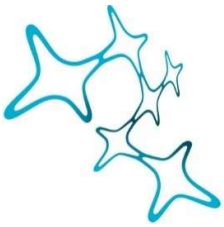
Moral Decisions in (and for) Groups
A Collective Approach

Anita Keshmirian



Graduate School of
Systemic Neurosciences
LMU Munich

Dissertation der Graduate School of Systemic Neurosciences der Ludwig-Maximilians-Universität München

5 May 2021

Reviewers

First Reviewer:
Prof. Dr. Ophelia Deroy
Chair of Philosophy of Mind
Ludwig–Maximilians–Universität München
Faculty of Philosophy, Philosophy of Science and the Study of Religion

Second Reviewer:
Prof. Dr. Stephan Sellmaier
Research Center for Neurophilosophy and Ethics of Neurosciences
Ludwig–Maximilians–Universität München
Faculty of Philosophy, Philosophy of Science and the Study of Religion

External Reviewer:
Prof. Dr. Morteza Dehghani
Brain and Creativity Institute (BCI)
University of Southern California
Department of Psychology

Date of Defense
21 July 2021

Acknowledgment

I would like to thank my supervisors, Prof. Dr. Ophelia Deroy, Prof. Dr. Bahador Bahrami, Prof. Dr. Stephan Sellmaier, and Prof. Dr. Fiery Cushman, for their support, advice, help, and patience through my Ph.D., as well as my collaborators, Babak Hemmatian and Sofia Bonicalzi for their contributions to the project.

I am profoundly thankful to the Graduate School of Neuroscience and CVBE for offering me a new family here in Munich, providing constant support and help throughout my Ph.D.

I am greatly grateful to my family and friends, in particular, to my mother Azar Moayedi and uncle Jay Keshmirian for their ever-continuing advice and sympathetic ears.

I thank Babak, Arash, Sofia, Azadeh, and Elham, and Nora for their moral support during the challenging days of my Ph.D.

I would like to express my very great appreciation to my husband, Mohsen Hamvardpour as he has been a constant source of support to me. This work is dedicated to him.

Summary

Moral cognition has a prominent social nature. We routinely discuss moral issues in our social groups, deliberate together on moral grounds, and collectively make moral decisions and judgments. This thesis investigates moral decisions and judgments in collective contexts in and for human groups.

In chapter 1, I review the relation between collective cognition and moral cognition. Defining the scope of this work to carve out its territory, I show how this thesis connects the early studies in collective moral psychology to the recent findings in the field of collective moral cognition. In particular, I investigate collective moral *cognition* (chapter 2), collective moral *judgments* (chapter 3), and collective moral *actions* (chapter 4) in three separate manuscripts.

In chapter 2 (manuscript one), I argue how *collective* moral cognition (e.g., moral judgments and decisions) differ from *individual* moral cognition. Furthermore, I discuss how collective *moral* decisions and judgments differ from *non-moral* decisions and judgments made by collectives.

After theoretical examination of these two questions in the literature, the thesis proceeds with the description of the empirical evidence conducted in two lines of work. In one line (chapter 3, manuscript two), I examine how collectives (i.e., groups of interacting individuals) arrive at consensus judgments for moral dilemmas. In particular, based on the theoretical models developed in chapter 2, I discuss how group-based emotions, social deliberation, and social motivations can shape collective moral judgments in small groups. Comparing collective judgments to those made by independent individuals, I examine three hypotheses showing how collective moral judgments can differ from the statistical

aggregate of individual judgments. Consistent with one hypothesis, the findings of my study show that collectives are more *utilitarian* than individuals. The underlying mechanisms of this collective utilitarian boost are discussed.

In the other line (chapter 4 - manuscript three), I examine the mechanisms through which people seek to punish individuals in collective moral transgressions *differently* than solo violations. I compare the punishment judgments in collective vs. individual moral transgressions by assessing how individuals seek to punish moral violations done *jointly* with others vs. *alone*. My findings show that individuals within a group receive *less* punishment for collective harmful actions compared to when they do them alone. Moreover, across several experiments, the role of intention, outcome, and moral domain (harm vs. purity) in the diffusion of punishment is explained. Exploiting discounting accounts in causal attribution, I discuss the mechanisms underlying the diffusion of punishment in collective actions.

In chapter 5, I explain how the collective dimension can be central to moral cognition. Concluding the findings of each chapter, the implications for cognitive science, moral psychology, and moral philosophy are discussed.

Contents

Note *1*

Chapter 1. General Introduction *3*

1.1 Collective moral decisions and judgments **3**
1.1.1 'We' vs. 'I': non-moral domain 4
1.1.2 'We' vs. 'I': a historical review of the moral domain 6
1.1.3 'We' vs. 'I': moral domain – current state 11
1.1.4 'We' vs. 'I': current thesis 12

1.2 Collective moral vs. non-moral decisions and judgments **15**

1.3 Collective moral transgressions **16**

Chapter 2. Making Moral Decisions and Judgments Together *25*

2.1 Introduction **25**

2.2 Why the cognitive science of morality has mainly focused on individuals **27**
2.2.1 Philosophical conceptualizations 27
2.2.2 Psychological conceptualizations 30
2.2.3 Neuroscientific conceptualizations 33

2.3 Collective vs. individual moral cognition **34**
2.3.1 Collective Dual System 35
2.3.2 Social Motivations 38

2.4 Collective moral vs. Collective non-moral cognition **39**
2.4.1 The psychology of metaethics in individuals 41
2.4.2 From *I* to *We*: intellective and judgmental tasks 44
2.4.3 Collective moral problem solving 49

2.5 Conclusion **51**

Chapter 3. Many Heads Are More Utilitarian Than One *79*

3.1 Introduction **81**
3.1.1 Moral dilemmas 82
3.1.2 Social interaction and modulators of moral judgment 86
3.1.3 Current study: moral judgment and social interactions 87

3.2 Experiment 1. ..... **92**
3.2.1 Material and Method 92
3.2.2 Procedure 94
3.2.3 Result 95
3.2.4 Discussion 97

3.3 Experiment 2 98
3.3.1 Material and Method 98
3.3.2 Procedure 98
3.3.3 Result 100
3.3.4 Discussion 102

3.4 General Discussion 103

3.5 Supplementary Material 119
3.5.1 A: Experiment1 119
3.5.2 B: Experiment2 135
3.5.3 C: Experiment 1 (Face to Face) vs. Experiment 2 (Online) 146
3.5.4 D: Scenarios in English 151
3.5.5 E: Added Measurements in German 155

Chapter 4. Diffusion of Punishment in Collective Norm Violations 159

4.1 Introduction 161
4.1.1 Intentionality 162
4.1.2 Causal responsibility 162
4.1.3 Existing research 162

4.2 Experiment 1 164
4.2.1 Methods 165
4.2.2 Results 166
4.2.3 Discussion 167

4.3 Experiment 2 167
4.3.1 Experiment 2.a 168
4.3.2 Experiment 2.b 171

4.4 General discussion 172

4.5 Supplementary Material 179
4.5.1 Experiment 1 179
4.5.2 Experiment 2 182
4.5.3 Scenarios used in Experiment 1 185
4.5.4 Scenarios used in Experiment 2 192

Chapter 5. Conclusion 197

Afterthought 200

List of publications 201

Eidesstattliche Versicherung/Affidavit 202

Authors Contribution 203

To the memory of ***Faraz Falsafi***
& all the other passengers of flight PS752

Note

While I was doing my Ph.D., an incident greatly influenced me, my family, my friends, and my country. On the 8th of January 2020, as widely reported by the international media, a passenger plane was shot down by two missiles in Iran, with the loss of all 176 travelers and crew. In the aftermath of the tragedy, the Iranian government incorrectly reported that a technical malfunction caused the plane crash. Following a litany of claims and denials, the government finally admitted that its own army shot down the aircraft due to days of heightened military and political tension between Iran and the U.S.

My 30-year-old friend was on that plane. He is dead now. We lost him twice: the first time when the plane crashed and the second time when we finally realized that the government had lied. To me, the government's strategic decision to hide the truth is a case of a *collective* moral violation based on a *collective* decision, which affected many people's lives and welfare.

History can attest to the claim that collective decisions and actions have played critical roles in atrocities around the world: from Rwanda mass killings to more recent crimes done by terrorist groups such as Al-Qaeda, Taliban, and ISIS, none was conceivable by single individuals in isolation, but many were involved.

I believe that understanding the mechanisms of collective moral decision-making and collective actions can help us understand the machinery that can lead to such violations. To provide preventive solutions for collective norm violations, we need a deep understanding of such mechanisms in groups.

Think of the global and existential challenges we face today, such as ecological and energy issues, global warming, weapons of mass destruction, nuclear agreements, etc. They all harbor vital moral concerns that can only be addressed at the collective level. If we ignore the collective aspects of morality, we may not reach a collective *solution* to solve those problems.

Being a woman born and raised in a totalitarian regime, surrounded by fundamentalist terror from neighbors to the east and west, I wrote this thesis in the hope of a dream: a better moral future. I invite you to read it in the same light.

Chapter 1. General Introduction

A family may discuss ending the life support for a comatose parent. An ethics committee may decide to approve a research proposal that involves animal pain. A group of strangers may argue whether abortion should be legalized on social media. Similar to these examples, moral decisions and judgments have a prominent social nature. But how do we assimilate morally relevant information in groups, discuss and deliberate together on moral grounds, and reach collective moral decisions and judgments? On a more general level, does collective moral cognition possess features that make it different from individual moral cognition? Similarly, do people judge collective immoral decisions or actions differently from individual ones?

Aiming to answer these questions posed above, in this thesis, I investigate moral cognition in collective contexts. In particular, I examine moral *decisions* and *judgments* in (and for) human groups. Thus, in this thesis, morality does not refer to certain attitudes (e.g., politeness, kindness, etc.) but the *content* of moral cognition and the *process* through which people integrate morally relevant information, emotions, biases, and intuitions in order to make moral decisions and judgments in (and for) groups. In addition, this thesis explores collective moral cognition at a *descriptive* level. It concerns what individuals find morally acceptable (or unacceptable) rather than what *is* or *ought to be* ethically, practically, or rationally right. Here, establishing the scope of the thesis, I argue how the collective dimension is central to moral decisions, judgments, and actions in the literature.

1.1 Collective moral decisions and judgments

Previous literature shows that there are fundamental differences between the decisions made by individuals vs. groups. In particular, a recent family of views underscores the central role of the interaction in human cognition. In the first part, reviewing the literature, I discuss the proposal of these recent views (1.1.1). In the second part, to evaluate the past in light of the present, I adopt a historical narrative to discuss the collective dimension in the *moral* realm. Rather than being exhaustive, my intention is to show how *collective* and *moral cognition* were

originally entangled in early works of moral psychology but examined independently in later years (1.1.2).

In the third part, highlighting the role of groups in moral cognition, I show how the theoretical concepts of collective cognition were recently imported into the moral realm. The main focus of this part is on contemporary works that stress the differences between *I vs. We as the agent of* moral decisions and judgments (1.1.3). Finally, putting together these three parts, I show how this thesis is the continuation of an emerging approach that connects *collective* cognition to *moral* cognition (1.1.4).

1.1.1 'We' vs. 'I': non-moral domain

We humans are not single minds encapsulated within isolated craniums, but we experience the world through our bodies while interacting with others in dynamic environments. Acknowledging this fact, a new turn in cognitive science has emerged, which examine human cognition not as abstract information-processing units in isolated minds (*individualism*), but rather as the extensions of bodies (embodied cognition, see Varela et al., 1991), with certain affective states (e.g., Izard, 1992; Zajonc, 1980), interacting with other minds and other bodies (*interactionism*, e.g., De Jaegher & Di Paolo, 2007). These accounts suggest that human information processing is a function of physical, affective, and situational signals (Roth & Jornet, 2013) in dynamic environments, coming from *interaction* with others (Gallotti & Frith, 2013).

A recent trend pushes this movement forward, highlighting the fundamental differences between mental processes in isolated *individuals* vs. interactive *groups*. According to this recent family of views, the cognitive mechanisms of collective actions *cannot* be *understood* merely by examining the individuals involved in them (e.g., Michael, 2011; Sebanz et al., 2006). In fact, when people interact, certain group dynamics may arise that make *collective actions* fundamentally different from the sum of the individuals' actions in groups (Gallotti & Frith, 2013).

During the interaction, people may align their minds (and bodies) and share their resources to bring about a change in the world *together* (Sebanz et al., 2006). In such cases, they may no longer process the information from a first-person perspective (as single observers), but they adopt a *shared perspective* to act and think collectively (Tuomela, 2007). It is argued that the shared perspective in collective actions can also modulate abstract thinking, decisions, and judgments (Herschbach, 2012; Thompson & Stapleton, 2009). Thus, when people are in groups, and they jointly make collective decisions, they may exhibit certain features that they may not show should they act as individuals *without* such shared perspective (Tuomela, 2007). Moreover, during interactions, people may have access to the information *differently* than what they would have as mere observers. Interactions, therefore, can lead to group-level information processing, which may be different from the statistical aggregate of individual mental processes (Higgins, 2020).

Scholars have referred to this approach by various terms such as 'we-perspective' 'we mode', 'we-ness' (see Tuomela, 2007), 'shared intentionality,' 'co-cognition ', 'interactionism' (Gallotti & Frith, 2013), '2nd person cognitive science' (Schilbach et al., 2013) 'interactive turn', 'participatory sense-making (see De Jaegher & Di Paolo, 2007), and 'collective psychology' (see Wilson, 2004). All of these terms revolve around one single idea: when a group thinks, decides, or acts as one single agent, it cannot be reduced to its individual parts (Bratman, 2014; Michael, 2011; Salmela, 2012; Sebanz et al., 2006; Tuomela, 2007; Wilson, 2004). The common denominator of these recent collective-interactive accounts is the emphasis on certain *interaction*s that lead to fundamental differences between the two modes of cognition, i.e., 'I' vs. 'We' mode: in the latter, the agent of the action or the decision is more than one person; thus, different mechanisms may emerge during the interaction, making the latter dissimilar to the former. Acknowledging this fact, the disagreement between different accounts within this family of views is mostly on the conditions (and the extent to which) this emergence may occur.

Not all cognitive mechanisms modulated by groups have such collective nature. For instance, basic forms of interaction may not require 'we' perspective. Such interactions introduce a novel social context that can affect *individual* behaviors. Examples of such effects are emotional contagion, facial imitation, or motor mimicry, which happen merely in the presence of others, or simply by observing the expressions of emotions or certain body states in them (see Hutto, 2004). What collective accounts suggest, however, is beyond these basic forms of interactions. The collectivist accounts propose that interactions are indeed critical settings that create a common ground, a shared sphere for collective cognition (Higgins, 2020). This shared sphere is primarily related to higher-level interactions (not merely due to observations or the presence of others), which may involve joint actions, coordination, communication, or shared intention (but see Gallotti et al., 2017).

Note that the emphasis here is not on the fact that mental processes are affected by high-level interactions. In other words, the mental states may be simply affected by other minds (or bodies), even in higher-level interactions, *without* the need for a *shared* perspective. For instance, when people collaborate, they may keep track of what others think, believe, intend, desire, etc. Therefore, they may attribute mental states to others, trying to infer the contents of these mental states to adjust their actions accordingly. Although these processes are the result of high-level interactions, they still happen at the individual level.

The argument of the 'collective' accounts, however, is that social dynamics in group interaction can create novel features that are different from what happens in single minds (Wilson, 2004). In such cases, the interactions seem to create distinct *access* to information (Tuomela, 2007) while it *enriches* cognition (Gallotti & Frith, 2013). Therefore, ascribed mental states that may prevent or encourage people to behave in certain ways in groups can occur at the individual level. But when people act *as groups*, such collective individual mental representations may create a shared perspective resulting from the interaction itself. Thus, due to the emergent properties arising at the collective level, group information processing cannot be

attributed to individuals. Put differently, when the perspective is shared, certain dynamics might emerge, which make individual experiences *qualitatively* and *conceptually* different from what every single individual can do or experience alone (Tuomela, 2007). Note that this view does not suggest a collective mind beyond the individual minds or brains (for the difference, see Wilson, 2004) but rather stresses the difference between individual vs. collective information processing.

Recently, across different domains, the collective dimension is shown to be of paramount significance in cognition. For instance, research in joint perceptual decisions in psychophysics (e.g., Bahrami et al., 2010), joint factual decisions (e.g., Mannes et al., 2014), collective dimension of information acquisition in social epistemology (Schmitt, 2017), and collaborative tasks in cognitive neuroscience (e.g., Sievers et al., 2020) show how interaction can shape collective decisions and judgments *differently* from individual decisions and judgments.

To sum up, according to a recent family of views, interactions in groups can create emergent phenomena, changing the content and the process of cognition at the collective level. When making collective decisions or acting as a group, people may adopt a shared perspective. This shared perspective is argued to change the process or the content of individual decisions in collective settings, which, in turn, lead to collective decisions that are different from the aggerate of the individual decisions. Several studies confirm that collective decisions and actions are qualitatively different from individual ones.

1.1.2 ‘We’ vs. ‘I’: a historical review of the moral domain

In the previous section, I reviewed an emerging trend in cognitive science, stressing the role of interactions in human cognition. Yet, the collective dimension is not new to the moral realm. In fact, the works of pioneers in psychology of morality show that the collective dimension was inextricably intertwined with studies of moral decisions, judgments and actions.

For instance, the founder of experimental psychology, Wundt, was preoccupied with the role of collectives in morality to the extent that he initiated a branch in psychology (parallel to his experimental psychology project) to address this issue. Between 1900 and 1920, he published ten volumes of Völkerpsychologie, the psychology of a community of individuals (literally ‘the psychology of peoples’) to address the role of interaction in morality and other domains (see Hogg & Williams, 2000).

To Wundt, thinking in isolation was different from sharing thoughts with others in verbal communication. For instance, he observed that people would sometimes revise their sentences when speaking about what they thought because their words could not properly communicate the content of their thought. Moreover, people could capture disagreement with others, sometimes very quickly, even before being able to think why they disagreed in the first place. These observations, to Wundt, were pieces of evidence showing that thoughts, as expressed to others with words,

could be qualitatively different from thoughts when experienced in silent thinking (Fancher & Rutherford, 2012).

Wundt also believed that moral principles, coming from religions, myths, and customs, have the same characteristic, as they could be hugely dependent on others external to us. In fact, morality to him was nothing but a reciprocal collective enterprise (Fancher & Rutherford, 2012). As a result, he objected to the individualist accounts in morality in his paper 'study of the facts and laws of moral life' while emphasizing the role of interaction and groups in the moral realm (see Klautke, 2013).

Wundt hugely influenced later thinkers, such as Émile Durkheim, a French social scientist, who also examined the collective representations of groups in the moral realm. Durkheim proposed that the unit of analysis in the studies of morality should be collectives rather than single individuals. However, unlike Wundt, he thought that the collective aspect of morality could be studied empirically as a natural phenomenon (Bellah, 1974; Klautke, 2013). He used the term "*collective psychology"* for this line of research as he believed that moral codes could be shaped via interactions with others (Hogg & Williams, 2000).

Gustave Le Bon was another French thinker influenced by Wundt's folk psychology (Hogg & Williams, 2000). Similar to Wundt, Le Bon believed that collectives could act differently than individuals in the moral realm. In fact, according to his observations, collectives were 'too impulsive and too mobile to be moral' (Le Bon, 1960, p 67) and often unable to reason properly. However, somewhat self-contradictory, he did not ascribe only negative moral characteristics to groups but also positive moral virtues such as altruism.

Perhaps even more pessimistic than Le Bon, Niebuhr, yet another influential scholar, argued the same. In 1932, Niebuhr published 'moral man and immoral society', in which he highlighted the differences between group and individual moral actions. Evident from the title of his book, he found collectives morally inferior to individuals, as groups could not acquire certain cognitive or affective capacities, such as empathic concern (Niebuhr, 1932).

McDougal, another early 20th-century psychologist, also had a significant impact on collective psychology (Farr, 1986; Hogg & Williams, 2000). To him, the cognitive, affective, and conative components of the mental characteristics could be regulated differently via interaction with others. He published *Group Mind* as the second volume of his influential book 'Introduction to social psychology', a burgeoning field of research at that time. In this book, he argued that, due to the interaction, the mental characteristics of groups are different from individuals. Yet, McDougall's ideas were sometimes *misunderstood* as ascribing a 'mind' to groups, perhaps due to the misleading title of his book (Hogg & Williams, 2000) or his eccentric personality (Fancher & Rutherford, 2012; Farr, 1986). He did not assign any (moral) agency to groups but to the individuals comprising the groups (Farr, 1986; Hogg & Williams, 2000; but see Wilson, 2004).

By and large, McDougal remains one of the pioneers who highlighted the difference between collective and individual psychology (Farr, 1986; Moscovici, 1986) and moral psychology specifically. For instance, he reacted against the individualistic assumptions underlying classic moral theories such as utilitarianism (Farr, 1986).

McDougall was criticized by many, including Alport (and many other scholars in the American camp), who were then greatly influenced by behaviorism (Farr, 1986) - a discipline focusing on behavioral changes as a function of reward and punishment in individuals. To behaviorists, the collective psychology was the psychology of individuals; thus, 'we' was simply an arithmetic aggregation of 'I's to them (Hogg & Williams, 2000).

As collectives and groups were not central to behaviorism at that time, Alport (and other behaviorists) succeeded in shifting the views towards individual behaviors, downplaying the influence of McDougall's works, as well as other collective approaches in understanding the human mind. Hence, the American camp (and the dominance of behaviorists at that time) were effective in diminishing collective morality as a research project which was initially started in Europe.

'By the late 1920s, the collectivist perspective of early social psychology—the view *that interaction produced emergent properties of collectives that could not be understood in terms of or reduced to individual psychology*—had all but disappeared from mainstream social psychology that focused on individual behavior. The disappearance was, of course, most troublesome for the social psychology of groups.' (from Hogg & Williams, 2000, p.82, italics mine).

In the European camp, however, the collective dimension was hugely under the influence of WW II. On the one hand, European psychology (as other domains of science) was economically and politically dependent on the US between 1940 and 1960 (Hogg & Williams, 2000) – hence under the influence of behaviorists. On the other hand, in the aftermath of the war, European thinkers were motivated to understand the collective mentality and mechanisms which led to the immoral actions leading to (and happening during) the war (Parkin-Gounelas, 2014). Therefore, when Europe finally gained its confidence as it was more economically and politically independent from the US, it revived the collective dimension in moral psychology. European social psychologists, then, highlighted the role of collective and group dynamics in understanding the mind and the behavior, once again (see Hogg & Williams, 2000). This influence passed over the Atlantic Ocean (Hogg & Williams, 2000) and was imported into the US social psychology (e.g., Milgram et al., 1969).

The experimental approaches to psychology eventually acknowledged the role of groups as the agents of decisions and focused on the interactions as crucial in shaping actions. The collective dimension finally got a chance to be studied experimentally in small groups and face-to-face interactions. Using empirical methods in controlled experiments, the collective dimension was studied rigorously, establishing the fundamental differences between decisions and actions in individuals vs. in groups.

Note that the difference between small ad-hoc groups vs. 'collectives' or 'crowds' is non-trivial. The scale of the collective moral decisions that Wundt, Durkheim, Le Bon, and other pioneers had in mind was much larger than that of small interactive groups in social psychological labs. For instance, when Niebuhr discussed group morality, he referred to formal groups rather than ad-hoc groups. Understanding morality in crowds, formal groups, and organizations is beyond the scope of this thesis, as I only investigate ad-hoc groups experimentally. The difference is, however, important since distinct mechanisms distinguish the social (formal) and ad-hoc (non-formal) groups (e.g., see Tang et al., 2020). Yet, in light of this caveat, it is even more noteworthy that emergent factors of group dynamics in collective settings were measured in small groups as an independent research line at that time.

As it was free from the authority of behaviorists, *collective or group decision-making* was finally rigorously studied across different labs. In fact, collective decision making, as an independent research project, was well received and studied as a separate research line in non-moral domains (e.g., in perceptual or factual joint decision making; c.f. chapter 2). [1]

As the experimental methods allowed researchers at social-psychological labs to address group decisions, scholars started to investigate *moral* decisions in groups as well. For instance, in a seminal work, Bandura et al. (1975) showed that people in a group are more prone to harm others via electric shocks than when they act as individuals. This effect was attributed to the diffusion of responsibility in groups (Bandura et al., 1975). [2]

However, this collective approach in the moral domain did not last long. Although non-moral psychologists were successful in importing the collective dimensions into their studies, the moral domain had a very different destiny. Studies of moral decisions and judgments, unlike other aspects of psychology, were soon dominated by another approach, whose views and methods were, once again, vastly individualistic. This new account, coming from developmental psychology, influenced later studies of morality while missing out the role of interactions and groups in the moral realm (see Leach et al., 2015).

Preoccupied with the investigations of rational *moral judgments* in children as a function of their developmental stages, pioneers like Piaget (1993) or Kohlberg (Kohlberg & Hersh, 1977) originated '*moral psychology'* as an independent discipline that concerned individuals rather than groups. For Kohlbergian moral psychologists, the interactive social aspects were critical, but only as the *origins* of

1 In chapter 2, I will return to collective decisions in non-moral realm and argue how they are different from moral realm.

2 Diffusion of responsibility is a recurring motif in this thesis, and I will return to it in chapters 2 and 3. It refers to the effect that even the mere presence of others can make individuals feel less responsible. This was originally observed in relation to the well-known bystander effect: when several observers witness a norm violation, each observer is less likely to intervene compared to what they would have done had they been alone (Darley & Latane, 1968).

moral judgments. Social dynamics or group agents were not the focus of developmental moral psychology. This will be established in detail in chapter 2. Reviewing the literature of moral psychology and the neuroscience of moral decisions, I will return to the fact that experimental moral psychology began as an individualistic research project, originally concerned with individuals' moral reasoning.

Yet, it should be noted that at this time, taking the same developmental approach, a few attempts were made to understand the role of groups and interactions in moral judgments. Research showed, for instance, that interaction could change moral judgments in adult groups. Using the '+1 manipulation' technique in dyadic discussions, a person whose moral stage was 1 stage above the target of manipulation (within the range of 1 to 6 moral stages) was matched with the target. This method was proposed as the optimum way to change the target's moral judgment after short discussions about moral issues (Berkowitz et al., 1980; Berkowitz & Gibbs, 1983). People were shown to change their moral judgments based on small differences (weak disagreement) they had with others during discussions. This change of mind led to a more advanced moral judgment in the Kohlbergian sense (see Keasey, 1973). Similarly, groups were shown to be more advanced than individuals in their moral reasoning (Damon & Killen, 1982; Maitland & Goldman, 1974), especially when they socially deliberated on moral dilemmas - e.g., stealing a drug in order to save the life of one's wife (Nichols & Day, 1982).

Although informative, these few attempts were sporadic. Developmental moral psychology was generally preoccupied with the process of deliberative reasoning within individuals, not between them (see Leach et al., 2015).

One should not forget that morality is a multidisciplinary field, and each field treats it differently. Another dominant approach here that fueled the later individualist trend and maintained it for some years was experimental moral philosophy (and philosophically inspired psychology). This influence was partly due to the nature of classic normative moral theories. In fact, despite their differences, classic moral theories had at least one thing in common: they all treated individuals and collective decisions in the same way (c.f. 2.2.1). For instance, being oversimplistic, for a Kantian, as long as the decision was committed to the categorical imperative, for a utilitarian, as long as it brought about the total good, and for a virtue ethicist, as long as it was faithful to certain virtues, there was no distinction, in principle, between 'I' or 'we' as the agent of the moral decision. Thus, as long as a decision fulfilled the required moral criteria, it was then regarded as morally right, no matter how the decision was made, either by one or many.

As moral psychology has a bidirectional relation with moral philosophy, moral psychologists seem to import this individualistic approach later from moral philosophy into the empirical studies of moral cognition in their research (c.f. 2.2.1, see also Leach et al., 2015).

To sum up, collectives were central to moral psychology in the early works, but the focus of moral psychology shifted on individuals in later years. On the one hand, the dominant developmental approaches in moral psychology concerned individuals rather than groups. On the other hand, moral theories did not distinguish collective decisions from individual decisions. Thus, a good deal of moral psychology over-relied on the individualized conceptualization of moral decisions and judgments. The dominance of these two approaches fueled individualistic trends in moral psychology in later years, which persisted until recently.

1.1.3 'We' vs. 'I': moral domain – current state

In section 1.1.1, I argued that the collective approach is not alien to contemporary cognitive science. I reviewed a family of recent views suggesting that collective actions and decisions cannot be understood by merely investigating individual actions and decisions. In section 1.1.2, I explained that the collective dimension was once critical to social and behavioral sciences in early works, especially in moral domains, but it received comparably little attention in the moral psychology of later years.

However, after decades of dominance of rationality, deliberation, and abstract thinking in individual moral reasoning, recently, this narrative gives way to theories that highlighted the role of emotions, quick automatic intuitions, and social interactions in moral reasoning. Similar to other domains of cognition, revolutionary turning points in moral cognition have helped scientists to move from traditional accounts (which study abstract isolated moral minds) towards social, affective and, interactive accounts (Colombetti & Torrance, 2009; Haidt, 2001; Métais & Villalobos, 2021; Prinz, 2010; Urban, 2014; Varela, 1999). Thus, a new trend in moral cognition seems to be emerging, bridging the gap between moral psychology and collective cognition.

For instance, many scholars have stressed the significance of emotions and intuitions in human moral decisions, judgments, and actions (Haidt, 2001; Prinz, 2010). One radical account - the social intuitionist model – pushes this notion forward, to the extent that it assumes human morality is primarily the outcome of automatic and emotional rapid intuitions rather than slow, deliberative reasoning (Haidt, 2007). Inspiring later researchers, this account argues that there is little room for deliberation in human moral judgments. When deliberation comes into play, it is relatively late, only after the moral judgments are already made. One important assumption in the social-intuitions model is that the rapid emotional intuitions have a *social* and *interactive* nature, acquired during social interactions (Haidt & Kesebir, 2010), suggesting that interactions are of paramount significance in making moral judgments.

The central role of emotion-laden *interactions* is now widely accepted in morality and has shifted scholars' attention from individual heads to real-life social interactive contexts in the moral realm (Colombetti & Torrance, 2009; Métais & Villalobos, 2021; Urban, 2014). Moral theorists do not see humans as rational moral

decision-makers in isolation anymore, but as minds with extended bodies with affective states (Prinz, 2006) in social contexts. Moreover, morality is argued to be integrated with our actions and not merely expressed as abstract thinking. Similar to other skills, moral reasoning can be learned via *interactions* with the environment (Varela, 1999) and other people (Colombetti & Torrance, 2009). Moral judgments are proposed as intuitive 'social doing' to build our reputation in social interactions (Haidt, 2007) or as 'social influence' in argumentation to convince others to accept our moral views (Mercier, 2011).

Focusing on the affective characteristics of interaction, new accounts show how interaction can be of paramount significance in moral decisions due to its social-affective aspects. The character of our social encounters is argued to be essentially moral, considering the important role of emotions in participatory sense-making in moral contexts (Colombetti & Torrance, 2009). Similarly, in the developmental psychology camp, moral cognition is examined as a function of communication and interaction during the first years of life (Dahl et al., 2013; Zahn-Waxler et al., 1992). Moreover, in the social-psychological camp, scholars endorse the interactions and group contexts as independent variables affecting moral decisions and actions in individuals. This approach shows, for instance, how people shape their identity according to the groups they belong to (see Leach et al., 2015) or how groups are perceived differently than individuals for their immoral actions (Brambilla et al., 2012; Sharvit et al., 2015). Even random assignments to groups are shown to lead to particular moral decisions (Goette et al., 2006). Moreover, groups can shape moral emotions and behaviors in individuals (Cikara et al., 2011; Cikara & Fiske, 2012), while the neural mechanisms of these moral decisions may depend on in-group vs. out-group division (Cikara et al., 2014).

To sum up, the revolutionary socio-emotional trends in cognitive science influenced moral psychology, shifting the attention from rational individual heads towards interactive, emotional, and embodied moral agents. A recent group-based approach appeared which examines moral decisions in relation to groups. These ground-breaking social-emotional interactive accounts look at moral cognition as a *social* enterprise.

1.1.4 'We' vs. 'I': current thesis

In the previous sections, I argued how the collective dimension was once central to the moral domain, overlooked for some years, but it became critical in recent years once again (see 1.1.3). Consequently, the effect of groups and collectives on individual moral decisions and judgments has become crucial in studies of moral cognition in recent years.

However, the recent group-based approach concerns the effect of groups on *individual moral reasoning* rather than exploring groups as a whole (c.f. 1.1.1). In this sense, a good deal of group-based moral psychology still misses out the collective aspect of morality. Put differently, previous research mainly investigated groups as the *context* (Us) or the *target* (Them) rather than the *agent* (We) of moral

decisions or judgments. By contrast, the mechanisms underlying social deliberation, moral argumentation, social influence, or persuasion in *collective moral* reasoning received relatively less attention. In particular, whether collective moral decisions and judgments are different from the aggregate of individual decisions and judgments is less clear.

Very recently, studying groups as the *agent* of moral decisions has become attractive to researchers (e.g., Navajas et al., 2019). A few studies have shown that group moral decisions can be different than what individuals decide in isolation. For instance, people find it more acceptable to sacrifice one person in order to save many when they are in groups vs. alone (Curşeu et al., 2020). The more people feel socially connected to their partners in such decisions, the more they find such utilitarian decisions morally acceptable (Lucas & Livingston, 2014). In resource allocation moral dilemmas, collectives distribute resources differently than individuals (Ueshima et al., 2021). Moreover, Individual and collective moral decisions seem to be different in certain age groups (Takezawa et al., 2006).

Following the same line of research, one objective of this thesis is to address the differences between groups and individuals in their moral judgments when the group is the agent of moral reasoning. The central question that will be repeatedly brought up is whether (and how) collective and individual moral decisions and judgments are different from each other. Combining insights from experimental social psychology, collective cognition, and moral philosophy, chapter 2 and chapter 3 aim at answering this question (c.f. chapters 2 and 3).

It should be noted that, with respect to collectivism in morality, my position is rather conservative. Therefore, in this thesis, I do not argue that collectives are *irreducible* to individual mental processes in the moral realm. Any argument about irreducibility is beyond the scope of this work. Yet, my position is not neutral either. I argue that when people come together in the social sphere to discuss moral issues or decide on moral grounds, specific emotional, cognitive, and social mechanisms may arise that do not come up when people reflect on the same issues *alone*. The simultaneous interplay between emotion and deliberation can contribute to *collective* moral decisions different than individual decisions, as I will discuss in chapter 2.

Drawing on recent literature in moral psychology, in chapter 2, I will explain the cognitive mechanisms that can act differently in a collective moral context, leading to an alteration of moral decisions and judgments in groups. Depending on the affective, deliberative, and social mechanisms at play - which emerge at the collective level - the content and process of collective moral decisions may vary. In chapter 2, the mechanisms such as virtue signaling (how we wish to be seen by others), diffusion of responsibility (how we join groups to reduce our responsibility), social deliberation (discussing moral issues together in order to solve them) will be discussed in more depth. These mechanisms can modulate moral decisions in groups, leading to substantial differences between individual and collective moral decisions and judgments.

After showing the difference between collective and individual moral cognition theoretically in chapter 2, the next logical step is to set up a study to examine them experimentally. Therefore, in chapter 3, I will answer three empirical questions: How are collective moral judgments different from individual ones? How are individual moral judgments different before and after social deliberation? And what is the underlying mechanisms at work in collective moral judgment which explain these differences?

To answer these questions, in a study of collective moral judgments (chapter 3), I use the models proposed in chapter 2 to build three hypotheses, each coming from one of the previously proposed mechanisms (emotional/reasoning/social). Testing these hypotheses, each predicting a different pattern of behavior in individuals vs. groups, this chapter will answer the three questions posed above. Therefore, consistent with the previous literature (chapter 2) and the empirical data (chapter 3), I show that the *process* and the *content* of collective decisions and judgments *differ* from individual decisions and judgments. In this sense, collective moral reasoning cannot be simply regarded as equal to the individual moral opinions in isolation, aggregated statistically.

I will explain the experiment in more depth in chapter 3, but a clarification seems necessary here. The collective decisions presented in chapter 3are particular types of decisions called '*moral judgments*. As I shall define it, a moral judgment is the moral evaluation of a third person's action (or inaction) after the action (or inaction) occurred. For instance, one might judge a friend's cheating on her/his exclusive partner by answering whether (and to what extent) that action was morally unacceptable. Similarly, in the experiment presented in chapter 3, I asked people to judge the moral acceptability of actions (or inactions) of hypothetical characters in short moral dilemmas.

Moral dilemmas are difficult situations with no apparent morally permissible answer. Some examples can be illuminative: think of the student, described by Sartre, torn between staying with his frail mother and fighting the Germans to avenge his only brother and defend France during WW II (Bastable, 1957); or the dramatic decision Sophie has to make in the book, later turned into a movie, *Sophie's Choice*. A lonely mother in distress, Sophie, is asked to choose which of her two children must be sent to death (Styron, 1979). The latter case belongs to a particular category of moral dilemmas called *sacrificial dilemmas*. Initially built by philosophers, sacrificial dilemmas show a conflict between sacrificing someone in order to save others. One of the most well-known sacrificial dilemmas imported from philosophy into psychology (apart from *Sophie's Choice*) is the trolley problem - a hypothetical situation in which an isolated bystander can kill one person to save five railway workers in different contexts (Foot, 1967; Thomson, 1976). Following their original appearance, the trolley problem, Sophie's Choice, and other sacrificial dilemmas were turned into experimental stimuli by cognitive (neuro)scientists. They were used to investigate the cognitive (Petrinovich & O'Neill, 1996) and neural basis (Greene et al., 2001) of people's moral inclination

and have since been predominantly used in moral psychology and neuroscience (for a review, see Christensen & Gomila, 2012).

In addition to sacrificial dilemmas (i.e., killing one to save many) - which are odd situations that are used uncritically in studies of moral judgments - the dilemmas presented in chapter 3 are more relevant to real-life daily moral challenges (to see the moral dilemmas, see chapter 3). They include short scenarios that, probably for the first time, involved omissions or inactions that led to certain utilitarian outcomes. These utilitarian outcomes shared a special feature: they were based on violating a norm to bring about the greatest total good; for instance, keeping a friend's cheating secret in order to help them save their marriage.

These scenarios were designed to test participants' utilitarian intuitions since a utilitarian approach would prioritize the consequences of the relevant action (e.g., preventing a divorce) over their abiding by moral duties (e.g., always telling the truth to a friend). As detailed in chapter 3, by asking people to evaluate these types of utilitarian actions (or inactions) first in private, then in groups (after short discussions), and later, in private again, I show how collective judgments are different from individual ones. I will explain how the observations in chapter 3 suggest certain mechanisms at play, driving collective judgments towards the utilitarian direction.

1.2 Collective moral vs. non-moral decisions and judgments

Another question I seek to address in this thesis is whether there are any differences between how people perceive moral vs. non-moral decisions and judgments, especially in collective contexts. This is important for two reasons:

1) If there is no difference between how we construe moral vs. non-moral domains, we can simply extend findings from studies of collective non-moral decision-making (a very well-explored research project) to the moral realm (a hugely underexplored research line). By contrast, in chapter 2, I argue that moral matters can be construed differently from non-moral matters. This difference can be especially significant in collective settings. The upshot is that we cannot simply outspread the non-moral realm into the collective moral realm by applying the findings in those domains directly to collective moral reasoning due to the difference between metaethical commitments in moral vs. non-moral domains.

2) Understanding whether, how, when, and why people interpret the moral domain differently from the non-moral domain can have practical implications. Clearly, if people think of their moral views as solid facts, it seems less likely that they change their opinions or compromise in group decisions and discussions (c.f. 2.4). Thus, getting along with others in moral discussions may depend on how we construe moral issues. Whether 'morally wrong or right' is metaethically subjective or objective to us can be of paramount significance in how we engage in collective

moral discussions (e.g., compare collective moral discussion to collective problem-solving in tasks with an objective solution or a well-defined expected utility). These metaethical perceptions, as I discuss in chapter 2, can affect moral disagreement and negotiations in morally relevant contexts *differently* than in other contexts. As I argue, moral opinions can be perceived on a continuum from matters of opinion to matters of fact. Then understanding what can shift this perception on this continuum towards one or the other end may help resolve moral disagreements and conflicts.

Based on these two reasons and building upon recent findings in moral psychology of metaethical commitments, in chapter 2, I will argue that certain vulnerabilities in the moral realm can penetrate collective moral reasoning. Such vulnerabilities are related to the peculiar nature of metaethical commitments in collective settings. As I argue in chapter 2, collective moral problem-solving can have properties of collective judgmental tasks (e.g., deciding together which piece of art is more beautiful) and intellective tasks (e.g., solving a mathematical equation in groups). Thus, collective moral decisions and judgments need to be examined as an independent group task.

To sum up, in chapter 2, I argue why collective moral decisions and judgments are peculiar based on *i*) how people construe moral and non-moral domains differently when it comes to collective contexts and *ii*) how collective and individual moral cognition can be conceptually different from each other. The upshot is that we cannot simply import the findings of other fields into collective moral decision-making precisely due to these significant differences. Group moral decisions and judgments, therefore, cannot be studied simply as an extension of existing research by, e.g., deploying existing resources and models in joint decision making. For instance, applying aggregation rules on decisions and judgments to statistically estimate the opinions of the group may be inapplicable in the moral realm. These substantial differences between individual and group moral decision-making give collective moral decisions a special nature that requires independent examinations.

1.3 Collective moral transgressions

One crucial finding of the collective experiment in chapter 3 is that people violate moral norms in groups more than in isolation. [3] Moral violations are not uncommon in collectives: groups are shown to be less obedient to norms (Fochmann et al., 2021), they develop more antisocial behaviors (Behnk et al., 2017), tend to be less generous (Bornstein and Yaniv, 1998), and lie more than individuals (Conrads et al., 2013, 2016; Kocher et al., 2018). Moreover,

[3] However, norm violations tested in chapter 3 are special cases since they maximize the total good.

collaboration increases corruption (Weisel & Shalvi, 2015), and similarity in groups increases collective cheating (Irlenbusch et al., 2020).

But why are groups more prone to moral violations? One reason, recently proposed, could be that by joining groups, people try to minimize the negative consequences of their decisions (El Zein et al., 2020), using groups as shields to protect themselves from the potential costs of their actions (El Zein et al., 2019). A critical cost for group moral violations is the punishment each member receives in moral violations. Therefore, seeking 'safety in numbers' by joining groups, each group member may expect to bear a lower cost than acting alone. Chapter 4 aims at addressing this hypothesis more directly: do individuals receive less punishment if they violate a norm together?

The findings of the prior research on punishment reduction in groups have been inconsistent. One major shortcoming of previous studies, which fuels the inconsistencies, is that they merely studied *financial cost* in *intentional* violations as the only form of immoral action. For instance, they studied collective robberies (Feldman & Rosen, 1978; Vainapel et al., 2019) or collective donations (El Zein et al., 2020). By contrast, in chapter 4, different cases of collective moral violations are used across three experiments. These violations are solid cases of moral transgressions, such as collective murders or group cannibalism.

Moreover, previous research showed that three factors could strongly affect moral judgments: *I. Intention:* was the action intentional or accidental? *II. Outcome*: what was the consequence? (see Cushman, 2008), and *III. Moral domain*: was the action immoral and harmful (e.g., murder), or immoral but harmless (e.g., cannibalism) (Dungan et al., 2017). Accordingly, to understand the processes underlying the reduction of punishment in groups, it would be crucial to investigate its sensitivity to these dimensions.

Thus, across three experiments, chapter 4 will address three factors (intention, outcome, domain) in both solo and collective actions from an impartial observer's perspective. I explain how these factors modulate the reduction of punishment in collective moral transgressions.

To understand the mechanisms of reduction of punishment in groups, across two experiments, participants seek to punish hypothetical characters who are involved in moral violations. These violations occur either individually or jointly with two other characters. I examine how different intentions (malign vs. innocent) and consequences (harmless vs. harmful) through intentional, attempted, and accidental cases of killings in these stories (experiment 1) can contribute to the reduction of punishment in group actions. Whether harmless (yet immoral) actions, such as eating human flesh in groups (in experiment 2), can modulate this effect is also explored. I will show how the result of these experiments can shed light on the mechanisms of punishment attribution in collective moral violations. Insights from causal theories of responsibility attribution will be used to explain the underlying mechanisms of reduction of punishment effect.

References

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science 329*(5995), 1081–1085. https://doi.org/10.1126/science.1185718

Bandura, A., Underwood, B., & Fromson, M. E. (1975). Disinhibition of aggression through diffusion of responsibility and dehumanization of victims. *Journal of Research in Personality*, *9*(4), 253–269. https://doi.org/10.1016/0092-6566(75)90001-X

Bastable, J. D. (1957). Existentialism from Dostoevsky to Sartre. *Philosophical Studies, 7*, 200–202. https://doi.org/10.5840/philstudies19577047

Behnk, S., Hao, L., & Reuben, E. (2017). Partners in Crime: Diffusion of Responsibility in Antisocial Behaviors. *Research Papers in Economics*. https://doi.org/10.5840/PHILSTUDIES19577047

Bellah, R. N. (1974). Emile Durkheim on Morality and Society. *Worldview*, *17*(6), 57–58. https://doi.org/10.1017/S0084255900024463

Berkowitz, M. W., & Gibbs, J. C. (1983). Measuring the Developmental Features of Moral Discussion. *Merrill-Palmer Quarterly*, *29*(4), 399–410. http://www.jstor.org/stable/23086309

Berkowitz, M. W., Gibbs, J. C., & Broughton, J. M. (1980). The relation of moral judgment stage disparity to developmental effects of peer dialogues. *Merrill-Palmer Quarterly of Behavior and Development*, *26*(4), 341–357. http://www.jstor.org/stable/23084042

Bornstein, G., & Yaniv, I. (1998). Individual and Group Behavior in the Ultimatum Game: Are Groups More "Rational" Players? *Experimental Economics 1***,** 101–108 https://doi.org/10.1023/A:1009914001822

Brambilla, M., Sacchi, S., Rusconi, P., Cherubini, P., & Yzerbyt, V. Y. (2012). You want to give a good impression? Be honest! Moral traits dominate group impression formation. *British Journal of Social Psychology*, *51*(1), 149–166. https://doi.org/10.1111/j.2044-8309.2010.02011.x

Bratman, M. E. (2014). Shared Agency. In *Shared Agency*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199897933.001.0001

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1249–1264. https://doi.org/10.1016/j.neubiorev.2012.02.008

Cikara, M., Jenkins, A. C., Dufour, N., & Saxe, R. (2014). Reduced self-referential neural response during intergroup competition predicts competitor harm. *NeuroImage*, *96*, 36–43. https://doi.org/10.1016/j.neuroimage.2014.03.080

Cikara, M., Botvinick, M. M., & Fiske, S. T. (2011). Us versus them: Social identity shapes neural responses to intergroup competition and harm. *Psychological Science*, *22*(3), 306–313. https://doi.org/10.1177/0956797610397667

Cikara, M., & Fiske, S. T. (2012). Stereotypes and schadenfreude: Affective and physiological markers of pleasure at outgroup misfortunes. *Social Psychological and Personality Science*, *3*(1), 63–71. https://doi.org/10.1177/1948550611409245

Colombetti, G., & Torrance, S. (2009). Emotion and ethics: An inter-(en)active approach. *Phenomenology and the Cognitive Sciences*, *8*(4), 505–526. https://doi.org/10.1007/s11097-009-9137-3

Conrads, J., Ellenberger, M., Irlenbusch, B., Ohms, E. N., Rilke, R. M., & Walkowitz, G. (2016). Team goal incentives and individual lying behavior. *Research Papers in Economics.*

Conrads, J., Irlenbusch, B., Rilke, R. M., & Walkowitz, G. (2013). Lying and team incentives. *Journal of Economic Psychology*, *34*, 1–7. https://doi.org/10.1016/j.joep.2012.10.011

Curşeu, P. L., Fodor, O. C., A. Pavelea, A., & Meslec, N. (2020). "Me" versus "We" in moral dilemmas: Group composition and social influence effects on group utilitarianism. *Business Ethics*, *29*(4), 810–823. https://doi.org/10.1111/beer.12292

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

Dahl, A., Schuck, R. K., & Campos, J. J. (2013). Do young toddlers act on their social preferences? *Developmental Psychology*, *49*(10), 1964–1970. https://doi.org/10.1037/a0031460

Damon, W., & Killen, M. (1982). Peer Interaction and the Process of Change in Children's Moral Reasoning. *Merrill-Palmer Quarterly*, *28*(3), 347–367.

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. Journal of Personality and Social Psychology, 8(4), 377–383. https://doi.org/10.1037/h0025589

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, *6*(4), 485–507. https://doi.org/10.1007/s11097-007-9076-9

Dungan, J. A., Chakroff, A., & Young, L. (2017). The relevance of moral norms in distinct relational contexts: Purity versus harm norms regulate self-directed actions. *PLoS ONE*, *12*(3). https://doi.org/10.1371/journal.pone.0173405

El Zein, M., Bahrami, B., & Hertwig, R. (2019). Shared responsibility in collective decisions. *Nature Human Behaviour*, *3*(6), 554–559. https://doi.org/10.1038/s41562-019-0596-4

El Zein, M., Seikus, C., De-Wit, L., & Bahrami, B. (2020). Punishing the individual or the group for norm violation. *Wellcome Open Research*, *4*, 139. https://doi.org/10.12688/wellcomeopenres.15474.2

Fancher, R. E., & Rutherford, A. C.(2012). *Pioneers of psychology: a history* (4th ed).

Farr, R. M. (1986). The Social Psychology of William McDougall. In *Changing Conceptions of Crowd Mind and Behavior* (pp. 83–95). Springer New York. https://doi.org/10.1007/978-1-4612-4858-3_6

Feldman, R. S., & Rosen, F. P. (1978). Diffusion of responsibility in crime, punishment, and other adversity. *Law and Human Behavior*, *2*(4), 313–322. https://doi.org/10.1007/BF01038984

Fochmann, M., Fochmann, N., Kocher, M. G., Müller, N., & Wolf, N. (2021). Dishonesty and risk-taking: Compliance decisions of individuals and groups. *Journal of Economic Behavior and Organization*, *185*, 250–286. https://doi.org/10.1016/j.jebo.2021.02.018

Foot, P. (1967). The Problem of Abortion and the Doctrine of the Double Effect. *Oxford Review*, *5*, 19–32. https://doi.org/10.1093/0199252866.003.0002

Gallotti, M., Fairhurst, M. T., & Frith, C. D. (2017). Alignment in social interactions. *Consciousness and Cognition*, *48*, 253–261. https://doi.org/10.1016/j.concog.2016.12.002

Gallotti, M., & Frith, C. D. (2013). Social cognition in the we-mode. *Trends in Cognitive Sciences, 17*(4), 160–165. https://doi.org/10.1016/j.tics.2013.02.002

Goette, L., Huffman, D., & Meier, S. (2006). The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups. *The American Economic Review, 96*(2), 212–216. https://doi.org/10.1257/000282806777211658

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*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*(5827), 998–1002. https://doi.org/10.1126/science.1137651

Haidt, J., & Kesebir, S. (2010). Morality. In *Handbook of Social Psychology*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470561119.socpsy002022

Herschbach, M. (2012). On the role of social interaction in social cognition: A mechanistic alternative to enactivism. *Phenomenology and the Cognitive Sciences*, *11*(4), 467–486. https://doi.org/10.1007/s11097-011-9209-z

Higgins, J. (2020). Cognising With Others in the We-Mode: a Defence of 'First-Person Plural' Social Cognition. *Review of Philosophy and Psychology*, 1–22. https://doi.org/10.1007/s13164-020-00509-2

Hogg, M. A., & Williams, K. D. (2000). From I to we: Social identity and the collective self. *Group Dynamics: Theory, Research, and Practice, 4(*1), 81–97. https://doi.org/10.1037/1089-2699.4.1.81

Hutto, D. D. (2004). The Limits of Spectatorial Folk Psychology. *Mind & Language, 19*(5), 548–573. https://doi.org/10.1111/j.0268-1064.2004.00272.x

Irlenbusch, B., Mussweiler, T., Saxler, D. J., Shalvi, S., & Weiss, A. (2020). Similarity increases collaborative cheating. *Journal of Economic Behavior and Organization*, *178*, 148–173. https://doi.org/10.1016/j.jebo.2020.06.022

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological Review*, *99*(3), 561–565. https://doi.org/10.1037/0033-295X.99.3.561

Keasey, C. B. (1973). Experimentally induced changes in moral opinions and reasoning. *Journal of Personality and Social Psychology*, *26*(1), 30–38. https://doi.org/10.1037/H0034210

Klautke, E. (2013). *The Mind of the Nation* (1st ed.). Berghahn Books.

Kocher, M. G., Schudy, S., & Spantig, L. (2018). I lie? We lie! Why? Experimental evidence on a dishonesty shift in groups. *Management Science*, *64*(9), 3995–4008. https://doi.org/10.1287/mnsc.2017.2800

Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. *Theory Into Practice, 16*(2), 53–59. https://doi.org/10.1080/00405847709542675

Le Bon, G. (1960). The crowd: A study of the popular-mind. New York: Viking Press.

Leach, C. W., Bilali, R., & Pagliaro, S. (2015). *Groups and morality.* In M. Mikulincer, P. R. Shaver, J. F. Dovidio, & J. A. Simpson (Eds.), *Group processes* (p. 123–149). American Psychological Association. https://doi.org/10.1037/14342-005

Lucas, B. J., & Livingston, R. W. (2014). Feeling socially connected increases utilitarian choices in moral dilemmas. *Journal of Experimental Social Psychology*, *53*, 1–4. https://doi.org/10.1016/j.jesp.2014.01.011

Maitland, K. A., & Goldman, J. R. (1974). Moral judgment as a function of peer group interaction. *Journal of Personality and Social Psychology*, *30*(5), 699–704. https://doi.org/10.1037/h0037454

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–299. https://doi.org/10.1037/a0036677

Mercier, H. (2011). What good is moral reasoning. *Mind & Society, 10*(2), 131–148. https://doi.org/10.1007/s11299-011-0085-6

Métais, F., & Villalobos, M. (2021). Embodied ethics: Levinas' gift for enactivism. *Phenomenology and the Cognitive Sciences*, *20*(1), 169–190. https://doi.org/10.1007/s11097-020-09692-0

Michael, J. (2011). Shared Emotions and Joint Action. *Review of Philosophy and Psychology*, *2*(2), 355–373. https://doi.org/10.1007/s13164-011-0055-2

Milgram, S., Bickman, L., & Berkowitz, L. (1969). Note on the drawing power of crowds of different size. *Journal of Personality and Social Psychology*, *13*(2), 79–82. https://doi.org/10.1037/h0028070

Moscovici, S. (1986). Introduction. In *Changing Conceptions of Crowd Mind and Behavior* (pp. 1–4). Springer New York. https://doi.org/10.1007/978-1-4612-4858-3_1

Navajas J, Álvarez Heduan F, Garrido JM, et al. Reaching Consensus in Polarized Moral Debates. *Current Biology* : CB. 2019 Dec;29(23):4124-4129.e6. https://doi.org/10.1016/J.CUB.2019.10.018

Nichols, M. L., & Day, V. E. (1982). a Comparison of Moral Reasoning of Groups and Individuals on the "Defining Issues Test." *Academy of Management Journal*, *25*(1), 201–208. https://doi.org/10.2307/256035

Niebuhr, R. (1932). *Moral Man and Immoral Society: A Study in Ethics and Politics*. Charles Scribner's Sons.

Parkin-Gounelas, R. (2014). The psychology and politics of the collective groups, crowds and mass identifications. Routledge, Taylor & Francis Group.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, *17*(3), 145–171. https://doi.org/10.1016/0162-3095(96)00041-6

Piaget, J. (1933). The Moral Judgement of the Child. *Philosophy* 8 (31):373-374.

Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations*, *9*(1), 29–43. https://doi.org/10.1080/13869790500492466

Prinz, J. (2010). The Moral Emotions. In P. Goldie (Ed.), *The Oxford Handbook of Philosophy of Emotion*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199235018.003.0024

Roth, W. M., & Jornet, A. (2013). Situated cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(5), 463–478. https://doi.org/10.1002/wcs.1242

Salmela, M. (2012). Shared emotions. *Philosophical Explorations*, *15*(1), 33–46. https://doi.org/10.1080/13869795.2012.647355

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(4), 393–414. https://doi.org/10.1017/S0140525X12000660

Schmitt, F. (2017). Social Epistemology. In *The Blackwell Guide to Epistemology* (pp. 354–382). Blackwell Publishing Ltd. https://doi.org/10.1002/9781405164863.ch15

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences, 10*(2), 70–76. https://doi.org/10.1016/j.tics.2005.12.009

Sharvit, K., Brambilla, M., Babush, M., & Colucci, F. P. (2015). To Feel or Not to Feel When My Group Harms Others? The Regulation of Collective Guilt as Motivated Reasoning. *Personality and Social Psychology Bulletin*, *41*(9), 1223–1235. https://doi.org/10.1177/0146167215592843

Sievers, B., Welker, C., Hasson, U., Kleinbaum, A., & Wheatley, T. (2020). *How consensus-building conversation changes our minds and aligns our brains*. https://doi.org/10.31234/osf.io/562z7

Styron, W. (1979). *Sophie's Choice*. Random House.

Takezawa, M., Gummerum, M., & Keller, M. (2006). A stage for the rational tail of the emotional dog: Roles of moral reasoning in group decision making. *Journal of Economic Psychology*, *27*(1), 117–139. https://doi.org/10.1016/j.joep.2005.06.012

Tang, S., Koval, C. Z., Larrick, R. P., & Harris, L. (2020). The morality of organization versus organized members: Organizations are attributed more control and responsibility for negative outcomes than are equivalent members. *Journal of Personality and Social Psychology*, *119*(4), 901–919. https://doi.org/10.1037/PSPI0000229

Thompson, E., & Stapleton, M. (2009). Making sense of sense-making: Reflections on enactive and extended mind theories. *Topoi*, *28*(1), 23–30. https://doi.org/10.1007/s11245-008-9043-2

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217. https://doi.org/10.5840/monist197659224

Tuomela, R. (2007). The Philosophy of Sociality: The Shared Point of View. In *The Philosophy of Sociality: The Shared Point of View*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195313390.001.0001

Ueshima, A., Mercier, H., & Kameda, T. (2021). Social deliberation systematically shifts resource allocation decisions by focusing on the fate of the least well-off. *Journal of Experimental Social Psychology*, *92*. https://doi.org/10.1016/j.jesp.2020.104067

Urban, P. (2014). Toward an expansion of an enactive ethics with the help of care ethics. *Frontiers in Psychology*, *5*(NOV), 1354. https://doi.org/10.3389/fpsyg.2014.01354

Vainapel, S., Weisel, O., Zultan, R., & Shalvi, S. (2019). Group moral discount: Diffusing blame when judging group members. *Journal of Behavioral Decision Making*, *32*(2), 212–228. https://doi.org/10.1002/bdm.2106

Varela, F. J. (1999). *Ethical know-how: action, wisdom, and cognition*. Stanford University Press.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, Mass: MIT Press.

Weisel, O., & Shalvi, S. (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(34), 10651–10656. https://doi.org/10.1073/pnas.1423035112

Wilson, R. A. (2004). *Boundaries of the Mind: The Individual in the Fragile Sciences - Cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511606847

Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., & Chapman, M. (1992). Development of Concern for Others. *Developmental Psychology*, *28*(1), 126–136. https://doi.org/10.1037/0012-1649.28.1.126

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*(2), 151–175. https://doi.org/10.1037/0003-066X.35.2.151

Chapter 2. Making Moral Decisions and Judgments *Together*

Anita Keshmirian[1,2], Sofia Bonicalzi [3]

(1) Graduate School for Neuroscience, Ludwig-Maximillian's-University, Munich, Germany.
(2) Faculty of Philosophy, Ludwig-Maximilian's University, Munich, Germany.
(3) Department of Philosophy, Communication and Performing Arts, Rome Tre University, Rome Italy.

Abstract: Moral issues are often the topics of extensive and lively interactions in social life. To date, however, research on moral decision and judgment has predominantly focused on individuals contemplating moral matters privately. We review the motivations underlying this approach and discuss why it is problematic: collective moral decisions and judgments (1) are characterized by specific dynamics that make them significantly different from individual decisions and judgments; and (2) are also different from collective non-moral decisions and judgments as they share properties with both *intellective* (matters of truth) and *judgmental* (matters of facts) tasks. We conclude that a thorough examination of the collective dimension is needed to promote a more inclusive understanding of the mechanisms underlying moral cognition.

2.1 Introduction

A good deal of our moral cognition is collective. Consider moral concerns intersecting with debated social matters: an ethics committee approving a research project involving animal pain; a family arguing over terminating life support for their comatose child; a panel of politicians discussing the ethics of drone warfare; a group of social media users debating online on the tradeoff between losing jobs and closing a polluting power plant. In these and many other cases, we make moral decisions *with others*, carry the weight of their consequences *together*, or judge

ethical issues *within our social groups*. Indeed, violations of moral norms encourage gossip, public shaming and sometimes lead to social or legal changes.

Given the ubiquity and impact of collective moral decisions and judgment, understanding their mechanisms is critical. But, in cognitive science, moral cognition has so far been very much treated as a private matter (Bloom, 2010; Ellemers, 2017; Ellemers et al., 2013, 2019; Fedyk, 2017; Haidt, 2007; Mercier, 2011). This paper focuses exactly on the question of how *groups* make moral decisions (what we ought to do) and judgments (what ought to be done) *together*, e.g., by processing (morally relevant) information in groups, persuading each other, signaling inclinations, casting votes, compromising, and/or reaching a consensus. With this, we aim to invite researchers in cognitive science to more actively focus on the *collective* dimension of moral cognition and hint at specific directions for future research.

To prove the importance of the collective dimension, we offer two main arguments. First, we argue that collective moral decisions and judgments are conceptually and cognitively *distinct* from the aggregate of the individual ones (taken in isolation). To this end, we draw on evidence in cognitive science stressing the role of emergent factors specifically arising in collective contexts, notably emotion-based and reason-based dynamics, virtue signaling, and diffusion of responsibility. We discuss how the process and content of the related decisions and judgments are affected by the interplay of these factors.

Second, we argue that collective moral decisions and judgments are also distinct from the collective *non-moral* ones. Most collective tasks can be neatly categorized as matters of objective truth (e.g., solving a mathematical equation) or of subjective opinion (e.g., casting votes in a beauty contest). By contrast, moral tasks do not have this clear-cut nature and cannot be straightforwardly ascribed to any of such categories. We then review existing evidence in moral psychology showing that people are neither fully objectivists nor fully subjectivists about the validity of their morals (as we later explain, they can be best qualified as *metaethical pluralists*). We discuss how this metaethical outlook bears relevance in collective discussions, especially when group members embrace divergent views or have to rely on (the expertise of) others.

Our paper concerns moral cognition at a descriptive level by examining what people *find* morally 'good' or 'bad' together. It does not concern what is practically or rationally 'right' or 'wrong.' Moreover, in this paper, morality does not refer to people's character (e.g., politeness, kindness, etc.) but to the *content* of people's moral cognition and to the *process* through which they integrate morally relevant information, emotions, biases, intuitions, etc., in order to make moral decisions and judgments in groups. Throughout our analysis, we assume that, overall, it makes sense to treat collective moral decisions and judgments as having some key commonalities. Focusing on those, an oversimplification is in order: we gloss over the differences concerning the *nature* (e.g., formal or informal, hierarchical or flat) and *size* (e.g., small or large) of the group; the discussion *topic* (e.g., bioethics, social justice, purity violations) and *format* (online, face-to-face); and the *consequences* of the decision or judgment (i.e., whether any relevant consequence stems from them).

The paper is structured as follows. In § 2, we review some motives why most models of moral cognition are centered on individuals (rather than groups) and discuss a more recent collective-oriented trend. In § 3, we explore the emergent features of collective moral decisions and judgments as compared to the aggregate of the individual ones taken in isolation. In § 4, we focus on collective moral decisions and judgments as compared to the non-moral ones. In § 5, we conclude that neglecting the collective dimension prevents us from having a comprehensive understanding of moral cognition. In this spirit, we invite cognitive scientists to further pursue this line of enquiry.

2.2 Why the cognitive science of morality has mainly focused on individuals

"How people make moral decisions and judgments *together*" is an empirical question that has previously received relatively little attention in the cognitive science of morality. This section identifies some potential motives underlying this neglect across three approaches to moral cognition: experimental moral philosophy, moral psychology, and the cognitive neuroscience of morality. We also discuss an emerging, more collective-oriented trend in the study of moral cognition.

2.2.1 Philosophical conceptualizations

Moral psychologists often glean ideas, predictions, and even methods from moral philosophy. And throughout its history moral philosophy has been interested in *normative* questions about how individuals ought to live. In this sense, deontological ethics, utilitarianism, and virtue ethics are the three dominant normative moral theories.

Centered on moral norms, deontology is paradigmatically associated with Kant's *categorical imperative* (Kant, 1785/1993). The categorical imperative is meant to be intrinsically obligating and universally valid: every rational person ought to act in accordance with the moral law, independently of contingent motives or cultural differences. Then, to the extent that the moral norm is applied, it can be relatively unimportant whether this ideal and abstract reasoner is an individual or a group. In turn, utilitarianism suggests that the moral rightness of an act (Mill, 1861/1998) or of a rule (Berkely, 1712/1972) consists in its promoting the maximum good: deciding what one ought to do consists in solving cost-benefit calculations between different options from an impartial observer's view. What counts is that the outcome impartially subserves the utilitarian interest. Thus, once again, whether the decision-maker is an individual or a collective might be of little relevance. Finally, Aristotelian virtue ethics pays attention to what virtues (i.e., character traits, dispositions) *individuals* must develop to live good lives or nourish their moral character (Anscombe, 1963).

While the individual or collective identity of the ideal reasoner is of little importance to decide what ought to be done, the individualistic point of view is rather crucial for other aspects of normative theories. For example, according to classic theories of moral responsibility, we must be able to blame single individuals who commit moral wrongs without allowing them to shake off their responsibilities by blending into the group, and humans must be able to develop their moral selves as individuals (Bonicalzi, 2019; Fischer and Ravizza, 1998). Furthermore, moral philosophers have also been interested in what motivates individuals to behave morally: different philosophical traditions see reasons (e.g., Kantian ethical rationalism) or rather feelings and emotions (e.g., Humean ethical sentimentalism) as intrinsic moral drives (Dancy, 2003).

What is problematic is that the individualistic trend, which waters down the distinction between the *I* and the *We*, often percolates tacitly into empirical moral psychology. Indeed, inspired by these various philosophical theories, empirical moral psychologists have been keen to understand how these philosophical notions (e.g., deontology, utilitarianism, rationalism, sentimentalism) fit actual moral intuitions (Białek et al., 2019; Greene et al., 2001, 2008; Kvaran et al., 2013; Strohminger et al., 2011; Valdesolo & Desteno, 2006). However, transitioning from this ideal and abstract reasoner (in philosophy) to concrete individuals (in psychology), researchers seem often to put aside an important fact: while the 'I vs. We' distinction may be of little relevance to standard normative theories, it suddenly becomes central to understand how people concretely gauge their morals in real life. As Bernard Williams (2011) puts it, if taken too literally, systematic moral theories will fail to capture the nuances of our everyday moral life. One of such nuances is indeed the collective dimension in which much moral cognition takes place.

The influence of individualistic moral philosophy on psychology is reflected vividly in how the latter has used classic philosophical dilemmas to test moral decisions and judgments taken by individuals in isolation. Philosophers originally devised moral dilemmas as thought experiments to assess the normative plausibility of deontological or utilitarian ethics (Foot, 1967; Thomson, 1976). So-called *sacrificial dilemmas*, such as the Trolley problem, describe hypothetical situations in which isolated bystanders must decide whether to sacrifice one to save many. These dilemmas were initially built to show, for example, whether utilitarianism is sufficiently respectful of justice and rights, and not to measure laypeople's intuitions. Moral dilemmas were then imported into empirical moral psychology to test whether subjects sitting for psychological experiments tend to be more utilitarian or deontological under various experimental manipulations (e.g., high vs. low attentional load (Tinghög et al., 2016)) or whether individuals are motivated by rational calculations or emotions (e.g. Greene et al., 2001; see Christensen & Gomila, 2012 for a review).[4] Although this research endeavor per se is intrinsically

[4] Previous researchers have raised critical concerns about the generalizability of such uncommon situations (Everett & Kahane, 2020; Kahane et al., 2015) or their proving that individuals are driven by authentically utilitarian or deontological intuitions (Bostyn et al., 2018; Kahane, 2015).

relevant, psychologists have mostly detracted from raising the question "how do we decide together?" in such situations.

Overall, the collective dimension is not alien to philosophical theories of practical reasoning and normativity. Discussions about how collectives act together are central to various philosophical fields or sub-fields, such as social epistemology (Fricker et al., 2021), collective intentionality (Sellars, 1974) and responsibility (Pettit, 2007),[5] group agency (List & Pettit, 2011), team reasoning (Gold & Sudgen, 2007), social choice theory (List, 2012) and joint commitment (Gilbert, 2014). For instance, the collective dimension of moral, and also political, decision-making is particularly central to the *social contract tradition* (including Hobbes, Locke, Rousseau, Kant, Rawls, and Scanlon) wherein moral and/or political obligations emerge from hypothetical agreements between idealized society members.

In particular, in his *A Theory of Justice* (Rawls, 1971/2017), Rawls devised an ideal scenario (the *original position*) from which people jointly decide about basic principles of justice regulating a fair society: the reasoners are asked to think of themselves as free and equal, endowed with fundamental interests, the general capacity to rationally form and commit to life plans, some competence about economics and the sciences, and some general psychological or biological features. Crucially, they should imagine themselves as ignorant about their own historical circumstances, personal interests, social roles, race, or gender (they make decisions under a *veil of ignorance*). By abstracting away from such personal details, the veil of ignorance promotes choices made from a neutral point of view, in accordance with the philosophical tradition of the moral reasoner as a *judicious* and *impartial spectator* discussed by Hume, Smith, or Sidgwick. Experimental attempts at modelling decisions under the Rawlsian veil of ignorance indicated that individuals are actually able to make impartial decisions (although in a utilitarian fashion) (Huang et al., 2019). However, when it comes to real life, such assumptions are more difficult to apply. In collective moral cognition, we deal with humans negotiating their preferences and making efforts to agree on shared options, without possibly ignoring their own preferences and histories.

We do not enter the debate concerning the extent to which the individualistic, idealized, ahistorical point of view must be taken as central to normativity.[6] Our point is rather that this perspective is not fully informative and can be potentially misleading when it is taken as a blueprint for a research program in moral cognition.

[5] Collective responsibility has also been at the heart of influential works in post-war 20th-century philosophy (e.g., Arendt, 1987), hugely inspiring the early days of social psychology (Festinger & Carlsmith, 1959; Milgram, 1963; Myers & Bishop, 1970; Myers & Kaplan, 1976; Myers & Lamm, 1976; Wallach et al., 1964; Wallach & Kogan, 1965). However, while philosophical models focus on whether attributing responsibility to collectives is plausible (for *collectivism*, see e.g., Gilbert, 2014; for *individualism*, see e.g., Narveson, 2002), little attention has been paid to how people collectively make the decisions for which they can then be held responsible.

[6] Against the oversimplification brought about by too idealized settings, some philosophers – notably Habermas (1995) with his *discourse ethics* – have highlighted the intrinsically dialogical dimension from which shared everyday moral practices arise. Well-known criticisms of impartialist, abstract and atomistic moralities have been raised also in the context of feminist (e.g., Benhabib, 1992) and communitarian (e.g., MacIntyre 1988) critiques, stressing the role that social groups have in nourishing our moral and political judgment.

When we examine how *real* people – with their gender, personal interests, and social roles – participate in moral discussions in real life, such idealized assumptions fall apart.

In sum, the close relationship between moral philosophy and empirical moral psychology may have some cost (see also Blasi, 1990). One of such costs, we argued, is that a good deal of philosophical moral theories is individualistic, and this individualistic approach has been often imported uncritically into empirical moral psychology.

2.2.2 Psychological conceptualizations

Psychological conceptualizations of moral cognition, with respect to how they conceive the collective dimension, can be roughly divided into two camps. We call them *standard* (§ 2.2.1) and *contemporary* views (§ 2.2.2). We first discuss how the former is partially responsible for the dominant individualistic narrative. Next, we discuss the latter and how the role of groups is progressively emerging in it.

2.2.2.1 The standard views

Standard views in moral psychology mainly consist of three major trends, each emerging from a different research area: *developmental*, *affective* and *personality* psychology.

The first trend is associated with the work of leading developmental psychologists. Preoccupied with the investigation of rational moral judgments in children as a function of their developmental stages, pioneers like Piaget (1993) or Kohlberg (Kohlberg & Hersh, 1977) established *moral psychology* as an independent discipline focusing on individuals rather than groups (see also Leach et al., 2015). For Kohlbergian moral psychologists (e.g., Damon & Killen, 1982; Maitland & Goldman, 1974; Nichols & Day, 1982), the interactive social aspects were critical, but only as the *context* explaining how individuals form specific moral judgments rather than as the *agent* collectively making such judgments.

An initial interest for the collective and interactive dimension of moral cognition started to rise within this development approach, but was unfortunately abandoned in later years. Following the methods developed by the Kohlbergian school, a few studies showed that groups are more *advanced* (in the Kohlbergian sense) than individuals in their moral reasoning: following discussion, groups of students gave proof of higher developmental stages compared to individuals, especially when they socially deliberated on moral dilemmas, e.g., stealing a drug in order to save someone's life (Damon & Killen, 1982; Maitland & Goldman, 1974; Nichols & Day, 1982). Using the '+1 manipulation' technique in dyadic discussions (with participants' being exposed to partners in the stage directly above theirs), some studies demonstrated that moral judgments in adults could change, leading individuals to more advanced views. In particular, people were shown to change their views based on small differences (weak disagreement) at the interpersonal level (see Keasey, 1973). As a result, this method was proposed as

the optimum way to change the target's moral views (Berkowitz et al., 1980; Berkowitz & Gibbs, 1983). Although informative, these attempts were sporadic: developmental moral psychology remained more interested in deliberative reasoning within individuals, not between them (see Leach et al., 2014).

As the years progressed, the rationalist view gave way to a different *Zeitgeist*, centered on emotions and automatic processes (Zajonc, 1980) as key drivers in moral cognition (Blasi, 1980). This *emotional/affective* approach, the second major trend in moral psychology, suggested that moral deliberation is primarily based on quick moral intuitions we might be unaware of, rather than on effortful reasoning (Haidt, 2001; 2007). Several related accounts underscored the role of fast intuitions, emotions, and automaticity as central to moral cognition (Prinz, 2010; but see Huebner et al., 2009; Pölzler, 2015). This trend started acknowledging the *social function* of moral decisions and judgments. In particular, Haidt assumed that these quick intuitions were not merely the outcome of our private thinking but developed primarily in the interpersonal context (Haidt 2001). Yet, the focus remained on the intuitive/affective aspects of *individual* moral cognition while the collective aspect remained underexplored (Haidt 2007).

The last trend, rooted in *personality* psychology, is the tendency to explore moral decisions and judgments as the outcome of stable idiosyncratic features – interindividual differences and personality traits – in individuals (Lifton, 1985), e.g., age and education (Rest et al., 1978), gender (Atari et al., 2020), political orientation (Graham et al., 2009), religiosity (Piazza & Sousa, 2014), genetics (Campbell et al., 2009) and personality (Leslau, 1994). These stable features contribute to forming the individual's *moral character* (Blasi, 2005), remaining steady across different situations (Vranas, 2004). Moreover, moral reasoning was argued to be affected by personality disorders such as psychopathy, narcissism, neuroticism, and sadism – as people with such disorders show more eccentric moral behaviors (see Arvan, 2013; Bartels & Pizarro, 2011; Djeriouat & Trémolière, 2014; Koenigs et al., 2012; Pailing et al., 2014; Pletti et al., 2017). While these features have an impact on moral cognition, it remains nonetheless unclear how they affect moral decisions and judgments in groups.

2.2.2.2 The contemporary views

Recent, but already established, trends in moral cognition have turned more decisively to social, affective and interactive approaches (Colombetti & Torrance, 2009; Métais & Villalobos, 2021; Urban, 2014; inspired by Varela, 1999). Notably, emotions-laden interactions are more and more accepted as key components of our moral life: as humans, we are not isolated atoms, but interacting minds with extended bodies and affective states, navigating the social contexts and learning from others (Colombetti & Torrance, 2009). Within such interactive social contexts, moral judgments become forms of 'social doing' to build one's own reputation (Haidt, 2007) or persuade others (Mercier, 2011). Correspondingly, even in the developmental psychology camp, moral cognition is more consistently examined as a function of communication and interaction during the first years of life (Dahl et al., 2013).

However, with some exceptions (see, e.g., Gallotti & Frith, 2013; Schilbach et al. 2012), even collective-oriented social psychologists have mostly looked at the group dimension as an independent variable affecting moral behaviors in single individuals. We know, for instance, that social norms and group values inform individual moral reasoning (Ellemers, 2017; Ellemers et al., 2013), moral identity (Leach et al., 2015) and the behavior of group members (Brambilla et al., 2012). Individual values are highly related to those of the social group (Sharvit et al., 2015), are susceptible to whether one is judging outgroups or ingroups (Cohen et al., 2006; Goldring & Heiphetz, 2020; Leidner et al., 2010; Van Laar et al., 2014), undergo the influence of creative members of society (Pizarro et al., 2006), and can in turn shape moral emotions towards ingroups and outgroups (Cikara et al., 2011; Cikara & Fiske, 2012). Even random assignment to arbitrary groups has been shown to drive specific moral actions (Goette et al., 2006).

In sum, these accounts look at moral cognition as a more social enterprise. However, one important limitation is that collective moral decisions and judgments per se (i.e., when the group is *the agent* and not *a context*) received relatively little attention. Other recent analytical reviews highlight this gap. For instance, Ellemers et al. (2019) provided a systematic review of the papers published in moral psychology journals in the period 1960-2017. Based on the review, only *one percent* of the reported studies concern some aspects of group-based morality, let alone collective decisions and judgments. Consequently, some scholars have warned us about how little we know about its underlying mechanisms by implying that "something is missing" in the current empirical research. This missing element is framed either as moral persuasion (Bloom, 2010), social argumentation (Mercier, 2011) or change in moral intuitions in interpersonal contexts (Haidt, 2007). However, even such critical approaches often do not stress the relevance of moral decisions and judgments as collective endeavors. Indeed, *none* of the reviewed four central ways in which groups affect morality (including the social nature of moral conventions), in Leach et al. 2015, concerns the dynamics of collective and interactive decisions and judgments.

Recently, empirical attempts have been made to fill the gap between collective and individual moral decisions and judgments. For instance, some studies showed that, when decisions are made collectively, people more easily accept breaches that increase the benefit of many (Curşeu et al., 2020; Keshmirian et al., 2021). One study identified *group rationality* as the mechanism shifting people's views towards accepting moral breaches (Curşeu et al., 2020). By contrast, another study proposed *reduction of stress and negative emotions* as the potentially underlying mechanism (Keshmirian et al., 2021). While both of them showed that collectives tend to be more utilitarian in moral dilemmas, another study yielded the opposite effect in resource allocation dilemmas: following a short discussion, the joint allocation of limited resources led to less utilitarian and less egalitarian distributions while benefitting the least well-off (Ueshima et al., 2021). Another study showed that, when social connections between group members are induced via experimental manipulations, dyads are more likely to jointly agree about sacrificing one to save many (Lucas & Livingston, 2014). Yet another study showed that differences

between individual and collective moral decisions can be limited to certain age groups (Takezawa et al., 2006).

2.2.3 Neuroscientific conceptualizations

Standard psychological models see moral reasoning as the outcome of individual heads. This is the case, and quite obviously so, also for the cognitive neuroscience of morality, which focuses on its brain basis (Ashcroft, 2005; Churchland, 2008; Decety & Wheatley, 2015; Greene, 2015; Moll et al., 2005. But see Schilbach et al. 2012).

Analogously to psychology, and perhaps due to its historical methodological limitations (e.g., lack of technologies that allow simultaneous brain imaging during interaction), this field treats moral decisions and judgments vastly as individual matters grounded in biological differences. Individual features such as the cortical thickness of specific brain areas (Patil et al., 2020), genetic specificities in the endogenous serotonin level (Marsh et al., 2011) or in the expression of oxytocin receptors (Bernhard et al., 2016), the activity of brain areas involved in individual decisions (Wiech et al., 2013), brain damages (Ciaramelli et al., 2007; Koenigs et al., 2007), the activity of the vagus nerve (Park et al., 2016) can affect moral decisions and judgments. The association between moral attitudes and brain disorders – e.g., breakdown of brain networks in psychopathy (Pujol et al., 2012), neural signatures of psychopathy (Glenn et al., 2009), and neuroticism (Harenski et al., 2009) – also contribute to the claim that morality is an individual matter.

Opposing the idea that moral cognition is exclusively rooted in biological differences at the individual level, a few steps have been made to understand its neural basis in relation to social settings. These studies, however, mostly focused on the social *origins* of moral and immoral behaviors in individuals, e.g., the impact of social norms and social feedback. For instance, prior understanding of the co-participant's moral character is shown to affect our reliance on feedback mechanisms in brain (see Ellemers & Van Nunspeet, 2020 for a review). By contrast, we know little about the brain basis of moral cognition in *interacting* individuals.

This deficiency can be now (partially) overcome given the availability of new techniques such as using hyper-scanning to test how two or more brains interact (Czeszumski et al., 2020; Dikker et al., 2017; Dumas et al., 2011; Hasson et al., 2012; Hu et al., 2018; Konvalinka & Roepstorff, 2012; Montague et al., 2002; see Wheatley et al., 2019 for a review). Using these methods, we can acquire evidence on how inter-brain neural activity can predict collective performance in groups (Reinero et al., 2021), how consensus in value-based decisions leads to the alignment in brain signals (Sievers et al., 2020), or how brainwave synchronization predicts finding solutions in collaborative problem solving (Balconi & Vanutelli, 2017; Hirata et al., 2014; Hu et al., 2018; Liu et al., 2016; Toppi et al., 2016).

More specifically, hyper-scanning has already proven to be helpful in the study of moral cognition, for instance in investigating punishing tendencies in collective settings (Ciaramidaro et al., 2018). However, finding a reliable signature of neural

synchronicity in collective moral cognition and defining how it contributes to actual moral decisions and judgments remain open challenges. For instance, whether the strength of the coupling/connectivity can predict consensus, whether and how it occurs when people converge on similar views, whether its pattern may differ from non-moral joint cognition or vary across moral domains (i.e., harmful actions vs. purity violations (Graham et al., 2013)) are questions that can potentially be addressed using these new methods.

To sum up, we proposed several reasons why moral cognition in individuals has been a most preferred research target across different domains, i.e., experimental moral philosophy, empirical moral psychology and the cognitive neuroscience of morality. We also observed some promising surging interest in the collective dimension of moral cognition. However, further studies are needed to investigate the process and content of collective moral cognition. In particular, more insight must be gained into how moral persuasion, negotiation, co-argumentation, and co-deliberation may work when groups act as the agents of moral decisions or judgments. Perhaps we are ready for a new turning point in moral cognition, shifting the research community's attention to the collective and interactive dimension of decisions and judgments.

2.3 Collective vs. individual moral cognition

As discussed in § 2, moral psychology and neuroscience have only recently developed an interest in the collective dimension of moral cognition. In this section, we move the discussion forwards: we rely on existing evidence about mechanisms that may help disentangling individual and group moral decisions and judgments. On this ground, the goal is to make more specific predictions about how their processes and outputs may turn out to be interestingly different. We argue that collective moral decisions and judgments, *when agents act as a group*, cannot be fully explained in terms of the mere aggregate of the attitudes of single group members taken in isolation: when collectives decide or judge together, arising affective, deliberative and social dynamics may shape the related processes and outputs.[7] This suffices to make collective moral cognition qualitatively different

[7] The thesis that collective mental processes are irreducible to the summation or aggregate of individual mental processes and can be attributed to groups in a collective way is central to theories of *collective intentionality* (Gallotti & Frith, 2013; Searle, 1995; Sellars, 1974; Tuomela, 2007; Higgins, 2020; Wilson, 2004). Here, we do not commit to a specific theory of collective intentionality, nor we argue in favor of the intrinsic irreducibility of collective mental processes. What we offer is rather a sketch of some distinctive features emerging in collective moral decisions and judgments. We argue that, when people come together to make moral decisions and judgments, specific emotional, deliberative, and social mechanisms may arise, i.e., where "specific" means that they tend not to come up when people debate on the same issues *alone*.

from individual moral cognition. Rather than striving for exhaustiveness, we aim to identify some key factors explaining such specificities. To review the mechanisms responsible for them, we deploy the resources of two different lines of research.

The first comes from a couple of often paired distinctions in psychology: affect/reason and intuition/deliberation. When people decide socially on moral issues, the interplay of these contrastive mechanisms may steer the underlying processes and outputs towards different directions: 1. The affective and intuitive mechanisms may take over, pushing the group towards acting more intuitively and/or emotionally; 2. Reason-based and effortful mechanisms may take over, shifting the group towards more argumentative and deliberative moral reasoning. We call this approach *collective dual system* and explain it in § 3.1.

The second comes from a line of research showing that people join moral discussions in view of forms of "social doing" (Haidt, 2007), i.e., to fulfill social motives. Among such motives, one can include influencing others, curating one's public image, reducing blame and guilt, or establishing a social status. In particular, in § 3.2, we consider two of such crucial social motives, i.e., *virtue signaling* (how virtuous we wish to be seen by others) and *diffusion of responsibility* (joining groups to reduce our responsibility).

2.3.1 Collective Dual System

There are two basic models in cognitive science accounting for the mechanisms underlying individual decision-making. The first one (affect/reason) suggests that individual decisions are alternatively driven by an emotion-based and a reason-based system. While the former is responsible for processing emotions and affects, the latter is responsible for abstract representations, logical thinking, and reasoning (Abelson & Carroll, 1965).

The second model (intuition/deliberation) highlights the role of two (other) interacting – if not competing – cognitive systems: one is automatic, involved in fast, intuitive, and effortless processes, and the other is deliberative, slower, and more effortful (Evans, 2003; Evans & Stanovich, 2013; Sloman, 1996).

Although these models have been developed independently (Darlow & Sloman, 2010), they are often used interchangeably (Haidt & Kesebir, 2010) and exhibit some conceptual and functional overlaps (Lodge & Taber, 2005) – i.e., the emotion-based system with the intuition-based one, often indicated as *system 1*; the reason-based system with the deliberation-based one, often indicated as *system 2*. As a result, their respective features are merged within *dual system* accounts. In our analysis, we will follow the same convention and discuss the potential impact of system 1 and system 2 at a collective scale.

Both systems are thought to shape moral cognition in individuals (Greene, 2009; Mallon & Nichols, 2011) but, when we move to groups, how does their interplay affect collective moral cognition? This question will be addressed in §§ 3.1.1, 3.1.2 and 3.1.3.

2.3.1.1 Fast, affective, and automatic cognition (system 1)

As mentioned in § 2.2.1, many cognitive scientists follow Hume's lesson in emphasizing the central role of emotion and automaticity in moral cognition (Blasi, 1980; Greene, 2007; Haidt, 2001; 2007; Haidt & Kesebir, 2010; Prinz, 2010; but see Huebner et al., 2009; Pölzler, 2015). In particular, *social intuitionist models* insist on the primacy of affective and intuitive states in individual moral cognition. These processes operate mainly at the unconscious level while remaining introspectively opaque. In this sense, they are the opposite of slow and effortful deliberative processes, which eventually play a role only in producing an *a posteriori* rationalization of fast moral reactions.

It is well established that individuals in groups experience emotions differently than those who are isolated (e.g., Barsade & Gibson, 1998; Smith, E. R. & Mackie, 2015; 2016). But how does this transfer to collective moral decisions and judgments? In other words, if we accept that the emotional/automatic system is so central to individuals' moral cognition, how does it influence information processing at the collective level?

In 1932, the leading U.S. theologian Reinhold Niebuhr (1932) wrote that emotional impulses express themselves more vividly in groups, leading to an increase in immoral actions. At the collective level, Niebuhr believed that emotions burst out free of the constraints that would be generally imposed by human reasoning, leading to egoistic decisions. Empirical research supports this idea that people violate moral norms more often when in groups: they lie more (Conrads et al., 2013; Kocher et al., 2018), are less compliant to norms (Fochmann et al., 2021), show more antisocial behaviors (Behnk et al., 2017), and are unfair in resource distribution (Bornstein & Yaniv, 1998; El Zein et al., 2020).

Previous research provides some potential explanation for these phenomena. Since norm violation is tied to emotional processing (Baron et al., 2018; Choe & Min, 2011; Greene, 2007; Wiech et al., 2013), groups can play a role in reducing the related emotional burden, i.e., regulating decision-related proximal (stress) or predicted (anticipated regret) emotions (for analogous effects in non-moral domains, see El Zein & Bahrami, 2020; El Zein et al., 2019). As a result, the negative affect resulting from norm violation can be alleviated. Indirectly supporting this hypothesis, a number of findings show the beneficial effect of the group dimension on emotional processing: group discussion reduces individuals' negative emotions even when these feelings are artificially induced (Kaplan & Miller, 1978), while group belongingness induce positive emotions (Van Kleef & Fischer, 2016). In turn, positive emotions tend to be correlated with more utilitarian decisions, which may lead to breaking norms or harming victims to maximize the general utility (Strohminger et al., 2011; Valdesolo & Desteno, 2006).

2.3.1.2 Slow, effortful, deliberative cognition (system 2)

Not all moral psychologists see emotions as so central to moral cognition (e.g., Bloom, 2010). Even if they do so, they often do not entirely dismiss the role of reasoning and deliberation (e.g., Haidt, 2001). Previous research has already emphasized that the decision-making setting (e.g., more time to deliberate) can modulate the respective contribution of emotions and reasons (see Greene, 2007). Here, we suggest that group moral discussion can also boost slow and effortful

deliberation, helping discussants to act more reasonably (i.e., according to commonsense).

In non-moral domains, several studies have emphasized that, when people discuss, they feel less uncertain (Bang & Frith, 2017; Fusaroli et al., 2012), provide arguments (Darmstadter, 2013), understand problems from a different angle and reach solutions they would not have endorsed had they been alone (Smith, E. R. & Collins, 2009). This suggests that deliberation and analytical reasoning get promoted in group contexts: people are forced to hear and provide arguments, engage in perspective-taking, and possibly converge on shared deliberative moral judgments (Mercier, 2011).

Consistently, collective contexts and social interactions are proposed to be exceptional settings in which the emotional and automatic mechanisms can be controlled and dominated by reasons (Haidt, 2001), especially in certain age groups (Takezawa et al., 2006). As already mentioned, some studies even show that moral cognition in small groups is developmentally more *advanced* than among individuals (Damon & Killen, 1982; Nichols & Day, 1982). Based on this, interacting people may enter a social-deliberative mode of thinking, countering the effect of automatic and emotional processes. This may contribute to shifting collective moral cognition towards dynamics and outputs that would not have been reached had reasoners been alone.

2.3.1.3 Putting the pieces together: reason and emotion in groups

Based on §§ 3.1.1 and 3.1.2, both automatic/affective and reason-based/deliberative mechanisms can be at play and even be boosted in group level interactions. At a further level, we can speculate that the thinking styles of interacting group members – more emotion-based or reason-based – can give rise to unprecedented processes and outputs, i.e., with one's emotional states triggering deliberative reasoning in another and vice versa (for how this may apply to political discussions, see Sloman & Rabb, 2019).

Obviously, the interplay between the two systems may occur at the individual level as well and be driven by different factors. For instance, concealing emotions and actively suppressing analytical or intuitive reasoning (Lee & Gino, 2015), solving mathematical puzzles to enter a deliberative mode (Kvaran et al., 2013), increasing emotional distance by presenting moral issues in a foreign language (Białek et al., 2019), decreasing cognitive fatigue (Timmons & Byrne, 2019), reducing negative feeling behaviorally (Strohminger et al., 2011; Valdesolo & Desteno, 2006) or with anti-anxiety drugs (Perkins et al., 2013; but also see Zhao et al., 2016) can all shift individual moral reasoning towards deliberative or intuitive processing. But whether, how, and to what extent these factors can swing collective moral cognition is far from being clear. Therefore, understanding the mechanisms by which a group, as a whole, acts deliberatively or emotionally remains an open pathway for future research.

In sum, we used dual system accounts to explain group-level dynamics in moral discussions and suggested that their processes and outputs may be driven by the interplay between the emotion-based and the reason-based system. Both manifest themselves at different degrees within individuals and groups, depending on

various factors and potentially leading to opposite outcomes. The upshot is that collective moral decisions and judgments can be different from the aggregate of individual opinions. The crucial task for future research is thus to examine the mechanisms underlying such differences.

2.3.2 Social Motivations

Dual system accounts are only part of the story of how collective moral cognition unfolds. Additionally, motivations specifically arising at the social level may affect the dynamics, and even the goal, of the interaction. Indeed, the purpose of engaging in moral discussions may exceed that of solving a specific moral issue. As individuals, we join groups also in the guise of 'intuitive' (Haidt, 2007) or strategic politicians. For instance, we can take advantage of the collective context as a *medium* to convey messages about our moral selves – 'who we are' and sometimes 'how we wish to be seen' morally (§ 3.2.1). Or we may implicitly or explicitly aim to reduce our responsibilities, by blending into the group, for difficult moral choices and their negative consequences (§ 3.2.2).

2.3.2.1 Virtue signaling

Morally connotated statements and actions communicate crucial social information about our inclinations but also, perhaps more importantly, about how we want to present ourselves to others (Bostyn & Roets, 2017; Everett et al., 2016; Kreps & Monin, 2014; Rom & Conway, 2018; Uhlmann et al., 2013). For instance, when one publicly says that eating meat is morally wrong, she might communicate two messages: her decision not to eat meat and the general moral stance by which she wants to be recognized by.

This tendency to signaling virtues, even pretending to be more virtuous than one actually is, may then affect the content and output of the group discussion. Indeed, signaling to be virtuous is generally associated with expected benefits: people who explicitly agree with harming few but benefiting many are seen as less agreeable than their deontological counterparts (Everett et al., 2018), are praised and chosen more often as social partners (Everett et al., 2016), are perceived as especially prosocial in economic games (Capraro et al., 2018) and are regarded as having integrity, empathy, and other valuable moral qualities (Uhlmann et al., 2009, 2013). As a result, people may publicly reject certain solutions to moral problems to present themselves as particularly 'virtuous', and thus to promote themselves as trustworthy and likable (Sacco et al., 2017).

In line with this view, people exhibit specific moral features when they are aware of being observed (Kurzban et al., 2007; Lee et al., 2018), feel more socially connected (Lucas & Livingston, 2014), are with a friend (Van Gils et al., 2020) or even watch themselves in a mirror (Reynolds & Conway, 2018). This self-representation is known to be strategic: people sometimes have an explicit meta-perception of the moral information they convey, think in advance about how others will receive it, and modify their behaviors accordingly (Rom & Conway, 2018). Alternatively, this process may even occur at a more implicit level. In both cases,

we predict that moral signaling would strongly influence group dynamics, with groups opting for solutions that single members would not opt for in private.

2.3.2.2 Diffusion of responsibility

Outsourcing knowledge and sharing effort for demanding tasks are among the main reasons people decide to join groups. The dark side of this is, however, that people may feel less responsible for their own share. Thus, another important social motive affecting collective moral cognition is the pernicious phenomenon of *diffusion of responsibility* whereby single members feel less responsible for the negative outcomes of their behaviors.

Research shows that even the mere presence of others can make individuals feel less responsible. This is typically observed in relation to the well-known *bystander effect*: when several observers witness a norm violation, each of them is less likely to intervene – compared to what they would have done had they been alone (Chekroun & Brauer, 2002; Darley et al., 1968). Diffusion of responsibility is proposed as one underlying reason for the increased incidence of norm violation in groups (Forsyth et al., 2002). Feeling less responsible, people are less generous in groups (Freeman et al., 1975), show more extreme behaviors (Mathes & Kahn, 1975), punish wrongdoers less if costly (Feng et al., 2016), but are generally more punitive (Bandura et al., 1975) and aggressive (Meier & Hinsz, 2004; for a review, see El Zein et al., 2019).

The presence of others affects both explicit (subjective reports) and implicit (neurophysiological correlates) markers of responsibility: people may feel less in control of, and thus less responsible for, the outcome of their actions when others are around (Beyer et al., 2017). The fact that participants verbally report responsibility reduction suggests that they may intentionally join groups to feel less accountable (El Zein et al., 2019). Indeed, participants themselves link their dishonesty to their 'feeling less responsible' when in groups (Conrads et al., 2013). These observations predict that diffusion of responsibility contribute explaining the peculiarities of moral cognition at the group level.

To sum up, in this section we reviewed several factors emerging at the social level and that may make groups' moral decisions and judgments interestingly different from the aggregate of the individuals' ones. In particular, we trace this back to the interaction between the automatic/emotional and the deliberative/reason-based system, and to specific social motives, notably virtue signaling and diffusion of responsibility. Further investigation is needed to determine how exactly these elements may contribute to collective moral cognition.

2.4 Collective moral vs. Collective non-moral cognition

In § 3, we emphasized the differences between *individual* and *collective* moral cognition. To delve further into how people jointly discuss what ought to be done

or should have been done, here we contrast collective *moral* and collective *nonmoral* decisions and judgments. In particular, we focus on people's understanding of, and commitment to, their morals as something that may deeply affect the dynamics and outputs of collective moral cognition. To this end, we first have to briefly introduce a branch of moral philosophy interested in the nature of morals, i.e., metaethics, and then explain its potential relevance for empirical moral psychology.

Metaethics is devoted to exploring the nature of moral beliefs and values, i.e., whether certain moral statements have a truth value (*cognitivism* (Sayre-McCord, 1986)) or just express non-cognitive attitudes, such as emotions (*non-cognitivism* (Stevenson, 1937. See also Gibbard, 1990)); whether they are true objectively (*objectivism* (Enoch, 2011; Railton, 1986; Sturgeon, 1985)) or subjectively (*subjectivism* (Blackburn, 1984)); whether they are true universally (*universalism*) or relatively to cultures, traditions or even individuals (*relativism*) (Harman, 1996). Unsettled disputes about ethical disagreements (Greene, 2002) or ethical expertise (McGrath, 2019) make cognitivism, objectivism and universalism vulnerable and puzzling, potentially resulting in forms of moral skepticism (Copp, 1991).[8]

Drawing on these philosophical categories, empirical moral psychologists have recently been keen to understand how people think of their ethical beliefs and values, in particular whether they tend to be objectivists or subjectivists, universalists or relativists (Beebe, 2014; Goodwin & Darley 2008; 2013; Hopster, 2019; Sarkissian, 2016; Wright et al., 2013). This research endeavor is grounded in the hypothesis that people have such meta-beliefs about (or at least an implicit commitment to) their morals, and that these meaningfully overlap with standard philosophical categories – although laypeople may lack understanding of the related fine-grained details.

To some extent, knowing how people think of their morals may be relevant even for philosophical metaethics. In particular, many metaethicists consider making sense of laypeople's metaethical beliefs as part of their job (Sarkissian, 2016). However, methodological concerns have been raised as to whether survey-based research is apt to test the match between philosophical and psychological categories (Hopster, 2019). Moreover, surveys might be more suited to capture quick, affect-laden and context-driven guesses rather than firmly held beliefs (Bengson, 2013; Ludwig, 2010). In any case, whereas their contribution to philosophical metaethics remain debated, answering such questions would provide some relevant insight on people's explicit or implicit commitment to their morals (Wright et al., 2013). Among other factors (e.g., personality, age, education), this should be investigated as a reliable predictor of how people will argue for their morals, or even fight for them, in collective contexts.

[8] For a more systematic review of the philosophical background, see Goodwin & Darley 2013. Following their standard metaethical categories, here we distinguish between questions about the source of the ethical beliefs (whether they are true objectively or subjectively) and their scope (whether they are true universally or relatively). However, we bear in mind that, both in the philosophical and the empirical literature, these distinctions are not universally accepted. In particular, universalism and objectivism are often grouped together and contrasted with relativism (e.g., Gowans, 2021; Wright et al., 2013).

Philosophers vastly hold the view that laypeople are metaethical objectivists (Blackburn, 1984; Smith, M. A. 1994. But see Wong, 2006). However, recent empirical research on the so-called *psychology of metaethics* (Goodwin & Darley, 2010) has yielded mixed results (§ 4.1), suggesting that laypeople are neither fully objectivists/universalists nor fully subjectivists/relativists, and can be best categorized as *metaethical pluralists* (Beebe, 2014; Sarkissian, 2016; Wright et al. 2013). Thus, a worth investigating question concerns what the effect of metaethical objectivism/universalism, subjectivism/relativism or pluralism is on group-level moral cognition. Indeed, this is seemingly a crucial angle to empirically examine how people will collectively gauge their morals, i.e., when they deliberate about what to do, convince others, aggregate opinions, or reach consensus on morally divisive issues.[9]

To discuss the psychology of metaethics at the group-level, we rely on an existing classification (§ 4.2), distinguishing (on a continuum) between collective tasks that are treated as matters of objective truth (*intellective*) and collective tasks that are matters of subjective preferences (*judgmental*). Then we consider where moral decisions and judgments fall on this continuum. In this respect, many non-moral collective tasks can be more neatly categorized as intellective (think of solving a mathematical equation) or judgmental (think of a beauty contest), and therefore approached with corresponding suitable strategies. By contrast, moral tasks have a less neat status – with people being neither fully objectivists nor fully subjectivists at the intraindividual and the interindividual level. On this ground, we focus on three typical difficulties people encounter in moral discussions and that are likely to affect its dynamics and outputs. These concern the value of group discussion (§ 4.2.1), ethical disagreement (§ 4.2.2) and ethical expertise (§ 4.2.3).

Whereas these difficulties may also emerge whenever people think about their morals in isolation, they become pressing when they have to intersubjectively accommodate their moral views. As such, we indicate them as three fertile research avenues for group-level empirical moral psychology. Building on this, we comment on the challenges that empirical moral psychology must face if the concrete dynamics of collective moral cognition are to be studied (§ 4.3).

2.4.1 The psychology of metaethics in individuals

In the last few years, the empirical research on the psychology of metaethics – i.e., on how people think of their morals – has started flourishing. In particular, objectivism (often jointly with universalism) is associated with the belief that moral statements are objectively true or false and that, in case of disagreement, one of the parties must be mistaken. Conversely, subjectivism (often jointly with relativism) is associated with the belief that morals are mind-dependent and tied to subjective preferences or conventions, and that, in case of disagreement, both opponents can

[9] Group reasoning about political matters may have analogous features. However, truly divisive political issues, e.g., about taxing the rich or welcoming immigrants, often incorporate moral concerns about how society must be organized.

be justified in holding their beliefs (Goodwin & Darley, 2008; Hopster, 2019; Wright et al., 2013).

Several classic studies, often in the field of developmental psychology, have argued that laypeople tend to be objectivists. Starting in early childhood (Nucci, 2001; Turiel, 2008; Wainryb et al., 2004), healthy individuals distinguish between moral convictions and less stringent social commitments, such as conventions and behavioral standards (Bucciarelli & Johnson-Laird, 2020; Heiphetz & Young, 2017; but see Kelly et al., 2007 for a criticism of the moral/conventional distinction). Moral convictions are seen as almost as objective as scientific facts (Goodwin & Darley, 2008) and not alterable even by God's intervention (Reinecke & Horne, 2018).

Along these lines, Skitka and colleagues (2021) reviewed a range of evidence that people are objectivists or universalists about their moral convictions, and that they strongly care about defending them. In particular, moral convictions are: metacognitively perceived as having universal, generalizable, and absolute validity, while being rooted in some fundamental factual truth (Van Bavel et al., 2012); experienced as having a solid link with emotions (Skitka & Wisneski, 2011), with moral agreements or disagreements eliciting specific emotional reactions (Ryan, 2014); considered as intrinsically obligatory, even in the absence of sanctions, and self-justifying rather than as imposed by some external authority. This to the extent that, if circumstances so require, people prioritize core moral norms over unfair juridical norms (Skitka et al., 2009; Smetana, 1983; Turiel, 1983).

Furthermore, compared to non-moral beliefs, moral convictions are relatively resistant to changes and majority influence and tied to intolerance for conflicting views (Aramovich et al., 2012; Skitka et al., 2005). Supporting this, research shows that moral diversity is perceived as socially more problematic than other diversities, including demographic diversity (Haidt, 2003). Strong confidence in one's morals is associated with unwillingness to compromise (Ryan, 2019) and the tendency to dissociate oneself from dissimilar others, demonstrating resilience to disenfranchisement fears (Wright et al., 2008). In this respect, it seems that laypeople's objectivist tendencies can peacefully co-exist with the inconsistencies of everyday moral life (Campbell, 2017), such as biased resistance to persuasion (Ahluwalia, 2000), affect-laden reactions (Tangney et al., 2011), context-dependent variations (FeldmanHall et al., 2018), self-serving and confirmation biases (Lin et al., 2017), introspection failure (Nisbett & Wilson, 1977) and framing effects (Hertwig & Gigerenzer, 2011).[10]

However, more recent research has provided a more nuanced, and somehow contrasting, picture of laypeople's metaethical commitments (Beebe, 2014; Hopster, 2019). At the interindividual level, the balance between objectivist and subjectivist tendencies has been shown to depend on individual traits, such as religiosity (Goodwin & Darley, 2010), age (Beebe & Sackris, 2016), or personality (Feltz & Cokely, 2008). Moreover, at the intraindividual level, people tend to be

[10] Not all inconsistences are consciously perceived as such. For instance, in an experiment by Hall and colleagues, participants failed to notice that their surveyed moral opinions were systematically altered by the experimenter and provided *post-hoc* justifications for defending views opposing their original positions (2012).

objectivists about certain moral topics and subjectivists about others (Davis, 2021). For instance, moral statements condemning moral transgressions are seen as more objective than positive statements praising good actions (Goodwin & Darley, 2012). While public consensus on a topic increases the tendency to see the supporting moral statements as objective (Goodwin & Darley, 2012), controversial matters elicit the relativist belief that no universally correct answers exist (Heiphetz & Young, 2017). Finally, contextual factors, such as in-group and out-group dynamics between the disagreeing parties, may increase or decrease laypeople's objectivism (Sarkissian et al., 2011).

One possible explanation of these mixed results is that people are actually full-fledged objectivists or universalists but vary in their ways of classifying what counts as a moral problem. In this case, people's commitments might not be intrinsically ambivalent. More simply, subjectivist/relativist tendencies may refer to topics that participants in experiments do not consider as authentically moral. If so, once people are given the opportunity to freely choose authentically moral problems, they should then reveal objectivist tendencies. However, research has shown that people express subjectivist/relativist tendencies even when they autonomously decide what counts as moral. As a result, many have concluded that laypeople actually are *metaethical pluralists*, i.e., their metaethical tendencies vary both at the interindividual and intraindividual level depending on various factors (Pölzler, 2017; Wright et al., 2013).

Brain imaging studies have lent indirect support to the claim that laypeople are not full-fledged objectivists. If anything, they seem to bend towards subjectivism (Theriault et al., 2017). Broadly speaking, the research on the *moral brain* has reliably distinguished between the neural response to moral vs. nonmoral stimuli (Moll et al., 2001) but has failed to single out neural substrates that uniquely support moral cognition (Young & Dugan, 2011). Indeed, moral cognition is seemingly made up of the contribution of domain-general neural substrates processing emotion and social cognition, such as the ventromedial prefrontal cortex, amygdala, superior temporal sulcus, bilateral temporoparietal junction, posterior cingulate cortex, and precuneus (Greene & Haidt, 2002). This domain-general nature of the neural processing contributing to moral cognition matches the view that moral decisions and judgments are generated by a combination of rational reasoning (Monin et al., 2007), affective inputs (Haidt, 2001), individual preferences (Yang et al., 2017), and social motivations (Everett et al., 2016).

However, looking for the neural correlates of moral reasoning is not the same as looking for the neural correlates of metaethical beliefs. In this more niche area of research, a study by Theriault and colleagues (2017) showed that the neural representations of morals and subjective preferences exhibit a significant overlapping within the dorsal-medial prefrontal cortex while no common pattern of activation was found between the representation of morals and objective facts. The authors concluded that laypeople's metaethical beliefs are more subjectivist than previously thought, with the underlying neural commonalities between morals and preferences potentially explained by their analogously eliciting representations of mental states.

Taken together, the evidence in § 4.1 suggests that people are likely to be metaethical pluralists. Asking whether they are objectivists or subjectivists full stop might therefore be meaningless. More interestingly, one may wonder under which conditions people express objectivist or subjectivist tendencies (Sarkissian, 2016). But how does this pluralist outlook affect the dynamics and output of collective moral cognition? We discuss this point in the following sections.

2.4.2 From *I* to *We*: intellective and judgmental tasks

In a number of works, Laughlin and colleagues provided a useful classification of collective or group tasks. This classification organizes group tasks on a demonstrability continuum, anchored by *intellective* and *judgmental* tasks (Laughlin, 2011; Laughlin & Adamopoulos, 1980; Laughlin & Ellis, 1986). In this section, we take advantage of this existing classification to discuss the features of collective moral tasks.

Intellective tasks are heterogeneous, but all have objective solutions. These are demonstrably correct within a scientific system or system of reference, i.e., mathematical, logical or verbal-conceptual. Consider, for example, a pool of engineers arguing about the dynamic of a car crash, a team of meteorologists forecasting the weather or a group of lawyers arguing about whether underage criminals can be prosecuted in a certain jurisdiction. To solve intellective tasks, people make evidence or reason-based decisions and judgments about what is true or false.[11] People who have access to sufficient information can profitably participate in the discussion, possibly detect (in)correct answers, and be convinced by evidence or arguments provided by (expert) group members: if someone points at the correct answer, the others can then converge on the same solution.

By contrast, judgmental tasks require people to make subjective evaluations for which no demonstrably correct answer can be provided, e.g., aesthetic and attitude-based judgments like preferences about food, art, physical attractiveness. In judgmental tasks, people might have more or less solid attitudes but are not equipped with reliable tools or a solid reference system that can be referred to in order to convince opponents or novices. Therefore, these tasks have to do with negotiating preferences with others, e.g., finding solutions that fulfill most group members' *desiderata* (Laughlin, 2011; Laughlin & Adamopoulos, 1980; Laughlin & Ellis, 1986).

Categorizing a whole task as judgmental or intellective is often an over-simplification, and people might diverge on how they subjectively interpret whether a task is intellective or judgmental. Most decisions and judgments are based on various parameters or include sub-tasks, some of which will be judgmental and some of which will be intellective, and all of which may weigh in on the final solution. Consider, for example, a family that is deliberating about whether to buy

[11] We remain agnostic here about how *truth* must be interpreted (for an overview, see Burgess & Burgess, 2011), i.e., as *correspondence* with reality (David, 2018); *coherence* between what is held as true and a systematic set of beliefs (Walker, 2018); or, as in pragmatist theories, as what *works* in practice and does not conflict with experience (Misak, 2018).

their children's clothes from a local shop or an online store. Typical considerations may concern what items are more durable or less expensive (intellective task) or more likable (judgmental task), the upshot being that the final solution will be a trade-off between the most valued parameters.

At the psychological level, depending on whether one perceives a task or a sub-task as more intellective or judgmental, one is expected to put forward different fitting strategies to argue for a given solution. For example, a pool of lawyers will likely use argumentative strategies based on previous records to discuss a criminal case (intellective task). Conversely, previous records become irrelevant when the jury in a beauty contest must crown a winner and each member has a vote to cast (judgmental task). However, assuming some metaethical pluralism at the psychological level, people will not uniformly see moral tasks or subtasks as intellective or judgmental and will carve out specific argumentative strategies depending on circumstances. Consider a daughter aiming to convince her reluctant parents that buying from the local shop is ethically praiseworthy: does she think that she is expressing an objective (as in intellective tasks) or a subjective (as in judgmental tasks) truth? Are her parents expected to simply converge on the same solution or can they legitimately have alternative views?

On this ground, we will now examine three specific difficulties that people may experience when discussing their morals, and that may thus affect the dynamics and outputs of collective moral cognition: value of group discussion (§ 4.2.1), ethical disagreement (§ 4.2.2), and ethical expertise (§ 4.2.3).

