, $p = .932$, $w = 0.001$.

Self-relevant consequences. The fit of the model estimating *C*-, *N*-, and *I*-parameters separately for self-relevance present and absent conditions did not deviate significantly from the observed category frequencies, $G^2(2) = 3{,}89$, $p = .143$, $w = 0.027$. Equating the *C*-parameters across self-relevance conditions caused a significant decrease in model fit, $\Delta G^2(1) = 60.35$, $p < .001$, $w = 0.104$, indicating that aggregate consequences played a larger role for judgment in the self-relevance present ($C_{present}$ = .36, 95% *CI* [.31, .40]) than in the self-relevance absent condition (C_{absent} = .08, 95% *CI* [.02, .13]). The *N*-parameters also differed between conditions, $\Delta G^2(1) = 50.67$, $p < .001$, $w = 096$, indicating stronger endorsement of norms in the absent (N_{absent} = .37, 95% *CI* [.34, .41]) than in the present condition ($N_{present}$ = .19, 95% *CI* [.16, .23]).

Personal involvement. The model considering *C*-, *N*-, and *I*-parameters separately per personal involvement conditions fit the data well, $G^2(2) = 0.62$, $p = .734$, $w = 0.011$. Setting the parameters equal across personal involvement conditions revealed a significant effect of the manipulation on the *N*-parameter, such that norms guided participants’ judgments more when the killing required high levels of personal involvement (N_{high} = .40, 95% *CI* [.37, .43]), than when it allowed for low levels of personal involvement (N_{low} = .17, 95% *CI* [.13, .20]),

$\Delta G^2(1) = 84.49$ $p < .001$, $w = 0.124$. There was a marginal effect of personal involvement on the *C*-parameter ($C_{high} = .27$, 95% *CI* [.22, .33], $C_{low} = 0.21$, 95% *CI* [.16, .25]), $\Delta G^2(1) = 3.58$, $p = .058$, $w = 0.025$.

Avoidability. The model estimating separate *C*-, *N*-, and *I*-parameters for the death-avoidable and death-inevitable conditions fit the data well, $G^2(2) = 0.13$, $p = .935$, $w = 0.005$. The avoidability manipulation affected the estimate of the *C*-parameter such that sensitivity to aggregate consequences was larger when the death of the victim was avoidable ($C_{avoidable} = .46$, 95% *CI* [.40, .53]) rather than inevitable ($C_{inevitable} = .14$, 95% *CI* [.10, .17]), $\Delta G^2(1) = 68.35$, $p < .001$, $w = 0.111$. However, the avoidability manipulation also had an effect on the *N*-parameter, $\Delta G^2(1) = 506.37$, $p < .001$, $w = 0.302$, such that responses endorsed norms more strongly when the death of the person to be sacrificed was avoidable ($N_{avoidable} = .56$, 95% *CI* [.53, .59]) rather than inevitable ($N_{inevitable} = .00$, 95% *CI* [.-.03, .04]). Furthermore, setting the *N*-parameter in the death-inevitable condition to zero did not reduce model fit, indicating that responses were not characterized by norm endorsement in this condition, $\Delta G^2(1) = 0.02$, $p = .880$, $w = 0.002$.

Discussion. All the effects of the proCNI model were replicated. Specifically, the presence of self-relevant consequences reduced the endorsement of norms and increased sensitivity to aggregate consequences. When personal involvement was high, the endorsement of norms was increased, while sensitivity to aggregate consequences was not significantly (although marginally) affected. In the case of inevitable death the endorsement of norms and consequences were reduced alike, and norms were not endorsed anymore. Thus, for Experiment 3a, results are identical across the two model versions.

Experiment 3b

Overall model. The overall model containing one *C*-parameter (*C* = .26, 95% *CI* [.22, .29]), one *N*-parameter (*N* = .21, 95% *CI* [.18, .24]), and one *I*-parameter provided a good fit to the data, $G^2(1) = 1.03$, $p = .310$, $w = 0.015$.

Self-relevant consequences. The fit of the model estimating *C*-, *N*-, and *I*-parameters separately for present and absent conditions did not deviate significantly from the observed category frequencies, $G^2(2) = 1.75$, $p = .417$, $w = 0.020$. Equating the *C*-parameters across self-relevance conditions caused a significant decrease in model fit, $\Delta G^2(1) = 90.24$, $p < .001$, $w = 0.140$, indicating that aggregate consequences played a larger role for judgment in the present ($C_{present}$ = .42, 95% *CI* [.37, .47]) than in the absent condition (C_{absent} = .08, 95% *CI* [.03, .13]). The *N*-parameters was not affected by the presence of self-relevant consequences, (N_{absent} = .22, 95% *CI* [.18, .26], $N_{present}$ = .19, 95% *CI* [.16, .23]), $\Delta G^2(1) = 1.28$, $p = .257$, $w =$ 017.

Personal involvement. The model considering *C*-, *N*-, and *I*-parameters separately per personal involvement conditions fit the data well, $G^2(2) = 1.19$, $p = .551$, $w = 0.016$. Setting the parameters equal across personal involvement conditions revealed a significant effect of the manipulation on the *N*-parameter, such that norms were endorsed more strongly when the killing required high levels of personal involvement (N_{high} = .56, 95% *CI* [.52, .61]), than when it allowed for low levels of personal involvement (N_{low} = .43, 95% *CI* [.38, .48]), $\Delta G^2(1)$ = 14.41 $p < .001$, $w = 0.055$. There was no effect of personal involvement on the *C*-parameter (C_{high} = .49, 95% *CI* [.39, .59], C_{low} = 0.44, 95% *CI* [.36, .53]), $\Delta G^2(1) = 0.47$, $p = .492$, $w =$ 0.010.

Avoidability. The model estimating separate *C*-, *N*-, and *I*-parameters for the death-avoidable and death-inevitable conditions did not show sufficient fit to the data, $G^2(2) =$ 17.40, $p < .001$, $w = 0.061$. In this model, the avoidability manipulation did exert a significant effect on the *C*-parameter, resulting in a stronger endorsement of consequences in the

avoidable ($C_{avoidable}$ = .46, 95% *CI* [.40, .52]) compared to the inevitable ($C_{inevitable}$ = .17, 95% *CI* [.13, .21]) condition, $\Delta G^2(1) = 50.43$, $p < .001$, $w = 0.105$. The avoidability manipulation also had an effect on the *N*-parameter, $\Delta G^2(1) = 422.30$, $p < .001$, $w = 0.303$, such that responses were more strongly characterized by norm endorsement when the death of the person to be sacrificed was avoidable ($N_{avoidable}$ = .50, 95% *CI* [.47, .53]) rather than inevitable ($N_{inevitable}$ = .00, 95% *CI* [-.04, .04]). Furthermore, setting the *N*-parameter in the death-inevitable condition to zero did not reduce model fit, indicating that norms were not endorsed in this condition, $\Delta G^2(1) = 0.00$, $p = 1$, $w = 0.000$.

Again, we also explored the source of model misfit. Given that the *N*-parameter in the death-inevitable condition was equal to the lower bound of the parameter space, we reversed the *N*-parameters coding in that condition to investigate the possibility of reversed effects. Note that this specification only influences estimates of the *N*-parameter and the according change in model-fit, while leaving the parameter estimates of *C*- (and *I*-) parameters reported above unaffected. With this modification, the model indeed provided a good fit to the data, $G^2(1) = 1.38$, $p = .507$, $w = 0.017$. The *N*-parameter in this condition was estimated at $N_{inevitable}$ = .08, 95% *CI* [.04, .12]), indicating that participants rejected norm endorsement, $\Delta G^2(1) = 16.04$, $p < .001$, $w = 0.059$.

Discussion. Virtually all effects obtained with the proCNI model were replicated, and no additional effects were found. The presence of self-relevant consequences reduced the endorsement of norms and increased sensitivity to aggregate consequences. When personal involvement was high, responses were more frequently consistent with norms, while endorsement of aggregate consequences remained unaffected. Thus, the marginal effect of personal involvement on the *C*-parameter found in Experiment 3a did not replicate. In the case of inevitable death, participants did not distinguish between minor and major consequences anymore. Furthermore, the endorsement of norms was reduced to zero. The additional analysis exploring the misfit of the avoidability model also yielded the same

results. Reversing the *N*-parameter alleviated the lack of fit, and indicated that participants rejected norms.