2.4.2.1 What's the value of discussing morals with others?

Groups demonstrably outperform individuals in several intellective tasks. This routinely happens in sensory domains (Sorkin et al., 2001), especially when participants' confidence is matched (Bahrami et al., 2010), numerical cognition (Bahrami et al., 2012), and non-moral problem solving (Mason & Watts, 2012). In these tasks, performances are measured by closeness to right/better answers according to standardized parameters (Jayles et al., 2017). For example, a pool of meteorologists performs better than one meteorologist depending on how close their respective predictions match the weather, and experts are expected to do better than novices. In this respect, people may join groups to distribute tasks and increase their chances of doing a good job.

In judgmental tasks, performance parameters cannot be quantitatively assessed with respect to numerical benchmarks. In such cases, the impact of group reasoning is more difficult to evaluate. Indeed, although we may value group discussions even in these contexts, it is often unclear whether a jury is better than single jurors (think of jurors in a beauty contest).

How do people assess the value of group discussions about morals? In what sense should we say that groups outperform (or underperform) individuals? Depending on the metaethical tendencies of group members, the discussion may be alternatively cast as a truth-seeking collective effort or rather as a work of mediation between individual preferences – with different people having different views about

this matter. In both cases, whenever contrasts arise, people cannot easily update their credence by relying on demonstrably correct or better answers.[12]

How do people tell right from wrong in controversial cases and convince others? Basic moral theories in Western philosophy – i.e., deontology, utilitarianism, virtue ethics – provide intersubjectively valid reference systems against which individual moral claims can be tested, with the result of being more or less justifiable with respect to that reference system. However, different moral theories often endorse contrastive views about the justifications of moral judgments, and sometimes even about the solutions to ethical dilemmas. No moral system has been canonically accepted as the correct reference system – even among ethicists, let alone non-experts, thus leaving the solution we should ultimately go for underspecified.

Independently of truth-seeking efforts, group reasoning can still be seen as advantageous based on different utility functions, e.g., homogeneity of results (Himmelroos & Christensen, 2013), informed understanding (Chambers, 2003) or social inclusivity (Tajfel & Turner, 2004). However, these values are not evidentially self-sufficient in the same way as solutions to intellective tasks are and can be valued as more or less relevant depending on the group members' inclinations. Consider, for example, the vaunted value of inclusivity: inclusivity in healthcare practices actually favors the patients' intellectual autonomy (Sandman & Munthe, 2010). Nonetheless, obsessing over finding an agreement between interested parties – particularly in pediatric contexts, where both doctors and parents are involved – has been challenged as potentially detrimental to the patients' interest in terms of granting them a better quality of life (Birchley, 2014).

These issues raise interesting empirical questions about how pluralist metaethical stances may affect how people evaluate the scope and value of group discussion. In turn, this may shape the strategies and effort group members will put in place to discuss with others and make their voices heard.

2.4.2.2 How can people solve ethical disagreements?

Compared to nonmoral disagreement, moral disagreement is often treated as especially intractable (Enoch, 2009; Wilkinson & Savulescu, 2018). On the one hand, this is obviously related to the lack of a unified benchmark against which moral claims can be tested (§ 4.2.1). On the other hand, this also depends on how variable morals are and have been throughout history and across people (Feinberg et al., 2019; Graham et al., 2009).

Indeed, moral standards do not display the same intertemporal stability that is shown by solutions to intellective tasks (Hermann, 2019). Even widely accepted moral standards can be presented as the byproduct of political, social, and cultural upheavals rather than as resulting from some demonstrably correct reasoning (Sloman & Rabb, 2019). For example, the Western idea that all humans are entitled to the same rights gained momentum after the French revolution, which saw the rise of the modern bourgeoisie's wisdom and economic interest, and previously was

[12] The popularity of conspiracy theories raises questions as to whether we actually have intersubjectively shared reference systems even for intellective domains (Douglas et al., 2019; Pennycook et al., 2021).

not so widespread (Brubaker, 1989). Analogously, the now diffused ethical preoccupations for non-human animals emerged relatively late in human history (Mayer, 2010).

Despite welcome homogeneity in moral values (e.g., most people would agree that genocide is morally wrong), hardly tractable moral diversity and disagreement still pervade our societies and are often clearly visible. In particular, people do not uniformly endorse the very same notions of fairness, purity, harm, authority, and loyalty (Graham et al., 2013), and have different inclusivity criteria about what creatures deserve moral consideration or what values should be prioritized in case of conflict (Hermann, 2017; Laham, 2009). Furthermore, people may have different views on how distributive justice should be fairly implemented (Ueshima et al., 2021) or on the extent to which a society should balance the pursuit of values such as equality, justice, or personal freedom (Giebler & Merkel, 2016).

To address moral problem solving constructively, we are often encouraged to take the perspective of others into account and to appreciate that individual values are tied to deep-seated personal and cultural sensitivities,[13] which is inappropriate to simply override. When moral disagreement is rooted in fundamentally different ways of seeing things, it is unclear what evidence or arguments can be legitimately used to convince others or whether we should even try to persuade them at all. Indeed, such attempts might be seen as inconsiderate to another person's values. In this respect, subjectivism at the individual level has been experimentally linked to more tolerant behaviors (Wright et al., 2008) and to the ability to explore alternative possibilities (Goodwin & Darley, 2010). Moral disagreement is associated with negative emotions that people tend to shield themselves from, sometimes purposedly avoiding engaging in debates with dissimilar others (Frimer et al., 2017). This marks an important distinction between moral and pure intellective tasks, where people, in principle at least, aim to seek the truth more than to be respectful of interindividual differences. Furthermore, in case of moral, compared to non-moral, disagreement people perceive personal experiences as more trustworthy than objective facts (Kubin et al., 2021).

Acknowledging that ethical disagreement is often intractable does not mean that people never have tools to convince others, as it is rather the case in judgmental tasks – it would be pretty meaningless to convince someone that apples are more delicious than pineapples based on argumentative reasoning. By contrast, engaging in moral discussion, people usually feel the pressure to provide intersubjectively acceptable (impartial) justificatory reasons for their actions (Pizarro et al., 2006), and tend to rely on non-moral facts supporting their moral views. For instance, non-moral evidence of animals' suffering typically grounds the argument that exploiting animals is immoral (Schwitzgebel et al., 2020; Singer, 2009). This appeal to justificatory reasons places moral tasks far from mere judgmental tasks.

At the societal level, moral disagreement can be pragmatically annihilated via the imposition of a shared code of conduct. This was typical of traditional societies where moral differences were forcefully synchronized through unified norms and

[13] Think of bioethical committees including people with religious and secular backgrounds (Jokowitz & Glick, 2009).

the rigid sanctioning of code violations (Chaves, 1994; Taylor, 2007). Religious or tribal leaders, in such cases, played the self-assigned role of moral experts and custodians of the shared moral knowledge (Orvis, 2001). However, the schema is unlikely to fit contemporary societies wherein a certain degree of tolerance towards diversity (Brown, 2008) is a welcome element of what (democratically) facing ethical disagreement is thought to imply. Democratic societies may tend to handle these problems pragmatically by letting people free to act in a way or another (e.g., in the case of abortion or organ donation). However, when autonomously chosen solutions are not permissible, a society must converge on shared ethical solutions that can also work as the basis for political or juridical deliberation (Farah & Heberlein, 2007; Greene & Cohen, 2004).

Contrasting moral and intellective tasks, as if disagreement in the latter could always be easily overcome and truth securely pursued, may sound overly idealistic. In particular, consider that intellective tasks are typical of scientific disciplines: the history and philosophy of science have shown that even science can hardly be seen as a linearly cumulative endeavor, indomitably progressing towards truth (Maxwell, 2017). Indeed, intractable disagreements (Dieckmann & Johnson, 2019), scientific pluralism (Kellert et al., 2006), incommensurability (Fayerabend, 1962), and radical paradigm shifts (Kuhn, 1962) are part of how science evolves as well. However, here we are not concerned with the status of science as a discipline – i.e., whether science, probabilistically and fallibilistically (Peirce, 1931-60), approximates an independent truth or is rather a social construct (Berger & Luckmann, 1967) –, but on how people think of the scientific knowledge that is central to pure intellective tasks. In this sense, psychological evidence suggests that people experience scientific knowledge as providing solid and uniform guiding principles while its fallibilistic nature remains a challenge to its public understanding (Bromme & Goldman, 2014). Differently from the moral domain, when disagreement in science manifests itself in the public eye, it is then perceived as strange and unsettling (Koehler & Pennycook, 2019): assuming that science solidly pursues the truth, it is no accident that people feel disconcerted about the tentativeness of empirical approaches to currently uncharted issues (Kreps & Kriner, 2020).

2.4.2.3 How can people rely on ethical experts?

The status of ethical expertise as a specialization is controversial both in theoretical (Singer, 1972) and applied ethics (Iltis & Sheehan, 2016) – as opposed to the scientific or technological expertise that proves useful in solving pure intellective tasks.

In general, knowledge outsourcing is a fundamental cog of human cognition and one of the most common reasons we join groups (Hemmatian & Sloman, 2020; Sloman & Fernbach, 2018; Sloman & Rabb, 2019). Indeed, relying on the information that one does not know but presumes that others can provide allows humans to exploit the community's representational and computational capacities in view of more sophisticated goals (Hemmatian & Sloman, 2020). However, this strategic outsourcing implies that people can reliably identify the individuals who are more likely to provide the required expertise. This routinely happens with

intellective tasks linked to professional fields, e.g., mechanical engineering, meteorology, or the law (but see Scharrer et al., 2016). But how do people experience the outsourcing of expertise, i.e., relying on group or community's knowledge (Sloman & Rabb, 2019), in moral decisions and judgments?

In general, creative members of society and public figures may be taken as moral models in virtue of their social influence skills (Pizarro et al., 2006). Iconic individuals like Martin Luther King or Gandhi are often invoked as ethical experts in the sense of being good at knowing what ought to be done in given settings, inspiring generations to come (Rudolph, 2010). However, they did not leave behind systematic manuals[14] we can consult to mechanically solve our everyday moral issues as we would do for solving a mathematical equation (but see Spinoza, 1677/1985-2016). Social epistemologists have emphasized the role of (expert) testimony to reach justified beliefs in social scenarios. Yet, the justificatory force of testimony or social influence has often been called into question since it remains unclear on what grounds people should decide to trust selected others (Hills, 2009).

Ethicists, occasionally or systematically, play the role of expert advisors on applied ethical matters, such as health care or environmental issues (McLean, 2007). Although they are recognizably good at pointing out what moral issues arise in a given context, no consensus exists about whether they have any relevant expertise at telling people how to solve their ethical worries (Baylis, 1989): on the one hand, current research surveying samples of ethics professors suggests that their field expertise does not positively correlate with moral action (Schönegger & Wagner, 2019; Schwitzgebel & Rust, 2009). On the other hand, people have a systematic tendency to see themselves, rather than others, as morally superior to the average population (Alicke et al., 2001; Gebauer et al., 2013; Tappin & McKay, 2016).

Taken together, the controversial status of ethical expertise and the widespread sense of self-righteousness raise interesting questions about how people may then defer to others in group discussion, eventually overcoming individual positions and idiosyncrasies.

2.4.3 Collective moral problem solving

The general question underlying this section is to what extent people's metaethical commitments affect the process and output of joint moral decisions and judgments. If laypeople are metaethical pluralists, they are then expected to approach moral tasks also in a pluralist manner, i.e., unlike tasks that can be more neatly thought of as intellective or judgmental. On this ground, we will now review some of the specific challenges that moral psychology must face in the study of collective moral cognition.

In psychology, group-level differences in moral and nonmoral cognition have been previously studied as standing alone independent variables. In particular,

[14] An exception seems to be that of religious codes providing moral rules. However, religious codes are rarely seen as morally authoritative outside the circle of those who already share the same religious beliefs.

previous work has focused on how group features affect the expected outputs, e.g., whether diversity (Gomez & Bernet, 2019), group size (Rezmer et al., 2011), confidence level (Bahrami et al., 2010) or expertise (Patel et al., 2000) affect collective performance. Research has also shown that being in a group affects our metaethical commitments: people tend to be objectivists in competitive frameworks and subjectivists in cooperative ones (Fisher et al., 2016), and they are more objectivists when fellow members share their moral views (Sarkissian, 2016). Other cognate research strands have investigated how rational agents incorporate social inputs when they have (a)symmetrical information (Blomqvist & Léger, 2005); how social reinforcement promotes options that others have previously selected (Mann, 2018); how intergroup differences (Ellemers et al., 1997), group membership (Tropp & Pettigrew, 2006), in-group and out-group mechanisms (Vives et al., 2021) and cultural and political affiliation (Ellemers et al., 2013) influence individual moral decisions and judgments. Our, related but distinguishable, concern is how groups make moral decisions and judgments given people's pluralist tendencies.

One more specific question concerns the aggregation rules underlying moral group reasoning. Consider a simple aggregation rule like the majority rule (Condorcet, 1785/2014). Overall, the truth-seeking value of the majority rule in group decision-making is the target of a long-lasting discussion (Austen-Smith & Banks, 1996). A simple rather than weighted majority rule can be detrimental to the group's fitness in intellective tasks where experts would outperform uninformed group members (Correa & Yildrim, 2021). By contrast, a simple majority rule might be considered the fairest solution when mediating between different preferences in judgmental tasks (Bang & Frith, 2017). Mechanically applying the majority rule to moral problem solving might favor compromise but also be experienced as outrageous when the engaged moral views are radically different. When the hearers are highly confident, attempts at steering their moral opinions easily backfire, and they may be scarcely open to hear and discuss divergent views (Wright et al., 2008; Wright et al., 2014).[15]

Even psychological solutions that prove (un)effective in intellective tasks may not work analogously in moral tasks. For instance, the so-called *equality bias* – allocating to all group members the same amount of time to discuss the solution to a problem – is detrimental to collective perceptual decisions. Indeed, in intellective tasks, the optimum is reached when participants are allocated discussion time based on their actual competence (Mahmoodi et al., 2015). Assuming people's self-righteousness in moral matters and the dubious status of ethical expertise, one problem is to define what the optimum would be in moral discussions, e.g., by recognizing competent moral decision-makers to allocate time accordingly.

How do then groups reach consensus and compromise along the line that goes from intolerance to openness? Existing research in non-moral domains has shown

[15] In particular, one interesting research direction obviously connects this empirical research with recent developments in the *theory of judgment aggregation*, looking at the structural properties and problems of the aggregation rules for reaching *consistent* collective judgments (for a review, see List, 2012).

that people willingly strive to make sure that their views are paid attention to (Hertz et al., 2017) and systematically devise arguments to persuade others (Mercier & Sperber, 2011). However, both personal effort and communication must be effectively targeted to convince others (Baek & Falk, 2018). As mentioned, evidence-based argumentative strategies may work well in intellective tasks but less so in judgmental tasks. What persuasion strategies are more likely to be successful, and what strategies are people likely to implement when discussing their morals? What is the specific weight given to truth-seeking rational argumentation compared to idiosyncratic preferences or emotions?

According to the well-known social intuitionist model, justificatory reasons are just debunkable *post hoc* rationalizations: emotion-based automatic patterns, preceding effortful deliberative reasoning, lie beneath our educated moral views (Haidt, 2001). Indeed, verbal and non-verbal communication within the group is affected by specific moral sentiments that may steer the deliberative process towards novel results. In addition to what discussed in § 3, the sentiment of empathy is nourished by mechanisms – such as mimicry (for a critical discussion, see Holland et al., 2020) or perspective-taking (Mata, 2019) – that can only emerge in ongoing or at least simulated interactions (Ruby & Decety, 2001). We do not dispute that this affect-based model might describe how moral decisions and judgments are often taken. But, even so, people's ways of experiencing and communicating their morals, as we aimed at illustrating in this section (and as the proponents of intuitionist models acknowledge (Haidt, 2001)), require a more encompassing explanation, discussing how they value the group's contribution to moral tasks, approach disagreement, and defer to the expertise of selected others.

In sum, we highlighted that collective moral cognition displays peculiar features that we should take into account if we are to study this field empirically. Unlike nonmoral decisions and judgments, the moral ones cannot be easily categorized as matters of truth or preference, nor ethical challenges can be treated as purely intellective or judgmental tasks. While the evidence about our individual metaethical commitments is mixed, the vulnerabilities of moral reasoning are likely to become especially vivid when we deliberate together with others and have to justify our moral standpoint or outsource decisions and judgments to (experts) others. The interplay between these aspects makes the study of collective moral cognition a fertile target for empirical investigation.

2.5 Conclusion

The scope of the paper was threefold. First, we showed that a good deal of the current research in moral cognition is individualistic, reviewed some of the underlying motivations for this, and acknowledged some more recent research trends that highlight the relevance of the collective dimension. Second, we discussed specific factors arising in collective moral cognition as having a worth investigating impact on the ensuing decisions and judgments. We would miss out

some vital information if we simply approached collective moral cognition in terms of the aggregate of individual moral views. Third, we proposed that group moral cognition differs significantly from collective non-moral cognition. Therefore, it must be addressed with targeted tools if essential details of how the related processes work are to be captured. In conclusion, this paper aimed to urge researchers in the field of moral cognition to pay more attention to the collective dimension, filling the gaps between individual and group dynamics.

The role of interactions in moral cognition ties together several timely questions about the societal and political impact of moral discussion, moral persuasion, social deliberation and moral change. In this light, across psychology, philosophy, and cognitive neuroscience, new models and frameworks are needed to foster understanding of the processes that make collective moral decisions and judgments possible. We hope that our theoretical overview of open issues in collective moral cognition will promote the investigation of this burgeoning research pathway.

References

Abelson, R. P., & Carroll, J. D. (1965). Computer Simulation of Individual Belief Systems. *American Behavioral Scientist*, *8*(9), 24–30. https://doi.org/10.1177/000276426500800908

Ahluwalia, R. (2000). Examination of Psychological Processes Underlying Resistance to Persuasion. *Journal of Consumer Research, 27*(2), 217-232. https://doi.org/10.1086/314321

Alicke, M. D., Vredenburg, D. S., Hiatt, M., & Govorun, O. (2001). The "Better Than Myself Effect." *Motivation and Emotion, 25*(1), 7–22. https://doi.org/10.1023/A:1010655705069

Anscombe, G. E. M. (1963). *Intention*. Blackwell.

Aramovich, N. P., Lytle, B. L., & Skitka, L. J. (2012). Opposing torture: Moral conviction and resistance to majority influence. *Social Influence, 7*(1), 21–34. https://doi.org/10.1080/15534510.2011.640199

Arendt, H. (1987). Collective Responsibility. In S. J. J. W. Bernauer (Ed.), *Amor Mundi* (pp. 43–50). Springer Netherlands. https://doi.org/10.1007/978-94-009-3565-5_3

Arvan, M. (2013). Bad news for conservatives? Moral judgments and the dark triad personality traits: A correlational study. *Neuroethics*, *6*(2), 307–318. https://doi.org/10.1007/s12152-011-9140-6

Ashcroft, R. E. (2005). The Ethical Brain. *Journal of the Royal Society of Medicine*, *98*(9), 433–434. https://doi.org/10.1258/jrsm.98.9.433

Atari, M., Lai, M. H. C., & Dehghani, M. (2020). Sex differences in moral judgements across 67 countries. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1937), 20201201. https://doi.org/10.1098/rspb.2020.1201

Austen-Smith, & D., Banks, J. S. (1996). Information Aggregation, Rationality, and the Condorcet Jury Theorem. *American Political Science Review, 90*(1), 34-45. https://doi.org/10.2307/2082796

Baek, E. C., & Falk, E. B. (2018). Persuasion and Influence: What Makes a Successful Persuader? *Current Opinion in Psychology, 24*, 53-57. https://doi.org/10.1016/J.COPSYC.2018.05.004

Bahrami, B., Didino, D., Frith, C., Butterworth, B., & Rees, G. (2012). Collective Enumeration. *J Exp Psychol Hum Percept Perform, 39*(2), 338-347. https://dx.doi.org/10.1037%2Fa0029717

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally Interacting Minds. *Science, 329*, 1081-1085. https://doi.org/10.1126/SCIENCE.1185718

Balconi, M., & Vanutelli, M. E. (2017). Cooperation and Competition with Hyperscanning Methods: Review and Future Application to Emotion Domain. *Frontiers in Computational Neuroscience*, *11*, 86–86. https://doi.org/10.3389/fncom.2017.00086

Bandura, A., Underwood, B., & Fromson, M. E. (1975). Disinhibition of aggression through diffusion of responsibility and dehumanization of victims. *Journal of Research in Personality*, *9*(4), 253–269. https://doi.org/10.1016/0092-6566(75)90001-X

Bang, D., & Frith, C. D. (2017). Making better decisions in groups. *Royal Society Open Science*, *4*(8), 170193. https://doi.org/10.1098/rsos.170193

Baron, J., Gürçay, B., & Luce, M. F. (2018). Correlations of trait and state emotions with utilitarian moral judgements. *Cognition and Emotion*, *32*(1), 116–129. https://doi.org/10.1080/02699931.2017.1295025

Barsade, S. G., & Gibson, D. (1998). Group emotion: A view from top and bottom. *In D. H Gruenfeld (Ed.), Research on Managing Groups and Teams.*, *1*, 81–102.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, *121*(1), 154–161. https://doi.org/10.1016/j.cognition.2011.05.010

Baylis, F. (1989). Persons with Moral Expertise and Moral Experts: Where in Lies the Difference? In B. Hoffmaster, B. Freedman, G. Fraser, *Clinical Ethics: Theory and Practice.* Clifton, NJ: Humana Press.

Beebe, J. R. (2014). How Different Kinds of Disagreement Impact Folk Metaethical Judgments. In H. Sarkissian & J.C. Wright (Eds.). *Advances in Experimental Moral Psychology* (pp. 167–187). Bloomsbury Academic. http://dx.doi.org/10.5040/9781472594150.ch-009

Beebe, J. R., & Sackris, D. (2016). Moral Objectivism Across the Lifespan. *Philosophical Psychology, 29*(6), 912-929. https://doi.org/10.1080/09515089.2016.1174843

Behnk, S., Hao, L., & Reuben, E. (2017). Partners in Crime: Diffusion of Responsibility in Antisocial Behaviors. *IZA Discussion Paper No. 11031*.

Bengson, J. (2013). Experimental Attacks on Intuitions and Answers. *Philosophy and Phenomenological Research, 86*, 495-532. https://doi.org/10.1111/J.1933-1592.2012.00578.X

Benhabib, S. (1992). *Situating the Self*. Routledge.

Berger, P. L., & Luckmann, T. (1967). The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday.

Berkely, G. (1712/1972). Passive Obedience, or the Christian Doctrine of Not Resisting the Supreme Power, Proved and Vindicated upon the Principles of the Law of Nature. Reprinted in D. H. Monro (ed.), A Guide to the British Moralists, Fontana, 217–27.

Berkowitz, M. W., & Gibbs, J. C. (1983). Measuring the Developmental Features of Moral Discussion. *Merrill-Palmer Quarterly*, *29*(4), 399–410. http://www.jstor.org/stable/23086309

Berkowitz, M. W., Gibbs, J. C., & Broughton, J. M. (1980). The relation of moral judgment stage disparity to developmental effects of peer dialogues. *Merrill-Palmer Quarterly of Behavior and Development*, *26*(4), 341–357. http://www.jstor.org/stable/23084042

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*(12), 1872–1881. https://doi.org/10.1093/scan/nsw103

Beyer, F., Sidarus, N., Bonicalzi, S., & Haggard, P. (2017). Beyond self-serving bias: diffusion of responsibility reduces sense of agency and outcome monitoring. *Social Cognitive and Affective Neuroscience*, *12*(1), 138–145. https://doi.org/10.1093/scan/nsw160

Białek, M., Paruzel-Czachura, M., & Gawronski, B. (2019). Foreign language effects on moral dilemma judgments: An analysis using the CNI model. *Journal of Experimental Social Psychology*, *85*, 103855. https://doi.org/10.1016/j.jesp.2019.103855

Birchley, G. (2014). Deciding Together? Best Interests and Shared Decision-Making in Paediatric Intensive Care. *Health Care Anal., 22*(3), 203-222. https://doi.org/10.1007/S10728-013-0267-Y

Blackburn, S. (1984). *Spreading the Word.* Clarendon Press.

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin*, *88*(1), 1–45. https://doi.org/10.1037/0033-2909.88.1.1

Blasi, A. (1990). How should psychologists define morality? or, The negative side effects of philosophy's influence on psychology. In *The moral domain: Essays on the on-going discussion between philosophy and the social sciences* (pp. 38–70). The MIT Press.

Blasi, A. (2005). *Moral character: A psychological approach.* In D. K. Lapsley & F. C. Power (Eds.), *Character psychology and character education* (p. 67–100). University of Notre Dame Press.

Blomqvist, A., & Léger, P. T. (2005). Information Asymmetry, Insurance, and the Decision to Hospitalize. *J Health Econ, 24*(4), 775-793. https://doi.org/10.1016/J.JHEALECO.2004.12.001

Bloom, P. (2010). How do morals change? *Nature*, *464*(7288), 490. https://doi.org/10.1038/464490a

Bonicalzi, S. (2019). *Rethinking Moral Responsibility*. Mimesis International

Bornstein, G., & Yaniv, I. (1998). Individual and Group Behavior in the Ultimatum Game: Are Groups More "Rational" Players?. *Experimental Economics* 1**,** 101–108 https://doi.org/10.1023/A:1009914001822

Bostyn, D. H., & Roets, A. (2017). Trust, trolleys and social dilemmas: A replication study. *Journal of Experimental Psychology: General*, *146*(5), e1–e7. https://doi.org/10.1037/xge0000295

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of Mice, Men, and Trolleys: Hypothetical Judgment Versus Real-Life Behavior in Trolley-Style Moral Dilemmas. *Psychological Science*, *29*(7), 1084–1093. https://doi.org/10.1177/0956797617752640

Brambilla, M., Sacchi, S., Rusconi, P., Cherubini, P., & Yzerbyt, V. Y. (2012). You want to give a good impression? Be honest! Moral traits dominate group impression formation. *British Journal of Social Psychology*, *51*(1), 149–166. https://doi.org/10.1111/j.2044-8309.2010.02011.x

Bromme, R., & Goldman, S. R. (2014). The Public's Bounded Understanding of Science. *Educational Psychologist, 49*(2), 59–69. https://doi.org/10.1080/00461520.2014.921572

Brown, W. (2008). Regulating Aversion: Tolerance in the Age of Identity and Empire. Princeton University Press.

Brubaker, W. R. (1989). The French Revolution and the Invention of Citizenship. *French Politics and Society, 7*(3), 30-49.

Bucciarelli, M., & Johnson-Laird, P. N. (2020). Beliefs and Emotions About Social Conventions. *Acta Psychologica, 210*, 103184. https://doi.org/10.1016/J.ACTPSY.2020.103184

Burgess, A. G., & Burgess, J. P. (2011). *Truth.* Princeton University Press.

Campbell, J., Schermer, J. A., Villani, V. C., Nguyen, B., Vickers, L., & Vernon, P. A. (2009). A behavioral genetic study of the dark triad of personality and moral development. *Twin Research and Human Genetics*, *12*(2), 132–136. https://doi.org/10.1375/twin.12.2.132

Campbell, R. (2017). Learning from Moral Inconsistency. *Cognition, 167*, 46-57. https://doi.org/10.1016/J.COGNITION.2017.05.006

Capraro, V., Sippel, J., Zhao, B., Hornischer, L., Savary, M., Terzopoulou, Z., Faucher, P., & Griffioen, S. F. (2018). People making deontological judgments in the Trapdoor dilemma are perceived to be more prosocial in economic games than they actually are. *PLoS ONE*, *14*(11). https://doi.org/10.1371/JOURNAL.PONE.0225850

Chambers, S. (2003). Deliberative Democratic Theory. *Annual Review of Political Science, 6*, 307-326. https://doi.org/10.1146/ANNUREV.POLISCI.6.121901.085538

Chaves, M. (1994). Secularization as Declining Religious Authority. *Social Forces, 72*(3), 749-774. https://doi.org/10.1093/SF/72.3.749

Chekroun, P., & Brauer, M. (2002). The bystander effect and social control behavior: The effect of the presence of others on people's reactions to norm violations. *European Journal of Social Psychology*, *32*(6), 853–867. https://doi.org/10.1002/ejsp.126

Choe, S. Y., & Min, K.H. (2011). Who makes utilitarian judgments? The influences of emotions on utilitarian judgments. *Judgment and Decision Making*, *6*(7), 580–592.

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: a principled review. *Neuroscience & Biobehavioral Reviews, 36*(4), 1249–1264. https://doi.org/10.1016/J.NEUBIOREV.2012.02.008

Churchland, P. S. (2008). The Impact of Neuroscience on Philosophy. *Neuron*, *60*(3), 409–411. https://doi.org/10.1016/j.neuron.2008.10.023

Ciaramelli, E., Muccioli, M., Làdavas, E., & Di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *2*(2), 84–92. https://doi.org/10.1093/scan/nsm001

Ciaramidaro, A., Toppi, J., Casper, C., Freitag, C. M., Siniatchkin, M., & Astolfi, L. (2018). Multiple-Brain Connectivity during Third Party Punishment: An EEG Hyperscanning Study. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-24416-w

Cikara, M., & Fiske, S. T. (2012). Stereotypes and schadenfreude: Affective and physiological markers of pleasure at outgroup misfortunes. *Social Psychological and Personality Science*, *3*(1), 63–71. https://doi.org/10.1177/1948550611409245

Cikara, M., Botvinick, M. M., & Fiske, S. T. (2011). Us versus them: Social identity shapes neural responses to intergroup competition and harm. *Psychological Science*, *22*(3), 306–313. https://doi.org/10.1177/0956797610397667

Cohen, T. R., Montoya, R. M., & Insko, C. A. (2006). Group morality and intergroup relations: Cross-cultural and experimental evidence. *Personality and Social Psychology Bulletin*, *32*(11), 1559–1572. https://doi.org/10.1177/0146167206291673

Colombetti, G., & Torrance, S. (2009). Emotion and ethics: An inter-(en)active approach. *Phenomenology and the Cognitive Sciences, 8*(4), 505–526. https://doi.org/10.1007/s11097-009-9137-3

Condorcet, N. (1785/2014). Essay sur l'Application de l'Analyse à la Probabilité des Décisions Rendue à la Pluralité des Voix. Cambridge University Press.

Conrads, J., Irlenbusch, B., Rilke, R. M., & Walkowitz, G. (2013). Lying and team incentives. *Journal of Economic Psychology*, *34*, 1–7. https://doi.org/10.1016/j.joep.2012.10.011

Copp, D. (1991). Moral Skepticism. *Philosophical Studies, 62*, 203–233.

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

Correa, A. J. N., & Yildirim, H. (2021). Biased experts, majority rule, and the optimal composition of committee. Games and Economic Behavior, 127, 1–27. https://doi.org/10.1016/j.geb.2021.01.010

Curşeu, P. L., Fodor, O. C., A. Pavelea, A., & Meslec, N. (2020). "Me" versus "We" in moral dilemmas: Group composition and social influence effects on group utilitarianism. *Business Ethics*, *29*(4), 810–823. https://doi.org/10.1111/beer.12292

Czeszumski, A., Eustergerling, S., Lang, A., Menrath, D., Gerstenberger, M., Schuberth, S., … König, P. (2020). Hyperscanning: A Valid Method to Study Neural Inter-brain Underpinnings of Social Interaction. *Frontiers in Human Neuroscience, 14*, 39. https://doi.org/10.3389/fnhum.2020.00039

Dahl, A., Schuck, R. K., & Campos, J. J. (2013). Do young toddlers act on their social preferences? *Developmental Psychology*, *49*(10), 1964–1970. https://doi.org/10.1037/a0031460

Damon, W., & Killen, M. (1982). Peer Interaction and the Process of Change in Children's Moral Reasoning. *Merrill-Palmer Quarterly*, *28*(3), 347–367.

Dancy, J. (2003). *Practical Reality*. Oxford University Press. https://doi.org/10.1093/0199253056.001.0001

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 8*(4), 377–383. https://doi.org/10.1037/h0025589

Darlow, A. L., & Sloman, S. (2010). Two systems of reasoning: Architecture and relation to emotion. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(3), 382–392. https://doi.org/10.1002/wcs.34

Darmstadter, H. (2013). Why do humans reason? A pragmatist supplement to an argumentative theory. *Thinking and Reasoning*, *19*(3–4), 472–487. https://doi.org/10.1080/13546783.2013.802256

David, M. (2018). The Correspondence Theory of Truth. In M. Glanzberg (ed.), *The Oxford Handbook of Truth* (pp. 238-258). Oxford: Oxford University Press. https://doi.org/10.1093/OXFORDHB/9780199557929.013.9

Davis, T. (2021). Beyond Objectivism: New Methods for Studying Metaethical Intuitions. *Philosophical Psychology, 34*(1), 125-153.

Decety, J., & Wheatley, T. (2015). The moral brain: A multidisciplinary perspective. In *The Moral Brain: A Multidisciplinary Perspective*. The MIT Press. https://doi.org/10.5860/choice.192126

Dieckmann, N. F., & Johnson, B. B. (2019). Why do Scientists Disagree? Explaining and Improving Measures of the Perceived Causes of Scientific Disputes. *PLoS ONE, 14*(2), e0211269.

Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J., Michalareas, G., Van Bavel, J. J., Ding, M., & Poeppel, D. (2017). Brain-to-Brain Synchrony Tracks Real-World Dynamic Group Interactions in the Classroom. *Current Biology*, *27*(9), 1375–1380. https://doi.org/10.1016/j.cub.2017.04.002

Djeriouat, H., & Trémolière, B. (2014). The Dark Triad of personality and utilitarian moral judgment: The mediating role of Honesty/Humility and Harm/Care. *Personality and Individual Differences*, *67*, 11–16. https://doi.org/10.1016/j.paid.2013.12.026

Douglas, K. M., & Uscinski, J. E., Sutton, R. M., Cichocka, A. (2019). Understanding Conspiracy Theories. *Political Psychology, 40*(S1), 3-35.

Dumas, G., Lachat, F., Martinerie, J., Nadel, J., & George, N. (2011). From social behaviour to brain synchronization: Review and perspectives in hyperscanning. *Irbm*, *32*(1), 48–53. https://doi.org/10.1016/j.irbm.2011.01.002

El Zein, M., & Bahrami, B. (2020). Joining a group diverts regret and responsibility away from the individual. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1922), 20192251. https://doi.org/10.1098/rspb.2019.2251

El Zein, M., Bahrami, B., & Hertwig, R. (2019). Shared responsibility in collective decisions. *Nature Human Behaviour*, *3*(6), 554–559. https://doi.org/10.1038/s41562-019-0596-4

El Zein, M., Seikus, C., De-Wit, L., & Bahrami, B. (2020). Punishing the individual or the group for norm violation. *Wellcome Open Research*, *4*, 139. https://doi.org/10.12688/wellcomeopenres.15474.2

Ellemers, N. (2017). Morality and the Regulation of Social Behavior: Groups as Moral Anchors (1st ed.), Routledge. https://doi.org/10.4324/9781315661322

Ellemers, N., & Van Nunspeet, F. (2020). Neuroscience and the Social Origins of Moral Behavior: How Neural Underpinnings of Social Categorization and Conformity Affect Everyday Moral and Immoral Behavior. *Current Directions in Psychological Science*, *29*(5), 513–520. https://doi.org/10.1177/0963721420951584

Ellemers, N., van Rijswijk, W., Roefs, M., & Simons, C. (1997). Bias in Intergroup Perceptions: Balancing Group Identity with Social Reality. *Personality and Social Psychology Bulletin*, *23*(2), 186–198. https://doi.org/10.1177/0146167297232007

Ellemers, N., Pagliaro, S., & Barreto, M. (2013). Morality and behavioural regulation in groups: A social identity approach. *European Review of Social Psychology*, *24*(1), 160–193. https://doi.org/10.1080/10463283.2013.841490

Ellemers, N., Van Der Toorn, J., Paunov, Y., & Van Leeuwen, T. (2019). The Psychology of Morality: A Review and Analysis of Empirical Studies Published From 1940 Through 2017. *Personality and Social Psychology Review*, *23*(4), 332–366. https://doi.org/10.1177/1088868318811759

Enoch, D. (2009). How is Moral Disagreement a Problem for Realism? *The Journal of Ethics, 13*(1), 15-50. https://doi.org/10.1007/S10892-008-9041-Z

Enoch, D. (2011). *Taking Morality Seriously: A Defense of Robust Realism.* Oxford University Press.

Evans, J. S. B. T. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454–459. https://doi.org/10.1016/j.tics.2003.08.012

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Everett, J. A. C., & Kahane, G. (2020). Switching Tracks? Towards a Multidimensional Model of Utilitarian Psychology. *Trends in Cognitive Sciences*, *24*(2), 124–134. https://doi.org/10.1016/j.tics.2019.11.012

Everett, J. A. C., 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. https://doi.org/10.1016/j.jesp.2018.07.004

Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of Trustworthiness From Intuitive Moral Judgments. *Journal of Experimental Psychology: General*, *145*(6), 772–787. https://doi.org/10.1037/xge0000165

Farah, M. J., & Heberlein, A. S. (2007). Personhood and Neuroscience: Naturalizing or Nihilating? *American Journal of Bioethics, 7*(1), 37–48. https://doi.org/10.1080/15265160601064199

Fayerabend, P. K. (1962). Explanation, Reduction and Empiricism. In H. Feigl, G. Maxwell, *Minnesota Studies in the Philosophy of Science* (Bd. 3, S. 28-97). University of Minnesota Press.

Fedyk, M. (2017). *The social turn in moral psychology*. The MIT Press. https://doi.org/10.1215/00318108-7213385

Feinberg, M., Kovacheff, C., Teper, R., & Inbar, Y. (2019). Understanding the Process of Moralization: How Rating Meat Becomes a Moral Issue. *Journal of Personality and Social Psychology, 117*(1), 50-72.

FeldmanHall, O., Son, J., & Heffner, J. (2018). Norms and the Flexibility of Moral Action. *Personality Neuroscience, 1*, E15. https://doi.org/10.1017/pen.2018.13

Feltz, A., & Cokely, E. T. (2008). The Fragmented Folk: More Evidence of Stable Individual Differences in Moral Judgments and Folk Intuitions. In Love, K. McRae V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1771-1776). Cognitive Science Society.

Feng, C., Deshpande, G., Liu, C., Gu, R., Luo, Y.-J., & Krueger, F. (2016). Diffusion of responsibility attenuates altruistic punishment: A functional magnetic resonance imaging effective connectivity study. *Human Brain Mapping*, *37*(2), 663–677. https://doi.org/10.1002/hbm.23057

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, *58*(2), 203–210. https://doi.org/10.1037/h0041593

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge University Press.

Fisher, M., Knobe, J., Strickland, B., & Keil, F. C. (2016). The Influence of Social Interaction on Intuitions of Objectivity and Subjectivity. *Cognitive Science, 41*(4), 1119-1134. https://doi.org/10.1111/COGS.12380

Fochmann, M., Fochmann, N., Kocher, M. G., Müller, N., & Wolf, N. (2021). Dishonesty and risk-taking: Compliance decisions of individuals and groups. *Journal of Economic Behavior and Organization*, *185*, 250–286. https://doi.org/10.1016/j.jebo.2021.02.018

Foot, P. (1967). The Problem of Abortion and the Doctrine of the Double Effect. *Oxford Review*, *5*, 19–32. https://doi.org/10.1093/0199252866.003.0002

Forsyth, D. R., Zyzniewski, L. E., & Giammanco, C. A. (2002). Responsibility diffusion in cooperative collectives. *Personality and Social Psychology Bulletin*, *28*(1), 54–65. https://doi.org/10.1177/0146167202281005

Freeman, S., Walker, M. R., Borden, R., & Latané, B. (1975). Diffusion of Responsibility and Restaurant Tipping: Cheaper by the Bunch. *Personality and Social Psychology Bulletin*, *1*(4), 584–587. https://doi.org/10.1177/014616727500100407

Fricker, M., Graham, P. J., Henderson, D., & Pedersen, N. J. L. L. (Eds.) (2019). *The Routledge Handbook of Social Epistemology*, Routledge.

Frimer, J. A., Skitka, L. A., & Motyl, M. (2017). Liberals and Conservatives Are Similarly Motivated to Avoid Exposure to One Another's Opinions. *Journal of Experimental Social Psychology, 72*, 1-12. https://doi.org/10.1016/J.JESP.2017.04.003

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science*, *23*(8), 931–939. https://doi.org/10.1177/0956797612436816

Gallotti, Mattia, & Frith, C. D. (2013). Social cognition in the we-mode. In *Trends in Cognitive Sciences*, *17*(4), 160–165. https://doi.org/10.1016/j.tics.2013.02.002

Gebauer, J. E., Wagner, J., Sedikides, C., & Neberich, W. (2013). Agency-Communion and Self-Esteem Relations Are Moderated by Culture, Religiosity, Age, and Sex: Evidence for the "Self-Centrality Breeds Self-Enhancement" Principle. *Journal of Personality, 81*(3), 261-275.

Gibbard, A. 1990. *Wise choices, Apt Feelings*. Harvard University Press.

Giebler, H., & Merkel, W. (2016). Freedom and Equality in Democracies: Is there a Trade-off? *International Political Science Review, 37*(5), 594-605.

Gilbert, M. (2014). *Joint Commitment: How We Make the Social World*. Oxford University Press.

Glenn, A. L., Raine, A., Schug, R. A., Young, L., & Hauser, M. (2009). Increased DLPFC activity during moral decision-making in psychopathy. *Molecular Psychiatry*, *14*, 909–911. https://doi.org/10.1038/mp.2009.76

Goette, L., Huffman, D., & Meier, S. (2006). The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social

groups. *American Economic Review*, *96*(2), 212–216. https://doi.org/10.1257/000282806777211658

Gold, N., & Sudgen, R. (2007). Collective Intentions and Team Agency. *The Journal of Philosophy, 104*(3), 109-137.

Goldring, M. R., & Heiphetz, L. (2020). Sensitivity to ingroup and outgroup norms in the association between commonality and morality. *Journal of Experimental Social Psychology*, *91*. https://doi.org/10.1016/j.jesp.2020.104025

Gomez, L. E., & Bernet, P. (2019). Diversity Improves Performance and Outcomes. *J Natl Med Assoc, 111*(4), 383-392.

Goodwin, G. P., & Darley, J. M. (2008). The Psychology of Meta-Ethics: Exploring Objectivism. *Cognition, 106*, 1339-1366. https://doi.org/10.1016/j.cognition.2007.06.007

Goodwin, G. P., & Darley, J. M. (2010). The Perceived Objectivity of Ethical Beliefs: Psychological Findings and Implications for Public Policy. *Review of Philosophy and Psychology, 1*(2), 161-188.

Goodwin, P., & Darley, J. M. (2012). Why Are Some Moral Beliefs Perceived to Be More Objective than Others? *Journal of Experimental Social Psychology, 48*, 250-256.

Gowans, C. (2012). Moral Relativism, *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2021/entries/moral-relativism/>.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and Conservatives Rely on Different Sets of Moral Foundations. *Journal of Personality and Social Psychology, 96*(5), 1029-1046.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S., & Ditto, P. (2013). Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. *Advances in Experimental Social Psychology. 47*. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Greene, J. D. (2002). The Terrible, Horrible, No Good, Very Bad Truth about Morality and What to Do About it. In *Unpublished doctoral dissertation Department of Philosophy Princeton University*.

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, *11*(8), 322–323. https://doi.org/10.1016/j.tics.2007.06.004

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*(3), 581–584. https://doi.org/10.1016/j.jesp.2009.01.003

Greene, J. D. (2015). The cognitive neuroscience of moral judgment and decision making. In *The Moral Brain: A Multidisciplinary Perspective* (pp. 197–220). Massachusetts Institute of Technology. https://doi.org/10.7551/mitpress/9988.003.0017

Greene, J. D., & Haidt, J. (2002). How (and Where) Does Moral Judgment Work? *Trends in Cognitive Science, 6*, 517-523.

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*(5537), 2105-2108.

Greene, J., & Cohen, J. (2004). For the Law, Neuroscience Changes Nothing and Everything. *Philos Trans R Soc Lond B Biol Sci., 359*(1451), 1775-1785.

Habermas, J. (1995). Reconciliation through the Public Use of Reason: Remarks on John Rawls's Political Liberalism. *The Journal of Philosophy, 92*(3), 109-131.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*(5827), 998–1002. https://doi.org/10.1126/science.1137651

Haidt, J., Rosenberg, E. and Hom, H. (2003), Differentiating Diversities: Moral Diversity Is Not Like Other Kinds1. *Journal of Applied Social Psychology, 33*: 1-36. https://doi.org/10.1111/j.1559-1816.2003.tb02071.x

Haidt, J., & Kesebir, S. (2010). Morality. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology, 5e*. John Wiley Sons, Inc. https://doi.org/10.1002/9780470561119.socpsy002022

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. *PLoS ONE, 7*(9), e45457.

Harman, G. (1996). Moral Relativism. In G. Harman, & J.J. Thompson (Eds.), *Moral Relativism and Moral Objectivity*, (3–64). Blackwell Publisher.

Harenski, C. L., Sang, H. K., & Hamann, S. (2009). Neuroticism and psychopathy predict brain activation during moral and nonmoral emotion regulation. *Cognitive, Affective and Behavioral Neuroscience*, *9*(1), 1-15. https://doi.org/10.3758/CABN.9.1.1

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-Brain coupling: A mechanism for creating and sharing a social world. *Trends in Cognitive Sciences, 16*(2), 114–121. https://doi.org/10.1016/J.TICS.2011.12.007

Heiphetz, L., & Young, L. L. (2017). Can Only One Person Be Right? The Development of Objectivism and Social Preferences Regarding Widely Shared and Controversial Moral Beliefs. *Cognition, 167*, 78-90.

Hemmatian, B., & Sloman, S. (2020). Two systems for thinking with a community. In S. Elqayam, I. Douven. J. St. B. T. Evans, & N. Cruz (Eds.), *Logic and Uncertainty in the Human Mind: A tribute to David E. Over* (pp. 102–115). Routledge. https://doi.org/10.4324/9781315111902-7

Hermann, J. (2017). Possibilities of Moral Progress in the Face of Evolution. *Ethical Theory and Moral Practice, 20*, 39-54. https://doi.org/10.1007/S10677-016-9737-2

Hermann, J. (2019). The Dynamics of Moral Progress. *Ratio, 32*(4), 300-311. https://doi.org/10.1111/RATI.12232

Hertwig, R., & Gigerenzer, G. (2011). Behavioral Inconsistencies Do Not Imply Inconsistent Strategies. *Frontiers in Psychology*, 292. https://doi.org/10.3389/fpsyg.2011.00292

Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C. D., & Bahrami, B. (2017). Neural computations Underpinning the Strategic Management of Influence in Advice Giving. *Nat. Commun., 8*, 2191. https://doi.org/10.1038/s41467-017-02314-5

Higgins, J. (2020). Cognising With Others in the We-Mode: a Defence of 'First-Person Plural' Social Cognition. *Review of Philosophy and Psychology*, 1–22. https://doi.org/10.1007/s13164-020-00509-2

Hills, A. (2009). Moral Testimony and Moral Epistemology. *Ethics, 120*(1), 94-127.

Himmelroos, S., & Christensen, H. S. (2013). Deliberation and Opinion Change: Evidence from a Deliberative Mini-public in Finland. *Scandinavian Political Studies, 37*(1), 41-60.

Hirata, M., Ikeda, T., Kikuchi, M., Kimura, T., Hiraishi, H., Yoshimura, Y., & Asada, M. (2014). Hyperscanning MEG for understanding mother-child cerebral interactions. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00118

Holland, A. C., O'Connell, G., & Dziobek, I. (2020). Facial Mimicry, Empathy, and Emotion Recognition: A Meta-Analysis of Correlations. *Cognition and Emotion*. 35:1, 150-168. https://doi.org/10.1080/02699931.2020.1815655

Hopster, J. (2019). The Meta-Ethical Significance of Experiments about Folk Moral Objectivism. *Philosophical Psychology, 32*(6), 831-852.

Hu, Y., Pan, Y., Shi, X., Cai, Q., Li, X., & Cheng, X. (2018). Inter-brain synchrony and cooperation context in interactive decision making. *Biological Psychology*, *133*, 54–62. https://doi.org/10.1016/j.biopsycho.2017.12.005

Huang, K., Greene, J. D., & Bazerman, M. (2019). Veil-of-ignorance reasoning favors the greater good. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(48), 23989–23995. https://doi.org/10.1073/pnas.1910125116

Huebner, B., Dwyer, S., & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, *13*(1), 1–6. https://doi.org/10.1016/j.tics.2008.09.006

Iltis, A. S., & Sheehan, M. (2016). Expertise, Ethics Expertise, and Clinical Ethics Consultation: Achieving Terminological Clarity. *J Med Philos., 41*(4), 416-433.

Jayles, B., Escobedo, R., Cezera, S., Blanchet, A., Kameda, T., Sire, C., & Theraulaz, G. (2017). How Social Information Can Improve Estimation Accuracy in Human Groups. *PNAS, 114*(47), 12620-12625.

Jokowitz, B. A., & Glick, S. (2009). Navigating the Chasm Between Religious and Secular Perspectives in Modern Bioethics. *Journal of Medical Ethics, 35*(6), 357-360.

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*, *10*(5), 551–560. https://doi.org/10.1080/17470919.2015.1023400

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. https://doi.org/10.1016/j.cognition.2014.10.005

Kant, I. (1785/1993). *Grounding for the Metaphysics of Morals: With on a Supposed Right to Lie Because of Philanthropic Concerns*. Hackett Publishing Company. https://doi.org/10.4324/9780203981948

Kaplan, M. F., & Miller, L. E. (1978). Reducing the effects of juror bias. *Journal of Personality and Social Psychology*, *36*(12), 1443–1455. https://doi.org/10.1037/0022-3514.36.12.1443

Karau, S. J., & Williams, K. D. (1993). Social Loafing: A Meta-Analytic Review and Theoretical Integration. *Journal of Personality and Social Psychology*, *65*(4), 681–706. https://doi.org/10.1037/0022-3514.65.4.681

Keasey, C. B. (1973). Experimentally induced changes in moral opinions and reasoning. *Journal of Personality and Social Psychology*, *26*(1), 30–38. https://doi.org/10.1037/H0034210

Kellert, S. H., Longino, H. E., & Waters, C. K. (2006). *Scientific Pluralism.* University of Minnesota Press.

Kelly, D., Stich, S., Haley, K. J., Eng, S. J., & Fessler, D. M. (2007). Harm, Affect, and the Moral/Conventional Distinction. *Mind Language, 22*(2), 117-131.

Keshmirian, A., Bahrami, B., & Deroy, O. (2021). Many Heads Are More Utilitarian Than One. https://doi.org/10.31234/OSF.IO/7E3DC

Kocher, M. G., Schudy, S., & Spantig, L. (2018). I lie? We lie! Why? Experimental evidence on a dishonesty shift in groups. *Management Science*, *64*(9), 3995–4008. https://doi.org/10.1287/mnsc.2017.2800

Koehler, D. K., & Pennycook, G. (2019). How the Public, and Scientists, Perceive Advancement of Knowledge from Conflicting Study Results. *Judgment & Decision Making, 14*, 671-682.

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*, *7*(6), 708–714. https://doi.org/10.1093/scan/nsr048

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*(7138), 908–911. https://doi.org/10.1038/nature05631

Kohlberg, L. (1969). Stage and sequence: The cognitivedevelopmental approach to socialization. Chicago: Rand McNally. *Handbook of Socialization Theory and Research*, 347–480.

Konvalinka, I., & Roepstorff, A. (2012). The two-brain approach: how can mutually interacting brains teach us something about social interaction? *Frontiers in Human Neuroscience, 6*, 215–215. https://doi.org/10.3389/fnhum.2012.00215

Kreps, S. E., & Kriner, D. L. (2020). Model Uncertainty, Political Contestation, and Public Trust in Science: Evidence from the COVID-19 Pandemic. *Sci Adv., 6*(43), eabd4563.

Kreps, T. A., & Monin, B. (2014). Core Values Versus Common Sense: Consequentialist Views Appear Less Rooted in Morality. *Personality and Social Psychology Bulletin*, *40*(11), 1529–1542. https://doi.org/10.1177/0146167214551154

Kubin, E., Puryear, C., Schein, C., & Gray, K. (2021), Personal Experiences Bridge Moral and Political Divides Better than Facts. *PNAS*, *118*(6), e2008389118. https://doi.org/10.1073/pnas.2008389118

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions.* University of Chicago Press.

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*(2), 75–84. https://doi.org/10.1016/j.evolhumbehav.2006.06.001

Kvaran, T., Nichols, S., & Sanfey, A. (2013). The effect of analytic and experiential modes of thought on moral judgment. *Progress in Brain Research*, *202*, 187–196. https://doi.org/10.1016/B978-0-444-62604-2.00011-3

Laham, S. M. (2009). Expanding the Moral Circle: Inclusion and Exclusion Mindsets and the Circle of Moral Regard. *Journal of Experimental Social Psychology, 45*(1), 250-253.

Latané, B., Williams, K., & Harkins, S. (2006). Many hands make light the work: The causes and consequences of social loafing. *Small Groups: Key Readings*, *37*(6), 297–308. https://doi.org/10.4324/9780203647585

Laughlin, P. R., & Adamopoulos, J. (1980). Social combination processes and individual learning for six-person cooperative groups on an intellective task. *Journal of Personality and Social Psychology, 38*(6), 941–947.

Laughlin, P. R. (2011). *Group Problem Solving.* Princeton University Press.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and Social Combination Processes on Mathematical Intellective Tasks. *Journal of Experimental Social Psychology*(22), 177-189.

Leach, C. W., Bilali, R., & Pagliaro, S. (2015). *Groups and morality.* In M. Mikulincer, P. R. Shaver, J. F. Dovidio, & J. A. Simpson (Eds.), *Group processes* (p. 123–149). American Psychological Association. https://doi.org/10.1037/14342-005

Lee, J. J., & Gino, F. (2015). Poker-faced morality: Concealing emotions leads to utilitarian decision making. *Organizational Behavior and Human Decision Processes*, *126*, 49–64. https://doi.org/10.1016/j.obhdp.2014.10.006

Lee, M., Sul, S., & Kim, H. (2018). Social observation increases deontological judgments in moral dilemmas. *Evolution and Human Behavior*, *39*(6), 611–621. https://doi.org/10.1016/j.evolhumbehav.2018.06.004

Leidner, B., Castano, E., Zaiser, E., & Giner-Sorolla, R. (2010). Ingroup glorification, moral disengagement, and justice in the context of collective violence. *Personality and Social Psychology Bulletin*, *36*(8), 1115–1129. https://doi.org/10.1177/0146167210376391

Leslau, A. (1994). Personality and moral judgment. Personality and Individual Differences, 16(5), 759–765. https://doi.org/10.1016/0191-8869(94)90217-8

Lifton, P. D. (1985). Individual differences in moral development: The relation of sex, gender, and personality to morality. *Journal of Personality*, *53*(2), 306–334. https://doi.org/10.1111/j.1467-6494.1985.tb00368.x

Lin, S. C., Zlatev, J. J., & Miller, D. T. (2017). Moral Traps: When Self-Serving Attributions Backfire in Prosocial Behavior. *Journal of Experimental Social Psychology, 70*, 198-203.

List, C. (2012). The Theory of Judgment Aggregation: An Introductory Review. *Synthese 187*, 179-207. https://doi.org/10.1007/s11229-011-0025-3

List, C., & Pettit, P. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents.* Oxford University Press.

Liu, N., Mok, C., Witt, E. E., Pradhan, A. H., Chen, J. E., & Reiss, A. L. (2016). Nirs-based hyperscanning reveals inter-brain neural synchronization during cooperative jenga game with face-to-face communication. *Frontiers in Human Neuroscience*, *10*, 82. https://doi.org/10.3389/fnhum.2016.00082

Lodge, M., & Taber, C. S. (2005). The automaticity of affect for political leaders, groups, and issues: An experimental test of the hot cognition hypothesis. *Political Psychology*, *26*(3), 455–482. https://doi.org/10.1111/j.1467-9221.2005.00426.x

Lucas, B. J., & Livingston, R. W. (2014). Feeling socially connected increases utilitarian choices in moral dilemmas. *Journal of Experimental Social Psychology*, *53*, 1–4. https://doi.org/10.1016/j.jesp.2014.01.011

Ludwig, K. (2010). Intuitions and Relativity. *Philosophical Psychology, 23*, 427-445.