IV. PD analyses

Additionally, we investigated the fit of the PD model (Conway & Gawronski, 2013) to our data. This model, however, can be interpreted in two different ways, both of which we investigated. As Conway and Gawronski (2013) specify the model, dilemma responses are assumed to be comprehensively explained by “deontological” (*N*) and “utilitarian” (*C*) response patterns. That is, the model assumes that when neither of these processes determines responses, participants always engage in *sacrificial killing*. Such a model thus contains the parameters *C* and *N* (Specification 1). Remember, however, that the PD model systematically equates “deontological” responding with inertia, and “utilitarian” responding with interference. That is, as the PD model assumes that inertia does not determine dilemma responses, the final assumption of the PD model could also be expressed as when neither of these processes determines responses, participants always engage in *interference*. Such a model would thus contain the parameters *C*, *N*, and an *I*-parameter equal to zero (Specification 2). Table S3 provides the results of both PD models.

Table S3.

Parameter Estimates and Fit Statistics for Both Specifications of the PD Model for Each Experiment

	Specification 1			Specification 2		
Exp.	**C**	**N**	**fit**	**C**	**N**	**fit**
1	.42	.76	$G^2(2) = 1.81$, $p = .404$, $w = 0.070$	.46	.65	$G^2(2) = 566.60$, $p < .001$, $w = 1.228$
2a	.39	.80	$G^2(4) = 4.29$, $p = .368$, $w = 0.106$	.43	.71	$G^2(4) = 540.11$, $p < .001$, $w = 1.186$
2b	.42	.65	$G^2(4) = 2.26$, $p = .687$, $w = 0.077$	.48	.53	$G^2(4) = 828.46$, $p < .001$, $w = 1.469$
3a	.17	.67	$G^2(2) = 8.08$, $p = .018$, $w = 0.038$	.30	.57	$G^2(2) = 14599.79$, $p < .001$, $w = 1.623$
3b	.20	.63	$G^2(2) = 10.59$, $p = .005$, $w = 0.048$	.34	.52	$G^2(2) = 12895.32$, $p < .001$, $w = 1.671$

Note. Specification 1 is fitted to a data set that was aggregated across default-state conditions. Specification 2 is fitted to the full data set accounting for default-state condition.

The size of the *C*- and *N*-parameters heavily depend on how the model deals with response tendencies. If the model assumes that participants kill whenever neither endorsement of consequences, nor endorsement of norms drive responses (first specification), the *C*-parameter is smaller than when the model assumes that participants act whenever neither endorsement of consequences, nor endorsement of norms drive responses (second specification). The *N*-parameter shows the reversed pattern. These patterns highlight that the parameter estimates can be distorted when response tendencies are not accounted for. The advantage of the proCNI model is that it freely estimates response tendencies rather than fixing them at a specific level (e.g., 1 - *N* in Specification 1 and $I = 0$ in Specification 2). Moreover, the proCNI model achieves good fit to all of our data. As the analysis of the two variants of the PD model shows, none of these specifications offers a good description of the data across all experiments. Finally, this comparison also illustrates that research into moral dilemma judgments needs to pay attention to the way in which response tendencies can affect results and their interpretation.

Appendix D (for Chapter IV)

Scenarios Implemented in Experiments 4a and 4b

For the construction of these scenarios, we adhered to a well-structured approach that is outlined in Appendix B of Hennig & Hütter (2019). The scenarios presented in Tables A1, A2, A5, and A6 were already used in Hennig & Hütter (2019), scenarios A3, A4, A7, A8, and A9 were created for the studies reported in Chapter IV, and were in part inspired by or adapted from the work of Moore et al. (2008).

Egoistic incentive present / death avoidable

Table A1.

Orphanage

High Involvement		**Low Involvement**	
Incongruent	**Congruent**	**Incongruent**	**Congruent**
You work in an orphanage in a war-torn country. Rebel soldiers have staged a successful takeover of the government. The soldiers have entered your town and are …			
killing …	dispossessing …	killing …	dispossessing …
all those who did not support their regime, or whose loyalty is in question. You are furthermore aware that the soldiers want to kill everyone of a particular ethnic group in the country, as part of a goal-directed genocide. You have one child of this ethnic group in your orphanage. It will shortly be picked up by some relatives, whereby it would escape the soldiers. However, because contrary to the relatives and the soldiers the orphanage does not own a vehicle, there is no way for you and the other children and workers in the orphanage to escape. You know that the soldiers will…			
kill everyone they find in the orphanage.	dispossess everyone they find in the orphanage.	kill everyone they find in the orphanage.	dispossess everyone they find in the orphanage.
You reach the conclusion that you could smother the child, which belongs to the ethnic group focused on by the soldiers, with a pillow while it is sleeping, …		You reach the conclusion that you could instruct a nurse to overdose the child, which belongs to the ethnic group focused on by the soldiers, on sleeping pills, …	
before its relatives arrive to pick it up. Since the soldiers hate the ethnic group it is a part of, they would see this as supporting their regime. This way you could make sure that the soldiers would spare you and the other children and workers instead of…			
killing you, although this would necessitate the death of the child.	dispossessing you, although this would	killing you, although this would necessitate the death of the child.	dispossessing you, although this would