MacIntyre, A. (1988). *Whose Justice? Which Rationality?*. University of Notre Dame Press.

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., & Bahrami, B. (2015). Equality Bias Impairs Collective Decision-Making across Cultures. *Proceedings of the National Academy of Sciences, 112*(12).

Mallon, R., & Nichols, S. (2011). Dual processes and moral rules. *Emotion Review*, *3*(3), 284–285. https://doi.org/10.1177/1754073911402376

Mann, R. P. (2018). Collective Decision Making by Rational Individuals. *PNAS, 115*(44), E10387-E10396.

Marsh, A. A., Crowe, S. L., Yu, H. H., Gorodetsky, E. K., Goldman, D., & Blair, R. J. R. (2011). Serotonin transporter genotype (5-HTTLPR) predicts utilitarian moral judgments. *PLoS ONE*, *6*(10), e25148. https://doi.org/10.1371/journal.pone.0025148

Mason, W., & Watts, D. J. (2012). Collaborative Learning in Networks. *PNAS, 109*(3), 764-769.

Mata, A. (2019). Social Metacognition in Moral Judgment: Decisional Conflict Promotes Perspective Taking. *J Pers Soc Psychol, 117*(6), 1061-1082.

Mathes, E. W., & Kahn, A. (1975). Diffusion of responsibility and extreme behavior. *Journal of Personality and Social Psychology*, *31*(5), 881–886. https://doi.org/10.1037/h0076695

Maxwell, N. (2017). *Understanding Scientific Progress: Aim-Oriented Empiricism*. Paragon House.

Mayer, J. (2010). Ways of Reading Animals in Victorian Literature, Culture and Science. *Literature Compass, 7*(5), 347-357.

McGrath, S. (2019). *Moral Knowledge.* Oxford University Press.

McLean, S. A. (2007). What and Who Are Clinical Ethics Committees For? *J Med Ethicsv., 33*(9), 497-500.

Meier, B. P., & Hinsz, V. B. (2004). A comparison of human aggression committed by groups and individuals: An interindividual-intergroup discontinuity. *Journal of Experimental Social Psychology*, *40*(4), 551–559. https://doi.org/10.1016/j.jesp.2003.11.002

Maitland, K. A., & Goldman, J. R. (1974). Moral judgment as a function of peer group interaction. *Journal of Personality and Social Psychology*, *30*(5), 699–704. https://doi.org/10.1037/h0037454

Mercier, H. (2011). What good is moral reasoning. *Mind & Society, 10*(2), 131–148. https://doi.org/10.1007/s11299-011-0085-6

Mercier, H., & Sperber, D. (2011). Why Do Humans Reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences, 34*(2), 57-74. https://doi.org/10.1017/S0140525X10000968

Métais, F., & Villalobos, M. (2021). Embodied ethics: Levinas' gift for enactivism. *Phenomenology and the Cognitive Sciences*, *20*(1), 169–190. https://doi.org/10.1007/s11097-020-09692-0

Milgram, S. (1963). Behavioral Study of obedience. *The Journal of Abnormal and Social Psychology*, *67*(4), 371–378. https://doi.org/10.1037/h0040525

Mill, J. S. (1861/1998). *Utilitarianism* (R. Crisp (ed.)), Oxford University Press.

Misak, C. J. (2018). The Pragmatist Theory of Truth. In M. Glanzberg (ed.), *The Oxford Handbook of Truth* (S. 283-303). Oxford: Oxford University Press.

Moll, J., Eslinger, P. J., & de Oliveira-Souza, R. (2001). Frontopolar and Anterior Temporal Cortex Activation in a Moral Judgment Task: Preliminary Functional MRI Results in Normal Subjects. *Arquivos de Neuro-Psiquiatria, 59*(3- B), 657-664.

Moll, J., Zahn, R., Oliveira-Souza, R. de, Krueger, F., & Grafman, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience, 6*(10), 799–809. https://doi.org/10.1038/nrn1768

Monin, B., Pizarro, D. A., & Beer, J. S. (2007). Deciding Versus Reacting: Conceptions of Moral Judgment and the Reason-Affect Debate. *Review of General Psychology, 11*(2), 99-111.

Montague, P. R., Berns, G. S., Cohen, J. D., McClure, S. M., Pagnoni, G., Dhamala, M., … Apple, N. (2002). Hyperscanning: Simultaneous fMRI during Linked Social Interactions. *NeuroImage, 16*(4), 1159–1164. https://doi.org/10.1006/nimg.2002.1150

Myers, D. G., & Bishop, G. D. (1970). Discussion effects on racial attitudes. *Science*, *169*(3947), 778–779. https://doi.org/10.1126/science.169.3947.778

Myers, D. G., & Kaplan, M. F. (1976). Group-Induced Polarization in Simulated Juries. *Personality and Social Psychology Bulletin*, *2*(1), 63–66. https://doi.org/10.1177/014616727600200114

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, *83*(4), 602–627. https://doi.org/10.1037/0033-2909.83.4.602

Narveson, J. (2002). Collective Responsibility. *The Journal of Ethics, 6*(2), 179-198.

Nichols, M. L., & Day, V. E. (1982). A Comparison of Moral Reasoning of Groups and Individuals on the "Defining Issues Test." *Academy of Management Journal*, *25*(1), 201–208. https://doi.org/10.2307/256035

Niebuhr, R. (1932). *Moral Man and Immoral Society: A Study in Ethics and Politics*. Charles Scribner's Sons.

Nisbett, R. E., & Wilson, T. E. (1977). Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review, 84*(3), 231-259.

Nucci, L. P. (2001). *Education in the Moral Domain.*Cambridge University Press.

Orvis, S. (2001). Moral Ethnicity and Political Tribalism in Kenya's "Virtual Democracy". *African Issues, 29*(1/2), 8-13.

Pailing, A., Boon, J., & Egan, V. (2014). Personality, the Dark Triad and Violence. *Personality and Individual Differences*, *67*, 81–86. https://doi.org/10.1016/j.paid.2013.11.018

Park, G., Kappes, A., Rho, Y., & Van Bavel, J. J. (2016). At the heart of morality lies neuro-visceral integration: Lower cardiac vagal tone predicts utilitarian moral judgment.

Social Cognitive and Affective Neuroscience, *11*(10), 1588–1596. https://doi.org/10.1093/scan/nsw077

Patel, V. L., Cytryn, E. H., Shortliffe, E. H., & Safran, C. (2000). The Collaborative Health Care Team: The Role of Individual and Group Expertise. *Teach Learn Med, 12*(3), 117-132.

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., Silani, G., Cikara, M., & Cushman, F. (2020). Reasoning Supports Utilitarian Resolutions to Moral Dilemmas Across Diverse Measures. *Journal of Personality and Social Psychology*, *120*(2), 443–460. https://doi.org/10.1037/pspp0000281

Peirce, C. S. (1931-60). *Collected Papers of Charles Sanders Peirce.* C. Hartshorne, P. Weiss, & A. Burks, (Eds.). Belknap Press of Harvard University Press.

Pennycook, G., Cheyne, J. A., Koehler, D., & Fugelsang, J. A. (2021). On the Belief that Beliefs Should Change According to Evidence: Implications for Conspiratorial, Moral, Paranormal, Political, Religious, and Science Beliefs. *Judgment and Decision Making, 15*, 476-498.

Perkins, A. M., Leonard, A. M., Weaver, K., Dalton, J. A., Mehta, M. A., Kumari, V., Williams, S. C. R., & Ettinger, U. (2013). A Dose of ruthlessness: Interpersonal moral judgment is hardened by the anti-anxiety drug lorazepam. *Journal of Experimental Psychology: General*, *142*(3), 612–620. https://doi.org/10.1037/a0030256

Pettit, P. (2007). Responsibility Incorporated. *Ethics, 117*(2), 171-201.

Piaget, J. (1933). The Moral Judgement of the Child. *Philosophy, 8* (31), 373-374.

Piazza, J., & Sousa, P. (2014). Religiosity, Political Orientation, and Consequentialist Moral Thinking. *Social Psychological and Personality Science*, *5*(3), 334–342. https://doi.org/10.1177/1948550613492826

Pizarro, D. A., Detweiler-Bedell, B., & Bloom, P. (2006). The Creativity of Everyday Moral Reasoning. Empathy, Disgust, and Moral Persuasion. In J. C. Kaufman, & J. Baer (Eds.), *Creativity and Reason in Cognitive Development* (81-98). Cambridge University Press.

Pletti, C., Lotto, L., Buodo, G., & Sarlo, M. (2017). It's immoral, but I'd do it! Psychopathy traits affect decision-making in sacrificial dilemmas and in everyday moral situations. *British Journal of Psychology*, *108*(2), 351–368. https://doi.org/10.1111/bjop.12205

Pölzler, T. (2015). Moral judgments and emotions: A less intimate relationship than recently claimed. *Journal of Theoretical and Philosophical Psychology*, *35*(3), 177–195. https://doi.org/10.1037/teo0000022

Pölzler, T. (2017). Revisiting Folk Moral Realism. *Review of Philosophy and Psychology, 8*, 455-476.

Prinz, J. (2010). The Moral Emotions. In P. Goldie (Ed.), *The Oxford Handbook of Philosophy of Emotion*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199235018.003.0024

Pujol, J., Batalla, I., Contreras-Rodríguez, O., Harrison, B. J., Pera, V., Hernández-Ribas, R., Real, E., Bosa, L., Soriano-Mas, C., Deus, J., López-Solà, M., Pifarré, J., Menchón, J. M., & Cardoner, N. (2012). Breakdown in the brain network subserving moral judgment in criminal psychopathy. *Social Cognitive and Affective Neuroscience*, *7*(8), 917–923. https://doi.org/10.1093/scan/nsr075

Railton, P. (1986). Moral Realism. *Philosophical Review, 95*(2), 163-207.

Rawls, J. (2017). A theory of justice. In *Applied Ethics: A Multicultural Approach: Sixth Edition*. Belknap Press of Harvard University Press. https://doi.org/10.4324/9781315097176

Reinecke, M., & Horne, Z. (2018). *Immutable morality: Even God could not change some moral facts*. https://doi.org/10.31234/osf.io/yqm48

Reinero, D. A., Dikker, S., & Van Bavel, J. J. (2021). Inter-brain synchrony in teams predicts collective performance. *Social Cognitive and Affective Neuroscience*, *16*(1–2), 43–57. https://doi.org/10.1093/scan/nsaa135

Rest, J. R., Davison, M. L., & Robbins, S. (1978). Age Trends in Judging Moral Issues: A Review of Cross-Sectional, Longitudinal, and Sequential Studies of the Defining Issues Test. *Child Development*, *49*(2), 263. https://doi.org/10.2307/1128688

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*(7), 1009–1023. https://doi.org/10.1037/emo0000413

Rezmer, J., Begaz, T., Treat, R., & Tews, M. (2011). Impact of Group Size on the Effectiveness of a Resuscitation Simulation Curriculum for Medical Students. *Teach Learn Med, 23*(3), 251-255.

Rom, S. C., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology*, *74*, 24–37. https://doi.org/10.1016/j.jesp.2017.08.003

Ruby, P., & Decety, J. (2001). Effect of Subjective Perspective Taking During Simulation of Action: A PET Investigation of Agency. *Nat Neurosci, 4*, 546-550.

Rudolph, L. (2010). Gandhi in the Mind of America. *Economic and Political Weekly, 45*(47), 23-26.

Ryan, T. J. (2014). Reconsidering Moral Issues in Politics. *J. Politics, 76*(2), 380-397.

Ryan, T. J. (2019). Actions Versus Consequences in Political Arguments: Insights from Moral Psychology. *J. Politics, 81*(2), 1-15.

Sacco, D. F., Brown, M., Lustgraaf, C. J. N., & Hugenberg, K. (2017). The Adaptive Utility of Deontology: Deontological Moral Decision-Making Fosters Perceptions of Trust and Likeability. *Evolutionary Psychological Science*, *3*(2), 125–132. https://doi.org/10.1007/s40806-016-0080-6

Sandman, L., & Munthe, C. (2010). Shared Decision Making, Paternalism and Patient Choice. *Health Care Anal., 18*(1), 60-84.

Sarkissian, H. (2016). Aspects of Folk Morality. Objectivism and Relativism. In J. Sytsma, & W. Buckwalter (Eds.), *A Companion to Experimental Philosophy* (212–224). John Wiley and Sons. https://doi.org/10.1002/9781118661666.CH14

Sarkissian, H., Parks, J., Tien, D., Wright, J. C., & Knobe, J. (2011). Folk Moral Relativism. *Mind & Language, 26*, 428-505. https://doi.org/10.1111/j.1468-0017.2011.01428.x

Sayre-McCord, G. (1986). The Many Moral Realisms. *The Southern Journal of Philosophy, 24*(Suppl), 1–22. https://doi.org/10.1111/j.2041-6962.1986.tb01593.x

Scharrer, L., Rupieper, Y., Stadtler, M., & Bromme, R. (2016). When Science Becomes Too Easy: Science Popularization Inclines Laypeople to Underrate Their Dependence on Experts. *Public Understanding of Science, 26*(8), 1003-1018.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a Second-person Neuroscience. *Behavioral and Brain Sciences, 36*(4), 393-414. doi:10.1017/S0140525X12000660

Schönegger, P., & Wagner, J. (2019). The Moral Behavior of Ethics Professors: A Replication-Extension in German-Speaking Countries. *Philosophical Psychology, 32*(4), 532-559.

Schwitzgebel, E., & Rust, J. (2009). The Moral Behavior of Ethicists: Peer Opinion. *Mind, 118*, 1043-1059.

Schwitzgebel, E., Cokelet, B., & Singer, P. (2020). Do Ethics Classes Influence Student Behavior? Case Study: Teaching the Ethics of Eating Meat. *Cognition, 203*, 104397.

Searle, J. (1995). *The Construction of Social Reality*.The Free Press.

Sellars, W. (1974). *Essays in Philosophy and its History.* Reidel.

Sharvit, K., Brambilla, M., Babush, M., & Colucci, F. P. (2015). To Feel or Not to Feel When My Group Harms Others? The Regulation of Collective Guilt as Motivated Reasoning. *Personality and Social Psychology Bulletin*, *41*(9), 1223–1235. https://doi.org/10.1177/0146167215592843

Sievers, B., Welker, C., Hasson, U., Kleinbaum, A., & Wheatley, T. (2020). How consensus-building conversation changes our minds and aligns our brains. https://doi.org/10.31234/osf.io/562z7

Singer, P. (1972). Moral Experts. *Analysis, 32*, 115-117.

Singer, P. (2009). *Animal Liberation: The Definitive Classic of the Animal Movement*. HarperCollins Publishers Inc.

Skitka, L. J., & Wisneski, D. C. (2011). Moral Conviction and Emotion. *Emotion Review, 3*(3), 328–330. https://doi.org/10.1177/1754073911402374

Skitka, L. J., Bauman, C. W., & Lytle, B. L. (2009). Limits on legitimacy: Moral and religious convictions as constraints on deference to authority. *Journal of Personality and Social Psychology, 97*(4), 567–578. https://doi.org/10.1037/A0015998

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*(6), 895–917. https://doi.org/10.1037/0022-3514.88.6.895

Skitka, L. J., Hanson, B. E., Morgan, G. S., & Wisneski, D. C. (2021). The Psychology of Moral Conviction. *Annual Review of Psychology, 72*.

Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Sloman, S., & Fernbach, P. (2018). Reasoning as Collaboration. *The American Journal of Psychology*, *131*(4), 493-496. https://doi.org/10.5406/AMERJPSYC.131.4.0493

Sloman, S., & Rabb, N. (2019). Thought as a determinant of political opinion. *Cognition*, *188*, 1–7. https://doi.org/10.1016/j.cognition.2019.02.014

Smetana, J. G. (1983). Social-Cognitive Development: Domain Distinctions and Coordinations. *Developmental Review, 3*, 131-147.

Smith, E. R., & Collins, E. C. (2009). Contextualizing Person Perception: Distributed Social Cognition. *Psychological Review*, *116*(2), 343–364. https://doi.org/10.1037/a0015072

Smith, E. R., & Mackie, D. M. (2015). Dynamics of Group-Based Emotions: Insights From Intergroup Emotions Theory. *Emotion Review*, *7*(4), 349–354. https://doi.org/10.1177/1754073915590614

Smith, E. R., & Mackie, D. M. (2016). Group-level emotions. *Current Opinion in Psychology, 11*(11), 15–19. https://doi.org/10.1016/j.copsyc.2016.04.005

Smith, M. A. (1994). *The Moral Problem.* Blackwell Publishers.

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-Detection Analysis of Group Decision Making. *Psychological Review, 108*(1), 183-203.

Spinoza, B. (1677/1985-2016). Ethics. In B. Spinoza, *The Collected Writings of Spinoza* (Bd. 1). Princeton University Press.

Stevenson, C. L. (1937). The Emotive Meaning of Ethical Terms. *Mind, 46*(181), 14–31. https://doi.org/10.1093/mind/XLVI.181.14

Strohminger, N., Lewis, R. L., & Meyer, D. E. (2011). Divergent effects of different positive emotions on moral judgment. *Cognition*, *119*(2), 295–300. https://doi.org/10.1016/j.cognition.2010.12.012

Sturgeon, N. (1985). Moral Explanations. In J. Rachels, *Ethical Theory 1: The Question of Objectivity.* Oxford University Press.

Tajfel, H., & Turner, J. C. (2004). The social identity theory of intergroup behavior. *Psychology of Intergroup Relations*, 276–293. https://doi.org/10.4324/9780203505984-16

Takezawa, M., Gummerum, M., & Keller, M. (2006). A stage for the rational tail of the emotional dog: Roles of moral reasoning in group decision making. *Journal of Economic Psychology*, *27*(1), 117–139. https://doi.org/10.1016/j.joep.2005.06.012

Tangney, J. P., Stuewig, J., & Hafez, L. (2011). Shame, Guilt, and Remorse: Implications for Offender Populations. *Journal of Forensic Psychiatry and Psychology, 22*(5), 706-723.

Tappin, B. M., & McKay, R. T. (2016). The Illusion of Moral Superiority. *Social Psychological and Personality Science, 8*(6), 623-631.

Taylor, C. (2007). *A Secular Age.* Belknap Press.

Theriault, J., Waytz, A., Heiphetz, L., & Young, L. (2017). Examining Overlap in Behavioral and Neural Representations of Morals, Facts, and Preferences. *Journal of Experimental Psychology: General, 146*(11), 1586–1605.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217. https://doi.org/10.5840/monist197659224

Timmons, S., & Byrne, R. M. J. (2019). Moral fatigue: The effects of cognitive fatigue on moral reasoning. *Quarterly Journal of Experimental Psychology*, *72*(4), 943–954. https://doi.org/10.1177/1747021818772045

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*(10), e0164012.

Toppi, J., Borghini, G., Petti, M., He, E. J., De Giusti, V., He, B., Astolfi, L., & Babiloni, F. (2016). Investigating cooperative behavior in ecological settings: An EEG hyperscanning study. *PLoS ONE*, *11*(4), e0154236. https://doi.org/10.1371/journal.pone.0154236

Tropp, L. R., & Pettigrew, T. S. (2006). Relationships Between Intergroup Contact and Prejudice Among Minority and Majority Status Groups. *Psychological Science, 16*(12), 951-957.

Tuomela, R. (2007). *The Philosophy of Sociality: The Shared Point of View*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195313390.001.0001

Turiel, E. (1983). *The Development of Social Knowledge. Morality and Convention.* Cambridge University Press.

Turiel, E. (2008). The Development of Children's Orientations toward Moral, Social, and Personal Orders: More than a Sequence in Development. *Human Development, 51*, 21-39.

Ueshima, A., Mercier, H., & Kameda, T. (2021). Social deliberation systematically shifts resource allocation decisions by focusing on the fate of the least well-off. *Journal of Experimental Social Psychology*, *92*. https://doi.org/10.1016/j.jesp.2020.104067

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making*, *4*(6), 479–491.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326–334. https://doi.org/10.1016/j.cognition.2012.10.005

Urban, P. (2014). Toward an expansion of an enactive ethics with the help of care ethics. *Frontiers in Psychology*, *5*(NOV), 1354. https://doi.org/10.3389/fpsyg.2014.01354

Valdesolo, P., & Desteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*(6), 476–477. https://doi.org/10.1111/j.1467-9280.2006.01731.x

Van Bavel, J. J., Packer, D. J., Johnson Haas, I., & Cunningham, W. C. (2012). The Importance of Moral Construal: Moral Versus Non-Moral Construal Elicits Faster, More Extreme, Universal Evaluations of the Same Actions. *PLoS ONE, 7*(11), e48693.

Van Gils, S., Otto, T., & Dinartika, N. (2020). Better together? The neural response to moral dilemmas is moderated by the presence of a close other. *Journal of Neuroscience, Psychology, and Economics*, *13*(3), 150–163. https://doi.org/10.1037/npe0000126

Van Kleef, G. A., & Fischer, A. H. (2016). Emotional collectives: How groups shape emotions and emotions shape groups. *Cognition and Emotion*, *30*(1), 3–19. https://doi.org/10.1080/02699931.2015.1081349

Van Laar, C., Bleeker, D., Ellemers, N., & Meijer, E. (2014). Ingroup and outgroup support for upward mobility: Divergent responses to ingroup identification in low status groups. *European Journal of Social Psychology*, *44*(6), 563–577. https://doi.org/10.1002/ejsp.2046

Varela, F. J. (1999). *Ethical know-how: action, wisdom, and cognition*. Stanford University Press.

Vives, M.L., Cikara, M., & FeldmanHall, O. (2021). Following Your Group or Your Morals? The In-Group Promotes Immoral Behavior While the Out-Group Buffers Against It. *Social Psychological and Personality Science*, 194855062110012. https://doi.org/10.1177/19485506211001217

Vranas, P. B. M. (2004). Lack of Character: Personality and Moral Behavior. *The Philosophical Review, 113*(2), 284–288. https://doi.org/10.1215/00318108-113-2-284

Wainryb, C., Shaw, L. A., Langley, M., Cottam, K., & Lewis, R. (2004). Children's Thinking about Diversity of Belief in the Early School Years: Judgments of Relativism, Tolerance, and Disagreeing Persons. *Child Development, 75*(3), 687-703.

Walker, R. C. (2018). The Coherence Theory of Truth. In M. Glanzberg (Ed.), *The Oxford Handbook of Truth* (pp. 219-237). Oxford University Press.

Wallach, M. A., & Kogan, N. (1965). The roles of information, discussion, and consensus in group risk taking. *Journal of Experimental Social Psychology*, *1*(1), 1–19. https://doi.org/10.1016/0022-1031(65)90034-X

Wallach, M. A., Kogan, N., & Bem, D. J. (1964). Diffusion of responsibility and level of risk taking in groups. *Journal of Abnormal and Social Psychology*, *68*(3), 263–274. https://doi.org/10.1037/h0042190

Wheatley, T., Boncz, A., Toni, I., & Stolk, A. (2019). Beyond the Isolated Brain: The Promise and Challenge of Interacting Minds. *Neuron*, *103*(2), 186–188. https://doi.org/10.1016/j.neuron.2019.05.009

Wiech, K., Kahane, G., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2013). Cold or calculating? Reduced activity in the subgenual cingulate cortex reflects decreased emotional aversion to harming in counterintuitive utilitarian judgment. *Cognition*, *126*(3), 364–372. https://doi.org/10.1016/j.cognition.2012.11.002

Wilkinson, D., & Savulescu, J. (2018). Ethics, Conflict and Medical Treatment for Children.

Williams, B. (2011). Ethics and the limits of philosophy. In *Ethics and the Limits of Philosophy*. Harvard University Press. https://doi.org/10.4324/9780203828281

Wilson, R. A. (2004). *Boundaries of the Mind: The Individual in the Fragile Sciences - Cognition*. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511606847

Wong, D. B. (2006). *Natural Moralities: A Defense of Pluralistic Relativism*. Oxford University Press.

Wright, J. C., Cullum, J., & Schwab, N. (2008). The Cognitive and Affective Dimensions of Moral Conviction: Implications for Attitudinal and Behavioral Measures of Interpersonal Tolerance. *Pers. Soc. Psychol. Bull., 34*(11), 1461-1476.

Wright, J. C., Grandjean, P. T., & McWhite, C. B. (2013). The Meta-Ethical Grounding of Our Moral Beliefs: Evidence for Metaethical Pluralism. *Philosophical Psychology, 26*, 336-361.

Wright, J. C., McWhite, C. B., & Grandjean, P. T. (2014). The Cognitive Mechanisms of Intolerance: Do our Meta-Ethical Commitments Matter? In T. Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford Studies in Experimental Philosophy* (Bd. 1). Oxford University Press.

Yang, Q., Luo, C., & Zhang, Y. (2017). Individual Differences in the Early Recognition of Moral Information in Lexical Processing: An Event-Related Potential Study. *Scientific Reports, 7*(1475).

Young, L., & Dugan, J. (2011). Where in the Brain is Morality? Everywhere and Maybe Nowhere. *Social Neuroscience, 7*(1), 1-10. https://doi.org/10.1080/17470919.2011.569146

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*(2), 151–175. https://doi.org/10.1037/0003-066X.35.2.151

Zhao, J., Harris, M., & Vigo, R. (2016). Anxiety and moral judgment: The shared deontological tendency of the behavioral inhibition system and the unique utilitarian

tendency of trait anxiety. *Personality and Individual Differences*, *95*, 29–33. https://doi.org/10.1016/j.paid.2016.02.024

Chapter 3. Many Heads Are More Utilitarian Than One

Anita Keshmirian*[1,2], Bahador Bahrami[3,4,5], Ophelia Deroy[2,6,7]

(1) Graduate School for Neuroscience, Ludwig-Maximillian's-University, Munich Germany.
(2) Faculty of Philosophy, Ludwig-Maximilian's University, Munich Germany.
(3) Department of General Psychology and Education, Ludwig Maximillian's University, Munich, Germany.
(4) Centre for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany.
(5) Department of Psychology, Royal Holloway University of London, Egham, Surrey, London UK.
(6) Munich Center for Neuroscience, Munich, Germany.
(7) Institute of Philosophy, School of Advanced Study, University of London, London UK
* Correspondence should be addressed to Anita Keshmirian, Ludwig-Maximillian's-University of Munich, Munich, Germany. Email: anita.keshmirian@campus.lmu.de

Author Note. This research was funded by the NOMIS Foundation (OD). BB was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (819040 - acronym: rid-O). BB was also supported by the Humboldt Foundation. **Open practices statement:** The experiment and analyses were preregistered at the Open Science Framework (https://osf.io/vxcp8/). The data generated and analyzed during the study, in addition to the script and the description of the analysis, are publicly available at DOI 10.17605/OSF.IO/JMKX5.

CRediT author statement: AK: Exp 1 Conceptualization, Methodology, Formal analysis, Investigation, Coding, Visualization, Writing-Original- Draft preparation **AK: Exp2:** Conceptualization, Methodology, Coding, Formal analysis, Investigation, Visualization, Writing-Original- Draft preparation **OD**: **Exp 1**: Methodology, Funding, Supervision **BB**: **Exp 1:** Methodology, Writing-Review & Editing, Supervision. **BB Exp2:** Writing-Review & Editing**,** funding, supervision. The authors declare that they have no known competing financial interests or personal

relationships that could have appeared to influence the work reported in this paper. We thank Nadine Fleischhut and Stephan Sellmaier for the helpful discussions.

Abstract: Moral judgments have a very prominent social nature, and in everyday life, they are continually shaped by discussions with others. Psychological investigations of these judgments, however, have rarely addressed the impact of social interactions. To examine the role of social interaction on moral judgments within small groups, we had groups of 4 to 5 participants judge moral dilemmas first individually and privately, then collectively and interactively, and finally individually a second time. We employed both real-life and sacrificial moral dilemmas in which the character's action or inaction violated a moral principle to benefit the greatest number of people. Participants decided if these utilitarian decisions were morally acceptable or not. In Experiment 1, we found that collective judgments in face-to-face interactions were more utilitarian than the statistical aggregate of their members compared to both first and second individual judgments. This observation supported the hypothesis that deliberation and consensus within a group transiently reduce the emotional burden of norm violation. In Experiment 2, we tested this hypothesis more directly: measuring participants' state anxiety in addition to their moral judgments before, during, and after online interactions, we found again that collectives were more utilitarian than those of individuals and that state anxiety level was reduced during and after social interaction. The utilitarian boost in collective moral judgments is probably due to the reduction of stress in the social setting.

Keywords Topic: Collective Moral Judgments, Group Moral Decisions, Moral Dilemmas, Moral Conformity, Moral Influence, Social Deliberation

Keywords Method: Logistic Mixed Effect Model, Bayesian Mixed Effect Models, Open Science, Open data.

Highlights:

- Collective consensual judgments made via face-to-face and online group interactions were more utilitarian compared to the private individual judgments.

- Group discussion did not change the individual judgments indicating a normative conformity effect.
- Individuals consented to a group judgment that they did not necessarily buy into personally.
- Collectives were less stressed than individuals after responding to moral dilemmas.
- Interactions reduced aversive emotions (e.g., stress, regrets) associated with violation of moral norms.

3.1 Introduction

Moral judgments are often collective. We discuss our individual opinions about the moral actions of friends, institutions, celebrities, and authorities within our social network. In fact, we spend most of our social conversations discussing others' moral failures (Dunbar, 2004). However, often different people have different moral opinions about the same moral issue. Consider the following scenario:

> "After a violent murder in Germany, a journalist who investigates the case found evidence that the government of a foreign country ordered the murder. That country is a long-time trade partner of Germany, with which the German state is about to conclude a large trade agreement. This agreement will create 10,000 new jobs in Germany. If the journalist blows the whistle, the trade deal will collapse. The journalist decides to ignore the evidence. The trade deal goes through successfully, bringing wealth and employment to thousands of people. Was the journalist's decision morally acceptable?"

Here, one might argue that upholding the principles of justice and journalistic duty requires pursuing and revealing the truth at any cost. This line of argument would conclude that what the journalist did was morally wrong. Others, who prefer to look at the outcome, may approve of the decision because it brought so much benefit and prosperity to many people.

Although fictional, these kinds of scenarios are not far from reality. Many decisions and actions involve breaking a norm, a promise, a rule, or a moral code to increase the utility for a larger group (e.g., active and passive euthanasia, abortion, white lies, restricting children's education to protect the elderly in the time of a global pandemic, discontinuing life support in comatose patients, etc.). The moral permissibility of such decisions may raise strong disputes in different people and lead to public and private discussions. In reality, however, within a group, a panel, among friends, or in families, when people discuss these decisions, how do they collectively decide about these moral issues? More generally, how collective

moral judgments are shaped by individual members' opinions? Conversely, do group interactions change the individual's private moral judgments?

Previous works have extensively examined the philosophical, social, cognitive, and neurobiological substrates of individual moral judgments and decisions (e.g., Greene et al., 2001; Greene et al., 2004; Haidt, 2001; Mallon & Nichols, 2011; Moll et al., 2008). However, most of the previous theories in moral psychology explain morality at the individual level. Several recent works have underscored this *overwhelming* focus on examining *individuals* making decisions or judgments *in isolation* (Bloom, 2010; Ellemers, 2017; Ellemers et al., 2019; Fedyk, 2019; Gert, 2005; Haidt, 2007; Leach et al., 2015). Here we set out to examine this overlooked but fundamental role of *social interaction* among individuals in groups that engage in moral discussions.

The relationship between the individual and the collective morality has been at the heart of some of the most influential works of the post-war 20th-century philosophy (Arendt, 1987) and social psychology (Festinger & Carlsmith, 1959; Milgram, 1963; Myers & Bishop, 1970; Myers & Kaplan, 1976; Myers & Lamm, 1976; Wallach et al., 1964; Wallach & Kogan, 1965). Our questions – posed above – invite the reader to evaluate the reciprocal interactive relationship between the group and individual moral judgments in the light of those previous influential works.

We first provide a brief overview of the literature on moral judgments in individuals. Rather than being exhaustive, we highlight the contextual or psychological factors that have been shown to drive moral judgments towards or away from one moral theory or another in individuals and in isolation. Then we turn to the literature on social interaction and majority influence to examine the effect of interaction on the same contextual and psychological factors that modulate moral judgments. Putting our review of the two fields together, we provide our theoretical synthesis, which is then tested empirically.

3.1.1 Moral dilemmas

Moral dilemmas describe situations where it is necessary to choose between alternative actions, each of which violates a moral principle (e.g., holding on to a secret or informing a happily-married friend that his/her partner has been cheating on them). Moral dilemmas are difficult to resolve because they admit two incompatible moral actions. Often each of these conflicting actions is related to a different moral theory. For instance, according to a broad family of *utilitarian* moral theories*,* an action is acceptable if it maximizes the utility for the greatest number of people (Mill, 1863; Rosen, 2006) even if securing that utility entails violating moral rules like disregarding promises, duties, norms, etc. *Utilitarianism*, therefore, is a consequentialist moral theory; because, in its moral evaluation of a given action,

it primarily cares about the consequences of that action. In contrast, *deontological* moral theories care primarily about upholding universal moral principles - what Kant (1948) called categorical imperatives- and give consequences a lower priority. These principles often make direct, inflexible, universal, and unequivocal moral rules such as 'Do not lie,' 'Do not kill' or 'Do not break a promise' (Kant, 1948; Scruton, 2001).

Borrowed from philosophy, a class of moral dilemmas known as 'sacrificial dilemmas' are commonly used in moral psychology, which entails instrumental harm to some in order to save others (see "Trolley Problem"; Foot, 1967; Thomson, 1976 for a review, see Christensen et al., 2014). In these dilemmas a utilitarian moral agent (whose does not benefit personally from the consequences) would harm one innocent person if the harm benefits many. Conversely, harming an innocent person is wrong for deontology regardless of the number of lives that the inflicted harm might save.

To study the psychological and neural processes underlying these conflicting motives in individuals, experimenters often have participants read scenarios that include different sacrificial dilemmas (or, more recently, experience the scenario in virtual reality), imagine themselves in the situation, and decide what they would choose to do. Similarly, moral judgments are measured by having the participant evaluate a scenario in which an action was taken by a protagonist and see if (or how much) the participant would endorse the protagonist's decision.

Moral Dilemmas and Deliberation

Over the last two decades, this research line has shown that moral judgments are not fixed in stone and can be modulated in individuals. For instance, several converging pieces of evidence support the effect of deliberation and reasoning in utilitarian judgments. Across diverse measurements, reasoning and deliberation led to more utilitarian responses (Patil et al., 2020). Reflection and deliberation encouraged more utilitarian views (Paxton et al., 2012). Giving participants analytical mathematical puzzles before reading the moral scenarios made them more utilitarian by 'activating their thinking mode' (Kvaran et al., 2013). Asking participants to be more deliberative and analytical had a similar effect (Li et al., 2018). Better performance in the cognitive reflection test (CRT) predicted more utilitarian decisions (Byrd & Conway, 2019). Conversely, increasing cognitive load decreased the utilitarian choices (Greene et al., 2008). Participants who experienced cognitive fatigue showed a similar result (Timmons & Byrne, 2019). Contrariwise, decreasing cognitive load by showing the ratio of 'killed' vs. 'saved' people in sacrificial scenarios increased utilitarian decisions (Trémolière & Bonnefon, 2014). Restricting the response time reduced the utilitarian responses in some studies as well (Cummins & Cummins, 2012; Suter & Hertwig, 2011).

Moral Dilemmas and Emotions

Moral judgments have not only been attributed to increased deliberation and reasoning but to emotional factors. Both the nature and intensity of the feelings that

one experiences when contemplating these dilemmas modulate moral judgments. For instance, difficulties in emotion regulation (Zhang et al., 2017) and emotional reappraisal (Feinberg et al., 2012) decreased deontological responses. Similarly, presenting the dilemmas in foreign languages increased the reported emotional distance and reduced deontological choices (Hayakawa et al., 2017). Active suppression of emotions (Lee & Gino, 2015) and administration of anti-anxiety drugs (i.e., Lorazepam) in normal participants increased utilitarian judgments (Perkins et al., 2013; but also see Zhao et al., 2016), and so did the induction of some (but not all) positive emotions (Strohminger et al., 2011; Valdesolo & Desteno, 2006). On the other hand, negative and aversive emotions reduced utilitarian preferences. For instance, socially induced physiological stress (e.g., by having the participants anticipate a rigorous social evaluation such as public speaking) which elevates the stress hormone Cortisol in humans (Kirschbaum et al., 1993), decreased utilitarian responses (Starcke et al., 2012; Youssef et al., 2012; Zhang et al., 2018).

Moral Dilemmas and Aversive Feelings Towards Norm Violations

One hypothesis for these complex links between emotions and moral judgments points to people having aversive feelings towards norm violations associated with the utilitarian branch of moral dilemmas. Since utilitarian actions in moral dilemmas entail norm violations such as instrumental harm in sacrificial dilemmas, it has been hypothesized that moral judgments are shaped by our sensitivity to norm violation and aversive emotional reaction to harm. In line with this hypothesis, reduced emotional responsiveness to the aversive nature of harm was associated with more utilitarian responses (Cushman & Greene, 2012; Greene, 2007). In fact, utilitarian judgments have been frequently found in patient groups who purportedly demonstrate hampered emotional responses, such as patients with ventromedial prefrontal cortex brain lesions (an area related to socio-emotional processing) (Ciaramelli et al., 2007; Koenigs et al., 2007) frontotemporal dementia (Mendez et al., 2005), and psychopaths (Koenigs et al., 2012). In healthy individuals, utilitarian judgments have been more frequently found in antisocial personality traits (Bartels & Pizarro, 2011) and psychopathy (Paytas, 2014; Pletti et al., 2017). A recent model that disentangled sensitivity to Consequence (or utilitarianism), Norm (or deontology), and Inaction showed that psychopaths had a weaker sensitivity to moral norms and therefore were less deontological in their moral decisions (Gawronski et al., 2017). Therefore, the previously reported utilitarian boost in psychopaths was probably not related to the higher-order reasoning or their concern for the greater good but less acceptance of norms (see Everett & Kahane, 2020; Kahane, 2015). Two interventional studies showed that experimentally increasing sensitivity to norm violation by induction of stress (Li et al., 2019) or by exogenously enhancing serotonin level (i.e., Citalopram administration) (Crockett et al., 2010) decreased utilitarian responses.

Moral Dilemmas and Post-decisional Emotions

The role of emotional valence in moral judgments has also been linked to post-decisional emotions such as regret. In fact, participants not only minimized their current distress at the time of the moral decision, but they also tried to minimize the post-decisional negative emotions such as regret (Tasso et al., 2017). Supporting this idea, one study showed that experiencing higher regret was negatively correlated with utilitarian choices (Szekely & Miu, 2015). Another work found that endorsing the utilitarian (vs. deontological) judgments induced more affective (rather than cognitive) regret (Goldstein-Greenwood et al., 2020). Experiencing other post-decisional negative emotions such as guilt, shame, anger, and disgust have also been reported in sacrificial moral dilemmas (Pletti et al., 2016).

Emotion vs. Deliberation in Moral Dilemmas: Dual Process Models

The role of emotion vs. deliberation has been has been at the heart of *understanding moral behavior*. For instance*,* the extensive body of empirical evidence for the role of emotions in moral judgments was preceded by much earlier works of Unamuno (1954), the Spanish philosopher who passionately argued that '*moral reasoning'* is nothing but the conscious, *ex-post,* phenomenal experience of some underlying, (emotional) unconscious process that has already made the agent's mind about the issue at hand before the agent starts to consider the reasons for or against it (Unamuno, 1954). Later, inspired by Unamuno's views, Blasi (1980) argued that moral decisions and actions motivate moral reasoning, not the other way around (Blasi, 1980). In line with this view, the social intuitionist account of morality (Haidt, 2001) described moral reasoning as post-hoc justifications of the unconscious, automatic and emotional processes that are only suitable for communicating one's moral position. The intuitionist account would argue that objective (e.g., less emotionally driven) moral reasoning might be possible but is very rare and happens under specific circumstances in which emotions and intuitions are kept under control such as in social interactions (Haidt, 2001).

Inspired by the dual-process models of cognition (Evans & Stanovich, 2013; Sloman, 1996; Stanovich, 2009, among others), dual-process approaches to morality framed the evaluation of moral dilemmas neither as purely automatic/emotional nor purely deliberative but as a competition between a. fast, intuitive processes that involve emotions and b. slower deliberative processes that involve reasoning. Dual-process models suggest that deontological judgments arise when the emotional-intuitive process overrides the cognitive system. Conversely, the more 'intellectual' utilitarian judgment is favored when the slower deliberative cognitive system overrides the emotional-intuitive one (Greene et al., 2001; 2004; 2008). Therefore, in deontological judgments, the emotional-intuitive system shapes the individual's conscious narrative of "why" they came to the deontological judgment. In the case of utilitarian judgments, the deliberated cost-benefit analysis of the cognitive system is communicated.

Put together, the above theoretical and empirical works show the role of emotions alongside deliberation and reasoning in moral judgments. They also suggest that moral judgments are not rigid. Rather than having to choose between

moral rationalists (e.g., Kohlberg, 1973), who claimed that moral judgments are the outcome of pure rational deliberation on the one hand, and the moral intuitionists (e.g., Unamuno, 1954; Blasi, 1980; Haidt, 2001), who prioritized emotions and intuitions exclusively on the other, more recent theories such as dual-process models (discussed above) suggest that the outcome of moral judgments in a specific situation is the inevitable result of the interplay between the deliberative and the emotional systems that are simultaneously present in human mental processes and lead the moral agent towards or away from utilitarian (or deontological) moral judgments.

3.1.2 Social interaction and modulators of moral judgment

In the previous section, we highlighted a number of factors that could modulate moral judgment. Next, we examine the existing evidence from social cognition about how these modulators may be affected by social interaction.

Social Interaction and Group Deliberation

Interpersonal communication of information in social contexts allows groups of people to surpass what each individual could have achieved in decision-making under uncertainty in sensory domains (Bahrami et al., 2016; Sorkin et al., 2001) in numerical cognition (Bahrami et al., 2012) and in problem-solving (Mason & Watts, 2012). Studies have shown that when people talk to one another, they could calibrate their uncertainty (Bang & Frith, 2017; Fusaroli et al., 2012), produce diverse arguments (Mercier & Sperber, 2011), understand the same problem from various viewpoints, and arrive at solutions that had not been available to any single member of the group (Smith & Collins, 2009). These studies strongly suggest that group interactions are likely to increase conscious deliberation, reasoning, and analytical thinking. We previously reviewed that deliberation could lead to more utilitarian decisions.

Social Interaction and Group Norm Violation

Utilitarian decisions are also often operationalized by breaking different norms, and individuals seem to break norms more often when they decide together. Niebuhr, in 1932, found the ‘limitations of human nature’ responsible for the moral failure of individuals in social groups to the extent that he thought man's collective moral behavior could never be dominated by reason. Thus groups always remain more immoral than their members (Niebuhr, 1932). Although Niebuhr’s ideas about individuals being more immoral in groups were related to social groups, different studies confirmed his predictions, even in informal groups. In fact, in diverse moral domains, immoral actions in forms of norm violations were more probable in groups: people lied more in groups (Conrads et al., 2013); communication within groups increased this group level dishonesty (Kocher et al., 2018); collaboration

made individuals excessively more lying (Weisel & Shalvi, 2015), free-riding and social loafing were more probable in groups (Heuzé & Brunel, 2003; Latane et al., 1979), groups showed less compliance to defined norms than individuals (Fochmann et al., 2021), and people tended to be less generous in groups. Bornstein & Yaniv (1998) and later, El Zein et al. (2020) showed that people violated the fairness norm more often when they are in groups of three, compared to when they were alone.

More recently, this increased incidence of norm violation in groups has been attributed to shared responsibility (Conrads et al., 2016; El Zein et al., 2019, 2020; El Zein & Bahrami, 2020). For instance, when participants were asked to provide reasons for being dishonest in groups, their arguments were based on 'feeling less responsible' rather than 'benefiting other group members by their lies' (Conrad et al., 2017). One clear demonstration of the role of this collectively shared responsibility in moral decision-making is the bystander effect: when several observers witnessed a norm violation, it was less likely that any one of them would intervene (Darley & Latané, 1968; Forsyth et al., 2002; Wallach et al., 1964). In a group, the responsibility for an action is not focused on anyone but is rather shared among all present (see El Zein et al., 2019). Norm violation (Wilson & O'Gorman, 2003) and their corresponding feeling of responsibility (Bell, 1982, 1985; Giorgetta et al., 2012; Loomes & Sugden, 1982; Zeelenberg, 1999) both are associated with a diverse range of negative emotions such as distress, disappointment, and regret. Sharing responsibility among members of a group decreases these emotions, such as feelings of anticipated regret in group economic decisions (El Zein & Bahrami, 2020). It also helped mitigate the negative emotions such as stress that accompanies difficult choices that may have long-lasting emotional repercussions (Botti et al., 2009; Frey & Tropp, 2006). Together, these studies span a diverse range of situations in which social context can reduce the emotional burdens of norm violation, both at the time of the decisions (e.g., stress) and the predicted emotions in future states (e.g., anticipated regret) in individuals.

3.1.3 Current study: moral judgment and social interactions

Study Overview

In the current study, we asked three questions: how are collective moral judgments different from individual ones? How are individual moral judgments different before and after a discussion? And finally, what is the underlying mechanism at work in collective moral judgment which explains these differences? To address these questions, we had small groups of interacting individuals, individually (in private) and collectively (after short discussions), rate the moral permissibility of actions (or inactions) described in different scenarios. In each scenario, a character's decision violated a moral norm to increase some utility for a greater number of people. For instance, in one scenario, the character had to lie to

collect money for people in need. In another, he had to kill someone to prevent more deaths or had to stay silent about the infidelity of a friend's partner to avoid disturbing a happy relationship. Examining these scenarios one at a time, participants started by reading each scenario privately and rated the acceptability of the character's choice in the scenario (First individual judgment). Then, for half of the scenarios, they proceeded to discuss the case with their fellow group members and rate the acceptability of choice as a group (Collective judgment). Finally, participants revisited all scenarios privately once again and rated the moral acceptability of what the character had done (Second individual judgment).

Moral Domain: Emergent Properties Related to Interaction

Earlier, we discussed some of the factors that modulate moral judgments and how those factors are affected by social context as examined by prior researchers. Here, we also note that collective deliberation is not equivalent to the aggregation of many individuals that deliberate independently. The interaction may shape the content and quality of deliberations producing emergent phenomena at the collective level that would not have been observed if many individuals' opinions were aggregated statistically (Karpowitz & Mendelberg, 2007). We focus on the interactions to see how their emergent properties differ from statistically aggregated groups.

Virtue Signaling via Interaction

One such emergent effect is the adaptive utility of taking a deontological position in public: previous research has shown that people who expressed deontological judgments were valued more, chosen more often as social partners (Capraro et al., 2018), and were perceived to be more prosocial in economic games (Everett et al., 2016). In a social context, people may express deontological judgments to advertise virtues and curate their social images by promoting perceptions of trust and likeability (Sacco et al., 2017). In line with this view, utilitarians were often regarded as lacking integrity, empathy, and moral character (Uhlmann et al., 2013). Conversely, it has been reported that utilitarian agents sometimes are regarded positively as well (as logical, competent, deliberative and intelligent, and leader-like; Uhlmann et al., 2013). However, in previous research, strategic self-presentation has been found more consistently in the deontological rather than utilitarian direction (Sacco et al., 2017; Everett et al., 2016). For instance, participants who were socially observed by a third person (Lee et al., 2018) or even by themselves in the mirror (Reynolds et al., 2019) preferred deontological judgments. By contrast, there is hardly any evidence to show that people may actually *practice* utilitarianism as a reputation management tactic. Therefore, we hypothesize that virtue signaling (VS) in groups may decrease collective utilitarian consensus. However, due to its social signaling function, we do not expect individuals to change their minds privately. Virtue Signaling hypothesis (VS) would also predict that group deliberation would not change individual (private) judgments (see Figure 1).

Deliberation via Interaction

The second emergent effect of social interaction is the deliberative role of social discussions. Social discussion promotes deliberation and analytical reasoning, permitting participants to spend more time, provide and hear more arguments, combine their perspectives, and share resources to reason about moral issues. One study that compared group vs. individual moral judgments showed that groups displayed more advanced moral reasoning than individuals (Nichols & Day, 1982). In addition, the discussion provides the participants with more pieces of information and arguments than individual deliberation. More deliberation and reasoning increase utilitarian judgments (Paxton et al., 2012; Paxton & Greene, 2010). Based on these findings, we offer our Social Deliberation (SD) hypothesis: collective judgments would be more utilitarian because they happen after social deliberation. Importantly, this utilitarian boost would be expected to permeate to the second individual judgment because deliberation facilitates better reasoning and sharing of information, helping to convince the individual participants to change their minds. In addition, in the second individual judgment, participants read the moral scenarios for the third time, allowing them to deliberate individually after the first and the collective social deliberations. This hypothesis, therefore, predicts that the second individual judgments would also be more utilitarian than the first individual judgments, but specifically for the scenarios discussed collectively. Indeed, a recent work that investigated moral reasoning in groups of students (Curşeu et al., 2020) found that discussion in groups led to more utilitarian decisions. This study evaluated majority influence, minority influence, and normative deviance as mechanisms that explain the association between individual and group level moral preferences using mediation analysis. However, it did not measure individual opinions after the discussion, leaving open any conclusions about the change of mind as the result of social deliberation in moral judgments. Myers & Kaplan (1976) previously examined the impact of social interactions on private opinions. They measured individual opinions after the discussion but did not include a consensus decision-making stage.

Stress Reduction via Interaction

A third hypothesis for how interactions may impact collective judgments is the reduction of negative feelings in groups. Group discussion decreased negative feelings in individuals on their decisions even when these negative feelings were artificially induced (Kaplan & Miller, 1978). We also saw earlier that social context could reduce the negative emotions associated with norm violation and facilitate the violation of different norms in groups. Previous research showed that reducing negative current emotions (e. g., stress) or future negative emotions (e.g., anticipated regret) could both increase utilitarian judgments. Therefore, once in the group, individuals may find it easier to endorse utilitarian actions that involve trading off violations of a norm such as "do not harm anyone" with the greater good for a larger number of people. Thus, this Stress Reduction (SR) hypothesis makes a similar prediction to Social Deliberation (SD). However, if more utilitarian

collective judgments arise from the transient regulation of emotional responses and temporary reduction of negative emotions such as stress and regret in groups, then SR would expect that when participants decide a second time individually, then group-induced stress reduction would no longer be available. SR hypothesis diverges from SD here and predicts that the second individual judgments would be no more utilitarian than the first. In the interest of transparency, we clarify here that in our OSF pre-registration, VS and SD hypotheses (but not SR) were included.

Conformity via Interaction

The role of the conformity effect must be carefully considered and clearly separated from interactions when trying to understand moral judgments in a social context. By conformity, we refer to opinion changes that follow from (merely) knowing other people's judgments. Conformity, therefore, does not involve any bilateral social interactive elements, and our group-interaction hypotheses discussed above (VS, SD, and SR) do not apply to it. However, previous conformity literature does offer important clues for drawing some testable predictions for our experiments. Those for whom the majority supported the utilitarian option did not differ from control. Using a social conformity paradigm similar to Asch's (1956), one study showed that conformity increased both permissible and impermissible moral actions depending on the majority opinion (Kundu & Cummins, 2013). In another study, participants who judged moral dilemmas were simultaneously presented with information about the percentage of other people favoring the deontological or utilitarian option. Those for whom the majority supported the deontological option deviated from the control condition (i.e., no social information) (Bostyn & Roets, 2017). Another study showed that conformity to publicly announced majority opinion was generally followed by a change of private attitudes (Cornwell et al., 2019). In yet another study, participants shifted their public (but not private) responses towards more or less utilitarian judgments depending on what a hypothetical, observing evaluator favored (Rom & Conway, 2018). Thus, although all four studies do show that conformity could indeed affect moral judgments, the latter two studies disagree on whether individual moral cognition could be genuinely susceptible to conformity. Whereas one study (Cornwell et al., 2019) showed evidence for change of private opinion in line with 'informational' conformity, the other (Rom & Conway, 2018) suggested that individual moral judgments are likely to superficially conform and demonstrate 'normative' influence without any change of private attitude (see Cialdini & Goldstein, 2004; See also Deutsch & Gerard, 1955).

Based on these findings, we propose two additional hypotheses. First, we expect that collective judgments would be different from individual judgments, and the direction of this change would follow the majority opinion of the group's individual members. Second, examining the individual opinions after discussion would allow us to distinguish between normative and informational conformity. If people only conform normatively, they change their opinion in public but hold on to their original individual judgments even after the discussion (Cialdini & Goldstein,

2004). If, on the other hand, informational conformity is at work, one would predict that the collective judgment should pull the second private judgment towards itself.

To sum up, we put forward two complementary sets of predictions to examine the possible psychological substrates of social interactive moral judgments (see Figure 1). Our conformity hypotheses do not make any directional predictions about whether groups favor utilitarian or deontological judgments, but the three interaction hypotheses (VS, SD, and SR) do. On the other hand, interaction hypotheses – but not conformity – are silent about any relationship between the distribution of individual opinions (e.g., the majority) and group consensus. Finally, both sets of theories make overlapping predictions about the change of second individual opinion: Informational Conformity and Social Deliberation hypotheses predict that second individual opinions would follow the consensus, but Normative Conformity, Virtue Signaling, and Stress Reduction hypotheses predict that individual opinions will not be affected by the consensus.

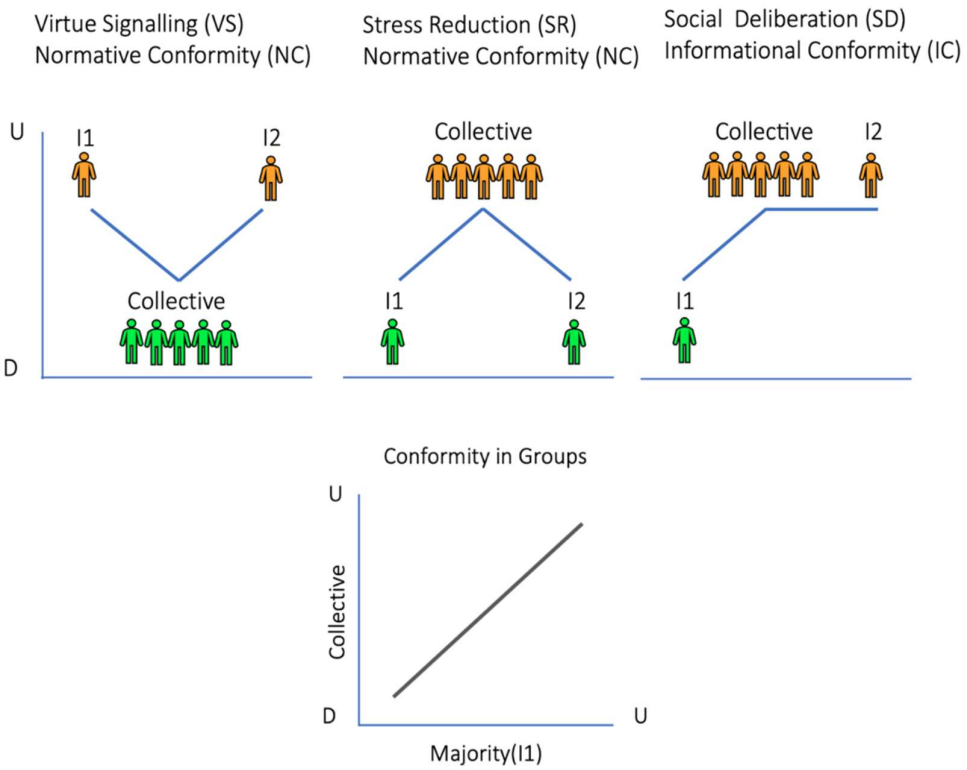


Figure 1. Top left: Virtue Signaling hypothesis (VS) predicts a deontological boost in collective judgments (i.e., fewer utilitarian judgments in collective condition). However, this deontological

boost is not expected to change individual judgments (i.e., no difference between I1 and I2), compatible with normative conformity (NC). Middle: Conversely, the Stress Reduction hypothesis (SR) predicts a utilitarian boost in collective judgments (i.e., less utilitarian deontological judgments in collective condition), but similar to VS, this deontological boost is not expected to change the individual judgments (i.e., no difference between I1 and I2), compatible with normative conformity (NC). Right: Similar to SR, the Social Deliberation hypothesis (SD) predicts a utilitarian boost in collective judgments. However, unlike in SR, this boost is expected to be followed by private individual judgments after the discussion (i.e., I2), compatible with informative conformity (IC). Bottom, conformity (both normative and informational) predicts that the collective judgments should correlate with the majority of the first individual opinions (i.e., before the discussion).