		necessitate the death of the child.		necessitate the death of the child.
	If you do not smother the child, though, the soldiers will …		If the child is not overdosed on sleeping pills, though, the soldiers will …	
	kill you and the other workers and children in the orphanage.	dispossess you and the other workers and children in the orphanage.	kill you and the other workers and children in the orphanage.	dispossess you and the other workers and children in the orphanage.
	In order to avoid that you yourself and the other workers and children are …			
	killed …	dispossessed …	killed …	dispossessed …
	by the rebels, you would have to smother the child to death with a pillow.		by the rebels, you would have to instruct a nurse to overdose the child on sleeping pills.	
Inaction Default	You have not yet attempted anything of this sort and are outside the room of the child. How would you behave in this situation? Would you get a pillow, enter the room of the child, and smother it with the pillow while it is sleeping?		You have not yet attempted anything of this sort and are outside the hospital ward. How would you behave in this situation? Would you enter the hospital ward, talk to the nurse, and instruct her to overdose the child on sleeping pills?	
Action Default	You are already inside the room of the child and stand with a raised pillow next to its bed, in order to smother it. How would you behave in this situation? Would you abort your plan, take down the pillow, and leave the room of the child again?		You have already entered the hospital ward and talked to the nurse, in order to instruct her to overdose the child on sleeping pills. How would you behave in this situation? Would you abort your plan, end the conversation prematurely, and leave the hospital ward again?	

Table A2.

Ebola

	High involvement		**Low involvement**	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	You are a Peace Corps health-worker who is volunteering in a rural African village. A man from a nearby village has contracted a dangerous virus that is extremely contagious and incurable. It is known to you that, apart from causing other mild symptoms, the virus regularly leads to …			
	death within a week. Miraculously, though he is showing the mild symptoms, the man has been spared of death for a month, and so he must be a rare carrier who is immune to the deadly effects of the virus.	several days of weakness and extreme nausea within a week. Miraculously, though he is showing the mild symptoms, the man has been spared of weakness and nausea for a month, and so he must be a rare carrier who is immune to	death within a week. Miraculously, though he is showing the mild symptoms, the man has been spared of death for a month, and so he must be a rare carrier who is immune to the deadly effects of the virus.	several days of weakness and extreme nausea within a week. Miraculously, though he is showing the mild symptoms, the man has been spared of weakness and nausea for a month, and so he must be a rare carrier who is immune to

		those effects of the virus.		those effects of the virus.
	However, he wrongly believes that your health center can cure his symptoms and approaches the village, without reacting to shouting or gesturing. Should he enter the village, he would pass the virus on to you and the other villagers so that, unlike him, all of you would …			
	die.	suffer several days of weakness and extreme nausea.	die.	suffer several days of weakness and extreme nausea.
	It comes to your mind that there is a loaded gun in the health center, which you could use to shoot him. This would stop him from entering the village. However, because due to the risk of infection nobody could treat his injuries, he would die from the bullet wound. If you do not shoot him, however, he will pass the virus on to you and the rest of the village.		It comes to your mind that you could warn the village council, which would lead to one of the members shooting the man. This would stop him from entering the village. However, because due to the risk of infection nobody could treat his injuries, he would die from the bullet wound. If you do not warn the village council, however, he will pass the virus on to you and the rest of the village.	
	In order to avoid that you yourself and the other villagers …			
	die from the virus, …	suffer several days of weakness and extreme nausea from the virus, …	die from the virus, …	suffer several days of weakness and extreme nausea from the virus, …
	you would have to shoot the man before he enters the village.		you would have to warn the village council before he enters the village.	
Inaction Default	You are standing outside the health center and have not yet picked up the gun that is located inside. How would you behave in this situation? Would you get the gun, level it, and shoot the man?		You have not yet informed the village council, in front of whose meeting place you are situated. How would you behave in this situation? Would you enter the meeting place, talk to the members of the village council, and announce the information?	
Action Default	You have already gotten the gun and levelled it outside, in order to shoot the man. How would you behave in this situation? Would you abort your plan, take the gun down, and bring it back to the health center?		You have already entered the meeting place of village council and started talking to the members, in order to announce the information. How would you behave in this situation? Would you abort your plan, end the conversation prematurely, and leave the meeting place again?	