3.2 Experiment 1.

3.2.1 Material and Method

Participants

Our sample size estimation was based on Myers and Kaplan (1976) c.f., pre-registration at https://osf.io/jmkx5/. We initially aimed for 12 groups with 5 participants. Due to a technical error, some collective responses were not recorded in groups 1 to 5 and not properly recorded in group 6. We, therefore, collected another 6 groups. Wherever possible, we have provided the analysis of all data together. We have also provided separate analyses in the Supplementary Material (Section **A**. Part 7.1, Table 6), showing no significant difference in the results when excluding the initial groups with missing data points, indicating that the results are not driven by missing values. One group's responses were removed because screening analysis prior to hypothesis testing proved their data to be gross outliers confirmed by the convergence of several different outlier detection methods (see Supplementary Material; Section **A**, Part 6). The final sample consisted of 73 participants (38 females; mean age 25.72, range: 18-56; SD=8.42) in 16 mixed-gender groups, each including 4 or 5 members. All participants were native Germans recruited via MELESSA (Munich Experimental Laboratory for Economic and Social Science). The study was approved by the School of Advanced Study, University of London Ethics committee (SASREC_1819-313).

Scenarios

To contrast utilitarian vs. deontological judgments, we adopted sacrificial moral scenarios used in previous research (Greene et al., 2001) in addition to our independently validated scenarios (see Supplementary Material, Section **D**). It is important to note that moral judgments' complexity is far from contrasting deontology vs. utilitarianism in hypothetical scenarios (Everett & Kahane, 2020; Kahane, 2015; Kahane et al., 2015). Utilitarianism and deontology depend more on

moral reasoning than on individual responses (Hennig & Hütter, 2020). Due to these limitations, we tried to overcome two important caveats which the sacrificial moral dilemmas suffer from:

1. Sacrificial moral dilemmas used in the previous literature (e.g., Greene et al., 2001) consist of utilitarian actions, but they miss utilitarian inactions (or omissions). Therefore, any tendency towards utilitarianism in these scenarios could be alternatively interpreted as a preference for action rather than inaction (Crone & Laham, 2017).

2. Sacrificial dilemmas operationalize utilitarianism exclusively by approving to 'kill' someone to prevent other people's death, which can be far from real-life situations, jeopardizing the external validity (Bauman et al., 2014; Schein, 2020) and the essence of utilitarian philosophy (Everett & Kahane, 2020; Kahane, 2015; Kahane et al., 2015).

Therefore, in addition to the sacrificial scenarios conventionally used in the literature, we constructed and validated our vignettes which **a.** included cases of omissions that maximized the greater good for many, **b.** were related to real-life situations, **c.** no direct instrumental harm such as killing was needed for utility maximization and **d.** the protagonist (or his/her family) would not generally benefit from the outcome of the utilitarian decision.

We used these scenarios to examine the role of discussion in the change of mind by answering three questions 1. Whether the collective judgment would differ from an average of individual judgments and, if so, in which direction 2. Whether discussion would change the individual moral judgments, and 3. To disentangle different mechanisms at play by taking into account the First individual private judgments before the discussion, the second individual judgments after the discussion, and the differences between these two. We used 8 scenarios adopted from Greene et al. (2001) in addition to our 8 scenarios (independently validated, see Supplementary Material, Section **A**, Part 4). All 16 items were translated to German. The translation was done by two German native speakers, back (double) translated by AI Assistance for Language ("DeepL") GmbH (Cologne, Germany). Later, the texts were also checked by Munich Experimental Laboratory for Economic and Social Sciences.

Each scenario included: (1) A situation: a protagonist was in a situation that required her decision. She had two mutually exclusive options **a**. to act according to a moral duty or an accepted norm although they are against a higher utility **b.** to break a norm to maximize the utility for the greater number (2) The decision: the protagonist always decided in favor of the utilitarian option at the expense of violating a norm (3) An outcome: the vignette confirmed that the outcome took place, with the expected greater utility.

Utilitarian Score: participants had to rate the moral acceptability of the protagonist's decision on the scale of 1 (not acceptable at all) to 7 (totally acceptable). As all decisions in the scenarios were utilitarian, we refer to this rating

as 'utilitarian score'. The 'utilitarian score' is only an experimental label, and the fact that these decisions were utilitarian (or not) was never made explicit to participants.

3.2.2 Procedure

The procedure and the design of this study were adopted from Myers & Kaplan (1976). Participants were invited to the lab in groups of 4-5 in separate sessions. Once the participants were seated, general instructions were presented to them orally in English. Then, each participant received a tablet (Surface Pro). The experiment was programmed in Testable (https://testable.org). On the tablet, the instructions were repeated in German and followed by one practice round. Throughout the experiment, a timer on top of the screen always showed the remaining time for each scenario's ratings. For a demo in German, see here.

The experiment started with each participant going through the 16 scenarios individually (First individual judgment; see Figure 2). The order of the scenarios was randomized across participants. Participants had 90 seconds to read each scenario privately and judge the moral acceptability of the action (or inaction) described in the scenario. When the last participant rated the last scenario, the first part of the experiment was completed.

Figure 2. Design of the experiment. Participants read each scenario alone and responded individually to all items (1st individual judgments), then were asked to discuss half of the items in groups to provide a collective judgment (collective judgments) and finally responded again to all of the items alone (2nd individual judgments).

Next, in the collective judgment stage (see Figure 2), the participants were asked to read 8 of the 16 scenarios (pseudo-randomly selected, see Supplementary Material: Section **A**, Part 5) with other group members in the room. Each scenario was projected on a screen for everyone to see. Participants were given 3 minutes to discuss each scenario and arrive at a collective judgment. They all had to enter their

collective agreement on their tablets. Discussed vs. Non-discussed scenarios were randomized and counterbalanced across groups. All discussions were conducted in German (i.e., the participant's native language).

Finally, each participant went through all 16 scenarios privately for 30 seconds in the Second individual judgment stage.

We used a repeated measure design with three levels for our main independent variable (Condition: First individual, Collective, Second individual). When examining the individual judgments, we had one additional factor (Type: Discussed, Undiscussed). The dependent variable (Utilitarian Score) was a Likert scale ranging from 1 (Not acceptable at all) to 7 (Totally acceptable), and the middle point was defined as 'morally neutral' in the instructions.

Data availability statement: experiment's pre-registration, materials, data, and analysis can be accessed at https://osf.io/jmkx5/

3.2.3 Result

Statistical analysis was conducted using R programming language (https://www.r-project.org/), employing generalized ordinal mixed-effects models appropriate for our design's hierarchical structure. Since our dependent variable was in the Likert scale, we employed ordinal mixed-effect models using the package 'ordinal' designed for ordinal logistic mixed models (Christensen, 2019). In addition, we adopt a Bayesian approach in the mixed-effect model using brms package in R (Bürkner, 2018).

We began by examining how collective and individual judgments differ (see the hypotheses illustrated schematically in Figure 1 top row). An ordinal logistic mixed-effect model was employed with one fixed factor of Condition (Collective, First individual, Second individual) while accounting for random effects of items (different scenarios) as well as participants within groups (as a random nested effect; see Supplementary Material Section **A** - Part 7.1 - Table 3, for model comparisons). Collective judgments were significantly more utilitarian than individual judgments (z = 3.688, b = .3689, SE = .100, p = .0002) both when comparing the Collective judgments to the First (z=3.688, b = .3689, SE = .1, p = .0007) as well as the Second individual judgment (z=3.657, b = .3630, SE = .0993, p = .0007), both in Action and Inaction scenarios (see Supplementary Material, Section **A**, Figure S2 and Table 8). In addition to this frequentist approach, a Bayesian mixed-effect model was performed which showed similar results (BF_{01}= .0005, b = .35, SE = .09, CI_{95}= [.18, .52]; see Supplementary Material, Section **A**, Figure S3.a and Table 12). For individual responses, we included all items, both discussed and undiscussed. However, a separate analysis only for the discussed items also confirmed the same results and is provided in the Supplementary Material (Section **A**, Part 7.1, Table 7).

Having demonstrated more utilitarian collective judgments conclusively, we then proceeded to examine if individual judgments before and after the discussion were different. In a second model, we excluded the collective judgments from the analysis and only considered the individual conditions (First individual, Second individual) and type (Discussed, Undiscussed) as fixed factors. Random factors were similar to the previous model described earlier. We did not observe any difference between individual judgments before and after the discussion in the previous model (z = .079, b = .006, p = .996, Figure 3a). This model, in addition, showed that there is no difference between the discussed and the undiscussed items in the first and the second individual condition (see Supplementary Material, Section **A**, Part 7.1, Table 10). In addition to this frequentist approach, a Bayesian mixed model as well showed the similar results (BF_{01} = 15.57, b = .01, SE = .06, CI_{95}= [-0.13, 0.12]; see Supplementary Material, Section **A**, Figure S3.b and Table 12).

Together, these results supported the Stress Reduction hypothesis (compare Figure 1 to Figure 3a). On the one hand, significantly more utilitarian Collective judgments (vs. the First condition) rejected the Virtue Signaling hypothesis. On the other hand, the significant reduction of utilitarian judgments in the Second condition (vs. Collective) and the lack of any difference between First and Second were inconsistent with the Social Deliberation hypothesis.

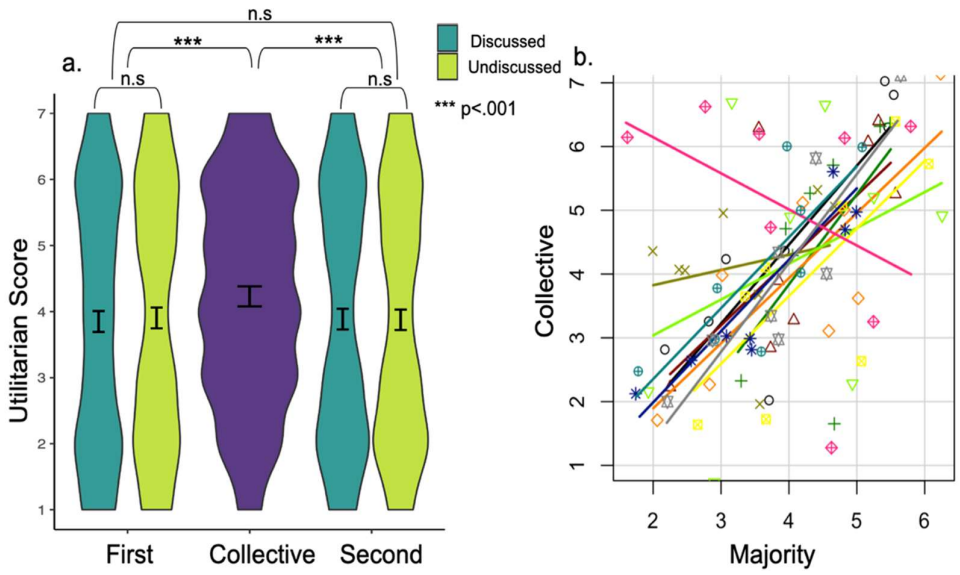


Figure 3 a. Distribution of responses over conditions. Collective judgments (in purple) are more utilitarian (higher utilitarian score) than individual judgments in the first and the second condition, but there is no difference between individual judgments before and after the discussion for discussed (dark green) and undiscussed (light green) items. Error bars: 95% confidence interval. **b.** There is a correlation between collective and majority (Discussed items before the discussion) at the level of

groups. To illustrate the effect observed in the mixed-model results, the correlation between different items within each group is separated and presented with different colors and symbols for each group (excluding the groups with missing data points). Data points with the same color and symbol refer to the same group. Lines are least-squares fits to the data within each group. Each group has 8 data points corresponding to the 8 discussed questions

We then proceeded to test our conformity hypotheses by first noting that the lack of a difference between First and Second (described above) was already more supportive of the normative (vs. informational) conformity hypothesis. We then set out to examine the effect of the group majority opinion on the collective judgment (Figure 1 bottom row). Note that the majority opinion was not experimentally manipulated or controlled in our study. However, given that we had elicited private individual opinions before discussions (First judgments), we could assess the group majority by averaging group members' individual judgments for each item before the discussion and examine the correlation between this metric and collective judgment to test our conformity hypothesis (Figure 1, bottom row). We describe conformity here as the relation between First judgments (before the discussion) and Collective judgments (in the ordinal logistic model; while controlling for the random effects of items and participants within each group). Consistent with our first conformity hypothesis, individual ratings prior to the discussion within each group were significantly correlated with the collective judgments (z = 2.987, b = .2920, SE = .1016, p = .004; see Figure 3b). See Supplementary Material, Section **A**, Part 8, Table 13, for details of the mixed-effect model.

We also examined the same hypothesis using a more intuitive definition of the majority vote. We first categorized the individual judgments by comparing them to a criterion based on the mean of the ratings of all First individual judgments over our entire sample across all to-be-discussed scenarios (Mean = 3.994) and classified each opinion as Utilitarian-inclined and Deontologically-inclined before the discussion. We then counted these opinions within each group for each to-be-discussed scenario and categorized each group as majority utilitarian or majority deontological for that item. In a new mixed-effect model with collective judgment as the dependent variable, we used this classification as a fixed factor (Majority) with two levels (Utilitarian, Deontological) while controlling for random effects of groups and items. We observed a significant main effect of majority (z = 4.145, b=1.62, SE = 0.392, p < .0001; See Supplementary Material, Section **A**, Part 8 for details and extra analysis).

3.2.4 Discussion

In Experiment 1, we discovered that groups, in comparison to individuals, are more utilitarian in their moral judgments. Importantly, this utilitarian boost was only observed at the collective level and not when participants rated the same questions privately later again. If the collective utilitarian boost was the result of deliberation and reasoning or due to conscious application of utilitarian principles

with genuine concern for the greater good, we expected that the effect would remain in the second private judgment. Consequently, our findings are inconsistent with the Social Deliberation (SD) and Virtue-Signaling (VS) hypotheses and in favor of the Stress Reduction (SR) hypothesis (Compare Figure3a to Figure 1).

3.3 Experiment 2

In order to examine the Stress Reduction (SR) hypothesis more directly, we performed Experiment 2, in which, in addition to moral judgments, we measured the stress level of each participant in each condition after responding to moral dilemmas.

3.3.1 Material and Method

Participants

The target sample size estimation was predetermined using a Monte Carlo simulation via the SIMR package (Green & MacLeod, 2016) to have 90% power (fixed factor effect size of Collective condition: 0.296). The final sample consisted of 70 participants (33 females, age: M = 25 years, SD = 4.9, range: 19 to 58) in 15 mixed-gender groups, each including 4 or 5 members (one group had 3 members). Due to the internet connection issue of two participants, two groups were excluded. All participants were native Germans recruited via MELESSA (Munich Experimental Laboratory for Economic and Social Science). The study was approved by the School of Advanced Study, University of London Ethics committee (SASREC_1819-313).

Scenarios

8 scenarios were chosen from Experiment 1 (Sacrificial Dilemmas and Real-life Dilemmas) based on their effect sizes. Utilitarian Scores were measured as in Experiment 1.

3.3.2 Procedure

The procedure and the design of this study were similar to Experiment 1, albeit performed online via Zoom (*Zoom.us*) due to the COVID19 pandemic. Participants were invited to separate Zoom sessions in small groups, their names were removed and replaced with experimental ID (given by the experimenter). General instructions were presented to them orally and written in English via screen share. Then, each participant received the internet link of the experiment via chat. The experiment has been programmed through the online software ('Qualtrics') (for a demo, see here; password: CMD2021)

After filling out a consent form, participants were requested to shut their phones off, close all of the windows on their computers (except Zoom), disable all notifications, mute their sound on Zoom, and start video sharing while hiding their own videos to not be seen by themselves (by turning the 'Self-view' function off and use 'Gallery View' afterward). This was done to exclude confounding factors such as the 'mirror' effect on moral dilemmas (see Reynolds et al., 2019).

Participants were then presented with the instruction in German via the questionnaire and asked to read and respond to 8 fully randomized moral scenarios chosen. This stage was followed by 10 questions measuring participants' affective states by the short version of the State-Trait Anxiety Inventory (STAI) validated for the German language (Jürgen, 2009). At this point, the First Individual stage was finished. The experimenter unmuted the participants and presented the same 8 scenarios – one at a time - via Screen Share in Zoom. In Collective condition, the order of presentation of the scenarios was fully randomized across groups via the 'randomizeR' (Uschner et al., 2018) package in R. In this stage, after reading each scenario, participants were asked to discuss the scenario with other participants at the Zoom session to reach a consensus. After this short discussion, they were asked to enter their consensus judgment in the questionnaire individually. Once all collective scenarios were discussed, participants answered the 10-item STAI individually once again. The experimenter muted all participants and instructed them to go through all 8 scenarios privately in the Second individual judgment stage. Similarly, after responding to the scenarios individually for the second time, STAI were presented to measure their stress level for one last time. The time limit of each condition was identical to Experiment 1.

To exclude other confounding factors in social settings, two sets of questions were asked at the end of the survey. The first set was employed to assess whether participants had a self-presentation strategy in mind in the collective stage. In this part, participants' metaethical perceptions were directly measured to evaluate how they thought they had *been seen by others* in their groups during the discussions. We measured the *warmth* and *competence* index using items adapted from Fiske et al. (2002) and Rom & Conway (2018). We measured coming across as logical, competent, intelligent, confident, and as a good leader during group discussions to measure *competence* and as warm, moral, good, tolerant, and trustworthy to measure *warmth* (items were translated to German; Questionnaires in the Supplementary Material Section **E**)

The second set assessed if participants felt socially connected to other members during the discussions. Previous research showed that feeling socially connected to others increases the utilitarian tendency in dyads. For instance, one study showed that participants endorsed utilitarian resolutions more often if they were more socially connected (Lucas & Livingston, 2014). In order to examine this possibility, adopted from Lucas & Livingston (2014), feeling socially connected, loneliness, and feeling of being accepted by others during discussions were also measured.

Two items of Affective Regret adopted from Goldstein-Greenwood et al. (2020) were also measured for exploratory analysis.

Similar to Experiment 1, we used a repeated measure design with three levels for the main independent variable (Condition: First individual, Collective, Second individual). The dependent variable (Utilitarian Score) was identical to Experiment 1. Data availability statement: experiment's pre-registration, materials, data, and analysis can be accessed at https://osf.io/jmkx5/

3.3.3 Result

Statistical analysis was conducted using R programming language (https://www.r-project.org/), employing generalized ordinal mixed-effects models appropriate for our design's hierarchical structure. Since our dependent variable was in the Likert scale, similar to Experiment 1, we employed ordinal mixed-effect models using the package 'ordinal' designed for ordinal logistic mixed models (Christensen, 2019). In addition, we adopt a Bayesian mixed-effect model using brms package in R (Bürkner, 2018).

We began by examining how collective and individual judgments differ. An ordinal logistic and a Bayesian mixed-effect model were employed with one fixed factor of Condition (Collective, First individual, Second individual) while accounting for random effects of items (different scenarios) as well as participants within groups (as a nested random effect). Similar to experiment 1, Collective judgments were significantly more utilitarian than individual judgments ($z = 2.437$, $b = .263$, $SE = .108$, $p = .0148$). Pairwise comparison tests corrected the p-value for multiple comparison showed that this difference was significant in the First judgments vs. Collective ($M_{Collective} = 4.11 > M_{First\ individual} = 3.87$; $z=3.688$, $b = .2689$, $SE = .1$, $p = .035$; $BF_{01} = 1.94$, $CI_{95}= [0.425, 0.041]$). Collective judgments were also more utilitarian that Second individual judgments ($M_{Collective} = 4.11 > M_{Second\ individual} = 4.01$) but this difference did not reach the statistical significance ($z=3.657$, $b = .3630$, $SE = .0993$, $p = .4$; $BF_{01} = 0.141$, $CI_{95}= [-0.085, 0.301]$; Power: 62.00%) (See Figure 4.a; c.f. General Discussion for Online vs. Face-To-Face interaction; Supplementary Material Section **C**). As in Experiment 1, our observation in Experiment 2 was consistent with the Stress Reduction hypothesis (Compare Figure 4.a to Figure 1, SR hypothesis).

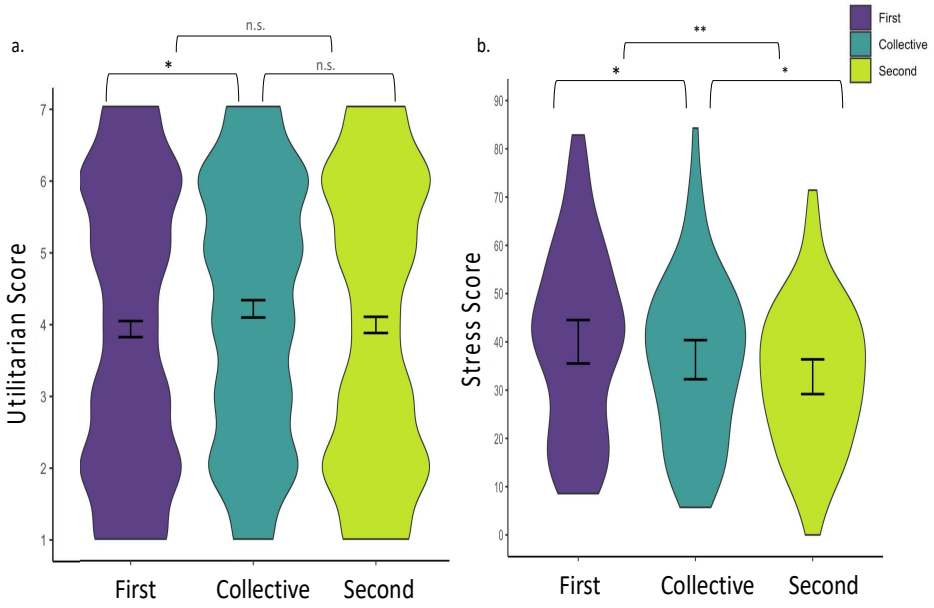


Figure 4 a. Distribution of responses to moral dilemmas over conditions. Collective judgments (in dark green) are more utilitarian (higher utilitarian score) than individual judgments in the first condition, but there is no difference between individual judgments before (in purple) and after the discussion for discussed (in light green). **b.** Participants reported less stress after Collective conditions (in dark green) than individual judgments in the first condition (in purple), but more than the second individual condition (in light green). Error bars: 95% confidence interval.

Having replicated the results of experiment 1 and demonstrated more utilitarian collective (vs. individual) judgments, we then proceeded to examine if individual judgments before and after the discussion were different. In a second model, we excluded the collective judgments from the analysis and only included the individual conditions (First individual, Second individual). Random factors were similar to the previous model described earlier. As in Experiment 1, we did not observe any difference between individual judgments before and after the discussion (z = 1.27, b = .139, p = .2, Figure 4a), another observation consistent with SR. In addition to this frequentist approach, a Bayesian mixed model showed the similar results (BF_{01} = 0.219, b = .12, SE = 0.10, CI_{95}= [-0.314, 0.072]). As in Experiment 1, the lack of difference between the First and Second conditions was inconsistent with the Social Deliberation hypothesis (Compare Figure 4.a to Figure 1, Social Deliberation hypothesis to see the difference) and in line with SR and VS hypotheses.

Next, to test the Stress Reduction (SR) hypothesis directly, we performed a linear mixed-effect model (package lme4; Bates et al., 2015). We examined the Stress measure across different conditions in individuals within groups with a fixed factor of Condition (Collective, First individual, Second individual) while accounting for random effects of participants within groups (as a random nested effect). As

expected, the Stress level in Collective condition was significantly lower than first individual judgments (pairwise comparison tests corrected for p-value (z = 2.6, b = 3.71, SE = .108, p = .027) (See figure 4b). The Stress level was even less in the Second individual condition vs. Collective condition (z = 2.47, b = 3.53, SE = .108, p = .038), probably due to a carry-over effect.

To examine the alternative explanations, we measured the correlation between ratings of warmth and competence scale. There was no significant correlation between these ratings and utilitarian scores at the group level using both linear regression (corrected for multiple comparisons; see Supplementary Material, Section **B**, Part 15) and Bayesian regression (Smartness: CI= [-0.902, 1.583], BF_{10} = .749; Reasonable: CI= [-0.902, 1.583]; BF_{10} = .696, Confident: CI= [-1.302, 0.332], BF_{10} = .696, Leader-like: CI= [-0.902, 1.583]; BF_{10} = .572, Competent: CI= [-1.568, 1.128], BF_{10} = .729). This result ruled out the possibility of *competence* signaling as an alternative explanation of our result (c.f. General Discussion).

To examine the effect of social connection, we measured the correlation between feeling lonely, accepted, and socially connected with the utilitarian score at the group level. In contrast to previous findings (Lucas & Livingston, 2014), we did not find evidence showing that feeling socially connected to others is related to more utilitarian scores at the group level using both logistic mixed effect model (corrected for multiple comparisons; see Supplementary Material, Section **B**, Part 15) as well as Bayesian analysis (Lonely: CI= [-0.50, 0.86], BF_{10} = .476; Accepted: CI= [-0.65, 0.98] , BF_{10} = ..441; Connected: CI= [-1.88, 0.11] , BF_{10} = .2.88) suggesting that the collective utilitarian boost has not driven by feeling socially connected. Together, these results supported the Stress Reduction hypothesis for the First vs. Collective conditions.

3.3.4 Discussion

In Experiment 2, with a sample size aimed at 90% power, we replicated the result of Experiment 1. Groups, in comparison to individuals, found utilitarian decisions more morally acceptable. Similar to Experiment 1, this utilitarian boost was short-lived and only observed at the collective level. We did not see this boost when participants rated the same questions privately later again. This was in line with the Stress Reduction hypothesis. To test this hypothesis more directly, we also measured the state anxiety level in each phase of the experiment. Consistent with the SR, the anxiety level was significantly reduced in the Collective phase compared to the First individual phase. Measuring the metaethical evaluations of participants as well as social connection excluded alternative explanations for this collective utilitarian boost. Hence, the findings of Experiment 2 were consistent with SR.

3.4 General Discussion

Moral judgments are ubiquitously social affairs that take place in the public sphere and are often shaped by the consensus in this sphere. Cognitive and psychological studies of moral judgment have so far predominantly focused on individuals and examined moral choices made privately. Across two experiments (N=143), we compared collective and individual moral judgments and examined the impact of face-to-face (Experiment 1) and Online (Experiment 2) interactions on these judgments. Participants, individually and collectively, examined actions (or inactions) of fictional characters that led to a higher utility for the greatest number at the expense of disregarding a moral norm in moral scenarios. We found that collective judgments are more utilitarian than individual judgments.

Utilitarian judgments are considered as the outcome of the inductive reasoning process; therefore, they entail a level of deliberation and abstraction that common-sense morality lacks. It has been argued that utilitarian decisions that necessitate certain norm violations (e.g., in moral dilemmas) require deliberation and reasoning to support utilitarian principles that entail greater happiness for the greater number of people (e.g., Greene et al., 2004). However, norm violation is emotionally aversive (as introduced above). Therefore, utilitarian choices often depend not just on cognitively costly deliberation but also on *overcoming negative aversive emotions*. Social context (e.g., group discussion) could change either of these factors, and our hypotheses (Figure 1) outlined three ways this could occur.

Our analysis revealed that groups, in comparison to individuals, are more utilitarian in their moral judgments. Thus, our findings are inconsistent with Virtue-Signaling (VS), which proposed the opposite effect. Crucially, the collective utilitarian boost was short-lived: it was only seen at the collective level and not when participants rated the same questions individually again. Previous research shows that moral change at the individual level, as the result of social deliberation, is rather long-lived and not transient (e.g., see Ueshima et al., 2021). Thus, this collective utilitarian boost could not have resulted from deliberation and reasoning or due to conscious application of utilitarian principles with authentic reasons to maximize the total good. If this was the case, the effect would have persisted in the second individual judgment as well. That was not what we observed. Consequently, our findings are inconsistent with the Social Deliberation (SD) hypotheses. Our observation was consistent with Stress Reduction (SR), which proposed a *transient* utilitarian boost. In the second experiment, measuring the stress level to test SR more directly, in addition to replicating the previous result, we observed that the stress level was significantly lower in collective conditions compared to the first individual conditions (see Figure 1; Figure 4), which was in line with the Stress Reduction (SR) hypothesis. This observation partially confirmed our SR hypothesis.

We showed that groups are more utilitarian than individuals. We proposed and tested the reduction of negative emotions (i.e., stress) as the most plausible mechanism for this boost. Other possibilities, however, were also examined. For

instance, one study showed that participants endorsed utilitarian resolutions more often in dyads if they were more socially connected (Lucas & Livingston, 2014). Since in our experiment the discussion might have made groups more socially connected and therefore increased collective utilitarian responses, we measured the feeling of being socially connected. We found no relationship between social connection and utilitarian score, discarding this alternative explanation.

Another alternative explanation for our result, the opposite of the 'virtue signaling', was also ruled out. This effect refers to the possibility that people may present themselves especially logical and intelligent in a group setting. For instance, Rom & Conway found that utilitarian decision-makers were viewed as especially logical, competent, deliberative, intelligent, and leader-like (Rom & Conway, 2018). People viewed scientists as more utilitarian than others (Sosa & Rios, 2019). Rom and colleagues moreover found that people were meta perceptively accurate gauging how others would view them upon making a utilitarian decision, and found that people select utilitarian decision-makers for important social roles such as running a hospital. Since it was equally plausible that people self-present strategically as more utilitarian than they privately prefer - to present themselves as logical and competent and leader-like and/or scientific - we measured the warmth and competence scale, following Rom and Conway (2018). However, no significant relation was observed between these meta-ethical perceptions and utilitarian judgments.

At another level of explanation, we tested the predictions of social conformity theory for the role of social context in moral judgments. At the group level, the within-group average of private utilitarian scores before discussion (as well as the majority private opinion) were positively correlated with the collective score. However, neither within-group average nor the majority of the first private opinion was correlated with the second private judgments. The combination of these two sets of results was more consistent with our normative but not informational conformity hypothesis. Importantly, we took care to set hypotheses such that they clearly separate conformity from the interaction. Conformity was defined as the majority influence which flows, unidirectionally, from the majority to the individual. The interaction effect, however, was defined as the bidirectional impact of group and individual on one another.

Some previous research shows that people in groups lie more, are less likely to offer help in urgent situations, are less generous, and more often engage in free riding and social loafing (c.f. 1. introduction). In this view, individuals are more egocentric in groups and more focused on maximizing their own utility. However, in our scenarios, the greater good for others never benefited the protagonist, let alone the participant. Therefore, the collective boost in utilitarian judgment cannot be explained by more *egoism* in the social context: the higher utility of norm violation in moral dilemmas had no utility consequences for the participant.

Our results could be interpreted as a sign of a transient shift in moral values when people make moral judgments *in groups*, creating a temporary (but actual) moment

of consensus during collective deliberation. Some previous works in social psychology (Heider, 1946) and neuroeconomics (Izuma & Adolphs, 2013) have also shown that social context can shift economic value preferences *transiently*. On the other hand, while noting that moral judgments are far from the economic value preferences examined in these studies, this collective utilitarian boost could be alternatively interpreted as a *compromise* rather than a genuine *consensus*, showing that the participants, individually, were not persuaded by the more utilitarian judgments that they rated collectively. We look forward to future works to examine this issue in depth.

Our attitude towards violating a norm consists of what we feel about the act as well as the consequences of the violation, e.g., guilt, blame, regret, empathic concern for the victim (Cushman et al., 2012; Miller et al., 2014; Reynolds & Conway, 2018). Negative emotions towards norm violation could therefore be aversive reactions to the act (McDonald et al., 2017) or its consequences (Miller & Cushman, 2013). The social context could trigger either of these mechanisms by diffusing responsibility related to norm violation (El Zein et al., 2019; Li et al., 2010) and thereby give rise to our observation that collectives were more utilitarian than individuals. We note here that our experimental design cannot distinguish between aversion to the norm-violating utilitarian *act* or its *consequences*. However, since we measured stress levels, we offer positive evidence that reduction of stress can contribute to this effect. We found that the stress was even more reduced in the second individual condition, probably due to a carry-over effect, which, by design, was inevitable. The question of why the reduction of stress in the second individual stage did not affect moral judgments needs further investigation.

We chose to employ a diverse range of moral scenarios, including different norms – some sacrificial and others not – to get close to the real-life heterogeneity of moral issues. To our knowledge, for the first time, we employed moral scenarios that included 'inactions' (e.g., staying silent) as well as actions that all led to a utilitarian outcome. Our participants were, correspondingly, richly heterogeneous. Contrary to the prescriptions of normative moral theories, participants showed flexibility in applying one or the other principle. This is an interesting and useful corollary of our study that previous researchers had predicted but had not been tested (Kahane, 2015). This variability of responses was evident in the distribution of scores across items (see Supplementary Material, Section **A**, Part 7). The diversity of responses across different norms calls into question the generalizability of the previous findings that operationalized utilitarianism *merely by accepting harm* which saves many in sacrificial dilemmas involving extremely unlikely (if not bizarre) conditions. Despite the heterogeneity of our scenarios and participants' responses to them, an item-based analysis confirmed that our key findings in Figure 3 are robust across items (see Supplementary Material, Section **A**, Part 7.2). Our novel stimuli could be used in future moral psychological researches to disentangle better the effect of different norms in moral judgments with greater ecological validity.

Comparing Experiment 1 (Face to Face) and Experiment 2 (Online) reveals a remarkably high correspondence between *private* responses in the two experiments. The shape of the distributions was very similar, indicating a very high level of replication in the private responses (see Supplementary Material Section **C**, Figure S11). Interestingly, the distribution of the consensus responses showed very different distributions between the Online and Face-to-Face versions of the experiment. This observation is in line with studies that have recently showed the differences between lab and online experiments comprising collective discussions (e.g., Tomprou et al. 2021; Hietanen et al., 2020; Schneider et al., 2002). Schneider et al., 2002 for instance, showed that online group discussions entail shorter comments (sometimes just a few words of agreement) compared to face-to-face interaction. Hence, online discussions are *uniform.* Similarly, we performed several exploratory analyses to examine if the variance of the consensus responses were different across the two experiments. Multiple measures of squared rank test of homogeneity revealed a significant difference in variance between online vs. face-to-face moral judgments scores: face-to-face (but not online) interaction promoted *more diverse consensus opinions*. No difference in *variance* was observed in the First or Second *private* responses between the two studies (Supplementary Material, Section **C**, Part 16). The effect size of collective vs. individual conditions in the online experiment was also smaller than Face to Face interaction, showing a smaller utilitarian boost in online interactions. More research is needed to understand the difference between online vs. in-person interactions in the moral domain.

Given the extended limitations imposed on laboratory work by the conditions of the global pandemic and the fact that many real-life meetings and social activities have been replaced by online video conferencing, we thought this would be a good opportunity to contribute to this very timely issue comparing the results of Experiments 1 and 2 in more detail, in addition to comparing the variance. A number of studies have examined the possible differences between face-to-face and in-person interactions and web-based video conferencing in various domains (e.g., Tomprou et al. 2021; Hietanen et al., 2020; Schneider et al., 2002). Therefore, we provided a number of further exploratory analyses that we have found instructive (Needless to say, these findings were not part of the original hypothesis) – See Supplementary Material, Section **C**.

Our data speak to a number of previous studies that examined the role of group synergy in moral decision-making. The synergy here refers to the possibility that interaction between group members during deliberation may result in a consensus that - rather than convergence to an opinion *within* the range privately held by the members - exceeds the maximally utilitarian opinion of the group. One reviewer's helpful advice, we examined this hypothesis in Experiment 1, we observed that the consensus score (M= 4.25) was lower than the highest individual utilitarian score within groups (M=4.76). Within Subject ANOVA revealed that this difference is significant ($t_{student(15)}$, p = .002; d_{cohen}=.96; see Supplementary Material, Section **D**, Part 19, Figure S13). A similar result was observed in Experiment 2: consensus

score (M= 4.11) was lower than the highest individual utilitarian score within groups (M=4.78). Within Subject ANOVA revealed that this difference is significant ($t_{student}$*(14), p* = .002; d_{cohen}=1.01; see Supplementary Material, Section **D**, Part 19, Figure S14). Interestingly, this pattern of results is consistent with Curşeu et al. (2013; 2020) and Meslec & Curşeu (2013).

In our experiment, participants had limited time for the discussion (i.e., three minutes) and responding to the questions. It is possible that by increasing the discussion time (or its frequency), we would have seen a different pattern in the second individual moral judgments. This is an important limitation of empirical approaches in assessing interactions in groups' judgments in the laboratory setting (for a critical review of such methods, see Weiten & Diamond, 1979). Being aware of this limitation, in two pilot groups, we had observed that participants needed around 90 seconds to read one dilemma (for the first time) and respond to it privately in a self-paced manner without the time pressure. In these pilot studies, we also observed that participants spent, on average, 3 minutes discussing each dilemma before moving on. In the final private judgment, as the participants had already seen the dilemmas and were going through them for a second or third time, self-paced responses were much quicker and did not take more than 30 seconds.

Another critical limitation was that our behavioral measure, conventionally, connected utilitarian and deontological judgment inextricably to one another: being less deontological or more utilitarian could not be distinguished from one another in moral dilemmas. Therefore, whether reduction of stress affected *utilitarian* responses or change the *deontological* tendencies is unclear, as utilitarian and deontological tendencies are often measured in the opposite directions of one single measure. Recent methods such as process dissociation (Conway & Gawronski, 2013) or computational models such as CNI (Consequence, Norm, Inaction; Gawronski et al. 2017) can be used in future studies to achieve this distinction. Future research that employs methodologies such as Natural Language Processing techniques to analyze the content of the arguments could also bring about a deeper understanding of social interaction in moral judgments.

To conclude, we found that collective consensual judgments made via face-to-face and online group interactions were more utilitarian compared to private individual judgments. Group discussion did not change the individual judgments, indicating a normative conformity effect whereby individuals consented to a group judgment that they did not necessarily buy into personally. We measured stress levels and showed that participants registered less state anxiety in solving moral dilemmas in groups than individually. The results were consistent with the hypothesis that interactions reduce aversive emotions (e.g., stress) associated with violation of moral norms, leading to more utilitarian judgments.

References

Arendt, H. (1987). Collective Responsibility. In S. J. J. W. Bernauer (Ed.), *Amor Mundi* (pp. 43–50). Springer Netherlands. https://doi.org/10.1007/978-94-009-3565-5_3

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70. https://doi.org/10.1037/h0093718

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). Together, slowly but surely: The role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 3–8. https://doi.org/10.1037/a0025708

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2016). Optimally interacting minds. *Discovering the Social Mind: Selected Works of Christopher D. Frith*, *329*(5995), 234–243. https://doi.org/10.4324/9781315630502

Bang, D., & Frith, C. D. (2017). Making better decisions in groups. *Royal Society Open Science*, *4*(8), 170193. https://doi.org/10.1098/rsos.170193

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, *121*(1), 154–161. https://doi.org/10.1016/j.cognition.2011.05.010

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*(9), 536–554. https://doi.org/10.1111/spc3.12131

Bell, D. E. (1982). Regret in Decision Making under Uncertainty. *Operations Research*, *30*(5), 961–981. http://www.jstor.org/stable/170353

Bell, D. E. (1985). Disappointment in Decision Making Under Uncertainty. *Operations Research*, *33*(1), 1–27. https://doi.org/10.1287/opre.33.1.1

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin*, *88*(1), 1–45. https://doi.org/10.1037/0033-2909.88.1.1

Bloom, P. (2010). How do morals change? *Nature*, *464*(7288), 490. https://doi.org/10.1038/464490a

Bornstein, G., & Yaniv, I. (1998). Individual and Group Behavior in the Ultimatum Game: Are Groups More "Rational" Players?. *Experimental Economics* 1**,** 101–108 https://doi.org/10.1023/A:1009914001822

Bostyn, D. H., & Roets, A. (2017). An Asymmetric Moral Conformity Effect: Subjects Conform to Deontological But Not Consequentialist Majorities. *Social Psychological and Personality Science*, *8*(3), 323–330. https://doi.org/10.1177/1948550616671999

Botti, S., Orfali, K., & Iyengar, S. S. (2009). Tragic choices: Autonomy and emotional responses to medical decisions. *Journal of Consumer Research*, *36*(3), 337–352. https://doi.org/10.1086/598969

Byrd, N., & Conway, P. (2019). Not all who ponder count costs: Arithmetic reflection predicts utilitarian tendencies, but logical reflection predicts both deontological and utilitarian tendencies. *Cognition*, *192*. https://doi.org/10.1016/j.cognition.2019.06.007

Capraro, V., Sippel, J., Zhao, B., Hornischer, L., Savary, M., Terzopoulou, Z., Faucher, P., & Griffioen, S. F. (2018). People making deontological judgments in the Trapdoor dilemma are perceived to be more prosocial in economic games than they actually are. *PLoS ONE*, *13*(10), e0205066. https://doi.org/10.1371/journal.pone.0205066

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*(JUL). https://doi.org/10.3389/fpsyg.2014.00607

Christensen, R. H. B. (2019). "ordinal—Regression Models for Ordinal Data." R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

Cialdini, R. B., & Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology*, *55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Ciaramelli, E., Muccioli, M., Làdavas, E., & Di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *2*(2), 84–92. https://doi.org/10.1093/scan/nsm001

Conrads, J., Ellenberger, M., Irlenbusch, B., Ohms, E. N., Rilke, R. M., & Walkowitz, G. (2016). Team goal incentives and individual lying behavior. *Research Papers in Economics.*

Conrads, J., Irlenbusch, B., Rilke, R. M., & Walkowitz, G. (2013). Lying and team incentives. *Journal of Economic Psychology*, *34*, 1–7. https://doi.org/10.1016/j.joep.2012.10.011

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology*, *104*(2), 216–235. https://doi.org/10.1037/a0031021

Cornwell, J. F. M., Jago, C. P., & Higgins, E. T. (2019). When Group Influence Is More or Less Likely: The Case of Moral Judgments. *Basic and Applied Social Psychology*, *41*(6), 386–395. https://doi.org/10.1080/01973533.2019.1666394

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(40), 17433–17438. https://doi.org/10.1073/pnas.1009396107

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. https://doi.org/10.1016/j.paid.2016.09.022

Cummins, D. D., & Cummins, R. C. (2012). Emotion and deliberative reasoning in moral judgment. *Frontiers in Psychology*, *3*(SEP), 328. https://doi.org/10.3389/fpsyg.2012.00328

Curşeu, P. L., Fodor, O. C., A. Pavelea, A., & Meslec, N. (2020). "Me" versus "We" in moral dilemmas: Group composition and social influence effects on group utilitarianism. *Business Ethics*, *29*(4), 810–823. https://doi.org/10.1111/beer.12292

Curşeu P. L., Jansen, R. J. G., & Chappin M. M. H. (2013) Decision Rules and Group Rationality: Cognitive Gain or Standstill? PLoS ONE 8(2): e56454. https://doi.org/10.1371/journal.pone.0056454

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion*, *12*(1), 2–7. https://doi.org/10.1037/a0025071

Cushman, F., & Greene, J. D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience*, *7*(3), 269–279. https://doi.org/10.1080/17470919.2011.614000

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. Journal of Personality and Social Psychology, 8(4), 377–383. https://doi.org/10.1037/h0025589

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology, 51*(3), 629–636. https://doi.org/10.1037/h0046408

Dunbar, R. I. M. (2004). Gossip in evolutionary perspective. *Review of General Psychology*, *8*(2), 100–110. https://doi.org/10.1037/1089-2680.8.2.100

El Zein, M., & Bahrami, B. (2020). Joining a group diverts regret and responsibility away from the individual. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1922), 20192251. https://doi.org/10.1098/rspb.2019.2251

El Zein, M., Bahrami, B., & Hertwig, R. (2019). Shared responsibility in collective decisions. *Nature Human Behaviour*, *3*(6), 554–559. https://doi.org/10.1038/s41562-019-0596-4

El Zein, M., Seikus, C., De-Wit, L., & Bahrami, B. (2020). Punishing the individual or the group for norm violation. *Wellcome Open Research*, *4*, 139. https://doi.org/10.12688/wellcomeopenres.15474.2

Ellemers, N. (2017). *Morality and the Regulation of Social Behavior: Groups as Moral Anchors (1st ed.)*. Routledge. https://doi.org/10.4324/9781315661322

Ellemers, N., Van Der Toorn, J., Paunov, Y., & Van Leeuwen, T. (2019). The Psychology of Morality: A Review and Analysis of Empirical Studies Published From 1940 Through 2017. *Personality and Social Psychology Review*, *23*(4), 332–366. https://doi.org/10.1177/1088868318811759

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Everett, J. A. C., & Kahane, G. (2020). Switching Tracks? Towards a Multidimensional Model of Utilitarian Psychology. *Trends in Cognitive Sciences*, *24*(2), 124–134. https://doi.org/10.1016/j.tics.2019.11.012

Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of Trustworthiness From Intuitive Moral Judgments. *Journal of Experimental Psychology: General*, *145*(6), 772–787. https://doi.org/10.1037/xge0000165

Fedyk, M. (2019). *The social turn in moral psychology*. The MIT Press.

Feinberg, M., Willer, R., Antonenko, O., & John, O. P. (2012). Liberating Reason From the Passions: Overriding Intuitionist Moral Judgments Through Emotion Reappraisal. *Psychological Science*, *23*(7), 788–795. https://doi.org/10.1177/0956797611434747

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, *58*(2), 203–210. https://doi.org/10.1037/h0041593

Fochmann, M., Fochmann, N., Kocher, M. G., Müller, N., & Wolf, N. (2021). Dishonesty and risk-taking: Compliance decisions of individuals and groups. *Journal of Economic Behavior and Organization*, *185*, 250–286. https://doi.org/10.1016/j.jebo.2021.02.018

Foot, P. (1967). The Problem of Abortion and the Doctrine of the Double Effect. *Oxford Review*, *5*, 19–32. https://doi.org/10.1093/0199252866.003.0002

Forsyth, D. R., Zyzniewski, L. E., & Giammanco, C. A. (2002). Responsibility diffusion in cooperative collectives. *Personality and Social Psychology Bulletin*, *28*(1), 54–65. https://doi.org/10.1177/0146167202281005

Frey, F. E., & Tropp, L. R. (2006). Being seen as individuals versus as group members: Extending research on metaperception to intergroup contexts. *Personality and Social Psychology Review*, *10*(3), 265–280. https://doi.org/10.1207/s15327957pspr1003_5

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science*, *23*(8), 931–939. https://doi.org/10.1177/0956797612436816

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*(3), 343–376. https://doi.org/10.1037/pspa0000086

Gert, B. (2005). *Morality: Its Nature and Justification*. Oxford University Press. https://doi.org/10.1093/0195176898.001.0001

Giorgetta, C., Zeelenberg, M., Ferlazzo, F., & D'Olimpio, F. (2012). Cultural variation in the role of responsibility in regret and disappointment: The Italian case. *Journal of Economic Psychology*, *33*(4), 726–737. https://doi.org/10.1016/j.joep.2012.02.003

Goldstein-Greenwood, J., Conway, P., Summerville, A., & Johnson, B. N. (2020). (How) Do You Regret Killing One to Save Five? Affective and Cognitive Regret Differ After Utilitarian and Deontological Decisions. *Personality and Social Psychology Bulletin*, *46*(9), 1303–1317. https://doi.org/10.1177/0146167219897662

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, *11*(8), 322–323. https://doi.org/10.1016/j.tics.2007.06.004

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*(3), 1144–1154. https://doi.org/10.1016/j.cognition.2007.11.004

Greene, J. D., Nystrom, L. E., Engell, D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*, 389–400. https://doi.org/10.1016/J.NEURON.2004.09.027

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*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Grimm, J. (Hg.) (2009): State-Trait-Anxiety Inventory nach Spielberger. Deutsche Lang- und Kurzversion.- Methodenforum der Universität Wien https://empcom.univie.ac.at/fileadmin/user_upload/p_empcom/pdfs/Grimm2009_StateTraitAngst_MFWorkPaper2009-02.pdf

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*(5827), 998–1002. https://doi.org/10.1126/science.1137651

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*(10), 1387–1397. https://doi.org/10.1177/0956797617720944

Heider, F. (1946). Attitudes and Cognitive Organization. *Journal of Psychology: Interdisciplinary and Applied*, *21*(1), 107–112. https://doi.org/10.1080/00223980.1946.9917275

Hietanen, J. O., Peltola, M. J., & Hietanen, J. K. (2020). Psychophysiological responses to eye contact in a live interaction and in video call. Psychophysiology, 57(6) https://doi.org/10.1111/PSYP.13587

Hennig, M., & Hütter, M. (2020). Revisiting the divide between deontology and utilitarianism in moral dilemma judgment: A multinomial modeling approach. *Journal of Personality and Social Psychology*, *118*(1), 22–56. https://doi.org/10.1037/pspa0000173

Heuzé, J., & Brunel, P. C. (2003). Social loafing in a competitive context. *International Journal of Sport and Exercise Psychology*, *1*(3), 246–263. https://doi.org/10.1080/1612197x.2003.9671717

Izuma, K., & Adolphs, R. (2013). Social manipulation of preference in the human brain. *Neuron*, *78*(3), 563–573. https://doi.org/10.1016/j.neuron.2013.03.023

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*, *10*(5), 551–560. https://doi.org/10.1080/17470919.2015.1023400

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. https://doi.org/10.1016/j.cognition.2014.10.005

Kant, I. (1948). Moral Law: Groundwork of the Metaphysics of Morals (1st ed.). Routledge. https://doi.org/10.4324/9780203981948

Kaplan, M. F., & Miller, L. E. (1978). Reducing the effects of juror bias. *Journal of Personality and Social Psychology*, *36*(12), 1443–1455. https://doi.org/10.1037/0022-3514.36.12.1443

Karpowitz, C. F., & Mendelberg, T. (2007). Groups and deliberation. *Swiss Political Science Review*, *13*(4), 645–662. https://doi.org/10.1002/j.1662-6370.2007.tb00092.x

Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The "Trier social stress test" - A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, *28*(1–2), 76–81. https://doi.org/10.1159/000119004

Kocher, M. G., Schudy, S., & Spantig, L. (2018). I lie? We lie! Why? Experimental evidence on a dishonesty shift in groups. *Management Science*, *64*(9), 3995–4008. https://doi.org/10.1287/mnsc.2017.2800

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*, *7*(6), 708–714. https://doi.org/10.1093/scan/nsr048

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*(7138), 908–911. https://doi.org/10.1038/nature05631

Kohlberg, L. (1973). The Claim to Moral Adequacy of a Highest Stage of Moral Judgment. *The Journal of Philosophy*, *70*(18), 630. https://doi.org/10.2307/2025030

Kundu, P., & Cummins, D. D. (2013). Morality and conformity: The Asch paradigm applied to moral decisions. *Social Influence*, *8*(4), 268–279. https://doi.org/10.1080/15534510.2012.727767

Kvaran, T., Nichols, S., & Sanfey, A. (2013). The effect of analytic and experiential modes of thought on moral judgment. *Progress in Brain Research*, *202*, 187–196. https://doi.org/10.1016/B978-0-444-62604-2.00011-3

Latane, B., Williams, K., Harkins, S., Diener, E., Har-Vey, J., Kerr, N., Kidd, R., Levinger, G., Ostrom, T., Petty, R., & Wheeler, L. (1979). Many Hands Make Light the Work: The Causes and Consequences of Social Loafing. *Journal of Personality and Social Psychology*, *37*(6), 822–832. https://doi.org/10.1037/0022-3514.37.6.822

Leach, C. W., Bilali, R., & Pagliaro, S. (2015). *Groups and morality.* In M. Mikulincer, P. R. Shaver, J. F. Dovidio, & J. A. Simpson (Eds.), *Group processes* (p. 123–149). American Psychological Association. https://doi.org/10.1037/14342-005

Lee, J. J., & Gino, F. (2015). Poker-faced morality: Concealing emotions leads to utilitarian decision making. *Organizational Behavior and Human Decision Processes*, *126*, 49–64. https://doi.org/10.1016/j.obhdp.2014.10.006

Lee, M., Sul, S., & Kim, H. (2018). Social observation increases deontological judgments in moral dilemmas. *Evolution and Human Behavior*, *39*(6), 611–621. https://doi.org/10.1016/j.evolhumbehav.2018.06.004

Li, P., Jia, S., Feng, T., Liu, Q., Suo, T., & Li, H. (2010). The influence of the diffusion of responsibility effect on outcome evaluations: Electrophysiological evidence from an ERP study. *NeuroImage*, *52*(4), 1727–1733. https://doi.org/10.1016/j.neuroimage.2010.04.275

Li, Z., Gao, L., Zhao, X., & Li, B. (2019). Deconfounding the effects of acute stress on abstract moral dilemma judgment. *Current Psychology*, 1–14. https://doi.org/10.1007/s12144-019-00453-0

Li, Z., Xia, S., Wu, X., & Chen, Z. (2018). Analytical thinking style leads to more utilitarian moral judgments: An exploration with a process-dissociation approach. *Personality and Individual Differences*, *131*, 180–184. https://doi.org/10.1016/j.paid.2018.04.046

Loomes, G., & Sugden, R. (1982). Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *The Economic Journal*, *92*(368), 805. https://doi.org/10.2307/2232669

Lucas, B. J., & Livingston, R. W. (2014). Feeling socially connected increases utilitarian choices in moral dilemmas. *Journal of Experimental Social Psychology*, *53*(53), 1–4. https://doi.org/10.1016/j.jesp.2014.01.011

Mallon, R., & Nichols, S. (2011). Dual Processes and Moral Rules. *Emotion Review*, *3*(3), 284–285. https://doi.org/10.1177/1754073911402376

Mason, W., & Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(3), 764–769. https://doi.org/10.1073/pnas.1110069108

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*(6), 770–778. https://doi.org/10.1016/j.evolhumbehav.2017.06.001

Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). An investigation of moral judgement in frontotemporal dementia. *Cognitive and Behavioral Neurology*, *18*(4), 193–197. https://doi.org/10.1097/01.wnn.0000191292.17964.bb

Mercier, H. (2011). What good is moral reasoning. *Mind & Society, 10*(2), 131–148. https://doi.org/10.1007/s11299-011-0085-6

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968ï

Meslec, N., & Curşeu, P. L. (2013). Too close or too far hurts: Cognitive distance and group cognitive synergy. *Small Group Research, 44*(5), 471–497. https://doi.org/10.1177/1046496413491988

Milgram, S. (1963). Behavioral Study of obedience. *Journal of Abnormal and Social Psychology*, *67*(4), 371–378. https://doi.org/10.1037/h0040525

Mill, J. S. (1863). *Utilitarianism* (R. Crisp (ed.)). Parker, Son and Bourn.

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*(10), 707–718. https://doi.org/10.1111/spc3.12066

Miller, R. M., Hannikainen, I. A., & Cushman, F. A. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion*, *14*(3), 573–587. https://doi.org/10.1037/a0035361

Moll, J., De Oliveira-Souza, R., & Zahn, R. (2008). *The Neural Basis of Moral Cognition*. *Annals of the New York Academy of Sciences*, *1124*(1), 161–180. https://doi.org/10.1196/annals.1440.005

Myers, D. G., & Bishop, G. D. (1970). Discussion effects on racial attitudes. *Science*, *169*(3947), 778–779. https://doi.org/10.1126/science.169.3947.778

Myers, D. G., & Kaplan, M. F. (1976). Group-Induced Polarization in Simulated Juries. *Personality and Social Psychology Bulletin*, *2*(1), 63–66. https://doi.org/10.1177/014616727600200114

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, *83*(4), 602–627. https://doi.org/10.1037/0033-2909.83.4.602

Nichols, M. L., & Day, V. E. (1982). A Comparison of Moral Reasoning of Groups and Individuals on the "Defining Issues Test." *Academy of Management Journal*, *25*(1), 201–208. https://doi.org/10.2307/256035

Niebuhr, R. (1932). *Moral Man and Immoral Society: A Study in Ethics and Politics*. Charles Scribner's Sons.