Table A3.

In the foxhole

	High involvement		**Low involvement**	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	You are the leader of a group of five soldiers in a country torn apart by civil war. Your group has become engaged in a firefight with hostile soldiers, in a mountainous area. Because their weaponry is superior to yours, you were pushed back and forced to take refuge in a cave. Your enemies decisively outnumber you and by now block the only exit, such that you know that there is no way for you to escape. Suddenly the leader of the other group makes you an offer. Your entire group would be allowed to retreat unharmed, under the condition that you would kill one of your enemies’ prisoners by …			

	cutting his throat.		shooting him from a distance.	
	Otherwise, they will use their heavy weaponry to bring down the only exit, before leaving you behind locked in the cave. You know that the leader of the other group is bound to his word as a soldier, and that he will hold up his end of the deal, as well as accept whatever decision you make. By taking up the offer and …			
	cutting the throat of the prisoner, …		shooting the prisoner from a distance, …	
	you would make sure that you and your group would be able to retreat. However, this would also mean the death of the prisoner. If you do not …			
	cut his throat, …		shoot him from a distance, …	
	though, your enemies’ heavy weaponry will bring down the exit to the cave. Because you are low on supplies and the cave is located in a remote part of the country, this would cause you yourself and the entire group to …			
	starve to death …	suffer some days of hunger …	starve to death …	suffer some days of hunger …
	before you will manage to dig your way out of the cave and make it back to your camp. In order to avoid that you yourself and the other soldiers …			
	starve to death, you would have to …	suffer some days of hunger, you would have to …	starve to death, you would have to …	suffer some days of hunger, you would have to …
	cut the throat of this prisoner.		shoot the prisoner from a distance.	
Inaction Default	You are currently inside the cave and have not answered to the offer yet. How would you behave in this situation? Would you agree to the offer, walk outside the cave, and cut the prisoner’s throat with a knife?		You are currently inside the cave and have not answered to the offer yet. How would you behave in this situation? Would you agree to the offer, walk outside the cave, and shoot the prisoner from a distance?	
Action Default	You have already agreed to the offer and walked outside the cave, in order to cut the prisoner’s throat with a knife. How would you behave in this situation? Would you stop, walk back into the cave, and withdraw your agreement to the offer?		You have already agreed to the offer and walked outside the cave, in order to shoot the prisoner from a distance. How would you behave in this situation? Would you stop, walk back into the cave, and withdraw your agreement to the offer?	

Table A4.

Crevasse Climbing

	High Involvement		**Low Involvement**	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	You and four other people are climbing a crevasse as a rope team. Suddenly the rope becomes wedged together across an ice ledge between the lead climber and you as the second climber of the rope team. This wedging makes it impossible for you and the three climbers behind you to continue climbing. The lead climber decides to secure the rope with an ice axe, to release himself, and to get help from the next rescue station. You know that this station is several miles off and that, before the rescue team arrives, you and the other three trapped climbers behind you will …			
	be frozen to death.	develop painful chilblains, which will last for some time.	be frozen to death.	develop painful chilblains, which will last for some time.
	The lead climber has already released himself from the rope, being currently unsecured. You come to the conclusion that you could …			
	tear at the leg of the lead climber, …		shoot at the lead climber with your flare gun, …	

	… , which would cause him to crash into the ice ledge. As a result, the ice ledge would break off, freeing you and the other three trapped climbers. The lead climber, however, would also die as a result of the crash.			
	If you do not …			
	tear at the leg of the lead climber, …		shoot at the lead climber with your flare gun, …	
	though, you and the other three climbers will remain trapped and exposed to the cold. In order to avoid that you yourself and the other three trapped climbers …			
	freeze to death, …	suffer from painful chilblains for some time, …	freeze to death, …	suffer from painful chilblains for some time, …
	you would have to tear at the leg of the lead climber.		you would have to shoot at the lead climber with your flare gun.	
Inaction Default	You are still clasping the rope and have not yet set one hand free. How would you behave in this situation? Would you set one hand free, grasp the leg of the lead climber, and tear at it?		You are still clasping the rope and have not yet set one hand free. How would you behave in this situation? Would you set one hand free, grasp your flare gun, and shoot at the lead climber?	
Action Default	You have already set one hand free from the rope and grasped the leg of the lead climber, in order to tear at it. How would you behave in this situation? Would you abort your plan, let go of the leg of the lead climber, and clasp the rope with both hands again?		You have already set one hand free from the rope and grasped your flare gun, in order to shoot at the lead climber. How would you behave in this situation? Would you abort your plan, put your flare gun away, and clasp the rope with both hands again?	

Egoistic incentive absent / death avoidable

Table A5.

Transplant

	High involvement		**Low involvement**	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	You are a doctor in a hospital. You have five patients, each of whom is …			
	about to die due to a failing organ, …	going to suffer from irreversible physical impairments like recurrent feelings of weakness, chronic pain or episodes of dizziness for the rest of their lives due to organ damage, …	about to die due to a failing organ, …	going to suffer from irreversible physical impairments like recurrent feelings of weakness, chronic pain or episodes of dizziness for the rest of their lives due to organ damage, …
	unless they have donor organs implanted within a very short period of time. A new patient is rushed into the hospital after a serious car accident. This patient is severely wounded and in need of immediate surgery. While performing a long and complicated surgery on the victim of the accident you realize that you could …			
	inconspicuously cut through one of his arteries, without anyone ever being able to prove that you …		instruct a nurse to administer the wrong medication to him, without anyone ever being able to prove that you or the nurse …	

	did anything on purpose. Hereby you would receive organ donations for the other five patients, the victim of the accident would die, though. If you do not cause the death of the accident victim and then transplant his organs, however, the other five patients will …			
	die.	live the rest of their lives with irreversible impairments.	die.	live the rest of their lives with irreversible impairments.
	In order to avoid that the other five patients die due to failing organs, …	In order to avoid that the other five patients will spend the rest of their lives with recurrent feelings of weakness, chronic pain, fits of dizziness or similar impairments due to organ damage, …	In order to avoid that the other five patients die due to failing organs, …	In order to avoid that the other five patients will spend the rest of their lives with recurrent feelings of weakness, chronic pain, fits of dizziness or similar impairments due to organ damage, …
	you would have to cut through one of the arteries of the accident victim, so that you could transplant his organs after his death.		you would have to instruct a nurse to administer the wrong medication to the accident victim, so that you could transplant his organs after his death.	
Inaction Default	You are currently performing the surgical operation and have until now followed the prescribed procedure. How would you behave in this situation? Would you deviate from the prescribed procedure, lead your scalpel to one of the big arteries of the accident victim, and cut it inconspicuously?		You are currently performing the surgical operation and have until now followed the prescribed procedure. How would you behave in this situation? Would you deviate from the prescribed procedure, start talking to a nurse, and instruct her to administer the wrong medication to the accident victim?	
Action Default	You are currently performing the surgical operation, have already deviated from the prescribed procedure and have led your scalpel to one of the big arteries of the accident victim, in order to cut it inconspicuously. How would you behave in this situation? Would you abort your plan, remove your scalpel from the artery, and return to the prescribed procedure?		You are currently performing the surgical operation, have already deviated from the prescribed procedure and have started talking to a nurse, in order to instruct her to administer the wrong medication. How would you behave in this situation? Would you abort your plan, end the conversation prematurely, and return to the prescribed procedure?	

Table A6.

Nobel Prize

	High involvement		**Low involvement**	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	You and a fellow researcher are working in a small laboratory on the development of new technologies. You have recently developed a powerful energy converter, which you consider advanced and efficient. Irrespective of potential other areas of application for this converter, your colleague, who is in charge of the research project, has already decided to sell the converter to a local arms dealer and split the profit evenly between the two of you. You know that he plans to contact the buyer, who would use the energy converter for the manufacturing of a weapon, through which several people would be …			
	killed, today.	irreversibly injured, today.	killed, today.	irreversibly injured, today.
	You could poison him by bringing him in contact with a chemical, you are in possession of, and which leads to immediate death in the case of		You could release an odorless gas in the laboratory, which would explode, when your colleague turns on the gas burner. Because this	