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., Silani, G., Cikara, M., & Cushman, F. (2020). Reasoning Supports Utilitarian Resolutions to Moral Dilemmas Across Diverse Measures. *Journal of Personality and Social Psychology*, *120*(2), 443–460. https://doi.org/10.1037/pspp0000281

Paxton, J. M., & Greene, J. D. (2010). Moral reasoning: hints and allegations. *Topics in Cognitive Science*, *2*(3), 511–527. https://doi.org/10.1111/J.1756-8765.2010.01096.X

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, *36*(1), 163–177. https://doi.org/10.1111/j.1551-6709.2011.01210.x

Paytas, T. (2014). Sometimes psychopaths get it right: A utilitarian response to "The Mismeasure of Morals." *Utilitas*, *26*(2), 178–191. https://doi.org/10.1017/S095382081400003X

Perkins, A. M., Leonard, A. M., Weaver, K., Dalton, J. A., Mehta, M. A., Kumari, V., Williams, S. C. R., & Ettinger, U. (2013). A Dose of ruthlessness: Interpersonal moral judgment is hardened by the anti-anxiety drug lorazepam. *Journal of Experimental Psychology: General*, *142*(3), 612–620. https://doi.org/10.1037/a0030256

Pletti, C., Lotto, L., Buodo, G., & Sarlo, M. (2017). It's immoral, but I'd do it! Psychopathy traits affect decision-making in sacrificial dilemmas and in everyday moral situations. *British Journal of Psychology*, *108*(2), 351–368. https://doi.org/10.1111/bjop.12205

Pletti, C., Lotto, L., Tasso, A., & Sarlo, M. (2016). Will I regret it? Anticipated negative emotions modulate choices in moral dilemmas. *Frontiers in Psychology*, *7*(DEC), 1918. https://doi.org/10.3389/fpsyg.2016.01918

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*(7), 1009–1023. https://doi.org/10.1037/emo0000413

Reynolds, C. J., Knighten, K. R., & Conway, P. (2019). Mirror, mirror, on the wall, who is deontological? Completing moral dilemmas in front of mirrors increases

deontological but not utilitarian response tendencies. *Cognition*, *192*. https://doi.org/10.1016/j.cognition.2019.06.005

Rom, S. C., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology*, *74*, 24–37. https://doi.org/10.1016/j.jesp.2017.08.003

Rosen, F. (2006). *Classical utilitarianism from Hume to Mill*. Routledge.

Sacco, D. F., Brown, M., Lustgraaf, C. J. N., & Hugenberg, K. (2017). The Adaptive Utility of Deontology: Deontological Moral Decision-Making Fosters Perceptions of Trust and Likeability. *Evolutionary Psychological Science*, *3*(2), 125–132. https://doi.org/10.1007/s40806-016-0080-6

Schein, C. (2020). The Importance of Context in Moral Judgments. *Perspectives on Psychological Science*, *15*(2), 207–215. https://doi.org/10.1177/1745691620904083

Schneider, S. J., Kerwin, J., Frechtling, J., & Vivari, B. A. (2002). Characteristics of the discussion in online and face-to-face focus groups. Social Science Computer Review, 20(1), 31–42. https://doi.org/10.1177/089443930202000104

Scruton, R. (2001). *Kant : A very short introduction*. Oxford University Press.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Smith, E. R., & Collins, E. C. (2009). Contextualizing Person Perception: Distributed Social Cognition. *Psychological Review*, *116*(2), 343–364. https://doi.org/10.1037/a0015072

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review*, *108*(1), 183–203. https://doi.org/10.1037/0033-295X.108.1.183

Sosa, N., & Rios, K. (2019). The utilitarian scientist: The humanization of scientists in moral dilemmas. Journal of Experimental Social Psychology, 84, Article 103818. https://doi.org/10.1016/j.jesp.2019.103818

Stanovich, K. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: is it time for a tri-process theory? In J. Evans & K. Frankish(Eds.), *In two minds: Dual processes and beyond* (pp. 55–88). Oxford University Press.

Starcke, K., Ludwig, A. C., & Brand, M. (2012). Anticipatory stress interferes with utilitarian moral judgment. *Judgment and Decision Making*, *7*(1), 61–68. https://doi.org/10.17185/duepublico/45052

Strohminger, N., Lewis, R. L., & Meyer, D. E. (2011). *Divergent effects of different positive emotions on moral judgment*. https://doi.org/10.1016/j.cognition.2010.12.012

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*(3), 454–458. https://doi.org/10.1016/j.cognition.2011.01.018

Szekely, R. D., & Miu, A. C. (2015). Incidental emotions in moral dilemmas: The influence of emotion regulation. *Cognition and Emotion*, *29*(1), 64–75. https://doi.org/10.1080/02699931.2014.895300

Tasso, A., Sarlo, M., & Lotto, L. (2017). Emotions associated with counterfactual comparisons drive decision-making in Footbridge-type moral dilemmas. *Motivation and Emotion*, *41*(3), 410–418. https://doi.org/10.1007/s11031-017-9607-9

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217. https://doi.org/10.5840/monist197659224

Timmons, S., & Byrne, R. M. J. (2019). Moral fatigue: The effects of cognitive fatigue on moral reasoning. *Quarterly Journal of Experimental Psychology*, *72*(4), 943–954. https://doi.org/10.1177/1747021818772045

Trémolière, B., & Bonnefon, J. F. (2014). Efficient kill-save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality and Social Psychology Bulletin*, *40*(7), 923–930. https://doi.org/10.1177/0146167214530436

Tomprou, M., Kim, Y. J., Chikersal, P., Woolley, A. W., & Dabbish, L. A. (2021). Speaking out of turn: How video conferencing reduces vocal synchrony and collective intelligence. PLOS ONE, 16(3). https://doi.org/10.1371/JOURNAL.PONE.0247655

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326–334. https://doi.org/10.1016/j.cognition.2012.10.005

Unamuno, M. (1954). *Tragic sense of life*. Dover Publications.

Uschner D, Schindler D, Hilgers R, Heussen N (2018). "randomizeR: An R Package for the Assessment and Implementation of Randomization in Clinical Trials." *Journal of Statistical Software*, **85**(8), 1–22. doi: 10.18637/jss.v085.i08.

Valdesolo, P., & Desteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*(6), 476–477. https://doi.org/10.1111/j.1467-9280.2006.01731.x

Wallach, M. A., & Kogan, N. (1965). The roles of information, discussion, and consensus in group risk taking. *Journal of Experimental Social Psychology*, *1*(1), 1–19. https://doi.org/10.1016/0022-1031(65)90034-X

Wallach, M. A., Kogan, N., & Bem, D. J. (1964). Diffusion of responsibility and level of risk taking in groups. *Journal of Abnormal and Social Psychology*, *68*(3), 263–274. https://doi.org/10.1037/h0042190

Weisel, O., & Shalvi, S. (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(34), 10651–10656. https://doi.org/10.1073/pnas.1423035112

Weiten, W., & Diamond, S. S. (1979). A critical review of the jury simulation paradigm - The case of defendant characteristics. Law and Human Behavior, 3(1), 71–93. https://doi.org/10.1007/BF01039149

Wiech, K., Kahane, G., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2013). Cold or calculating? Reduced activity in the subgenual cingulate cortex reflects decreased emotional aversion to harming in counterintuitive utilitarian judgment. *Cognition*, *126*(3), 364–372. https://doi.org/10.1016/j.cognition.2012.11.002

Wilson, D. S., & O'Gorman, R. (2003). Emotions and actions associated with norm-breaking events. *Human Nature*, *14*(3), 277–304. https://doi.org/10.1007/s12110-003-1007-z

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*(4), 491–498. https://doi.org/10.1016/j.psyneuen.2011.07.017

Zeelenberg, M. (1999). Anticipated regret, expected feedback, and behavioral decision making. *Journal of Behavioral Decision Making*, *12*(2), 93–106. https://doi.org/10.1002/(SICI)1099-0771(199906)12:2<93::AID-BDM311>3.0.CO;2-S

Zhang, L., Kong, M., & Li, Z. (2017). Emotion regulation difficulties and moral judgment in different domains: The mediation of emotional valence and arousal. *Personality and Individual Differences*, *109*, 56–60. https://doi.org/10.1016/j.paid.2016.12.049

Zhang, L., Kong, M., Li, Z., Zhao, X., & Gao, L. (2018). Chronic stress and moral decision-making: An exploration with the CNI model. *Frontiers in Psychology*, *9*(SEP), 1702. https://doi.org/10.3389/fpsyg.2018.01702

Zhao, J., Harris, M., & Vigo, R. (2016). Anxiety and moral judgment: The shared deontological tendency of the behavioral inhibition system and the unique utilitarian tendency of trait anxiety. *Personality and Individual Differences*, *95*, 29–33. https://doi.org/10.1016/j.paid.2016.02.024

3.5 Supplementary Material

3.5.1 A: Experiment1

1 Consent and data management

Each participant received detailed information on the conduct of the study and signed the informed consent letter. They were informed that their participation was voluntary, that they could stop at any time and without having to justify themselves. Participants knew that they would be able to access their own data and research results if they wish. All data (relative to participants' performance in

the computer-based task) were anonymized by using an arbitrary code (one code per participant). These codes did not contain any information to identify the participant (such as first name, surname, or date of birth). Participant identifiable information (name, surname, date of birth) was not collected.

2 Variables

We manipulated two independent variables: 1. Condition: First (Individually before the discussion), Collective, Second (Individually after the discussion), and 2. Mode (Discussed, Undiscussed) and measured the dependent variable (ratings related to the moral acceptability of each scenario) measured in the Likert scale 1 (Not at all acceptable) to 7 (absolutely acceptable).

3 Instruction

The following instructions were provided in a written form in German and English. There was no deception involved; the instructions reflected the whole process of the experiment.

This experiment has three stages. In the first one, you will read 16 moral scenarios. We are interested in your honest opinion about the moral acceptability of a given choice in these scenarios. For instance, you may be provided with a scenario about a doctor and his patient, asked, "Is it morally permissible for Dr. Herzog to lie to his patient?" Depending on whether you think the action is not morally right or perfectly right, you will need to enter your response on the scale provided after each scenario. You cannot go back to your previous responses. In the second phase, you will be asked to discuss some scenarios with your group. Here, after discussion, you will be asked to provide a collective judgment and enter it on the tablet provided for each of you. In the third phase, you will get the opportunity to read all the scenarios again and provide a judgment about them individually for the second time.

The written instructions were accompanied by a visual aid and further explanation of the experimenter to secure comprehension.

After general instructions were given, participants familiarized themselves with the task and the relevant moral stimuli by performing one practice round. Participants then performed the task of responding to moral scenarios on the tablet screen individually for 1.5 minutes for each question. A timer on top of the screen showed the time. Responses that took more time than 1.5 minutes were not excluded. Participants were then asked questions together (8 items). The experimenter presented one scenario on the screen using the projector in the room, asked them to discuss each question with their groups in order to reach a consensus within 3 minutes. Each participant submitted their collective answers in the tablets for each question at the same time. The participants collectively reached a consensus, and each and every one entered the collective response on their

respective devices. After the collective conditions, they responded again individually to each question for 0.5 minutes each. The instructions made it clear that participants had a second chance to express their individual opinion, which could deviate from consensus. In nearly all trials, our participants followed these instructions. Participants were then debriefed, compensated, and thanked for their participation.

4 Scenario Validation

Flesch-Reading-Ease-Score was calculated (adapted to the German language) for all scenarios using the online Flesch value calculator (https://fleschindex.de/). The final score across all the scenarios was 61, which indicates "Plain English" that is easily understood by 13- to 15-year-old students (Thomas et al., 1975).

The German names used in each scenario were carefully chosen based on the most common German names from the lists of the first 100 hits from 1970-1990, excluding the ones that might have negative connotations (e.g., Kevin or Jaqueline).

We validated the above scenarios in a separate pilot study by asking n = 50 participants (recruited via Academic Prolific; www.prolific.co) to measure whether the utilitarian responses were perceived as the choices which *maximized the utility for many* in a two-alternative forced-choice task. Twenty scenarios were given to each participant, and each participant answered which choice has the highest benefit for the many (for 20 scenarios).

One example is as follows:

A German investigation journalist has been violently murdered. Anna is in charge of the investigation, and there is mounting evidence that the murder was ordered by a foreign country, which is a long-time trade partner for Germany and with which the German state is about to conclude a large sale deal. The new sale deal would create 10,000 well-paid jobs in Germany over the next two years and take 10,000 people out of unemployment. The trade deal with the foreign country will be compromised if Anna processes the evidence that foreign power is involved.

The participants then were asked:

Which decision has a higher benefit for many people:

1. For Anna not to process the evidence that the foreign power is involved.

2. For Anna to process the evidence that the foreign power is involved.

One of the responses was always related to utilitarian inaction (here 1) and the other to non-utilitarian action (here 2). The order of the scenarios and choices was random across individuals.

We also asked the participants whether they needed other information to answer the questions. We modified the final version of our scenarios according to their responses. Finally, we chose the scenarios (8 out of 20) with significantly higher responses in choices of utilitarian inaction than a non-utilitarian (deontological) action across all participants (we set p-value ≤ 0.07 since the study was underpowered).

5 Randomization

Discussed items were pseudo-randomized across groups according to Table 1. Items across Group11 to Group 17 were the same as 1 to 5 (See Table1).

Table 1.
Randomization of discussed items

	Action.	*Inaction.*
Group1	1,2,3,4	9,10,11,12
Group2	2,3,4,5	10,11,12,13
Group3	3,4,5,6	11,12,13,14
Group4	4,5,6,7	12,13,14,15
Group5	5,6,7,8	13.14,15,16
Group6	6,7,8,1	14,15,16,9
Group7	7,8,1,2	15,16,9,10
Group8	8,1,2,3	16,9,10,11
Group9	1,3,5,7	2,4,6,8

6 Outlier detection

To perform tests of outlier detection, we used <outlier_function.r> from the **Performance** package (Lüdecke et al., 2021), which uses a composite measure combined of several distance and/or clustering methods (see Figure S1 and Table 2). Using a multivariate approach of outlier detection, we calculated the mean of moral judgments in First, Collective and Second conditions for both discussed and undiscussed items for each group. The composite measure detected one group as an outlier (see Figure S2).

Table 2
Distance score for the outlier group (i.e., gr

	Distance.
zscore	7.4193
iqr	0.6
mahalanobis	13.8526
robust	273.6640
mcd	405.8142
ics	13.8314
optics	3.7763
lof	1.4256

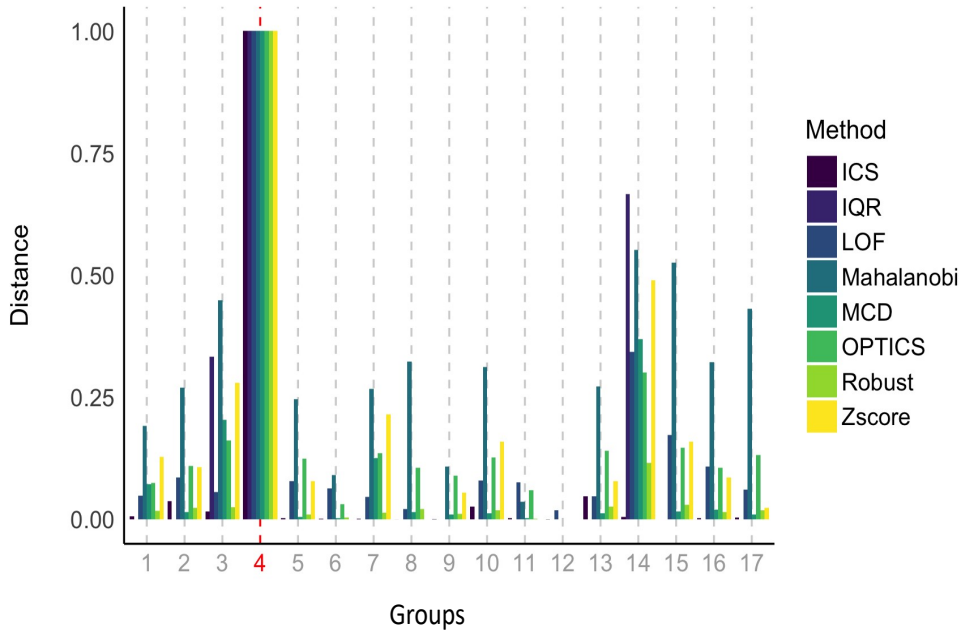


Figure Figure S1. Multiple methods of outlier detection detected group 4 (in red) as an outlier according to the mean of the individual and collective responses.

In addition to the composite method above, different univariate methods from the **outliers** R package (Komsta, 2011) were used. Grubbs test (recommended for small samples) (Grubbs et al., 1950) - G = 3.43389, U = 0.21697, p<.0001, in addition to Dixon test (Q = 0.75, p<.0001) showed the same group (i.e. group 4) as an outlier.

7 Mixed effect models

7.1 Logistic mixed effect model

We performed different ordinal logistic mixed-effect models by using the package **ordinal** (Christensen, 2019). In model 1, Items in model 2, groups and individuals, and in model 3 items, groups and individuals were considered as random factors of the model. Model comparisons favored model 3 (see Table 3). The pairwise comparison performed for model 3 is shown in table 5.

Table 3
Model comparison for different random slopes.

Row	***no.par***	***AIC***	***logLik***	***LR.stat***	***df***	***Pr.Chisq.***
model1	9	9799.4	4890.7			
model2	10	10351.8	5165.9	549.40	1	1.00
model3	11	9614.1	4796.1	739.65	1	**0.00**

Table 4
Logistic regression results for the winning model

	Rating		
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.07	0.04 – 0.13	**<0.001**
2|3	0.36	0.21 – 0.63	**<0.001**
3|4	0.87	0.50 – 1.52	0.616
4|5	1.48	0.85 – 2.59	0.169

5|6	3.70	2.11 – 6.48	**<0.001**
6|7	22.71	12.78 – 40.3	**<0.001**
		3	
Condition [Collective]	1.45	1.19 – 1.76	**<0.001**
Condition [Second]	1.01	0.87 – 1.16	0.936
Random Effects			
σ^2	3.29		
$\tau_{00\ id:\ group}$	0.41		
$\tau_{00\ group}$	0.00		
$\tau_{00\ Item}$	1.16		
N_{Item}	16		
N_{id}	73		
N_{group}	16		
σ^2	3.29		
Observations	2768		

Pairwise comparison between conditions for model 3 is presented in Table 5

Table 5
Pairwise comparison between conditions

contrast	***estimate***	***SE***	**z.ratio**	**p.value**
First - Collective	-0.3689	0.100	-3.686	**.0007**
First - Second	-0.006	0.074	-0.080	.9965
Collective - Second	0.3630	0.0993	3.657	**.0007**

P-value adjustment: Tukey method for comparing a family of 3 estimates

The same model was performed on a subset of the data, excluding the groups with missing data (see Table 6) and only on discussed items (see Table 7).

Table 6
Pairwise comparison between conditions (excluding groups with missing data points)

contrast	*estimate*	*SE*	z.ratio	p.value
First - Collective	-0.3577	0.1091	-3.278	**.003**
First - Second	-0.0373	0.0913	-0.410	.911
Collective - Second	0.3201	0.1081	2.962	**.0086**

P-value adjustment: Tukey method for comparing a family of 3 estimates

Table 7
Pairwise comparison between conditions only for discussed items

contrast	*estimate*	*SE*	z.ratio	p.value
First - Collective	-0.3688	0.114	-3.226	**.0036**
First - Second	-0.0512	0.106	-0.482	.8797
Collective - Second	0.3176	0.112	2.826	**.0131**

P-value adjustment: Tukey method for comparing a family of 3 estimates

In addition, we examined the effect of conditions on items which include actions (Sacrificial Dilemmas) and inactions (Real-life Dilemmas) (see Table 8 and Figure S2).

Table 8
Pairwise comparison between conditions for Action and Inaction items

contrast	estimate	SE	z.ratio	p.value
First Action - Collective Action	-0.3685	0.1000	-3.685	**.0031**
First Action - Second Action	0.0059	0.0747	0.080	.1
Collective Action - Second Action	0.3626	0.0993	3.653	**.0035**
First InAction - Collective InAction	-0.3685	0.1000	-3.685	**.0031**
First InAction - Second InAction	-0.0059	0.0747	-0.080	.9999
Collective InAction - Second InAction	0.3624832	0.0993	3.653	**.0035**

P-value adjustment: Tukey method for comparing a family of 3 estimates

Table 9

Logistic regression results of the interaction between item type and condition

	Rating		
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.06	0.03 – 0.12	**<0.001**
2|3	0.29	0.13 – 0.62	**0.001**
3|4	0.69	0.32 – 1.48	0.341
4|5	1.18	0.55 – 2.53	0.670
5|6	2.95	1.38 – 6.33	**0.005**
6|7	18.12	8.36 – 39.26	**<0.001**
Condition [Collective]	1.44	1.11 – 1.87	**0.005**
Condition [Second]	0.94	0.77 – 1.16	0.582
type [InAction]	0.63	0.22 – 1.82	0.395
Condition [Collective] * type [InAction]	1.00	0.68 – 1.47	0.992
Condition [Second] * type [InAction]	1.14	0.85 – 1.53	0.377
Random Effects			
σ^2	3.29		
$\tau_{00\ id:group}$	0.41		
$\tau_{00\ group}$	0.00		
$\tau_{00\ Item}$	1.12		
$N_{\ Item}$	16		
$N_{\ id}$	73		
$N_{\ group}$	16		
Observations	2768		

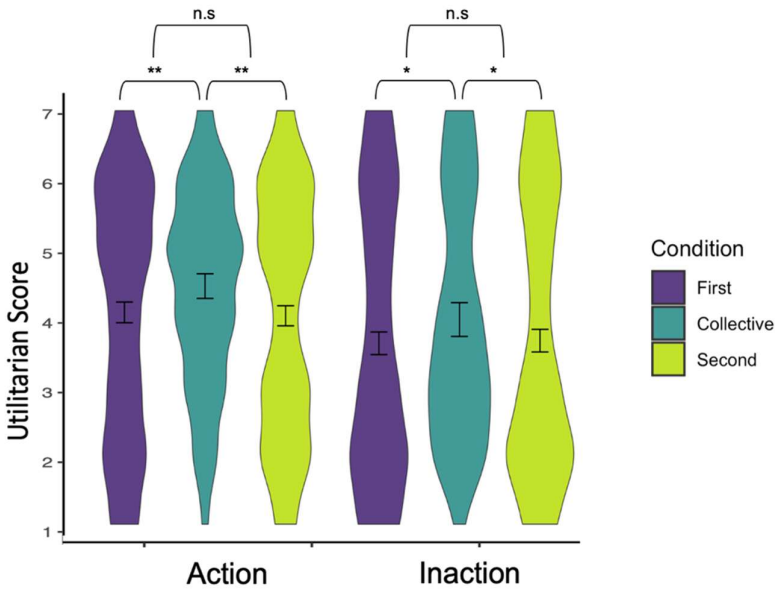


Figure S2. In both items, which involved utilitarian action and utilitarian inaction, the utilitarian score was higher in the collective condition than the first and second condition.

Finally, to compare the utilitarian score of Discussed and Undiscussed items before and after the discussion, we excluded collective judgments and only included First and Second conditions in a new model. In this model, we compared the change of utilitarian score as a function of the interaction between Condition (First vs. Second) and Mode (discussed vs. undiscussed). The result of the pairwise comparison of different conditions in this model is shown in Table 10.

Table 10

Pairwise comparison between discussed and undiscussed items in First and Second condition

contrast	*estimate*	*SE*	z.ratio	p.value
First Discussed - Second Discussed	0.0548	0.105	-0.521	0.9540
First Discussed - First Undiscussed	0.0208	0.107	-0.194	0.9974
First Discussed - Second Undiscussed	0.0117	0.106	0.109	0.9995
Second Discussed - First Undiscussed	0.0339	0.106	0.319	0.9888

Second Discussed - Second Undiscussed	0.0665	0.105	0.630	0.9223
First Undiscussed - Second Undiscussed	0.0326	0.105	0.309	0.9898

P-value adjustment: Tukey method for comparing a family of 3 estimates

7.2 Linear mixed models

Excess mass test (using **multimode** package; Ameijeiras-Alonso, 2021) confirmed that the distribution of responses across conditions is bimodal when we include both item and individual variability (excess mass = 0.05123, p-value <.0001; alternative hypothesis: true number of modes is greater than 1). In order to assure that this did not affect our result in the logistic model described in 7.1, we adopted an alternative analysis: instead of considering items as random effects, we averaged them across individuals in different conditions. The outcome distribution is no longer bimodal (p-value =.99; alternative hypothesis: true number of modes is greater than 1). In our new model, we considered groups and subjects as random factors and Conditions as fixed factors. Since the average rating for items is no longer on the Likert scale, we performed a linear mixed-effect model (package **lme4**; Bates et al., 2015) instead of an ordinal logistic mixed model. The result of this analysis (pairwise comparison) is shown in Table 11. Figure S3 shows that this model captures the distribution as well.

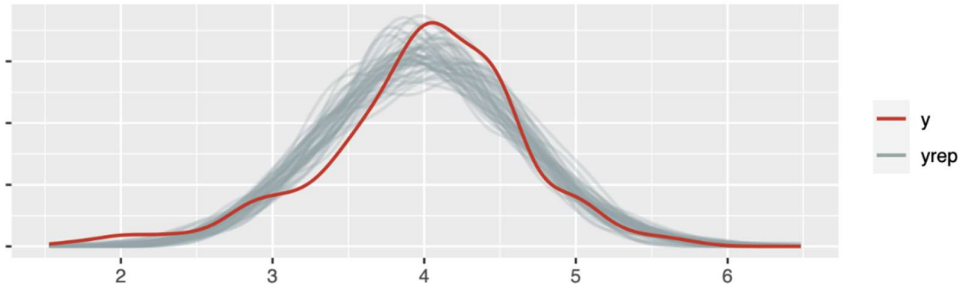


Figure S3. Different simulations with the properties extracted from the mixed effect model showed that the model (in grey) captured the distribution of actual data (in red).

Table 11
Pairwise comparison between conditions in the linear mixed model

contrast	*estimate*	*SE*	z.ratio	p.value
First - Collective	-0.4017123	0.0818339	-4.9088762	**.0000073**
First - Second	0.0037427	0.0818339	0.0457349	.9988475
Collective - Second	0.4054550	0.0818339	4.9546111	**.0000060**

df= 144. P-value adjustment: Tukey method for comparing a family of 3 estimates

7.3 Bayesian Mixed Model

In addition to the frequentist approach above, here, we also performed Bayesian mixed effect models. This is especially important since two of our hypotheses mentioned in the manuscript, SD and VS, were based on the null effect between the First and the Second judgments. We used **brms** package in R (Bürkner, 2018), with 5000 iterations, five chains, and weakly informative prior (model betas drawn from normal distribution; mean = 0 and SD =1).

Table12
Bayesian Mixed Effect Model Results

	Rating	
Predictors	*Estimates*	*CI (95%)*
Intercept	3.86	3.34 – 4.37
Condition: Collective	0.35	**0.18 – 0.53***
Condition: Second	-0.01	-0.13 – 0.12
Mode: Undiscussed	0.04	-0.16 – 0.25

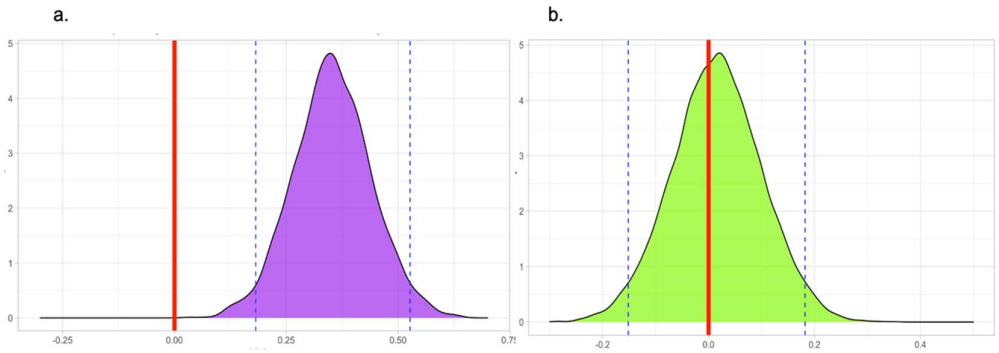


Figure S3. Posterior distribution of fixed factors of the Bayesian mixed effect model for a. collective condition (in purple) and b. the second condition (in green). The vertical red line crossed zero, and the blue dotted lines show the confidence intervals *(95%)*.

8 Conformity models

We performed different logistic mixed-effect models by using the package **ordinal** (Christensen, 2019). to see the relation between Collective judgments (JC) and individual judgments and while controlling for the random effects of items and participants within each group, summarized in Table 13.

Table13
Conformity Mixed Effect Model

		JC	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.01	0.00 – 0.03	**<0.001**
2|3	0.11	0.05 – 0.24	**<0.001**
3|4	0.44	0.22 – 0.90	**0.024**
4|5	1.35	0.67 – 2.74	0.401
5|6	4.09	2.01 – 8.42	**<0.001**
6|7	24.70	11.22 – 54.36	**<0.001**
J1D	1.34	1.10 – 1.63	**0.004**
Random Effects			
σ^2	3.29		
$\tau_{00\ id:group}$	0.00		
$\tau_{00\ group}$	1.42		
$\tau_{00\ Item}$	0.28		
N_{Item}	16		

N_{id}	49
N_{group}	11

In a new mixed-effect model with collective judgment as the dependent variable, we classified groups with utilitarian majority vs. deontological majority according to their first judgments. This classification is used as a fixed factor (Majority) with two levels (Utilitarian, Deontological) while controlling for random effects of groups and items. The result is shown in Table 14.

Table14
Majority Mixed Effect Model

		JC	
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.04	0.01 – 0.15	**<0.001**
2|3	0.35	0.19 – 0.64	**0.001**
3|4	0.38	0.21 – 0.69	**0.001**
4|5	0.41	0.23 – 0.74	**0.003**
5|6	1.15	0.67 – 1.97	0.623
6|7	1.21	0.70 – 2.07	0.497
Majority (U)	2.99	1.64 – 5.46	**<0.001**
Random Effects			
σ^2	3.29		
$\tau_{00\ group}$	0.00		
$\tau_{00\ Item}$	0.00		
N_{Item}	16		
N_{group}	16		

9 Item-based analysis

An item-based analysis was performed to compare the ratings for each item in different conditions. Repeated measure ANOVA test shows a significant difference between conditions across items. The result of the analysis is shown in Figure S4.

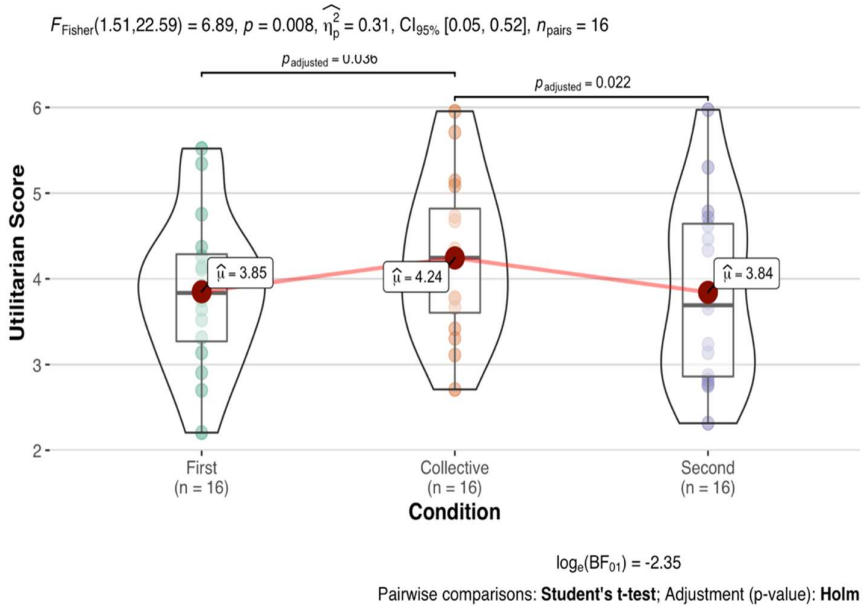


Figure S4. Across 16 items, the ANOVA test shows a significant difference between different conditions. Items are rated more utilitarian in Collective condition in comparison to First and Second condition. The statistical test result is shown in the figure using the **ggstatssplot** package in R (Patil, 2021).

10 ANOVA

The primary analysis performed was the Logistic Mixed-Effect Models (in addition to the Bayesian Mixed Effect Models). Mixed-effects Models are the more suitable tools for the nested designs (items within participants within groups) and another study has previously demonstrated that they can be usefully applied to data on moral dilemmas (Patil et al. 2020 c.f. Methodological Issues in the moral task). This approach is a more appropriate than simple ANOVAs because it takes into account the design structure to explain the sources of variance such as item and participants and groups. In addition, this method is better equipped to handle the missing data points, as well as modeling Likert-Scale measurements. However, the ANOVA test was also performed which shows a significant difference between the moral judgements in the three stages of Experiment 1, corrected for p-values adjustment using Holm test in multiple comparison of conditions (See Figure S5).

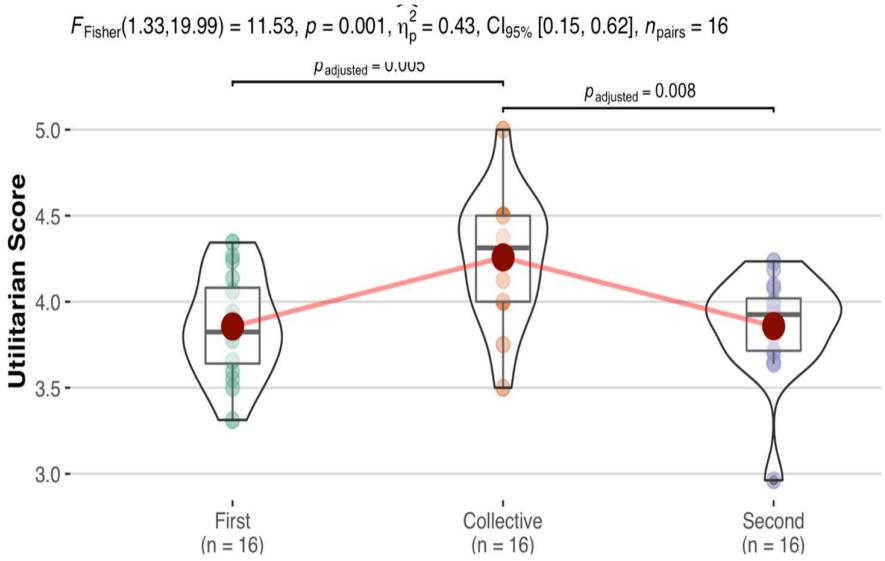


Figure S5. ANOVA test result for Experiment 1 - utilitarian scores (vertical axis) and different conditions (First – Collective – Second) across groups (each data point). Collective judgments are significantly higher than the first and the second conditions, using **ggstastplot** package in R (Patil, 2021).

3.5.2 B: Experiment2

11. Sample size estimation

Our sample size estimation for our first experiment was based on Myers and Kaplan (1976) c.f., pre-registration at https://osf.io/jmkx5/. We initially aimed for 12 groups with 5 participants. For the second experiment, we have reported the result of a replication study performed in Zoom software (n=70). The target sample size was predetermined using a Monte Carlo simulation, via the **SIMR** package (Green & MacLeod, 2016) to have 90% power using the parameters obtained via Experiment1's Mixed Effect Model (fixed factor effect size was set to 0.296). The final sample consisted of 70 participants (33 females, age: $M = 25$ years, $SD = 4.9$, range: 19 to 58) in 15 mixed-gender groups. (The simulation was informed by all the parameters obtained via Experiment1 Mixed Effect Model-not only the effect size). – see Figure S6 (outcome of **SMIR** package).

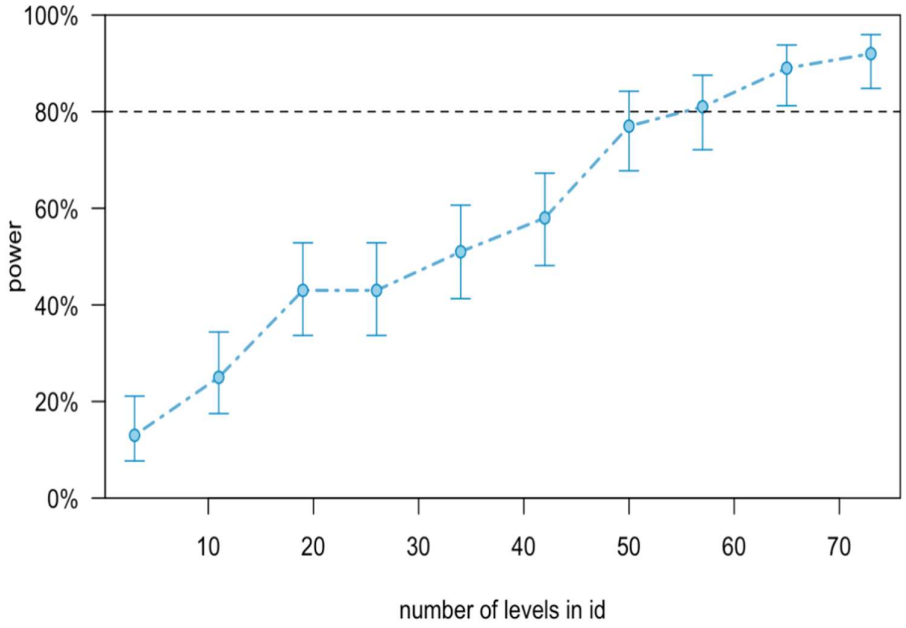


Figure S6. The target sample size was predetermined using a Monte Carlo simulation, using the **SIMR** package (Green & MacLeod, 2016) to have 90% power using the parameters obtained via Experiment1 Mixed Effect Model (fixed factor effect size: 0.296). The obtained effect size in experiment 2 is 0.2439782, and the power was 62.00% (51.75, 71.52)

12. Mixed effect models

12.1 Logistic mixed effect model

We performed ordinal logistic mixed-effect models by using the package **ordinal** (Christensen, 2019). In this model, items, groups, and individuals were considered as random factors of the model and condition as the fixed factors with 3 levels (First individual, Second Individual, Collective) – see table 15. The pairwise comparison performed for this model is shown in table 16.

Table 15
Logistic regression mixed effect model results.

	Rating		
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.09	0.05 – 0.19	**<0.001**
2|3	0.36	0.18 – 0.74	**0.006**
3|4	0.73	0.36 – 1.50	0.394
4|5	1.31	0.64 – 2.68	0.463
5|6	3.32	1.62 – 6.81	**0.001**
6|7	17.96	8.62 – 37.39	**<0.001**
Condition [Collective]	1.30	1.05 – 1.61	**0.015**
Condition [Second]	1.14	0.92 – 1.40	0.223
Random Effects			
σ^2	3.29		
$\tau_{00\ CODE}$	0.82		
$\tau_{00\ Item}$	0.95		
$\tau_{11\ CODE.Group}$	0.01		
$\rho_{01\ CODE}$	-0.71		
ICC	0.35		
N_{CODE}	70		
N_{Item}	8		
Observations	1680		

Pairwise comparison between conditions in the above model.

Table 16
Pairwise comparison between conditions

contrast	***estimate***	***SE***	**z.ratio**	**p.value**

First - Collective	-0.2637	0.1082	-2.436	**0.03**
First - Second	-0.1286	0.1055	-1.218	0.44
Collective - Second	0.1351	0.1074	1.256	0.41

P-value adjustment: Tukey method for comparing a family of 3 estimates

In addition, we examined the effect of conditions on items which include actions (Sacrificial Dilemmas) and inactions (Real-life Dilemmas) (see Table 17 and Figure S4).

Table 17

Pairwise comparison between conditions for Action and Inaction items

contrast	**estimate**	**SE**	**z.ratio**	**p.value**
First Action - Collective Action	0.7164	0.1510	4.744	**0.000**
First Action - Second Action	0.0214	0.149	0.142	0.999
Collective Action - Second Action	1.713	0.480	3.570	**0.004**
First InAction - Collective InAction	0.205	0.156	1.319	0.774
First InAction - Second InAction	-0.222	0.149	-1.48	0.677
Collective InAction - Second InAction	-0.428	0.153	-2.78	0.060

P-value adjustment: Tukey method for comparing a family of 3 estimate

Table 18
Logistic regression results of the interaction of item type and condition

	Rating		
Predictors	*Odds Ratios*	*CI*	*p*
1|2	0.21	0.11 – 0.43	**<0.001**
2|3	0.84	0.42 – 1.67	0.617
3|4	1.74	0.87 – 3.48	0.114
4|5	3.11	1.56 – 6.21	**0.001**
5|6	7.93	3.95 – 15.91	**<0.001**
6|7	43.01	21.04 – 87.92	**<0.001**
Condition [Collective]	2.05	1.52 – 2.75	**<0.001**
Condition [Second]	1.02	0.76 – 1.37	0.886
type [InAction]	5.55	2.17 – 14.22	**<0.001**
Condition [Collective] * type [InAction]	0.40	0.26 – 0.61	**<0.001**
Condition [Second] * type [InAction]	1.22	0.81 – 1.85	0.343
Random Effects			
σ^2	3.29		
$\tau_{00\ CODE}$	0.82		
$\tau_{00\ Item}$	0.41		
$\tau_{11\ CODE.Group}$	0.01		
$\rho_{01\ CODE}$	-0.70		
ICC	0.27		
N_{CODE}	70		
N_{Item}	8		

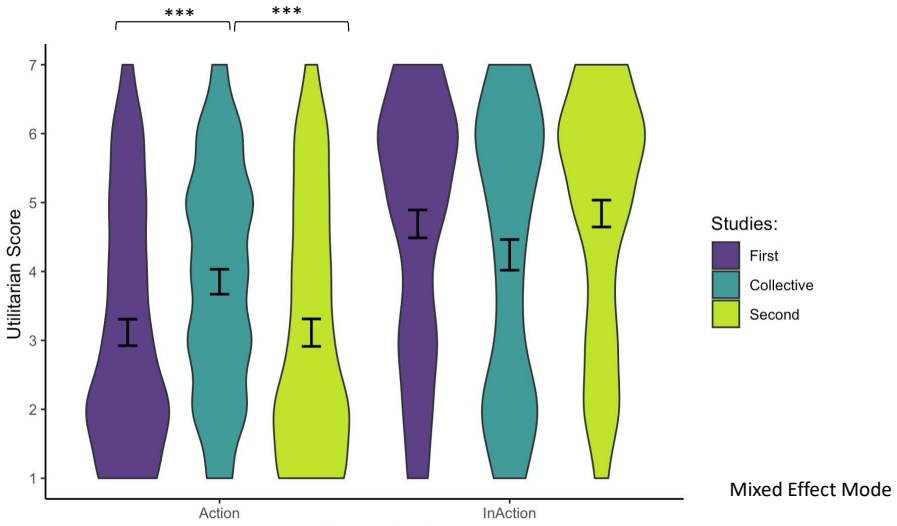


Figure S7. In Action items, which involved Sacrificial Dilemmas the utilitarian score was higher in the collective condition than the first and second condition.

13. Stress Models

To examine the Stress score across different conditions, we first measured the stress score then we performed linear mixed-effect models by using the package (package **lme4**; Bates et al., 2015). In this model, groups and individuals were considered as random factors of the model and condition as the fixed factors with 3 levels (First individual, Second individual, Collective) – see table 19. The pairwise comparison performed for this model is shown in table 20.

Table 19.
Logistic Mixed Effect Model of Stress Responses

		USTRESS	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	40.03	35.94 – 44.12	**<0.001**
Condition [Collective]	-3.71	-6.51 – -0.91	**0.009**
Condition [Second]	-7.24	-10.04 – -4.45	**<0.001**

Random Effects	
σ^2	71.41
$\tau_{00\ CODE:Group}$	215.48
$\tau_{00\ Group}$	3.87
ICC	0.75
N_{CODE}	70
N_{Group}	15
Observations	210
Marginal R^2 / Conditional R^2	0.029 / 0.762

Table 20.
Pairwise comparison between conditions in the linear mixed model of stress

contrast	***estimate***	***SE***	**z.ratio**	**p.value**
First - Collective	3.714286	1.428417	138	2.600282
First - Second	7.244898	1.428417	138	5.071978
Collective - Second	3.530612	1.428417	138	2.471696

14. Bayesian Mixed Model for Moral Judgments

In addition to the frequentist approach above, here, we also performed Bayesian mixed effect models. This is especially important since two of our hypotheses mentioned in the manuscript SR and VS, were based on the null effect between the First and the Second judgments. We used **brms** package in R (Bürkner, 2018), with 5000 iterations, 5 chains, and weakly informative prior (model betas drawn from normal distribution; mean = 0 and SD =1) The result of the model is shown in Table 21. Model detail is shown in Table 22.

Table 21.
Bayesian Mixed Model

	Rating	
Predictors	*Estimates*	*CI (95%)*
Intercept	3.91	3.04 – 4.77
Condition: Collective	**0.23**	**0.04 – 0.42**

Condition: Second	0.12	-0.07 – 0.31
Random Effects		
σ^2	2.65	
$\tau_{00\ Group}$	0.31	
$\tau_{00\ Group:CODE}$	0.16	
$\tau_{00\ Item}$	1.15	
ICC	0.38	
N_{CODE}	70	
N_{Group}	15	
N_{Item}	8	
Observations	1680	
Marginal R^2 / Conditional R^2	0.003 / 0.299	

Table22.
Bayesian Mixed Model

	Parameter	CI_low	CI_high	BF
2	First - Collective	-0.424	-0.040	1.8372
3	First - Second	-0.313	0.071	0.2181
1	Collective - Second	-0.087	0.300	0.1335

15. Added Measurements

In Experiment 2, we included a number of questions in the survey to directly examine the possibility that expressing more moral judgments could be motivated by 1. the intention to appear *competent* or *warm* and/or 2. feeling *socially* connected to others in group discussions.

Given that Rom and colleagues (2018) showed that people are meta-perceptively mindful of how others may view them upon making a moral judgment, we tested if people had metaethical access to such information. Therefore, we measured the Warmth and Competence scale (items were translated to German) using items from Rom & Conway 2018.

In addition, adopted from Lucas & Livingston (2014), feeling socially connected, loneliness, and feeling of being accepted by others during discussions

were also measured. We did not find any difference between these added measurements and group utilitarian score - using a Mixed Effect Model (Table 22; 23; 24) - a null result which was observed in the Bayesian mixed effect model. The correlation between each item and the utilitarian score is presented the Figures S8, S9, S10. Each dot represents one group.

15.1 Warmth Scale

15.1.1 WARMTH SCALE LINEAR REGRESSION

Table 22.
Linear regression of Warmth Scale

		USCORE	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.33	-3.37 – 10.02	0.290
warm	0.01	-1.41 – 1.42	0.993
good	-0.32	-2.00 – 1.35	0.673
moral	0.12	-1.20 – 1.43	0.845
trustworthy	-0.12	-2.23 – 2.00	0.904
tolerant	0.41	-0.92 – 1.74	0.506
Observations		15	
R^2 / R^2 adjusted		0.072 / -0.443	

15.1.2 WARMTH SCALE ITEM-BASED CORRELATIONS:

The correlation between each item and the utilitarian score presented in Figure S8. Each dot represents one group.

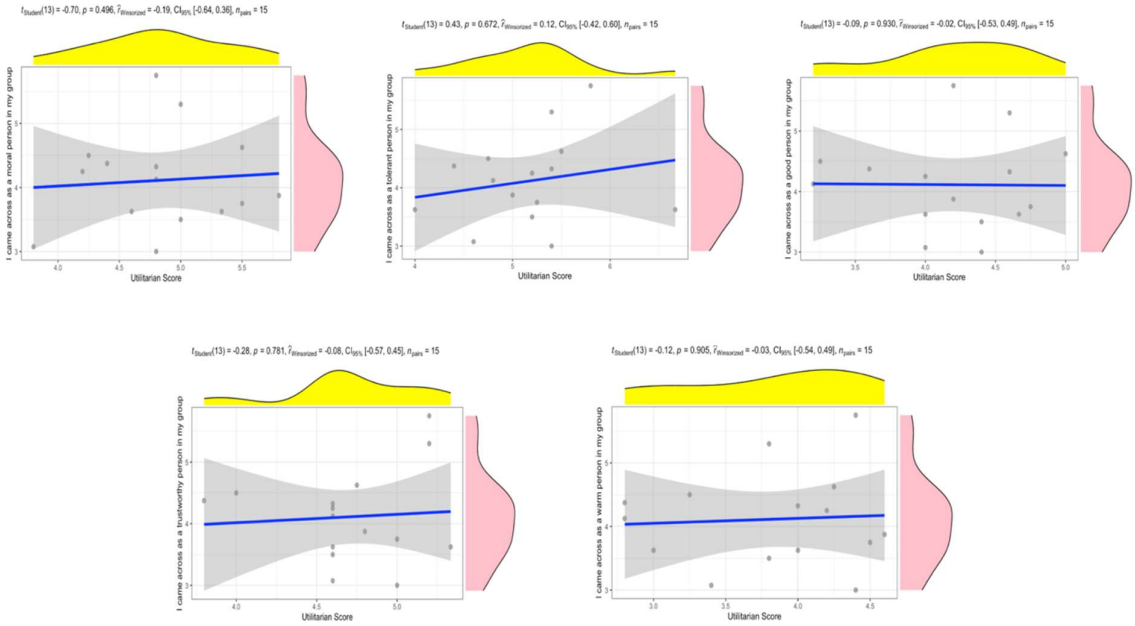


Figure S8. Correlation between each item of Warmth Scale (vertical axis) and the utilitarian score (horizontal axis). Each dot represents one group.

15.2 Competence Scale

15.2.1 Competence Scale Linear Regression:

Table 23.
Linear Regression of Competence Scale

		USCORE	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	6.14	0.93 – 11.35	0.026
intelligent	0.88	-1.64 – 3.41	0.448
reasonable	-0.48	-2.06 – 1.11	0.513
confident	-0.42	-1.53 – 0.69	0.413
Leader-like	0.34	-0.67 – 1.35	0.465
competent	-0.65	-3.13 – 1.83	0.569

Observations	15
R^2 / R^2 adjusted	0.342 / -0.024

15.2.2 Competence Scale Item-based correlation:

The correlation between each item and the utilitarian score presented in Figure S9. Each dot represents one group.

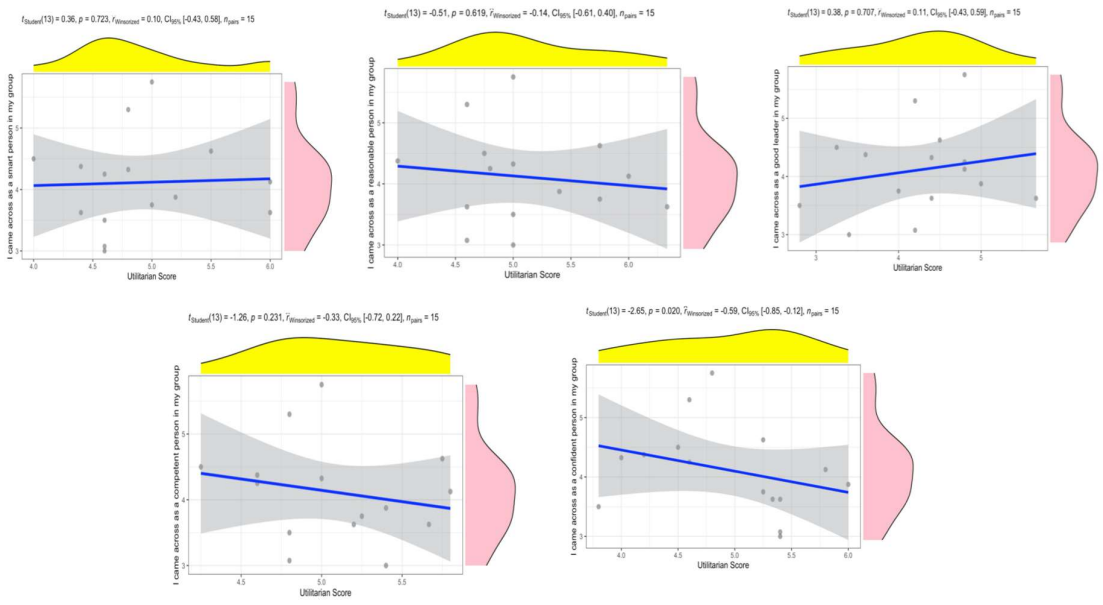


Figure S9. Correlation between each item of Competence Scale (vertical axis) and the utilitarian score (horizontal axis). Each dot represents one group.

15.3 Social Connection Scale

15.3.1 Social Connection Scale Linear Regression Models:

Here we performed regressions between different items of the Social Connection Scale and U-Score.

Table 24.
Linear Regression Model of Social Connection Scale

		USCORE	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	7.41	0.99 – 13.84	0.028
Lonely	0.30	-0.41 – 1.02	0.370
Accepted	0.32	-0.65 – 1.29	0.487
Connected	-1.23	-2.38 – -0.07	**0.039**
Observations	15		
R^2 / R^2 adjusted	0.377 / 0.206		

Here, in contrast to previous findings (Lucas & Livingston, 2014), one of the items of feeling socially connected to others was correlated with *less* utilitarian scores at the group level using linear regression model. However, when p-valued corrected for multiple comparisons following Benjamini & Hochberg (1995) method (i.e."BH" or its alias "fdr" via p.adjust function in R). We did not observe such effect after correcting for multiple comparisons. Here we report the corrected p values for Socially Connected: $p = .1173451$.

15.3.2 Social Connection Item-based correlations:

The correlation between each item and the utilitarian score is presented in Figure S10. Each dot represents one group.

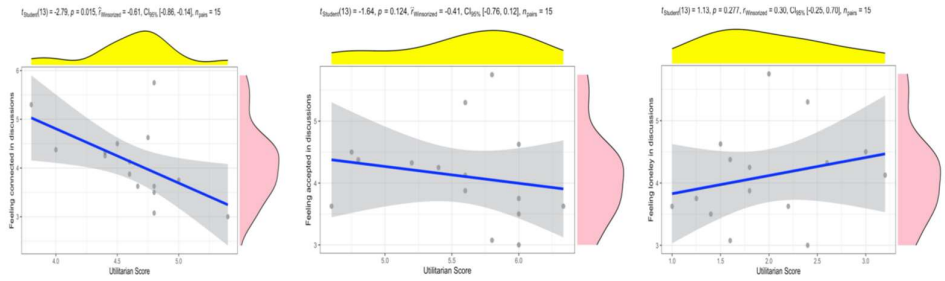


Figure S10. Correlation between each item of Social Connection Scale (vertical axis) and the utilitarian score (horizontal axis). Each dot represents one group.

3.5.3 C: Experiment 1 (Face to Face) vs. Experiment 2 (Online)

16. Test of Variance

We compare the distribution of responses to the same scenarios in Experiment 2 (top row) and Experiment 1 (bottom row). Following our convention in other figures, the left panel in each row corresponds to first private response, the middle panel to the consensus and the right panel to the second private response. As can be visually observed here, there was a remarkably high correspondence between private responses in the two experiments. The shape of the distributions was very similar indicating a very high level of replication in the private responses. Interestingly, the distribution of the consensus responses showed very different distributions between the Online (top) and face-to-face (bottom) version of the experiment (See Figure S11).

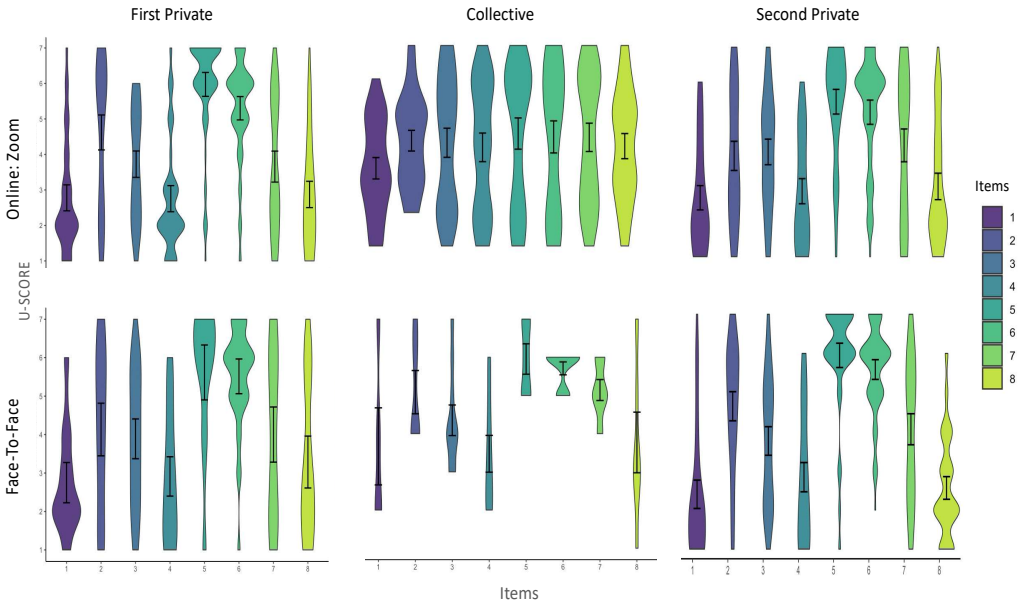


Figure S11. The distribution of responses to the same scenarios (Utilitarian Scores – Vertical axis) in experiment 2 (top row) and experiment 1 (bottom row). The left panel in each row corresponds to first private response, the middle panel to the consensus and the right panel to the second private response. Items are colour coded.

Exploratory analysis using multiple measures of squared rank test of homogeneity difference revealed a significant difference in the variance of Online vs. Face-to-Face moral judgments scores: face to face (but not online) interaction promoted more diverse consensus opinions on exactly the same dilemmas while the distribution of individual opinions remains identical. The Collective variance was significantly different in Experiment 1 from Experiment 2. By contrast, no

difference of variance was seen in the First or Second conditions between the two studies. The squared rank test details for collective responses are as below:

1) F test to compare two variances' data: Rating by EXPF = 0.7, num df = 223, denom df = 639, p-value = 0.003 (alternative hypothesis: true ratio of variances is not equal to 95 percent confidence interval: 0.577- 0.889) sample estimates: ratio of variances 0.712.

2) Levene's Test for Homogeneity of Variance (center = median) Levene's Test for Homogeneity of Variance (center = median) Pr(>F) = 18.7 1.7e-05 ***

3) Fligner-Killeen test of homogeneity of variances data: Rating by EXPFligner-Killeen:med chi-squared = 14, df = 1, p-value = 2e-04

17. Sacrificial Dilemmas (action) vs. Real-Life Dilemmas (inaction)

In addition, we also report another feature of our design which turned out to show a marked difference between the face-to-face and online experiments. Note that in our study, we combined a number of sacrificial dilemmas adapted from previous literature (Greene et al., 2001) with a number of new independently validated items that we had specifically developed for our purpose to have real-life scenarios. In the sacrificial dilemmas, the participants would judge a protagonist's action, for example, to kill one in order to save many. In the non-sacrificial dilemmas, the protagonist avoids an action and thereby violates a norm in order to maximize the benefit for many, e.g., stay silent about a cheating friend to avoid disintegrating the friend's marriage and family. As such, our dilemmas can be classified as Action (sacrificial) vs. Inaction (non-sacrificial). In addition, we also report another feature of our design which turned out to show a marked difference between the face-to-face and online experiments. Note that in our study, we combined a number of sacrificial dilemmas adapted from previous literature (Greene et al., 2001) with a number of new independently validated items that we had specifically developed for our purpose to have real-life scenarios. In the sacrificial dilemmas, the participants would judge a protagonist's action, for example to kill one in order to save many. In the non-sacrificial dilemmas, the protagonist avoids an action and thereby violates a norm in order to maximize the benefit for many, e.g., stay silent about a cheating friend to avoid disintegrating the friend's marriage and family. As such, our dilemmas can be classified as Action (sacrificial) vs. Inaction (non-sacrificial).

In Experiment 1, i.e., face-to-face interaction, we had observed similar moral judgments for the two categories. In Experiment 2, i.e., online interaction, we observed a marked difference between judgments of Action and Inaction

dilemmas (see table below). For Action dilemmas, the results replicated Experiment 1. Impressively, the effect size was almost identical between experiments 1-2. For Inaction dilemmas, however, a significant difference was observed between moral judgments in Experiment 1 and 2. See Figure S12.

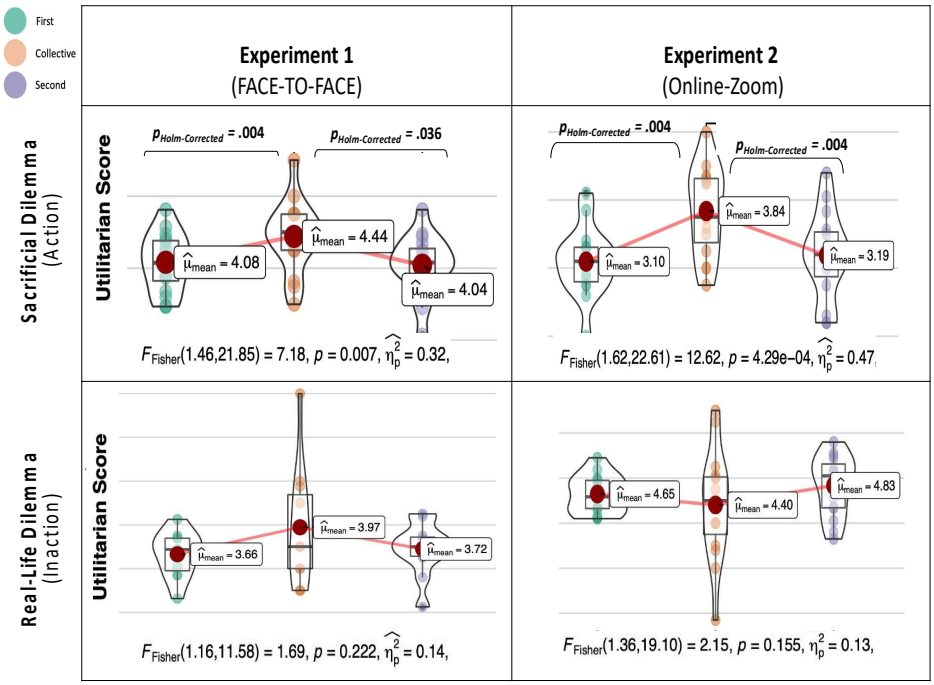


Figure S12. For Action dilemmas, the results of Experiment 2 replicated Experiment 1. Impressively, the effect size was almost identical between Experiments 1-2 and closely replicated another previously published study which used the similar items (Cercu et al. 2020). For Inaction dilemmas, however, a significant difference was observed between moral judgements in Experiment 1 and 2.

18. Rationale for time limits

The time limit in our experiment was mainly driven by practical observations based on two pilot groups. Participants needed around 90 seconds to read one dilemma (for the first time) and respond to it privately in a self-paced manner without the time pressure. Participants spent, on average, 3 minutes discussing each dilemma before moving on. In the final private judgment, as the participants had already seen the dilemmas and were going through them for a second or third time, self-paced responses were much quicker and did not take more than 30 seconds. The effect of time pressure is debated in the literature (e.g., see Cummins & Cummins, 2012). Besides, the previous works (e.g., see Suter & Hertwig, 2011)

showed that the time pressure that impacted participants' responses (towards more *deontological* judgment) was 3to *8 seconds* which were *considerably shorter* than our 90-, 180- and 30-seconds time windows.

19. Synergy in Groups

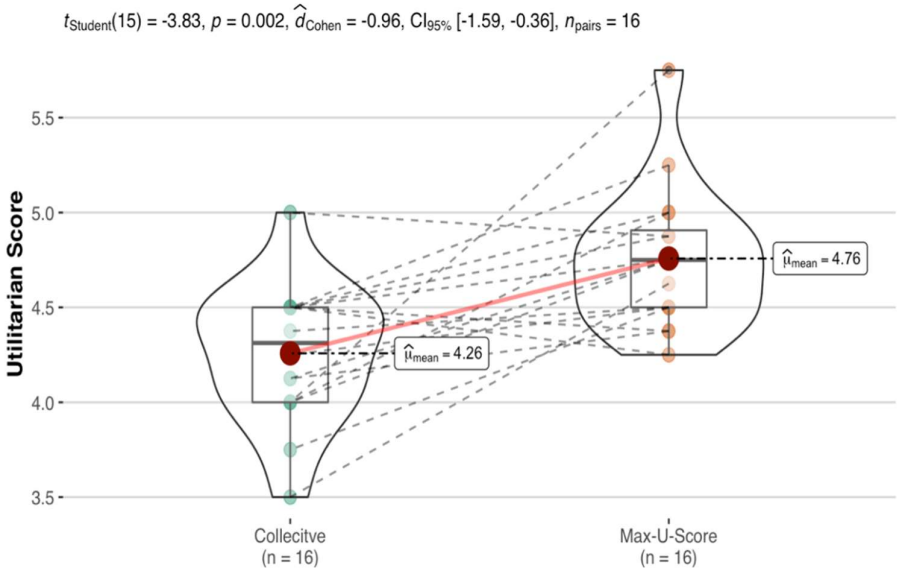


Figure S13. Collective utilitarian scores (in green) are lower than the highest individual utilitarian scores within groups (In orange) In Experiment 1, pattern of results that fully replicates the results reported in Curseu et al (2020). Each dot represents one group. The statistical test result is shown in the figure using the **ggstatssplot** package in R (Patil, 2021).

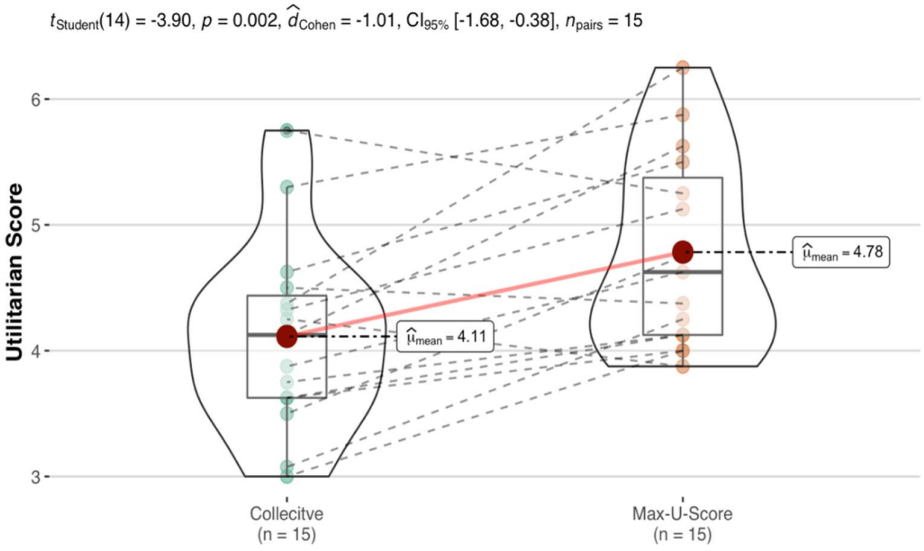


Figure S14. Collective utilitarian scores (in green) are lower than the highest individual utilitarian scores within groups (In orange) In Experiment 2, pattern of results that fully replicates the results reported in Curseu et al (2020). Each dot represents one group. The statistical test result is shown in the figure using the **ggstatssplot** package in R (Patil, 2021).

3.5.4 D: Scenarios in English

1- Emilia is negotiating with a powerful and determined terrorist who is about to set off a bomb in a crowded area. Her advantage is that she has his innocent teenage son in her custody. There is only one thing that she can do to stop him from detonating his bomb, which will kill thousands of people if detonated. To stop him, she must contact him over the satellite hook-up that she has established, and, in front of the camera, she breaks one of his son's arms and then threatens to break the other one if he does not give himself up. Emilia accepts to do this action. On a scale of 1 to 7, how morally acceptable was Emilia's decision?

2- Azzo is the captain of a military submarine traveling underneath a large iceberg. An onboard explosion has caused him to lose a large part of the oxygen supply and has injured one of his crew, who is quickly losing blood. The injured crew member is going to die from his wounds no matter what happens. The remaining oxygen supply is not sufficient for the entire crew to make it to the surface. The only way to save the other crew members is to shoot dead the injured crew member so that there will be just enough oxygen for the rest of the crew to survive. Azzo decides to kill the fatally injured crew member. On a scale of 1 to 7, how morally acceptable was Azzo's decision?

3- Ann is the late-night-watch nurse in a hospital. Due to an accident in the building next door, there are deadly fumes rising up through the hospital's ventilation system. In a certain room of the hospital are three patients. In another room, there is a single patient. If Anna does nothing, the fumes will rise up into the room containing the three patients and cause their deaths. The only way to avoid the deaths of these patients is to hit a certain switch, which will cause the fumes to bypass the room containing the three patients. As a result of doing this, the fumes will enter the room containing the single patient, causing his death. Anna decides to hit the switch to kill this patient. On a scale of 1 to 7, how morally acceptable was Anna's decision?

4- A runaway trolley is heading down the tracks toward five workmen who will be killed if the trolley proceeds on its present course. Friedrich is on a footbridge over the tracks, in between the approaching trolley and the five workmen. Next to him on this footbridge is a stranger who happens to be very large. The only way to save the lives of the five workmen is to push this stranger off the bridge and onto the tracks below, where his large body will stop the trolley. The stranger will die if Friedrich does this, but the five workmen will be saved. Friedrich pushes the large stranger. The stranger dies, but the trolley stops, and the five workmen do not die. On a scale of 1 to 7, how morally acceptable was Friedrich's decision?

5- Alex is the leader of a small group of soldiers. He is on his way back from a completed mission deep in enemy territory when one of his men has stepped into a trap that has been set by the enemy and is badly injured. The trap is connected to a radio device that by now has alerted the enemy to your presence. They will soon be on their way. If the enemy finds his injured man, they will torture him and kill him. He begs Alex not to leave him behind, but if he tries to take the injured man, his entire group will be captured. The only way to prevent this injured soldier from being tortured is to shoot him. Alex shoots the soldier, who dies. On a scale of 1 to 7, how morally acceptable was Alex's decision?

6- Lena is a police officer, and she recently caught a criminal she has been hunting for some time. The criminal is allegedly responsible for rigging a series of explosive devices: some that have already gone off and some that have yet to detonate. He places explosives outside city cafes and sets them to go off at a time when people are drinking coffee on the patios. In this manner, he has injured many people and might injure many more. Now that the criminal is in custody, Lena wants to know where the unexploded bombs are so she can defuse them. He refuses to talk, so Lena decides to use "aggressive interrogation techniques" like holding his head under water and beating him. She decides to use "aggressive interrogation techniques". She knows she will be able to find and defuse the unexploded bombs. On a scale of 1 to 7, how morally acceptable was Lena's decision?

7- Mark is an expert motorcyclist participating in a Bike Week. As he is driving down the road in front of a large group of ten bikers, he suddenly notices that a

biker upfront is losing control over his machine. Mark speeds up to pull alongside the biker, who is going to crash any moment. If the biker crashes on the road, this will result in a large pile-up, and the group of ten bikers behind him will die. If Mark forces this biker off the road, the biker will crash into the trees. This will prevent the pile-up, but at his current speed, the biker will die in the crash. Mark decides to prevent the large pile-up by forcing this biker off the road and kills him. On a scale of 1 to 7, how morally acceptable was Mark's decision?

8- Eric is a miner. The only way out of the mine is to ride up with a simple cable elevator with small cubicles. It is the end of his turn, and everybody is riding up to the surface in the cubicles. While he is in the queue, he sees that the cable supporting all the cubicles is about to snap. 10 miners are riding the cubicles at this time. If he pushes an emergency switch, the last cubicle will be detached from the cable. This will reduce enough weight to maintain the cubicles of the other 10 miners. This will kill the miner in the detached cubicle, but it will save the other ten. Eric decides to reduce the weight by pushing the emergency switch. This detaches the last cubicle, and the one miner in that cubicle dies, the other ten miners can reach the exit. On a scale of 1 to 7, how morally acceptable was Eric's decision?

9- Lucia is terminally ill and dying from cancer. Her doctor knows that she has just started using heroin, cocaine, and other illegal drugs as a way of distracting herself from her illness, but no one else knows. Lucia plans to consume many drugs over the next two months to enjoy the last moments of her life. Her doctor knows that the drugs are bad for Lucia's health but using drugs will not shorten or extend her life at all. He knows that she will die for sure in two months. The doctor decides not to tell Lucia to stop using drugs even though it is illegal and bad for her health. Lucia enjoys consuming illegal drugs over two months and dies afterward. On a scale of 1 to 7, how morally acceptable was the doctor's decision.

10- Nina is a good friend of Martin, who is married to Linda. One day Nina discovers that Linda is having a love affair with another man for one year. It is clear that Linda does not want to leave Martin or tell him the truth but wishes the affair to continue secretly. Nina knows that for Martin, it is important that he and Linda are faithful to each other and do not sleep with anyone else. Nina also knows that John would leave Linda if he discovers that she has been lying to him. Nina decides not to tell Martin that Linda has an affair. Martin and Linda stay together, and Linda continues to have a secret affair. On a scale of 1 to 7, how morally acceptable was Nina's decision?

11- Dr. Beckert is a German doctor. He is working for a medical charity in Congo and gets infected with Ebola. The virus is spreading in the body fast, and he will die for sure unless he is returned to Germany for treatment. Bringing him back would save him, but it involves a high risk that the virus infects between one to four people: the nurse, driver, and two doctors who will need to attend to him. The risk that more than one person is infected is 50%. There are no alternative options, and it would be impossible to mitigate the risks of traveling back. Johann

is the head of the medical charity that employs Dr. Beckert. He has to decide whether to repatriate Dr. Beckert and risk the life of the four people needed to attend to him during the trip from Congo to Germany. The standard process would be to send only the nurse, driver, and two doctors voluntarily but not precisely tell them about the risks involved. Johann, the officer in charge of this case, decides not to return Dr. Beckert to Germany due to the risks involved. As a result, Dr. Beckert cannot receive life-saving treatment in Germany and dies in Congo. On a scale of 1 to 7, how morally acceptable was Johann's decision?

12- Simon is an engineer leading a project to build a water supply for a poor city in a developing country. One day, they discover the remains of an unknown ancient culture when drilling for water. Simon has to decide whether to inform the authorities, which would lead to archaeological excavations and greatly enrich the cultural world heritage and knowledge. But it would also put an end to the water supply project, leaving the 1000 inhabitants of the city in poverty. Simon decides to keep silent about the discovery. As a result, the cultural heritage is lost, but the wellbeing of the poor inhabitants is greatly increased. On a scale of 1 to 7, how morally acceptable was Simon's decision?

13- Amelie overhears her colleague at work, explaining on the phone that he has managed to hack a private pension saving. He has stolen the funds saved by 10 elderly people who have reached their pension age, but instead of keeping the money for himself, he donated it to a poor orphanage that can now afford to feed, clothe, and care for 100 children until they reach adulthood. The colleague also explains that this is a one-off and that he will not do this again. Amelie knows that this type of private pension savings is not insured. If she goes to the authorities to denounce her colleague, the money will be returned to the pension scheme, but the 100 children will not be fed or dressed properly. Amelie decides not to denounce the man to the police, which means that the money from the private pension scheme of the 10 elderly people is gone forever, but 100 children will be fed and dressed properly until they reach adulthood. On a scale of 1 to 7, how morally acceptable was Amelie's decision?

14- Otto is a politician in a country of 3 million people. He has to decide whether to increase the tax on tobacco by 20% or keep the current tax policy. Medical experts can reliably predict that increasing the tax would prevent 10,000 deaths every year due to lung cancer. However, from a purely economic point of view, the early deaths of 10,000 smokers have economic benefits: even with the new tax income, the government does not have to pay for pensions for 10,000 smokers. Instead, it can invest in infrastructure and building schools so 1 million people would benefit from the early deaths of 10,000 smokers. Otto realizes that leaving the tax as it means that smoking will just continue as it is, and government money can be used in infrastructure and building schools. Otto decides to keep the current tax policy on tobacco, which means that 10,000 smokers will continue to smoke and die earlier because of lung cancer, but this keeps funding available to build schools and infrastructure for 1 million inhabitants. On a scale of 1 to 7, how morally acceptable was Otto's decision?

15- Joachim and Ingrid are a couple hosting a farewell party before emigrating to Australia the next day. At the farewell party, Ingrid meets Paul and is very much attracted to him. She asks Paul if he wants to sleep with her. Paul knows that Ingrid is in a relationship with Joachim and promises him never to sleep with anyone but him. Paul, too, is in a relationship and has promised his partner not to have sex with anyone but her. Both Ingrid and Paul know that if they secretly left the party and had sex together, nobody would ever find out, and they would have a very enjoyable night together. Paul does not decline to sleep with Ingrid. Both of them thus enjoy great pleasure but violate their promises.

16- A German investigation journalist has been violently murdered. Anna is in charge of the investigation, and there is mounting evidence that the murder was ordered by a foreign country in Africa, which is a long-time trade partner for Germany and with which the German state is about to conclude a large sale deal. The new sale deal would create 10,000 well-paid jobs in Germany over the next 2 years and take 10,000 people out of unemployment. Anna decides not to process the evidence that foreign power is involved. The case remains unresolved, and no one else is accused of it. The trade deal with the foreign country is not compromised by the results of the investigation, and wealth and 10,000 jobs are created. On a scale of 1 to 7, how morally acceptable was Anna's decision?

3.5.5 E: Added Measurements in German

- Während der Diskussionen fühlte ich mich sozial mit anderen verbunden.
- Während der Diskussionen habe ich mich von anderen akzeptiert gefühlt.
- Während der Diskussionen fühlte ich mich einsam.
- Ich denke, dass ich während den Diskussionen als warmherzige Person wahrgenommen wurde.
- Ich denke, dass ich während den Diskussionen als selbstbewusst Person wahrgenommen wurde.
- Ich denke, dass ich während den Diskussionen als moralisch Person wahrgenommen wurde.
- Ich denke, dass ich während den Diskussionen als gutmütig Person wahrgenommen wurde.
- Ich denke, dass ich während der Diskussionen als eine vernünftige Person wahrgenommen wurde.
- Ich denke, dass ich während den Diskussionen als intelligent Person wahrgenommen wurde.

- Ich denke, dass ich während den Diskussionen als guter Anführer Person wahrgenommen wurde.
- Ich denke, dass ich während den Diskussionen als vertrauenswürdig Person wahrgenommen wurde.
- Ich denke, dass ich bei den Interviews als tolerante Person wahrgenommen wurde.
- Ich denke, dass ich bei den Gesprächen als kompetente Person wahrgenommen wurde.

References

Ameijeiras-Alonso, J., Crujeiras, R., & Rodriguez-Casal, A. (2021). multimode: An R Package for Mode Assessment. *Journal of Statistical Software, 97*(9), 1 - 32. doi:http://dx.doi.org/10.18637/jss.v097.i09

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using {lme4}. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series b-Methodological, 57(1), 289–300.

Bürkner, P.-C. (2018). Advanced {Bayesian} Multilevel Modeling with the {R} Package {brms}. *The R Journal*, *10*(1), 395–411.

Christensen, R. H. B. (2019). "ordinal—Regression Models for Ordinal Data." R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

Cummins, D. D., & Cummins, R. C. (2012). Emotion and deliberative reasoning in moral judgment. *Frontiers in Psychology*, *3*(SEP), 328. https://doi.org/10.3389/fpsyg.2012.00328

Curşeu, P. L., Fodor, O. C., A. Pavelea, A., & Meslec, N. (2020). "Me" versus "We" in moral dilemmas: Group composition and social influence effects on group utilitarianism. *Business Ethics*, *29*(4), 810–823. https://doi.org/10.1111/beer.12292

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. Methods in Ecology and Evolution, 7(4), 493–498. https://doi.org/10.1111/2041-210X.12504

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*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Grubbs, F. (1950). Sample Criteria for Testing Outlying Observations. *The Annals of Mathematical Statistics, 21*(1), 27-58. Retrieved May 2, 2021, from http://www.jstor.org/stable/2236553

Komsta, L. (2011). outliers: Tests for outliers. *CRAN*. https://cran.r-project.org/web/packages/outliers/outliers.pdf

Lüdecke D, Ben-Shachar M, Patil I, Waggoner P, & Makowski D (2021). "Assessment, Testing, and Comparison of Statistical Models using R." *Journal of Open Source Software*, **6**(59), 3112. DOI: 10.31234/osf.io/vtq8f.

Patil, I., (2021). Visualizations with statistical details: The 'ggstatsplot' approach. Journal of Open Source Software, 6(61), 3167, https://doi.org/10.21105/joss.03167

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., … Cushman, F. (2021). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. Journal of Personality and Social Psychology, 120(2), 443–460. https://doi.org/10.1037/PSPP0000281

Thomas, G., Kincaid, J. P., & Hartley, R. D. (1975). Test-Retest and Inter-Analyst Reliability of the Automated Readability Index, Flesch Reading Ease Score, and the Fog Count. *Journal of Literacy Research*, *7*(2), 149–154. https://doi.org/10.1080/10862967509547131

Chapter 4. Diffusion of Punishment in Collective Norm Violations

Anita Keshmirian*, Babak Hemmatian, Bahador Bahrami, Ophelia Deroy, Fiery Cushman

Anita Keshmirian
[Graduate School for Neuroscience, Ludwig-Maximilians-University]
[Faculty of Philosophy, Ludwig-Maximilians University]

Babak Hemmatian
[Department of Cognitive, Linguistic and Psychological Sciences, Brown University]

Bahador Bahrami
[Faculty of Psychology, Ludwig-Maximilians-University]
[Department for Psychology, Royal Holloway University of London]
[Centre for Adaptive Rationality, Max Planck Institute for Human Development]

Ophelia Deroy
[Faculty of Philosophy, Ludwig-Maximilians University]
[Munich Center for Neuroscience]
[Institute of Philosophy, School of Advanced Study, University of London]

Fiery Cushman
[Department of Psychology, Harvard University]

Correspondence should be addressed to Anita Keshmirian, Ludwig-Maximilians-University of Munich, Munich, Germany. Email: anita.keshmirian@campus.lmu.de

Author Note

This research was funded by a grant from SEED (to FC) and by the NOMIS foundation (to OD). An overview of Experiments 2 was presented at Oxford University at The Ninth International Symposium on "Biology of Decision Making" in May 2019. BB was supported by the Humboldt

Foundation, the NOMIS Foundation, and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 819040 - acronym: rid-O).

Author Contributions

A. Keshmirian generated the study concept. Data collection performed by A. Keshmirian for experiment 1 and A. Keshmirian and B. Hemmatian for experiment 2. A. Keshmirian performed interpretation and visualization of both experiments. A. Keshmirian prepared materials of Experiment 1. A. Keshmirian and O. Deroy prepared materials of Experiment 2 while B.Bahrami and B.Hemmatian provided critical revisions. Experiment 1 was performed under F.Cushman and Experiment 2 under F. Cushman, B.Bahrami and O.Deroy and supervision. A. Keshmirian analyzed the data of both experiments and drafted the original manuscript, while B. Hemmatian, B. Bahrami, O. Deroy, and F. Cushman provided critical revisions. F.Cushman, O. Deroy and B. Bahrami provided the funding.

The Declaration of Conflicting Interests

The author(s) declare no conflicts of interest concerning the authorship or the publication of this article.

Open Practices Statement

The preregistration for experiment two can be accessed at https://osf.io/hjnxm . De-identified data for all experiments, along with a codebook and materials, are openly available at https://osf.io/m3f47/

Word count: 2165 words (excluding cover page, abstract, method and results, statement of relevance, author contributions, references, tables, figures, and figure legends)

.

Abstract: We show that people assign less punishment to individuals who cause harm together with others, compared to those who act alone. In Experiment 1, participants (N=1002) assigned less punishment to individuals involved in collective violations leading to intentional and/or accidental deaths, but not failed attempts, emphasizing that harmful outcomes, but not malicious intentions, were necessary and sufficient for the diffusion of punishment. Experiment 2a compared the diffusion of punishment across harmful actions and 'victimless' purity violations (e.g., eating human flesh in groups; N=752). Punishments were reduced in group context more strongly for harmful actions vs. purity violations.

Experiment 2b (N= 500) exclusively examined purity violations and, as expected, found no diffusion. We propose discounting in causal attribution as the underlying cognitive mechanism for reducing punishment in collective norm violations.

3-5 Keywords regarding Methods used: Online data acquisition, Mixed Effects Models, Bayesian Regression, Open science, Open data

3-5 Keywords regarding the scientific topic: Collective moral transgression, diffusion of punishment, Moral Foundations Theory, Causal Attribution, Discounting principle.

Statement of Relevance

We use cognitive and social psychology concepts to examine an overlooked contextual influence on moral judgments, with implications for forensic psychology. We not only show that the punishment reduces in collective harmful violations for each perpetrator but confirm discounting in causal attribution as a potential cognitive mechanism. The asymmetries found in the reduction of punishment based on intention and harmfulness have ramifications for individuals' propensity to commit moral transgressions as parts of groups and for any situation in which moral judgment is passed on such transgressions. Many of the actions represented by our stimuli show illegal behaviors where being part of a group should not affect punishment according to many justice systems (e.g., that of the United States). Our findings, therefore, raise the possibility that this intended aspect of legal code may not be easily implemented in practice, or if implemented, may clash with public perceptions of blameworthiness and deserved punishment.

4.1 Introduction

In 44 BC, 60 Roman senators conspired to murder Julius Caesar at a senate meeting. They collectively stabbed him 23 times, leading to his death. But who, exactly, was to blame — and how much? Many crimes are committed by groups, including gang rapes, collective hate crimes, co-offending, and conspiracies. Understanding how people assign blame and punishment in such ‘group crimes’ is important in its own right, and can also help resolve discrepancies between current theories of moral judgment.

Generally, people judge an actor as fully blameworthy if they intentionally cause harm (Guglielmo et al., 2009; Malle et al., 2014; Malle & Knobe, 1997; Shaver, 1985; Shultz & Wright, 1985). Much research suggests that these two factors — intentionality and causal responsibility for harm — play dissociable roles in moral judgment (Cushman, 2008; Young et al., 2007, 2010). But they may influence the

judgment of group crimes in different, even contradictory ways. Thus, our approach is to differentiate their relative contributions to the judgment of group crimes.

4.1.1 Intentionality

How do intent-based judgments of group actors compare to solo actors? One natural possibility is that group actors are held just as responsible, given that each volitionally decides to engage in transgressive behavior. Alternatively, they may also be held less responsible on the belief that they got "caught up" in something they would not otherwise have done (Malle et al., 2001; Malle & Knobe, 1997). In this case, group actors would receive less blame and punishment than solo actors committing equivalent crimes.

4.1.2 Causal responsibility

How will judgments of the causal responsibility of group actors compare with solo actors? One possibility is that causal responsibility is categorical— either one is responsible or not (see also Moore, 2009). Since participants in group crimes are causal *contributors* to the outcome, they would be held causally responsible to the same extent as a solo actor (e.g., felony murder, see Binder et al., 2016). For instance, in many US states (e.g., Connecticut General Statutes. tit. 53a-54c, Chapter 952; 2012), in the case of collective murder, it is argued that since the felony itself *causes* death, every participant in the felony is *causally* responsible for the death (Figure 1 – left).

Alternatively, causal responsibility may be diminished as it is distributed across a number of people (Figure 1 - right). This comports with the well-studied phenomenon of causal discounting (Kelley, 1973; Morris & Larrick, 1995). Additionally, when harm is "overdetermined"–i.e., it would have occurred even without the action of any sole individual—people are likely to perceive each individual as less causally responsible (Lagnado et al., 2013). Similarly, the degree to which the individual has causal control over the outcome may be diminished in collective violations, and so causal power theory would suggest diminished attributions of causal responsibility (Cheng, 1997). Because punishment judgments are sensitive to attributions of causal responsibility for harm (Cushman, 2008), these theories predict diminished punishment for group actors.

4.1.3 Existing research

Existing research offers mixed evidence on the punishment of group compared to solo crimes. An archival study of sentences given out to individual and collective violations offered some evidence for reduction of punishment as judges gave harsher sentences to lone offenders controlling for the crime (Feldman & Rosen, 1978). However, a follow-up experiment on hypothetical robberies failed to find corroborating evidence, a result ascribed to the small sample size (Feldman & Rosen, 1978). More recently, El Zein et al. (2020) investigated second-party

punishment in fairness-based group games but found no difference between proposed punishment for lone fairness violators compared with collective ones (El Zein et al., 2020). In another study, comparing suspicious outcomes in a coin-toss task, although violators in a group were considered more honest (controlling for outcome), the difference in punishment judgments of dishonest individuals in isolation vs. in groups was only marginally significant (p = .08; Vainapel et al., 2019).

A key limitation of prior studies is that they cannot dissociate the potentially divergent role of intent-based versus responsibility-based processes in judgments of deserved punishment. To disentangle these, in Experiment 1, we compared accidental (where there is no intent, but causal responsibility is preserved) and attempted harm (where the intent is preserved, but no harm is caused). In Experiment 2, we investigated cases of collective "victimless" transgressions, such as disrespecting the deceased, and compared them to collective harmful transgressions. Like attempted harms, these preserve the element of transgressive volitional action while eliminating any relevant question of causal responsibility.

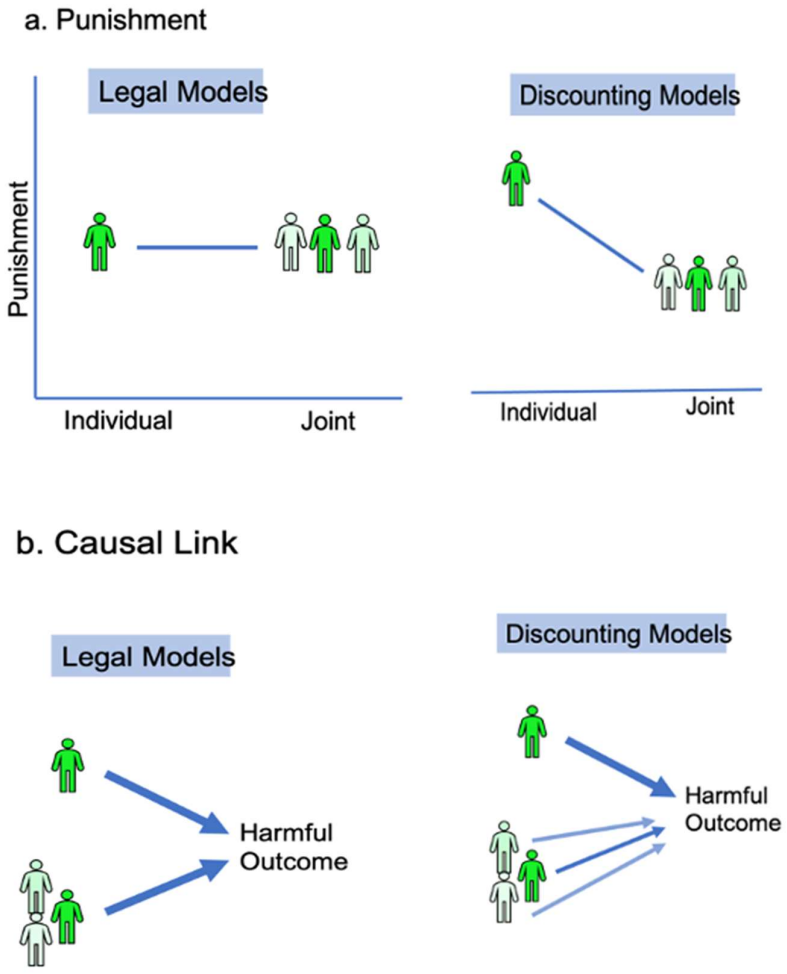


Figure 1. a. Two models of punishment in individuals and joint actions: legal models suggest similar punishment for joint and individual violations. Discounting models predict less punishment in joint than individual violations. b. Causal links in two models of punishment. In legal models, all perpetrators in joint violations are causally responsible for the harmful outcome to the same degree as in solo violations. In the discounting models, each individual in the group is less causality responsible for the outcome than solo violations.

4.2 Experiment 1

Experiment 1 tested whether a third party punishes an individual less if she inflicts a harmful outcome on a victim as part of a group, rather than acting alone. We employed a 2x2x2 design with three factors: Collectivity, intention, outcome. Collectivity was treated as a between-subject, and intention and outcome as within-subject factors. By independently manipulating the agents' intentions (malicious vs. neutral) and the actions' outcomes (harmful vs. harmless), we can differentiate the

effect of collectivity on intent- and responsibility-based processes of moral judgment.

4.2.1 Methods

4.2.1.1 Participants

One thousand and seventy-five participants were recruited via Amazon's Mechanical Turk. Thirty-seven participants were excluded for having duplicated IDs. We used a data-driven Mahalanobis Distance measure (Dupuis et al., 2019) to identify non-human participants and inconsistent or inattentive responses in our data (see Supplementary Material). This step resulted in excluding 36 participants. The final sample of 1002 US residents (452 males, eight choosing the "other" option) had an average age of 24.5 years (*SD* = 10.5, range: 18 to 64). We replicated the main results, including those who failed the Mahalanobis exclusion criterion (see Table3, Supplementary Material).

4.2.1.2 Material and procedure

Each participant was randomly assigned to one of the two collectivity conditions (joint or individual scenarios) and read four moral scenarios in which a character committed an act either as part of a group (joint action) or alone (individual action). The dependent measure was always the deserved punishment for a given character on a 7-point scale (1 labeled as "not at all", 4 as "somewhat", and 7 as "a lot").

Intention (innocent, malign) and outcome (harmless, harmful) were crossed within subjects across the four scenarios. In neutral conditions, the agent(s) acted with no malign intention, and no harm ensued. Accidental conditions involved an unintended death following the described action. In the attempted and intentional cases, the agent(s) acted with malign intent, either failing or succeeding in murdering another person. The following is an intentional, individual violation scenario adapted from Young et al. (2010) (see Supplementary Material for full scenario texts):

> *Stacey* and Kate are friends and decide to go rock climbing. They are going to use new harnesses to scale a gigantic cliff.
>
> Kate starts to put on one of the new harnesses. The clamp on the new harness is subtly flawed, so the whole harness is incredibly unsafe to use.
>
> Because the clamp on the harness does not audibly click into place, *Stacey realizes* that the new harness is malfunctioning and may not be safe to use.
>
> *She straps Kate into the harness and asks Kate to go first.* Partway up the cliff, the harness gives way, causing Kate to fall and die.

The three sentences in italics were substituted in joint action conditions with statements about "Stacey, Anita, James, and Kate" instead, implicating the first three characters in the harm inflicted on the last-named individual.

A random pairing of stories was first created for within-subject manipulations and then counterbalanced across participants. The order of scenarios was randomized. Demographics followed the last vignette, including age, gender, political orientation (from 1 denoted as "very liberal" to 7 marked as "very conservative"), ethnicity, and education level.

4.2.2 Results

Figure 3 shows the results of Experiment 1. Given that our moral scenario tasks followed a multilevel pattern (across items, conditions, and participants), and our dependent variable were measured in Likert scale, we used ordinal mixed-effect models in R (https://www.r-project.org/) to account for individual differences.

Punishment ratings for intentional harm were significantly higher than accidental and attempted harm, and ratings in the mentioned conditions all exceeded those for the neutral condition, showing that the intention and outcome manipulations worked as intended (see Table 2 - Supplementary Material). To test the diffusion of punishment hypothesis, we included different combinations of Collectivity, intention, and outcome as factors in mixed-effect models. Model comparison favored the variant including all three factors (see Supplementary Material). The interaction of outcome and collectivity was significant (b = 1.4125, SE= .2, z =7.245, p <.001), while that of intention and collectivity was not (b = .138, SE = .163, z = 7.245, p = .3). Pairwise comparison showed less assigned punishment for characters when involved in joint actions compared to individual actions for intentional (b = .745, SE = .129, z = 4.451, p = .0002) and accidental killings (b =.4363, SE = .128, z = 3.411, p = .014), but not in failed attempts (b = .029, SE=.129, z =. 229, p = 1.0). Since we predicted a null effect for attempted murder, following Aczel et al.'s (2018) method, a Bayesian mixed-effects pairwise comparison was performed to confirm the pattern of results (intentional: BF_{10}= 498.32, CI_{95}= [.29, .69]; accidental: BF_{10} = 21.74, CI_{95}= [.12, .52]; attempted: BF_{10}= .05, CI_{95}= [-.17, .22]; see Table 4 in Supplementary Material). Surprisingly, protagonists in neutral conditions received harsher proposed punishment for joint compared to individual actions (b =.9761, SE=.166, p<.0001; see Figure 1 and Table 2 in Supplementary Material). This effect was not predicted and could be due to a few outliers. Comparisons of specific items can be found in Supplementary Material, Figure S1.

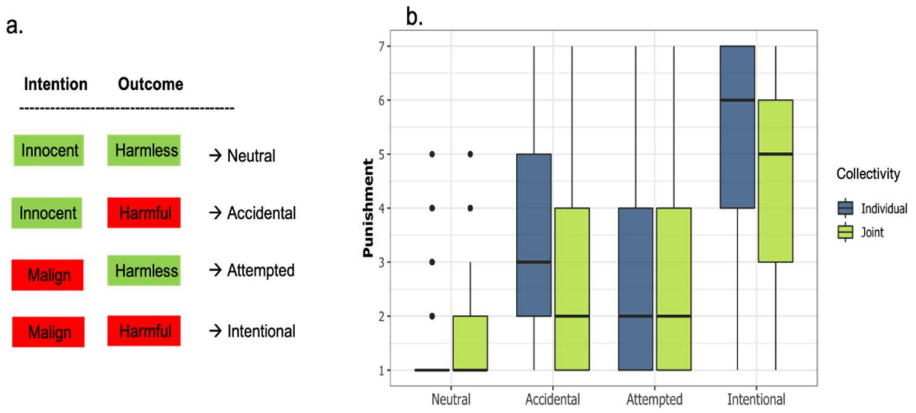


Figure 2. a. Four experimental conditions as the outcome of 2 by 2 design: Intention (Innocent, Malign) and outcome (Harmless, Harmful) b. Box-and-whisker plot of punishment ratings as a function of Collectivity (different colors) across neutral, accidental, attempted, and intentional actions (horizontal axis). The box = middle 50% of scores. The thick horizontal line within each box represents the median. Upper and lower whiskers show the range of scores in the highest and lowest quartiles. The dots represent outliers.

4.2.3 Discussion

We found a robust reduction in proposed punishment across instances of intended and accidental harm when perpetrators acted as part of a group rather than lone agents. The contrast between these results and previous studies (Feldman & Rosen, 1978; El Zein et al., 2020) may be attributed to the more representative range of clearly and more strongly harmful outcomes (i.e., death) represented in our materials. That no diffusion of punishment was observed for attempted harm suggests that diffusion of punishment depends on discounting principle in causal attribution of harmful outcomes.

4.3 Experiment 2

Not all immoral acts involve harmful outcomes. 'Victimless' purity violations are condemned on the basis of a transgressive action rather than the outcome (McHugh et al., 2017). For instance, research shows that they are judged based on perpetrators' impact on themselves rather than victims (Chakroff et al., 2013; Dungan et al., 2017) and they elicit disgust only when people judge immoral characters, and not the outcomes (Giner-Sorolla & Chapman, 2017).

In Experiment 2, while directly measuring harmfulness and grossness of scenarios, we extended the context to impure victimless acts. This allowed us to investigate the role of discounting in causal attribution in more depth. If the

diffusion of punishment results from a discounting principle in causal attribution, it would only apply to violations that cause harmful outcomes. Therefore, we expected it to be weaker for judgments of purity violations.

4.3.1 Experiment 2.a

In Experiment 2, we extended items to 'victimless' purity violations which involved no prominent harm. We hypothesized that the diffusion of punishment results from discounting in causal attribution, where more than one sufficient cause means less causal responsibility assigned to each agent for the harmful outcome. Therefore, if agents are not perceived as causing significant harm, we expect impartial observers to treat collective purity violations similar to the individual violation, resulting in no punishment reduction.

Experiment 2a compared judgments for less grave *harm* than in Experiment 1 (e.g., intentionally breaking someone's leg) with *purity* violations (e.g., masturbating over a grave). Collectivity was manipulated as before. Unlike most previous studies, instead of assuming that harm scenarios induce a sense of harmfulness alone and the purity vignettes only a sense of disgust, we asked our participants to rate how harmful or gross they found the protagonist(s)' action in all scenarios, in order to examine their evaluation of the outcome more directly. The same judgments allow us to compare the Moral Foundation Theory's predictions (Graham et al., 2011), which considers harm and purity judgments as fundamentally different from those of Dyadic Harm theory (Gray et al., 2014), which casts them both as reflecting judgments of harm.

4.3.1.1 Methods

4.3.1.1.1 Participants

A target sample size was predetermined using a Monte Carlo simulation, following guidelines provided by DeBruine & Barr (2021; see the OSF repository). The final sample consisted of 752 US and UK residents (331 females; three others, age: M = 28.08 years, SD = 6.5, range: 18 to 60) recruited through Prolific Academic (https://www.prolific.co/) and compensated for their time. Twenty-six participants were excluded for having Prolific IDs that duplicated those from a pilot study. To increase precision, we used data-driven methods (as in Exp 1) and questionnaires for attention checks to exclude inattentive responders. No difference was observed between the two methods. Another 39 were excluded for failing attention checks—they assigned 0 to 49 (on a 100-point scale) blame to a person who "destroys the entire planet" (n = 18), or 51 to 100 for someone who "gives money to a charitable organization" (n = 21).

4.3.1.1.2 Materials and procedure

The Collectivity was manipulated as in Experiment 1. The moral domain (harm vs. purity) was additionally manipulated within subjects. Hence, each subject responded to four scenarios, presented to her in two blocks (harm block and purity

block). The scenarios were randomly chosen from a battery of 8 items, counterbalanced across participants. The order of blocks and the items within each block were randomized and counterbalanced across subjects. Four items were adapted from a previous study on Harm (Young et al., 2010, see Supplementary Material), and four items were original materials (some inspired by Rottman & Young, 2019) representing purity violations. The items were matched for severity of joint vs. individual action in a pilot study.

For instance, a purity violation in the individual transgression condition would read:

> Dan's favorite singer has died and has been buried in a nearby cemetery. He always had wild fantasies about the singer, and one night, he forms the following plan: He enters the cemetery late at night and goes to masturbate over the singer's grave, making sure he cannot be seen. After that, he ensures that the grave is clean and exactly as it was before and leaves.

The same scenario in the joint condition would introduce Dan, Ray, and Carl as friends who collectively committed the act.

To provide more precision regarding variability in judgments by providing more options along the spectrum, we measured the punishment scenario on a 100-point Likert scale. Zero to 50 was labeled as mild and 50 to 100 as severe punishment. Perceived harmfulness and grossness were measured on similar scales. Other design aspects were identical to Experiment 1.

4.3.1.2 Results

We first tested whether our conceptual distinction between harm and purity is successfully reflected in our item construction. We included ratings of perceived harmfulness and grossness as factors in a mixed-effects model. As expected, a significant main effect of Domain was found for both harmfulness (b = 38.76, SE = 4.68, t = -8.29, df = 6.32, $p < .0001$) and grossness (b =41.45, SE = 2.43, t = 17.06, df = 7.37, p <.0001). No significant effect of collectivity was found for either Harmfulness ratings (b = .012; p = .1; BF_{10}= .754, CI_{95} = [-1.68, 1.443]) and Grossness ratings (b = .0018; p =.4; BF_{10} = .768, CI_{95} = [-1.577, 2.268]), whether using linear mixed effects analysis or it's Bayesian counterpart (see Figure 3). In addition to establishing the adequacy of our item construction, these results help address a corollary question for which Dyadic Harm (Gray et al., 2014; Schein & Gray, 2018) and Moral Foundation (Graham et al., 2013) theories make contrasting predictions regarding the determinants of moral judgments. Our results (Figure 3) supported the Moral Foundation account of disparate moral domains of harm and purity.

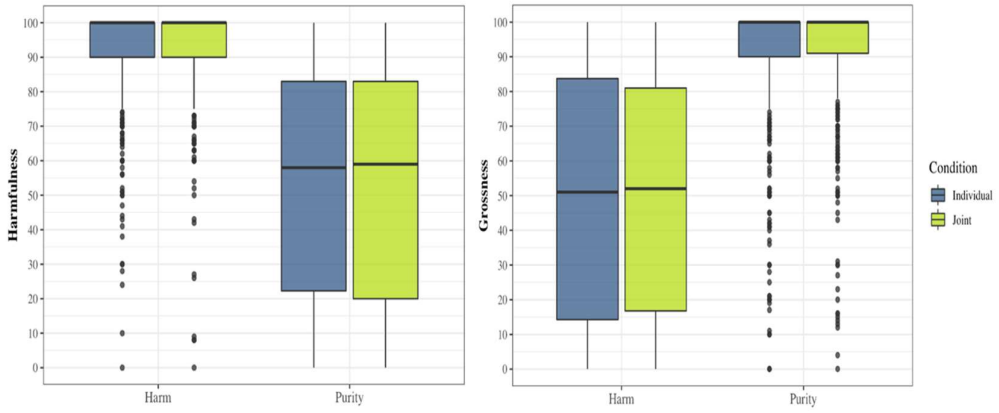


Figure 3. Harmfulness (left) and Grossness (right) ratings are matched across Joint and Individual actions but significantly different across domains: Harmfulness is higher in the Harm domain(left), and grossness is higher in the Purity domain (right). Graph conventions are the same as Figure 2.

We investigated the interaction between collectivity (individual vs. joint) and domain (purity vs. harm) in a mixed-effect model. The predicted interaction with domain was not significant (b = 2.31, SE = 1.23, t = -1.88, df = 2053.67, p = .06). Performing a within subject analysis, we found that overall effect of diffusion was significant in both harm and purity domains (b = 8.81, SE = 1.3, df = 1103.83, t = 6.79, p < .0001).

However, an exploratory analysis revealed a possible carryover effect of purity on harm. A between-subject analysis on the first trials and the first blocks, in which, by design - no carryover effect occurred - showed the diffusion of the punishment only in harm (b = 6.99, SE = 3, p = .02, df = 230.55) and not in purity violations (b = 3.67, SE = 3.78, p = .33, df = 208.336). Similarly, a mixed effect model on the first blocks, accounting for item variability, showed that diffusion of punishment, as predicted, was only significant in harm blocks (b = 5.93, SE = 2.04, p = .003, df = 466) and not in purity blocks (b = 1.78, SE = 2.73, p = .514, df = 421) (see Figure 4; for more details see Supplementary Material).

These analyses suggested that the significant overall effect in purity could be due to a carryover between conditions due to the within-subject design. To examine the conclusions from our exploratory analysis, we conducted a pre-registered replication Experiment 2b, where we focused on purity violation in a between-subject design.

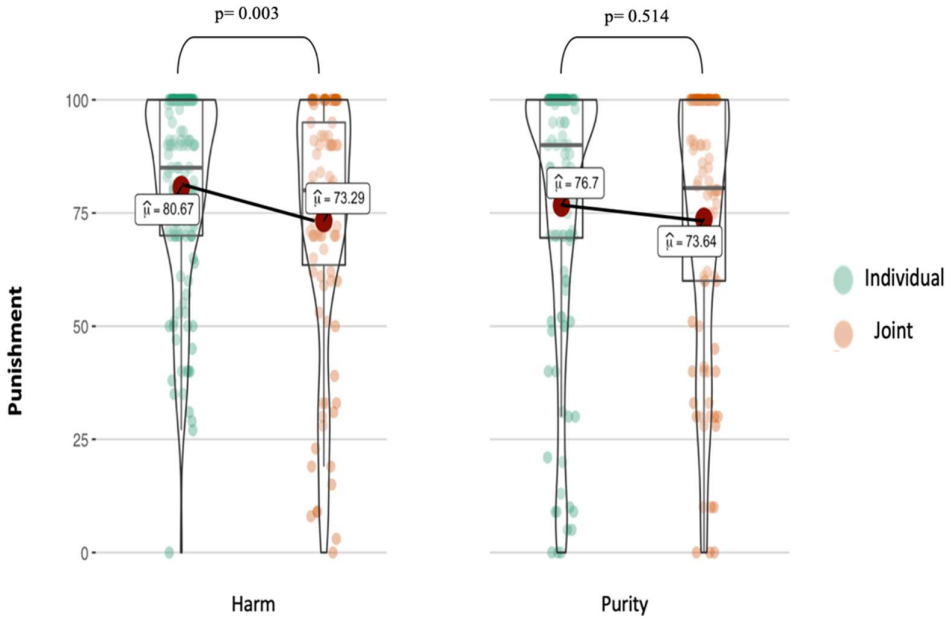


Figure 4. In the first blocks, subjects punished individuals alone (green) more than in joint action (orange) across all items in harm (left) but not in the purity (right) domain.

4.3.2 Experiment 2.b

4.3.2.1 Methods Exp 2.b

4.3.2.1.1 Participants

A target sample size was predetermined using a Monte Carlo simulation, using SIMR package (Green & MacLeod, 2016). R script is accessible at OSF. The final sample consisted of 500 US and UK residents (250 females; three others, age: M = 30.1 years, SD = 5, range: 18 to 58) recruited through Prolific Academic (https://www.prolific.co/) and compensated for their time. Five participants were excluded for having Prolific IDs that duplicated those from a pilot study. To increase precision, we used both data-driven methods (as in Exp 1) and questionnaires for attention checks (as in Exp 2.a). No difference was observed between the two methods (see Supplementary Material). Another 32 were excluded for failing attention checks.

4.3.2.1.2 Materials and procedure

The number of protagonists was manipulated as in Experiment 2.a while only one moral domain (purity) was provided to the subjects. Each participant responded to four fully randomized scenarios, all from the purity domain. The scenarios were the same as experiment 2.a. We measured punishment after each scenario on a 7-point Likert scale, the same as Experiment 1.

4.3.2.2 Results

A linear mixed-effect analysis was performed with collectivity (joint vs. individual) as fixed, and participants and items as random factors (see Supplementary Material for more details). Pairwise comparison indicated that judgments were similar in group and individual transgressions (b = .1853, SE = .1190, t =1.557), which was confirmed by Bayesian mixed-effects analysis (BF_{10} = 0.392, CI_{95}= [-.429, .041]).

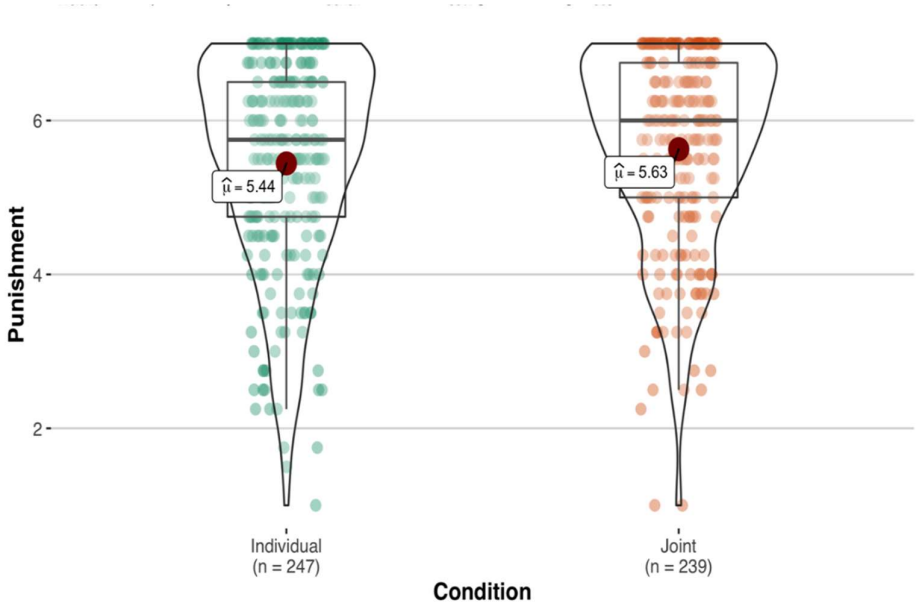


Figure 5. No difference in punishment judgments was observed between joint and individual purity violations.

4.3.2.3 Discussion

Experiment 2b found, despite ample power, no diffusion of punishment for actions deemed impure but harmless. In other words, it appears that punishment is diffused only when the outcome of collective action is harmful.

4.4 General discussion

Group crimes are commonly performed, but punitive reactions to them are rarely studied. We know from research on solo crimes that punishment depends on two general processes: judgments of causal responsibility for harm and intent to harm

(Cushman, 2008). Drawing on two complementary and well-established methods for dissociating causal and intent-based processes, we studied how the two responds to collective violations.