	direct skin contact. Because this chemical is normally found in laboratories like yours, everybody would think that it was just a lab accident. You merely would have to trickle a drop of the chemical on his skin in an unobserved moment. This way you would ensure that the energy converter is not used as a weapon, though you would also cause the death of your colleague. If you do not poison your colleague, however, he will sell the energy converter to the arms dealer, whereby several people will …		gas is normally found in laboratories like yours, everybody would think that it was just a lab accident. You merely would have to turn open the container with the gas. This way you would ensure that the energy converter is not used as a weapon, though you would also cause the death of your colleague. If you do not release the flammable gas, however, he will sell the energy converter to the arms dealer, whereby several people will …	
	lose their lives. In order to avoid that due to the sale of the converter to the arms dealer several people die, …	suffer injuries. In order to avoid that due to the sale of the converter to the arms dealer several people are irreversibly injured, …	lose their lives. In order to avoid that due to the sale of the converter to the arms dealer several people die, …	suffer injuries. In order to avoid that due to the sale of the converter to the arms dealer several people are irreversibly injured, …
	you would have to poison your colleague.		you would have to release the flammable gas.	
Inaction Default	You have not undertaken anything yet and are currently in the laboratory close to the container with the deadly chemical. How would you behave in this situation? Would you put on protective gloves, inconspicuously approach your colleague with a small amount of the chemical, and trickle a drop of it on his skin?		You have not undertaken anything yet and are currently alone in the laboratory close to the container with the flammable gas. How would you behave in this situation? Would you approach the container, turn open the valve, and release the gas in the laboratory?	
Action Default	You are currently in the laboratory, have already put on protective gloves and have inconspicuously approached your colleague with a small amount of the chemical, in order to trickle a drop of it on his skin. How would you behave in this situation? Would you abort your plan, distance yourself from your colleague, and dispose of the chemical and the protective gloves again?		You are currently alone in the laboratory, have already approached the container with the flammable gas and grabbed the valve, in order to release the gas in the laboratory. How would you behave in this situation? Would you abort your plan, let go of the valve, and distance yourself from the container again?	

Table A7.

Modified vaccine

	High involvement		**Low involvement**	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	A viral epidemic has spread across the globe causing …			
	the death of …	severe stomach cramps in …	the death of …	severe stomach cramps in …
	many people. You are a medical researcher and have manufactured a substance in your laboratory, which could be used to develop a cure for the epidemic. Although your research has advanced well, you know that you will have to conduct tests investigating how the substance reacts when it comes into contact with a living human organism, in order to be able to finalize the antidote. At the current stage of development, however, this would also cause the death of the person, on which it is tested. Once you figure out how to finalize the substance, you will be able to create more of it and consequently prevent that multiple people will…			

	die.	suffer from severe temporary stomach cramps.	die.	suffer from severe temporary stomach cramps.
	You come to the conclusion that you could forcibly inject some of the substance into the body of your lab assistant, who is working with you.		You come to the conclusion that you could pour some of the substance into the coffee cup of your lab assistant, who is working with you.	
	This would enable you to conduct the final tests on the substance and manufacture an antidote. However, your lab assistant would die from getting in contact with the substance. If you do not …			
	inject the substance into your lab assistant, however, people will continue to die due to the epidemic. In order to avoid that people keep dying …	inject the substance into your lab assistant, however, people will continue to temporarily suffer from severe stomach cramps due to the epidemic. In order to avoid that people keep temporarily suffering from severe stomach cramps …	pour the substance into the coffee cup of your lab assistant, however, people will continue to die due to the epidemic. In order to avoid that people keep dying …	pour the substance into the coffee cup of your lab assistant, however, people will continue to temporarily suffer from severe stomach cramps due to the epidemic. In order to avoid that people keep temporarily suffering from severe stomach cramps …
	because of the epidemic, you would have to inject some of the substance into your lab assistant.		because of the epidemic, you would have to pour some of the substance into the coffee cup of your lab assistant.	
Inaction Default	You are currently at work in the laboratory and several meters away from your lab assistant, without having done anything of this sort. How would you behave in this situation? Would you take a syringe, approach your lab assistant from behind, and inject the substance into his body against his will?		You are currently at work in the laboratory and several meters away from your lab assistant, without having done anything of this sort. How would you behave in this situation? Would you take a syringe, approach your lab assistant‘s coffee cup, and unobtrusively pour the substance into it?	
Action Default	You have already taken a syringe and approached your lab assistant from behind, in order to inject the substance into his body against his will. How would you behave in this situation? Would you stop what you are doing, move away from your lab assistant, and dispose of the syringe again?		You have already taken a syringe and approached your lab assistant’s coffee cup, in order to unobtrusively pour the substance into it. How would you behave in this situation? Would you stop what you are doing, move away from your lab assistant’s coffee cup, and dispose of the syringe again?	

Egoistic incentive absent / death avoidable

Table A8.

Police spy

	High Involvement		**Low Involvement**	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	You are a police spy investigating a radical underground organization. This underground organization fights against a group of activists that they consider their enemies. To you it is well known that the underground organization is planning an assault on the activist group. Since the underground organization is highly prone to violence, the assault will …			
	kill …	injure …	kill …	injure …

	several members of the activist group. At the current stage, you are lacking precise information about the upcoming assault and are therefore unable to prevent it. You just know that it is going to happen in a couple of days. You decide to join the underground organization as an undercover investigator, in order to get access to a storage medium with internal information about the planned assault. Since the leader of the underground organization is sceptical towards new members, he demands a sign of loyalty. You could accept his demand, which consists in …			
	strangling one of the activists.		shooting one of the activists.	
	This activist is about to leave the country for some weeks and would therefore not be affected by the assault. This would persuade the leader of the underground organization of your loyalty, enabling you to get access to the storage medium. However, it would also mean the death of the activist that is about to leave the country.			
	If you do not strangle the activist, though, …		If you do not shoot the activist, though, …	
	the assault on the activist group will be executed as planned. In order to avoid that several members of the activist group are…			
	killed …	injured …	killed …	injured …
	in the assault, you would have to strangle the activist that is about to leave the country.		in the assault, you would have to shoot the activist that is about to leave the country.	
Inaction Default	You are currently observing the activist, who has not noticed you, from inside your car. How would you behave in this situation? Would you wait for an unobserved moment, approach the activist, and strangle him?		You are currently observing the activist, who has not noticed you, from inside your car. How would you behave in this situation? Would you wait for an unobserved moment, find a safe position, and shoot him from there?	
Action Default	You have already waited for an unobserved moment and approached the activist, in order to strangle him. How would you behave in this situation? Would you abort your plan, distance yourself from the activist, and get back to your car again?		You have already waited for an unobserved moment and found a safe position, in order to shoot the activist from there. How would you behave in this situation? Would you abort your plan, leave your position, and get back to your car again?	

Table A9.

Trolley/Footbridge

	High involvement		**Low involvement**	
	Incongruent	**Congruent**	**Incongruent**	**Congruent**
	A runaway trolley is heading down the tracks toward five people who it will collide with if it proceeds on its present course. Due to the size and weight of the trolley, this would cause …			
	the death of …	injuries to …	the death of …	injuries to …
	those five people. You are standing next to the track on which the trolley is travelling, but you are too far away from the people to warn them of the impending danger.			
	Next to you there is a very heavy stranger minding his own business, who you could push onto the tracks in front of the trolley.		Next to you there is a control switch for the tracks, which you could use to reroute the trolley onto another part of the track.	
	This would stop the trolley from continuing on its current course. However, it would also mean the death …			
	of the heavy man. If you do not push the heavy man in front of the trolley, …		of another man, who is on the track the trolley would be diverted on and who would be overrun as a result. If you do not flip the switch, …	
	though, the trolley will continue on its present course, resulting in …			

	the death of …	injuries for …	the death of …	injuries for …
	the five people. In order to avoid that the five people on the end of the track are hit and …			
	killed, …	injured, …	killed, …	injured, …
	you would have to push the heavy man in front of the trolley.		you would have to flip the switch.	
Inaction Default	You are currently standing close to the heavy man, without having walked further towards him. How would you behave in this situation? Would you walk towards the heavy man, extend your arms, and push him in front of the trolley?		You are currently standing close to the switch, without having walked further towards it. How would you behave in this situation? Would you walk towards the switch, extend your arms, and flip it?	
Action Default	You have already walked further towards the heavy man and extended your arms, in order to push him in front of the trolley. How would you behave in this situation? Would you stop the action, take down your arms, and get away from the heavy man again?		You have already walked further towards the switch and extended your arms, in order to flip it. How would you behave in this situation? Would you stop the action, take down your arms, and get away from the switch again?	