In Experiment 1, a reduction in the punishment of group crimes was attributable to the causal process of moral judgment—a diffusion of causal responsibility. This finding is consistent with discounting theories which argue that assigning punishment follows from a causal attribution of harmful outcomes, whereby having more than one sufficient cause results in lower responsibility assigned to (Gerstenberg & Lagnado, 2010; Halpern, 2016; Kelley, 1987; 1973; Lagnado et al., 2013; Morris & Larrick, 1995; Shaver, 1985). In contrast, we found no reduction in punishment attributable to the intent-based process of moral judgment.

Two different methods provided convergent evidence for the dissociation between causal and intent-based contributions to the judgment of group crimes. First, we found that accidental harm-doers (who bear causal responsibility for harm without intent) were punished less when in the group compared to when solo. Yet, attempted harm-doers (who act with harmful intent but bear no causal responsibility) were punished identically across these contexts. Second, we found that having an identifiable harmed victim was necessary for the diffusion: victimless violations were punished equivalently across group and solo contexts.

Perhaps for this reason, individuals use group membership to minimize the negative consequences of their actions (e.g., regret and responsibility see El Zein et al., 2020) and protect themselves from the costs of violation such as punishment (El Zein et al., 2019). Our results reinforce that when seeking 'safety in numbers' by acting as part of groups, each perpetrator may expect to mitigate punishment and blame. The diffusion of punishment may, therefore, promote collective transgressions (Bandura et al., 1975; Darley & Latane, 1968; Latane et al., 1979).

Our findings also bear on theories of moral judgment. First, they support the idea that causal and mental-state processes play dissociable roles in moral judgment (Cushman, 2008; Rottman & Young, 2019; Young et al., 2007, 2010). Second, they support important dissociations between the moral judgment processes regarding harmful versus "victimless" crimes (Chakroff et al., 2013, 2017; Dungan et al., 2017; Giner-Sorolla & Chapman, 2017; Rottman & Young, 2019). Third, they reinforce the idea that punishment often involves a "backward-looking" retributive focus on responsibility, rather than a "forwards-looking" focus on rehabilitation, incapacitation, deterrence, etc. (which, we presume, would generally favor treating individual and group actors equivalently). Punishers' own future-oriented self-serving motives and their evolutionary roots, as an alternative hypothesis for punishment diffusion, need further investigation. For instance, punishing joint violators can produce more enemies for the punisher, reducing the motivation for a severe reaction.

Whether the diffusion of punishment and our causal explanation for it extends to other moral domains (e.g., fairness; Graham et al., 2011) is a topic for future

research. Another interesting extension would be to ask whether different kinds of causal structures reliably produce different effects. Our vignettes were intentionally ambiguous about the causal chains and whether multiple agents overdetermined the harmful outcome. Contrasting diffusion in conjunctive moral transgressions (when collaboration is *necessary* for norm violation) with disjunctive ones (when only one individual would *suffice)* would be informative, since attributions of responsibility would generally be higher in the former case than the latter (Gerstenberg & Lagnado, 2010; Kelley, 1973; Lagnado et al., 2013; Morris & Larrick, 1995; Shaver, 1985; Zultan et al., 2012).

Our findings highlight a divergence between legal theories of justice and laypeople's perceptions of apt punishment in cases of severe collective harm. As such, they shed light on the cognitive underpinnings of collective atrocities in the hopes of a more moral future. Whether and how the discrepancy can be addressed may have implications for society at large.

References

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2018). Quantifying Support for the Null Hypothesis in Psychology: *An Empirical Investigation. Advances in Methods and Practices in Psychological Science,* 357–366. doi:10.1177/2515245918773742

Bandura, A., Underwood, B., & Fromson, M. E. (1975). Disinhibition of aggression through diffusion of responsibility and dehumanization of victims. *Journal of Research in Personality, 9*(4), 253–269. https://doi.org/10.1016/0092-6566(75)90001-X

Binder, G., Weisberg, R., & Fissell, B. M. (2017). Capital Punishment of Unintentional Felony Murder. *Social Science Research Network.*

Chakroff, A., Dungan, J., & Young, L. (2013). Harming ourselves and defiling others: what determines a moral domain? *PloS One, 8*(9). https://doi.org/10.1371/journal.pone.0074434

Chakroff, A., Russell, P. S., Piazza, J., & Young, L. (2017). From impure to harmful: Asymmetric expectations about immoral agents. *Journal of Experimental Social Psychology, 69*, 201–209. https://doi.org/10.1016/j.jesp.2016.08.001

Cheng, P. W. (1997). From Covariation to Causation: A Causal Power Theory. *Psychological Review, 104*(2), 367–405. https://doi.org/10.1037/0033-295X.104.2.367

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

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 8*(4), 377–383. https://doi.org/10.1037/h0025589

DeBruine, L. M., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through Data Simulation: *Advances in Methods and Practices in Psychological Science, 4*(1). https://doi.org/10.1177/2515245920965119

Dungan, J. A., Chakroff, A., & Young, L. (2017). The relevance of moral norms in distinct relational contexts: Purity versus harm norms regulate self-directed actions. *PLoS ONE, 12*(3). https://doi.org/10.1371/journal.pone.0173405

Dupuis, M., Meier, E., & Cuneo, F. (2019). Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices. *Behavior Research Methods, 51*(5), 2228–2237. https://doi.org/10.3758/s13428-018-1103-y

El Zein, M., Bahrami, B., & Hertwig, R. (2019). Shared responsibility in collective decisions. *Nature Human Behaviour, 3*(6), 554–559. https://doi.org/10.1038/s41562-019-0596-4

El Zein, M., Seikus, C., De-Wit, L., & Bahrami, B. (2020). Punishing the individual or the group for norm violation. *Wellcome Open Research, 4*, 139. https://doi.org/10.12688/wellcomeopenres.15474.2

Feldman, R. S., & Rosen, F. P. (1978). Diffusion of responsibility in crime, punishment, and other adversity. *Law and Human Behavior, 2*(4), 313–322. https://doi.org/10.1007/BF01038984

Gerstenberg, T., & Lagnado, D. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition, 115*(1), 166–171. https://doi.org/10.1016/j.cognition.2009.12.011

Giner-Sorolla, R., & Chapman, H. A. (2017). Beyond Purity: Moral Disgust Toward Bad Character. *Psychological Science, 28*(1), 80–91. https://doi.org/10.1177/0956797616673193

Graham, Jesse & Haidt, Jonathan & Koleva, Sena & Motyl, Matt & Iyer, Ravi & Wojcik, Sean & Ditto, Peter. (2012). Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. *Advances in Experimental Social Psychology. 47*. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the Moral Domain. *Journal of Personality and Social Psychology, 101*(2), 366–385. https://doi.org/10.1037/a0021847

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*(4), 1600–1615. https://doi.org/10.1037/a0036149

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry, 52*(5), 449–466. https://doi.org/10.1080/00201740903302600

Halpern, J. (2016). *Actual causality*. The MIT Press.

Kelley, H. H. (1987). *Causal schemata and the attribution process*. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (p. 151–174). Lawrence Erlbaum Associates, Inc.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist, 28*(2), 107–128. https://doi.org/10.1037/h0034225

Lagnado, D., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science, 37*(6), 1036–1073. https://doi.org/10.1111/cogs.12054

Latane, B., Williams, K., Harkins, S., Diener, E., Har-Vey, J., Kerr, N., Kidd, R., Levinger, G., Ostrom, T., Petty, R., & Wheeler, L. (1979). Many Hands Make Light the Work: The Causes and Consequences of Social Loafing. *Journal of Personality and Social Psychology, 37*(6), 822–832.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A Theory of Blame. *Psychological Inquiry, 25(*2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

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

Kelley, H. H. (1972). Causal schemata and the attribution process.

Malle, B. F., Moses, L. J., & Baldwin, D. A. (2001). *Intentions and Intentionality: Foundations of Social Cognition.*

McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. L. (2017). Searching for moral dumbfounding: Identifying measurable indicators of moral dumbfounding. Collabra: *Psychology, 3*(1). https://doi.org/10.1525/collabra.79

Moore, M. S. (2009). *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics.* https://doi.org/10.1093/acprof:oso/9780199256860.001.0001

Morris, M. W., & Larrick, R. P. (1995). When One Cause Casts Doubt on Another: A Normative Analysis of Discounting in Causal Attribution. *Psychological Review, 102*(2), 331–355. https://doi.org/10.1037/0033-295X.102.2.331

Rottman, J., & Young, L. (2019). Specks of Dirt and Tons of Pain: Dosage Distinguishes Impurity From Harm. *Psychological Science, 30*(8), 1151–1160. https://doi.org/10.1177/0956797619855382

Schein, C., & Gray, K. (2018). The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. *Personality and Social Psychology Review, 22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Shaver, K. G. (1985). *The Attribution of Blame*. https://doi.org/10.1007/978-1-4612-5094-4

Shultz, T. R., & Wright, K. (1985). Concepts of negligence and intention in the assignment of moral responsibility. *Canadian Journal of Behavioural Science 17*(2), 97–108. https://doi.org/10.1037/h0080138

Vainapel, S., Weisel, O., Zultan, R., & Shalvi, S. (2019). Group moral discount: Diffusing blame when judging group members. *Journal of Behavioral Decision Making, 32*(2), 212–228. https://doi.org/10.1002/bdm.2106

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America, 107*(15), 6753–6758. https://doi.org/10.1073/pnas.0914826107

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America, 104*(20), 8235–8240. https://doi.org/10.1073/pnas.0701408104

Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: causality and counterfactuals in group attributions. *Cognition, 125*(3), 429–440. https://doi.org/10.1016/J.COGNITION.2012.07.014

4.5 Supplementary Material

4.5.1 Experiment 1

4.5.1.1 Outlier detection

To perform tests of outlier detection, we used <outlier_function.r> from **Performance** package (Lüdecke et al., 2021), which uses Mahalanobis distance. We calculated the mean of punishment in all conditions of Outcome by Intention and used a multivariate approach to exclude the inattentive subjects with the alpha threshold is set to 0.025 (corresponding to the 2.5% most extreme observations).

4.5.1.2 Mixed effect models

4.5.1.2.1 Logistic mixed effect model

We performed different ordinal logistic mixed-effect models by using package 'ordinal' (Christensen et al., 2019). In all models we accounted for subjects and Items variability by adding them as random slopes to the model.

4.5.1.2.2 Sanity check

In **model 0** we include Outcome and Intention as main factors for sanity check. Pairwise comparison between conditions for model 0 is presented in Table S 1 as well as Figure 1.

Table S 1.
Pairwise comparison of punishment for different levels of Intention by Outcome

contrast	*estimate*	*SE*	z.ratio	p.value
Innocent Death - Malign Death	-1.8760254	0.0883773	21.227456	**<.0001**
Innocent Death - Innocent Harmless	2.2000354	0.1007347	21.839906	**<.0001**
Innocent Death - Malign Harmless	0.5258203	0.0818840	6.421525	**<.0001**
Malign Death - Innocent Harmless	4.0760608	0.1151210	35.406760	**<.0001**
Malign Death - Malign Harmless	2.4018457	0.0923410	26.010611	**<.0001**
Innocent Harmless - Malign Harmless	-1.6742151	0.0985121	16.995017	**<.0001**

P-value adjustment: Tukey method for comparing a family of 3 estimates

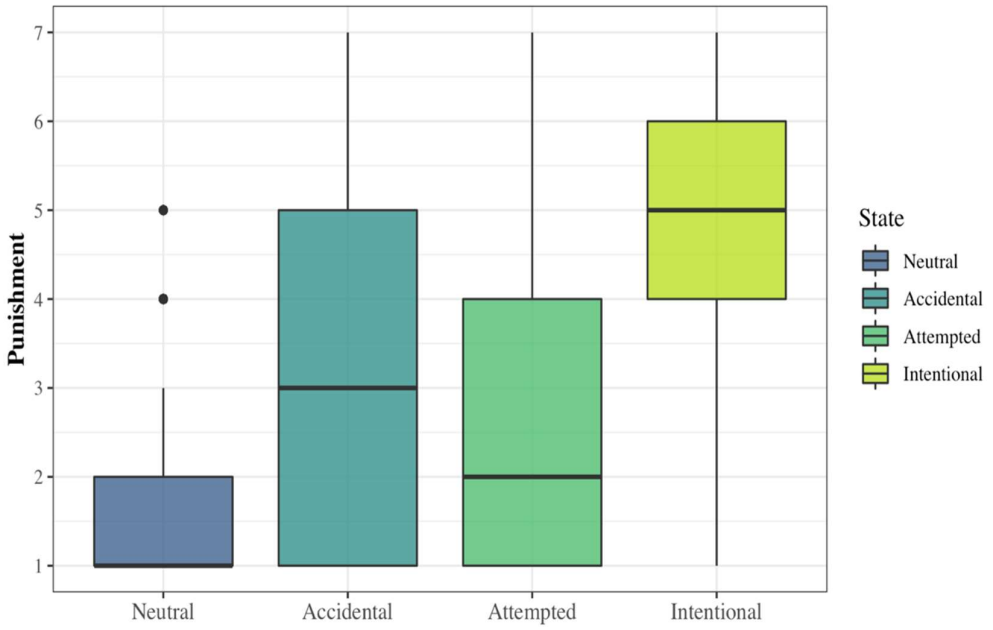


Figure S1. Intentional murder received the highest level of punishment, more than accidental killings and attempted murders, while neutral actions received the lowest amount of punishment.

4.5.1.2.3 Main models

In **model 1**, we added outcome in **model 2** Intention and in **model 3** both to our effect of interest Collectivity as an interaction effect, while accounting for subjects (TurkIDs) and Items as random factors in all models. Model comparisons favored model 3 ($p<.0001$, AIC=12358). Pairwise comparison between conditions in model 3 is presented in Table S 2.

Table S 2.
Pairwise comparison of punishment for different levels of Intention by Outcome by Collectivity.

contrast	***estimate***	***SE***	**z.ratio**	**p.value**
Individual Innocent Death – Joint Innocent Death	0.4363	0.1278	3.4115	**0.0149**
Individual Malign Death –Joint Malign Death	0.5746	0.1290	4.4514	**0.0002**
Individual Innocent Harmless – Joint Innocent Harmless	0.9761	0.1660	-5.8785	**<.0001**
Individual Malign Harmless – Joint Malign Harmless	0.0294	0.1286	0.2291	0.9999

P-value adjustment: Tukey method for comparing a family of 3 estimate

Table S 3.
Pairwise comparison of punishment as in Table S2 but including outliers

contrast	*estimate*	*SE*	z.ratio	p.value
Individual Innocent Death – Joint Innocent Death	0.3500	0.126	2.782	0.0949
Individual Malign Death –Joint Malign Death	0.5746	0.1290	4.4514	**0.0274**
Individual Innocent Harmless – Joint Innocent Harmless	0.9761	0.1660	-5.8785	**<.0001**
Individual Malign Harmless – Joint Malign Harmless	0. 0325	0. 127	0. 2561	1.0000

P-value adjustment: Tukey method for comparing a family of 3 estimates

4.5.1.3 Bayesian mixed model

In addition to the frequentist approach above, we also performed Bayesian mixed effect analysis. This is an important analysis since our effect of interest was not significant in attempted cases. In order to examine this null effect, we used **brms** package in R (Bürkner, 2018), with 5000 iterations, 5 chains and weakly informative prior (model betas drawn from normal distribution; mean = 0 and SD = 1). Result highlights the interaction Collectivity by Outcome (BF_{10}= 1,720,000, b = .71, *SE* = .13, CI. $_{Lower}$ = .13, CI. $_{Upper}$ = .46) more than Intention (BF_{10}= 0.32, *b* = .64, *SE* = .13, CI. $_{Lower}$ = - .439 CI. $_{Upper}$ = .18). Pairwise comparison result of this model is shown in Table S 4.

Table S 4.
Pairwise comparison of Bayesian mixed effect model.

Parameter	*CI*	*CI_low*	*CI_high*	BF_{10}
Individual Innocent Death - Joint Innocent Death	95	0.12	0.52	21.74
Individual Malign Death – Joint Malign Death	95	0.29	0.69	498.32
Individual Innocent Harmless - Joint Innocent Harmless	95	-0.61	-0.22	63.12
Individual Malign Harmless - Joint Malign Harmless	95	-0.17	0.22	0.05

4.5.1.4 Item-based analysis

An item-based analysis was performed to compare the ratings for each item in different conditions. Repeated measure ANOVA test shows a significant difference between conditions across items in intentional and accidental cases. The result of the analysis is shown in Figure S1.

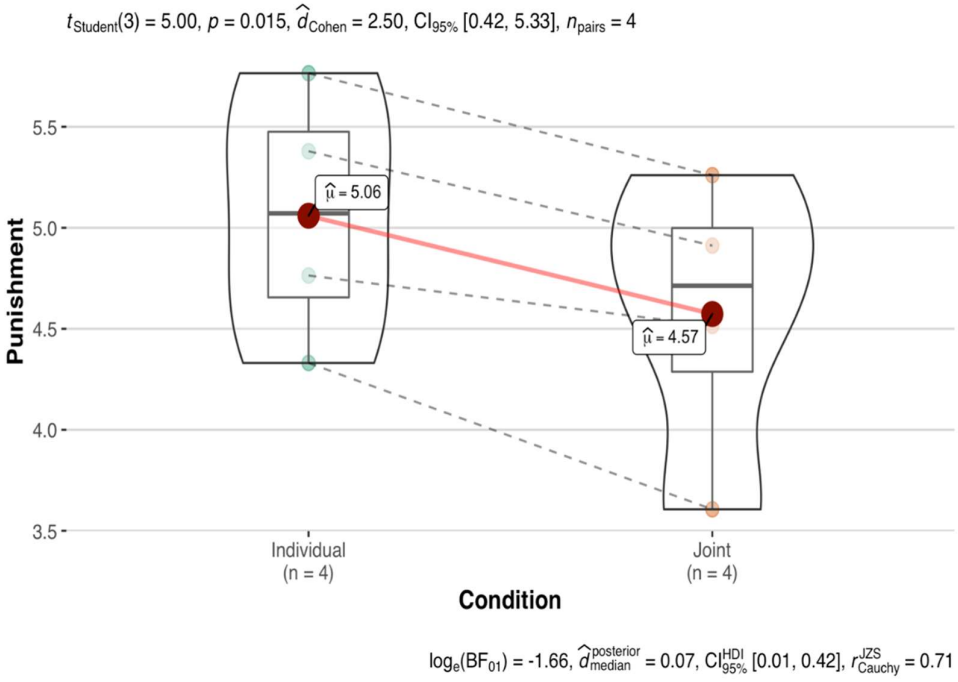


Figure S2. Across 4 items in intentional murder and accidental killings, ANOVA test shows a significant difference between different conditions. Items are rated more punishable in Individual Condition in comparison to Joint Condition. The statistical test result is shown on the figure, using **ggstatssplot** package (Patil, 2021).

4.5.2 Experiment 2

4.5.2.1 Mixed effect models

4.5.2.1.1 Linear mixed effect model experiment

We performed different linear mixed-effect models by using package **LME4** (package lme4; Bates et al., 2015). In all models we accounted for subjects and Items variability by adding them as random slopes to the model.

For experiment 2.1 we used three models, with condition as fixed factor in **model1**, its interaction with Domain **model2** and without the interaction in **model3**. Model comparison showed no difference (Table S 5).

Table S 5.
Model comparison between two models of punishment

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
M0_punishment	5	24930.31	24959.91	-12460.16	24920.31	NA	NA	NA
M2_punishment	6	24930.00	24965.51	-12459.00	24918.00	2.3150673	1	0.1281258
M1_punishment	7	24931.98	24973.41	-12458.99	24917.98	0.0150753	1	0.9022800

For Harmfulness and Grossness, we used a model with Condition and Domain (with interaction term) as fixed factors. The result is shown in Table S 6 and Table S 7, respectively.

Table S 6.
Pairwise comparison of harmfulness for different levels of Domain by Collectivity.

contrast	estimate	SE	df	t.ratio	p.value
Individual Harm - Joint Harm	-0.5315015	1.497927	1325.211229	-0.3548247	0.9846839
Individual Harm - Individual Purity	38.7623136	4.676596	6.321956	8.2885735	**0.0005367**
Individual Harm - Joint Purity	39.5008489	4.774951	6.870517	8.2725147	**0.0003567**
Joint Harm - Individual Purity	39.2938151	4.774964	6.870587	8.2291340	**0.0003686**
Joint Harm - Joint Purity	40.0323503	4.686346	6.374827	8.5423373	**0.0004316**
Individual Purity - Joint Purity	0.7385352	1.497848	1325.015195	0.4930643	0.9606476

P-value adjustment: Tukey method for comparing a family of 3 estimates

For purity we adopted the same approach as above. The result is shown in Table S 7.

Table S 7.
Pairwise comparison of grossness for different levels of Domain by Collectivity.

contrast	estimate	SE	df	t.ratio	p.value
Individual Harm - Joint Harm	0.57	1.7	1150.0	0.33	0.99
Individual Harm - Individual Purity	-41.45	2.4	7.4	-17.06	**0.00**
Individual Harm - Joint Purity	-42.29	2.7	11.9	-15.44	**0.00**
Joint Harm - Individual Purity	-42.03	2.7	11.9	-15.35	**0.00**
Joint Harm - Joint Purity	-42.86	2.5	7.6	-17.50	**0.00**
Individual Purity - Joint Purity	-0.84	1.7	1149.9	-0.48	0.96

P-value adjustment: Tukey method for comparing a family of 3 estimates

We later performed a between subject analysis for two different domains: Harm and Purity. This exploratory analysis was done to check any carryover effect. In doing so, we checked the punishment for first Blocks (either Harm or Purity), averaged across each participant in a separate mixed effect with fixed effect of Domain and Collectivity accounted for Item variability (Table S8). We also used two different mixed effect models using the same method for purity and harm blocks separately, averaged for each participant but accounting for item variability (Table S9).

Table S 8.
Pairwise comparison punishment in harm and purity - between subject analysis of first blocks.

contrast	estimate	SE	df	t.ratio	p.value
Individual Harm - Joint Harm	5.90	2.34	886.88	2.523	0.0571
Individual Harm - Individual Purity	8.03	6.24	6.74	1.288	0.5981
Individual Harm - Joint Purity	9.81	6.27	6.91	1.564	0.4540
Joint Harm - Individual Purity	2.13	6.31	7.05	0.338	0.9857
Joint Harm - Joint Purity	3.91	6.34	7.22	0.616	0.9238
Individual Purity - Joint Purity	1.78	2.43	886.14	0.733	0.8838

P-value adjustment: Tukey method for comparing a family of 3 estimates

Table S 9.
Comparison punishment for harm vs, purity items - between subject analysis of the first blocks to exclude the carryover effect.

contrast	estimate	SE	df	t.ratio	p.value
Individual Harm - Joint Harm	5.93	2.04	466	2.908	0.0038
Individual Purity - Joint Purity	1.78	2.73	421	0.653	0.5144

Degrees-of-freedom method: Kenward-Roger

4.5.2.2 Bayesian mixed models

In addition to the frequentist approach above, we also performed Bayesian mixed effect models for experiment 2.b. This is especially important since our effect of interest was a null result in purity violations. In order to examine this null effect, we used **brms** package in R (Bürkner, 2018), with 5000 iterations, 5 chains and weakly informative prior (model betas drawn from normal distribution; mean = 0 and SD = 1). Result highlighted little evidence to reject the null hypothesis (BF_{10}= 1,720,000, b = .71, SE = .13, CI. $_{Lower}$= .13 CI. $_{Upper}$= .46). Pairwise comparison result of this model is shown in Table S10.

4.5.3 Scenarios used in Experiment 1

4.5.3.1 Individual neutral conditions:

Rafting

1- Monica and her friend are rafting down the Colorado River. Monica's friend is a novice. They stop by the bank to have a swim. Monica's friend starts to swim downstream. The next segment of the river is very tame this year. It is easy to swim through. Because Monica has rafted down this part of the Colorado River last year, she believes the next segment is very tame and gentle. She thinks her friend will be able to enjoy the scenery. Monica silently watches as her friend swims downstream. Monica's friend swims safely down the next segment of the river.

Climbing

2- Stacey and her friend are going rock climbing, and they are going to use harnesses to scale a gigantic cliff. Stacey's friend starts to put on the new harness. The clamp on the new harness is subtly flawed, so the whole harness is unsafe to use. Because Stacey bought the harness from a quality sports store, Stacey believes that the harness is in prime functioning

condition. Stacy's friend wears the harness. Halfway up the cliff, the harness gives way, and her friend plummets to her death.

Swimming

3- Peter is traveling in Africa with a friend. His friend sees a pond and wants to go wading in it because it is very hot. His friend begins to walk toward the pond. The pond is a good place for tourists. It does not contain any disease-carrying organisms. The water is unusually clean, so it is safe to wade in. Peter believes that it is not safe to wade in the pond because he heard stories about Malarial mosquitoes lives in the pond. Peter encourages his friend to wade in the pond. His friend loves the cool water and has a great time splashing around.

Watching the dolphin

4- Ryan is at the zoo with his nephew. They are watching the dolphin show when the nephew complains that his stomach hurts. Ryan's nephew is really sick. He is suffering from severe appendicitis. It may cause the appendix to burst, spreading infection throughout the abdomen. Ryan believes that his nephew's stomach hurts because of severe appendicitis. Ryan thinks that his nephew needs medical attention immediately. Ryan takes his nephew to see the monkeys next. His nephew starts feeling worse and soon blacks out and dies because of severe internal bleeding.

4.5.3.2 Individual accidental conditions:

Rafting

5- Monica and her friend are rafting down the Colorado River. Monica's friend is a novice. They stop by the bank to have a swim. Monica's friend starts to swim downstream. The next segment of the river is very rough and fast this year. It is full of boulders that make it dangerous to swim through. Because Monica has rafted down this part of the Colorado River last year, she believes the next segment is very tame and gentle. She thinks her friend will be able to enjoy the scenery. Monica silently watches as her friend swims downstream. Monica's friend gets thrown by the current and crashes into a boulder and dies.

Climbing

6- Stacey and her friend are going rock climbing, and they are going to use harnesses to scale a gigantic cliff. Stacey's friend starts to put on the new harness. The new harness is a top-of-the-line model, in fine working condition, and completely safe to use. Because the clamp on the harness does not audibly click into place, Stacey believes that the harness is malfunctioning and not safe to use. Stacy's friend wears the harness, scales the cliff safely, and enjoys the exhilarating experience.

Swimming

7- Peter is traveling in Africa with a friend. His friend sees a pond and wants to go wading in it because it is very hot. His friend begins to walk toward the pond. Malarial mosquitoes actually live in the pond. A single bite is enough to create an infection, so the pond is unsafe to wade in. Peter believes that it is not safe to wade in the pond because he heard stories

about Malarial mosquitoes lives in the pond. Peter encourages his friend to wade in the pond. His friend is bitten by several mosquitoes and contracts malaria, which leads to his death.

Watching the dolphin show

8- Ryan is at the zoo with his nephew. They are watching the dolphin show when the nephew complains that his stomach hurts. Ryan's nephew is really fine. His stomach sometimes hurts when he eats too much junk food, as on that day, but he usually feels a lot better after an hour or so. Ryan believes that his nephew's stomach hurts because he ate too much cotton candy and fried dough that afternoon. Ryan thinks his nephew just needs to walk it off. Ryan takes his nephew to see the monkeys next. His nephew starts feeling better. They end up seeing nearly all the exhibits at the zoo.

4.5.3.3 Individual attempted conditions:

Rafting

9- Monica and her friend are rafting down the Colorado River. Monica's friend is a novice. They stop by the bank to have a swim. Monica's friend starts to swim downstream. The next segment of the river is very tame this year. It is easy to swim through. Because Monica has rafted down this part of the Colorado River last year, she believes that the next segment is very rough and dangerous. she thinks that the current will be too strong for her friend. Monica silently watches as her friend swims downstream. Monica's friend swims safely down the next segment of the river.

Climbing

10- Stacey and her friend are going rock climbing, and they are going to use harnesses to scale a gigantic cliff. Stacey's friend starts to put on the new harness. The clamp on the new harness is subtly flawed, so the whole harness is incredibly unsafe to use. Because the clamp on the harness does not audibly click into place, Stacey believes that the harness is malfunctioning and not safe to use. Stacy's friend wears the harness. Halfway up the cliff, the harness gives way, and her friend plummets to her death.

Swimming

11- Peter is traveling in Africa with a friend. His friend sees a pond and wants to go wading in it because it is very hot. His friend begins to walk toward the pond. The pond is a good place for tourists. It does not contain any disease-carrying organisms. The water is unusually clean, so it is safe to wade in. Peter believes that it is safe to wade in the pond because other tourists around them are doing it too and are obviously having fun. Peter encourages his friend to wade in the pond. His friend loves the cool water and has a great time splashing around.

Watching the dolphin show

12 - Ryan is at the zoo with his nephew. They are watching the dolphin show when the nephew complains that his stomach hurts. Ryan's nephew is really sick. He is suffering from severe appendicitis. It may cause the

appendix to burst, spreading infection throughout the abdomen. Ryan believes that his nephew's stomach hurts because he ate too much cotton candy and fried dough that afternoon. Ryan thinks his nephew just needs to walk it off. Ryan takes his nephew to see the monkeys next. His nephew starts feeling worse and soon blacks out and dies because of severe internal bleeding.

4.5.3.4 Individual intentional conditions:

Rafting

13- Monica and her friend are rafting down the Colorado River. Monica's friend is a novice. They stop by the bank to have a swim. Monica's friend starts to swim downstream. The next segment of the river is very rough and fast this year. It is full of boulders that make it dangerous to swim through. Because Monica has rafted down this part of the Colorado River last year, she believes that the next segment is very rough and dangerous. She thinks that the current will be too strong for her friend. Monica silently watches as her friend swims downstream. Monica's friend gets thrown by the current and crashes into a boulder and dies.

Climbing

14-Stacey and her friend are going rock climbing, and they are going to use harnesses to scale a gigantic cliff. Stacey's friend starts to put on the new harness. The new harness is a top-of-the-line model, in fine working condition, and completely safe to use. Because Stacey bought the harness from a quality sports store, Stacey believes that the harness is in prime functioning condition. Stacy's friend wears the harness, scales the cliff safely, and enjoys the exhilarating experience.

Swimming

15 - Peter is traveling in Africa with a friend. His friend sees a pond and wants to go wading in it because it is very hot. His friend begins to walk toward the pond. Malarial mosquitoes actually live in the pond. A single bite is enough to create an infection, so the pond is unsafe to wade in. Peter believes that it is safe to wade in the pond because other tourists around them are doing it too and are obviously having fun. Peter encourages his friend to wade in the pond. His friend is bitten by several mosquitoes and contracts malaria, which leads to his death.

Watching the dolphin show

16- Ryan is at the zoo with his nephew. They are watching the dolphin show when the nephew complains that his stomach hurts. Ryan's nephew is really fine. His stomach sometimes hurts when he eats too much junk food, as on that day, but he usually feels a lot better after an hour or so. Ryan believes that his nephew's stomach hurts because of severe appendicitis. Ryan thinks that his nephew needs medical attention immediately. Ryan takes his nephew to see the monkeys next. His nephew starts feeling better. They end up seeing nearly all the exhibits at the zoo.

4.5.3.5 Group neutral conditions:

Rafting

17- Monica, Kate, Josh, and their friend Tom are rafting down the Colorado River. Tom is a novice. They stop by the bank to have a swim. Tom starts to swim downstream. The next segment of the river is very tame this year. It is very easy to swim through. Because Monica, Kate, and Josh have rafted down this part of the Colorado River last year together, they believe the next segment is very tame and gentle. They think Tom will be able to enjoy the scenery. They all silently watch as Tom swims downstream. Tom swims safely down the next segment of the river.

Climbing

18 - Stacey, Anita, James, and Kate are going rock climbing, and they are going to use harnesses to scale a gigantic cliff. Kate starts to put on the new harness. The clamp on the new harness is actually subtly flawed, so the whole harness is incredibly unsafe to use. Because Stacey, Anita, James bought the harness from a quality sports store together, they all believe that the harness is in prime functioning condition. Kate wears the harness. Halfway up the cliff, the harness gives way, and Kate plummets to her death.

Swimming

19 - Peter, Sophia, David, and Anna are traveling in Africa. Anna sees a pond and wants to go wading in it because it is very hot. She begins to walk toward the pond. A pond is a good place for tourists. It does not contain any disease-carrying organisms. The water is unusually clean, so it is safe to wade in. Peter, Sophia, David all believe that it is not safe to wade in the pond because they have heard stories about the Malarial mosquitoes in the pond. Peter, Sophia, and David encourage Anna to wade in the pond. Anna loves the cool water and has a great time splashing around.

Watching the dolphin show

20 - Ryan, Emma, and Joe are at the zoo with their nephew. They are watching the dolphin show when the nephew complains that his stomach hurts. Their nephew is really sick. He is suffering from severe appendicitis and. It may cause the appendix to burst, spreading infection throughout the abdomen. Ryan, Sophia, and Joe all believe that their nephew's stomach hurts because of severe appendicitis. Ryan, James, and Joe all think that their nephew needs medical attention immediately. Ryan, Sophia, and Joe take their nephew to see the monkeys next. Their nephew starts feeling worse and soon blacks out and dies because of severe internal inflammation.

4.5.3.6 Group accidental conditions:

Rafting

21 - Monica, Kate, Josh, and their friend Tom are rafting down the Colorado River. Tom is a novice. They stop by the bank to have a swim. Tom starts to swim downstream. The next segment of the river is very rough and fast this year. It is full of boulders that make it dangerous to

swim through. Because Monica, Kate, and Josh have rafted down this part of the Colorado River last year together, they all believe the next segment is very tame and gentle. They think Tom will be able to enjoy the scenery. They all silently watch as Tom swims downstream. Tom gets thrown by the current and crashes into a gigantic boulder and dies.

Climbing

22- Stacey, Anita, James, and Kate are going rock climbing, and they are going to use harnesses to scale a gigantic cliff. Kate starts to put on the new harness. The new harness is a top-of-the-line model, in fine working condition, and completely safe to use. Because the clamp on the harness does not audibly click into place, Stacey, Anita, and James all believe that the harness is malfunctioning and not safe to use. Kate wears the harness, scales the cliff safely, and enjoys the exhilarating experience.

Swimming

23-Peter, Sophia, David, and Anna are traveling in Africa with a friend. Anna sees a pond and wants to go wading in it because it is very hot. Anna begins to walk toward the pond. Malarial mosquitoes actually live in the pond. A single bite is enough to create an infection, so the pond is unsafe to wade in. Peter, Sophia, and David believe that it is not safe to wade in the pond because they have heard stories about the Malarial mosquitoes in the pond. Peter, Sophia, and David encourage Anna to wade in the pond. Anna is bitten by several mosquitoes and contracts malaria, which leads to her death.

Watching the dolphin show

24- Ryan, Emma, and Joe are at the zoo with their nephew. They are watching the dolphin show when the nephew complains that his stomach hurts. Their nephew is really fine. His stomach sometimes hurts when he eats too much junk food, as on that day, but he usually feels a lot better after an hour or so. Ryan, Emma, and Joe believe that their nephew's stomach hurts because he ate too much cotton candy and fried dough that afternoon. Ryan, Emma, and Joe think their nephew just needs to walk it off. Ryan, Emma, and Joe take their nephew to see the monkeys next. Their nephew starts feeling better. They end up seeing nearly all the exhibits at the zoo.

4.5.3.7 Group attempted conditions:

Rafting

25- Monica, Kate, Josh, and their friend Tom are rafting down the Colorado River. Tom is a novice. They stop by the bank to have a swim. Tom starts to swim downstream. The next segment of the river is very tame this year. It is very easy to swim through. Because Monica, Kate, and Josh have rafted down this part of the Colorado River last year together, they all believe that the next segment is very rough and dangerous. They think that the current will be too strong for Tom. They all silently watch as Tom swims downstream. Tom swims safely down the next segment of the river.

Climbing

26- Stacey, Anita, James, and Kate are going rock climbing, and they are going to use harnesses to scale a gigantic cliff. Kate starts to put on the new harness. The clamp on the new harness is actually subtly flawed, so the whole harness is incredibly unsafe to use. Because the clamp on the harness does not audibly click into place, Stacey, Anita, and James all believe that the harness is malfunctioning and not safe to use. Kate wears the harness. Halfway up the cliff, the harness gives way, and Kate plummets to her death.

Swimming

27- Peter, Sophia, David, and Anna are traveling in Africa. Anna sees a pond and wants to go wading in it because it is very hot. She begins to walk toward the pond. The pond is a good place for tourists. It does not contain any disease-carrying organisms. The water is unusually clean, so it is safe to wade in. Peter, Sophia, and David all believe that it is safe to wade in the pond because other tourists around them are doing it too and are obviously having fun. Peter, Sophia, and David encourage Anna to wade in the pond. Anna loves the cool water and has a great time splashing around.

Watching the dolphin show

28- Ryan, Emma, and Joe are at the zoo with their nephew. They are watching the dolphin show when the nephew complains that his stomach hurts. Their nephew is really sick. He is suffering from severe appendicitis, and it may cause the appendix to burst, spreading infection throughout the abdomen. Ryan, Emma, and Joe believe that their nephew's stomach hurts because he ate too much cotton candy and fried dough that afternoon. Ryan, Emma, and Joe think their nephew just needs to walk it off. Ryan, Emma, and Joe take their nephew to see the monkeys next. Their nephew starts feeling worse and soon blacks out and dies because of severe internal inflammation.

4.5.3.8 Group intentional conditions:

Rafting

29- Monica, Kate, Josh, and their friend Tom are rafting down the Colorado River. Tom is a novice. They stop by the bank to have a swim. Tom starts to swim downstream. The next segment of the river is very rough and fast this year. It is full of boulders that make it dangerous to swim through. Because Monica, Kate, and Josh have rafted down this part of the Colorado River before, they all believe that the next segment is very rough and dangerous. They think that the current will be too strong for Tom. They all silently watch as Tom swims downstream. Tom gets thrown by the current and crashes into a gigantic boulder and dies.

Climbing

30- Stacey, Anita, James, and Kate are going rock climbing, and they are going to use harnesses to scale a gigantic cliff. Kate starts to put on the new harness. The new harness is a top-of-the-line model, in fine working

condition, and completely safe to use. Because Stacey, Anita, and James bought the harness together from a quality sports store, they all believe that the harness is in prime functioning condition. Kate wears the harness, scales the cliff safely, and enjoys the exhilarating experience.

Swimming

31- Peter, Sophia, David, and Anna are traveling in Africa. Anna sees a pond and wants to go wading in it because it is very hot. She begins to walk toward the pond. Malarial mosquitoes actually live in the pond. A single bite is enough to create an infection, so the pond is unsafe to wade in. Peter, Sophia, and David believe that it is safe to wade in the pond because other tourists around them are doing it too and are obviously having fun. Peter, Sophia, and David encourage Anna to wade in the pond. Anna is bitten by several mosquitoes and contracts malaria, which leads to her death.

Watching the dolphin show

32- Ryan, Emma, and Joe are at the zoo with their nephew. They are watching the dolphin show when the nephew complains that his stomach hurts. Their nephew is really fine. His stomach sometimes hurts when he eats too much junk food, as on that day, but he usually feels a lot better after an hour or so. Ryan, Emma, and Joe believe that their nephew's stomach hurts because of severe appendicitis. Ryan, Emma, and Joe all think that their nephew needs medical attention immediately. Ryan, Emma, and Joe take their nephew to see the monkeys next. Their nephew starts feeling better. They end up seeing nearly all the exhibits at the zoo.

4.5.4 Scenarios used in Experiment 2

Domain: Harm

Zoo- Group

Ryan, Emma, and Joe are siblings and at the zoo with their nephew. They are watching the dolphin show when their nephew complains that his stomach hurts. Their nephew is sick. He is suffering from severe appendicitis, and it may cause the appendix to burst, spreading the infection throughout the abdomen. Ryan, Emma, and Joe know that their nephew needs immediate medical attention. Still, they take him to see the monkeys instead of going to a hospital, although they know it is dangerous. Their nephew starts feeling worse and soon blacks out because of severe internal inflammation.

Zoo-Alone

Ryan is at the zoo with his nephew. They are watching the dolphin show when his nephew complains that his stomach hurts. His nephew is sick. He is suffering from severe appendicitis, and it may cause the appendix to burst, spreading the infection throughout the abdomen. Ryan knows that his nephew needs immediate medical attention. Still, he

takes him to see the monkeys instead of going to the hospital, although he knows it is dangerous. His nephew starts feeling worse and soon blacks out because of severe internal inflammation.

Rock Climbing- Group

Stacey, Anita, James, and Kate are all friends. One day they decide to go rock climbing. They are going to use new harnesses to scale a gigantic cliff. Kate starts to put on one of the harnesses. The clamp on the harness is actually flawed, so the whole harness is incredibly unsafe to use. Because the clamp on the harness does not audibly click into place, Stacey, James, and Anita realize that the harness is malfunctioning and not safe to use. Still, they strap Kate into the harness and ask Kate to go first, although they know it is dangerous. Partway up the cliff, the harness gives way, causing Kate to fall and break both her legs.

Rock Climbing- Alone

Stacey and Kate are friends and decide to go rock climbing. They are going to use harnesses to scale a gigantic cliff. Kate starts to put on one of the harnesses. The clamp on the harness is actually flawed, so the whole harness is incredibly unsafe to use. Because the clamp on the harness does not audibly click into place, Stacey realizes that the harness is malfunctioning and not safe to use. Still, she straps Kate into the harness and asks Kate to go first, although she knows it is dangerous. Partway up the cliff, the harness gives way, causing Kate to fall and break both her legs.

Party - Group

Charles, Grace, and Dan are classmates. They are at a party where people are drinking lots of beer. They are having fun mingling when they notice a friend on a couch with his eyes closed. Their friend had just failed his examinations and was drinking heavily to take his mind off them. He has passed out now. Charles, Grace, and Dan believe that their friend drank too much and needs immediate help. Charles, Grace, and Dan decide to hide their friend under some pillows, so nobody sees him, although they know it is dangerous for him. Their friend gets severe alcohol poisoning and remains in a coma for a day.

Party – Alone

Charles is at a party with a friend, where people are drinking lots of beer. He is having fun mingling when he notices his friend on a couch with his eyes closed. His friend had just failed his examinations and was drinking heavily to take his mind off them. He has passed out now. Charles believes that his friend drank too much and needs immediate help. Charles decides to hide his friend under some pillows, so nobody sees him, although he knows it is dangerous for him. His friend gets severe alcohol poisoning and remains in a coma for a day.

Rafting - Group

Susan, Steve, and Jane are gym buddies. They are rafting down the Colorado River with Tom, who is a novice. They stop by the bank to swim. Tom starts to swim downstream. The next segment of the river is very

rough and fast this year. It is full of gigantic boulders that make it dangerous to swim through. Susan, Steve, and Jane have rafted down this part of the river before; they believe that the next segment is very rough. They think that the current will be too strong for Tom. Susan, Steve, and Jane still decide to encourage Tom to swim in the river. They watch as their friend swims downstream, knowing it is dangerous for him. Tom gets thrown by the current and crashes into a gigantic boulder, and breaks his head.

Rafting – Alone

Susan is rafting down the Colorado River with Tom, who is a novice. They stop by the bank to swim. Tom starts to swim downstream. The next segment of the river is very rough and fast this year. It is full of gigantic boulders that make it dangerous to swim through. Susan has rafted down this part of the river before; she believes that the next segment is very rough. She thinks that the current will be too strong for Tom. Susan still decides to encourage Tom to swim in the river. She watches as her friend swims downstream, knowing it is dangerous for him. Tom gets thrown by the current and crashes into a gigantic boulder, and breaks his head.

Domain: Purity

Corpse - Group

Linda, Moira, and Regina are all colleagues and work in a morgue. They all have had wild sexual fantasies about touching corpses. One night, Linda, Moira, and Regina decide to enter the morgue late at night. They touch the genitals of a corpse in the dark to their hearts' desire, making sure they cannot be seen, not even by each other. After that, they ensure that the corpse is intact and the morgue is exactly as it was before and leave.

Corpse - Alone

Linda works in a morgue. She has had wild sexual fantasies about touching corpses. One night, Linda decides to enter the morgue late at night. She touches the genitals of a corpse in the dark to her heart's desire, making sure she cannot be seen. After that, she ensures that the corpse is intact and the morgue is exactly as it was before and leaves.

Singer - Group

Dan, Ray, and Carl are friends. Their favorite singer has died and has been buried in a nearby cemetery. They have all had wild fantasies about the singer. One night according to a previous plan, Dan, Ray, and Carl enter the cemetery late at night. They masturbate over the singer's grave, making sure they cannot be seen, not even by themselves. After that, they make sure that the grave is clean and exactly as it was before and leaves.

Singer – Alone

Dan's favorite singer has died and has been buried in a nearby cemetery. Dan has always had wild fantasies about the singer. One night, according to a previous plan, Dan enters the cemetery late at night. He masturbates over the singer's grave, making sure he cannot be seen, not even by himself.

After that, he makes sure that the grave is clean and exactly as it was before and leaves.

Human- Group

Anne, Monica, and Janet are colleagues. They work in a research institute where people donate their organs for scientific purposes after their death. They have been keen to eat human flesh once for the sake of experiencing its taste. One night, Anne, Monica, and Janet decide to enter the lab late at night. They cut from one of the corpses a body part that had already been studied and is no longer useful for any scientific purpose. After they make sure it is completely free of any contamination, they bring the small piece of human flesh home, cook it and eat it with bread and wine. They enjoy it, and nothing bad happens later.

Human – Alone

Anne works in a research institute where people donate their organs for scientific purposes after their death. She has been keen to eat human flesh once for the sake of experiencing its taste. One night, Anne decides to enter the lab late at night and cut from one of the corpses from a body part that had already been studied and is no longer useful for any scientific purpose. After she makes sure it is completely free of any contamination, she brings the small piece of human flesh home, cooks it, and eats it with bread and wine. She enjoys it, and nothing bad happens later.

Ash – Group

Joe, Eli, and Liz are young siblings. Even though their grandmother is not alive anymore, they still hate her. Before her death, their grandmother asked them to have her remains cremated after her death. She also asked them to keep the ashes in a beautiful urn. Joe, Eli, and Liz do exactly what their grandmother asked them to do, but also, they add a large amount of dog faeces into an urn of their grandmother's ashes.

Ash – Alone

Joe is a young adult. Even though his grandmother is not alive anymore, he still hates her. Before her death, her grandmother asked him to have her remains cremated after her death. She also asked him to keep her ashes in a beautiful urn. Joe does exactly what his grandmother asked him to do, but also, he adds a large amount of dog faeces into the urn of his grandmother's ashes.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1 - 48. doi:http://dx.doi.org/10.18637/jss.v067.i01

Bürkner, P.-C. (2018). Advanced {Bayesian} Multilevel Modeling with the {R} Package {brms}. *The R Journal*, *10*(1), 395–411.

Christensen, R. H. B. (2019). "ordinal—Regression Models for Ordinal Data." R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

Lüdecke D, Ben-Shachar M, Patil I, Waggoner P, & Makowski D (2021). "Assessment, Testing and Comparison of Statistical Models using R." *Journal of Open Source Software*, *6*(59), 3112. doi: 10.31234/osf.io/vtq8f.

Patil, I. (2018). *ggstatsplot: "ggplot2" Based Plots with Statistical Details*. https://doi.org/10.5281/zenodo.2074621

Chapter 5. Conclusion

In this thesis, I aimed at tackling the collective aspect of cognition, examining different moral decisions and judgments in (and for) groups. After historically approaching the burgeoning field of moral decisions and judgments in groups (chapter 1), which was conceptually clarified under the unifying term of *collective turn* in moral cognition (chapter 2), I empirically examined collective moral judgments in small interactive groups (chapter 3) and in individuals seeking to punish collective vs. solo moral violations (chapter 4) across several experiments.

In the first chapter, I showed how this thesis would be tied to a family of recent collectivist views that aim at exploring the process and the content of cognition in joint tasks. Introducing the objectives and scopes of the thesis, I argued that the collective dimension was historically entrenched to understanding morality, attracted relatively little attention later, but recently became central to the study of moral cognition, once again.

In the second chapter, I developed a theoretical framework that described the mechanisms of moral decisions and judgments in groups. I argued that collective moral decisions and judgments could be different from both moral decisions made by individuals and non-moral decisions made by collectives. I discussed why these two aspects could make collective moral cognition a distinct target of investigations.

In particular, I reviewed the social dynamics that surface in collective moral settings, making group decisions and judgments different from the statistical aggregates of individual moral opinions. Drawing on recent work in moral philosophy, moral psychology, and cognitive neuroscience, I argued that emotions, automatic and deliberative processes, virtue-signaling, and diffusion of responsibility could be modulated in collective moral contexts.

Next, reviewing the psychology of meta-ethical commitments (i.e., the degree to which we see moral matters as solid mind-independent facts or personal mind-dependent opinions), I argued that moral issues would not be construed as facts or opinions *uniformly* across individuals. In fact, meta-ethical commitments might vary depending on the topic, individual differences, and social factors, leading to specific challenges in collective moral discussions. Reviewing these challenges, I

argued why this metaethical heterogeneousness could make collective moral tasks different from collective non-moral tasks. Proposing that the machinery of moral cognition should be tackled at the collective level, this chapter ended with the implications of collective moral discussions in real-life issues. Certain difficulties when discussing moral issues at the group level were also discussed.

In the third chapter, building upon the group dynamics suggested in chapter two, I proposed and tested three hypotheses concerning collective judgments vs. individual judgments. The analysis showed that participants were more utilitarian *collectively* after short social discussions. In other words, groups, after short discussions, found it more acceptable to violate moral norms that increased the total good, compared to conditions in which the participants evaluated these violations individually. Hence, the collective consensuses on these moral dilemmas were more utilitarian than the individual moral judgments aggerated statistically.

When participants were asked to provide their individual judgments for the second time, but this time after the discussions, their private judgments remained unchanged. Put differently, people did not keep the collective utilitarian boost in their private judgments. This indicated 'no change of mind' in private moral evaluations of moral dilemmas. As the participants did not change their private moral judgments after the discussions, the collective utilitarian boost in groups could not be the result of social deliberation (utilitarian boost via social deliberation would have affected the second individual judgments). This observation was more consistent with the hypothesis that suggested stress would be reduced in collective settings: when in groups, the emotional burden of moral judgments could be shared, potentially resulting in higher utilitarian scores only in groups.

In the last chapter, I studied moral violations to investigate how people seek to punish collective moral violations compared to solo moral violations. Accounting for *intention*, *outcome*, and moral *domain*, across three experiments, I showed that people punished individual characters in accidental and intentional harmful violations less when they committed these violations jointly with others. Put it differently, characters who violated moral norms with others received less punishment than lone perpetrators in harmful actions.

Consistent with a pre-registered hypothesis, further analysis showed that people punished individuals *less* in collective moral violations (compared to solo violations) only when these violations entailed harmful *outcomes*. In other words, the reduction of punishment was observed when the collective actions were harmful to a (hypothetical) victim. By contrast, when this criterion was not satisfied (i.e., in victimless actions), the deserved punishment level for solo vs. collective violations remained unchanged. In particular, as predicted, when moral violations did not lead to harmful outcomes (e.g., eating human flesh collectively) or when they were intended to be harmful but failed (e.g., failed attempts of group murders), no reduction of punishment was observed in collective norm violations.

I argued how this result could be consistent with causal accounts of punishment attribution, which proposed discounted punishments in collective harmful transgressions. I explained how this finding could be explained by causal discounting in responsibility attribution. Using 'punishment' as a proxy of moral responsibility and extending the diffusion of responsibility to attributions thereof, the hypothesis - that a person who violates a norm with others could be perceived as less responsible, hence deserving less punishment - was tested more directly. The analysis explained how the interaction between intention, outcome, and domain in group violations could support the causal attribution of responsibility. Having a causal attribution approach, I argued why we did not see this effect in collective victimless moral violations such as cannibalism or even failed attempts of murders.

The contribution of this thesis to the field of cognitive science is fourfold. First, revitalizing the collectivist approach in moral decisions and judgments, this thesis ties the collective empirical studies to pioneering theoretical work in collective morality. Second, it extends the recent collectivist views in non-moral joint tasks (e.g., joint attention) to collective moral tasks (e.g., joint moral judgments – chapter 3) while proposing that joint moral tasks are qualitatively and cognitively different from individual moral tasks (chapter 2). Third, it connects the recently emerging trend in group-based morality to interactive and collective morality, when the *agents* of the moral decision or judgments are a group (chapter 2 and 3). Finally, it proposes a mechanistic explanation for the reduction of punishment in collective immoral actions. Hence, this thesis specified three different aspects of the *collective* dimension in moral cognition: collective moral decisions (chapter 2), collective moral judgments (chapter 3), and collective moral actions (chapter 4).

All chapters of this thesis were descriptive, aimed at showing how moral judgments in collectives (and for collectives) *work*, rather than showing how they *should* work. Given the interactive nature of moral judgments and decisions in real life, the broader ramifications of this work for scientific research are that we need to treat collective moral cognition as an independent target of investigation in order to have a comprehensive understanding of the machinery of interactive moral decisions and judgments in real life.

Accordingly, collective moral cognition is not only central to understanding the mechanics of moral cognition but also to understanding real-life collective moral decisions and judgments. The insight from research in moral decisions and judgments for and in groups could then inform moral theories in philosophy and social sciences to propose solutions for real-life collective moral issues.

Afterthought

Given that collective moral judgments and decisions often have profound political and societal consequences, the collective approach in the moral domain seems critical. Positive solutions to today's morally relevant challenges – from environmental problems to moral conflicts– cannot be anything other than the product of collective endeavors. Although the primary purpose of this thesis was to understand moral decisions and judgments in collective contexts, studies in collective moral cognition might also be beneficial in understanding real-life moral challenges.

For instance, at the time of writing this thesis, policymakers face numerous collective morally relevant challenges during the COVID19 pandemic, such as the decisions to vaccinate citizens to save their lives, while vaccinations might have adverse side-effects in minority groups; to offer health-care facilities to younger populations but sacrifice older individuals when the resources are scarce, or to mitigate the animal rights to accelerate the process of the vaccine development. These real-life moral dilemmas are not solved by single individuals in isolation. Their solutions are the outcome of several discussions in medical, political, and ethics committees. Though not purely moral, such decisions have solid moral components and need certain collective moral considerations before any further applications.

Collective moral decisions are not confined to extreme epochs such as pandemics. Similarly, the critical challenges we humans face today comprise an important share of our social and political sphere while entailing solid moral components. In many of these cases, our existence may depend on the solutions we may find *collectively* to resolve them. In this light, the lack of a scientific understanding of the mechanism underlying the collective aspects of moral decisions and judgments may result in high human, environmental, and existential costs. While I acknowledge the oversimplicity of the normative solutions that can be proposed to solve social challenges via empirical interventions, overlooking the complexity of the topic, I believe that *understanding* the collective aspects of moral cognition, which was the primary goal of this thesis, could also contribute to *understanding* the mechanisms of such morally relevant social challenges we human face today.

List of publications

Published:

Li, L., Kumano, S., **Keshmirian, A**., Bahrami, B., Lee, J, Wright, N. (2018). Parsing cultural impacts on regret and risk in Iran, China and the United Kingdom. *Scientific Reports,* Vol. 8, 13862

Kruschwitz, J., Kausch, A., Brovkin, A., **Keshmirian, A.**, Walter, H. (2019). Self-control is linked to interoceptive inference: craving regulation and the prediction of aversive interoceptive states induced with inspiratory breathing load. *Cognition*. Vol 193,104028.

Under review:

Keshmirian, A., Bahrami, B., Deroy, O. (2021). Many Heads Are More Utilitarian Than One. (preprint).

Keshmirian, A., Bonicalzi, S. (in-prep). Collective Turn in Moral Cognition (Submited).

In-preparation:

Keshmirian, A., Hematian, B., Bahrami, B., Deroy, O., Cushman, F. (in-prep). Diffusion of punishment in collective norm violations (working title).

Eidesstattliche Versicherung/Affidavit

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation "Moral Decisions In (And For) Groups: A Collective Approach" selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation "Moral Decisions In (And For) Groups: A Collective Approach" is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

München, den 05 May 2021 _____

Munich, date 05 May 2021 Anita Keshmirian

Authors Contribution

Manuscript 1 - Chapter 2

AK: Generation of the Opinion Concept, Conceptualization, Writing- Original draft, Writing-Review & Editing.

SK: Conceptualization, Writing- Original Draft, Writing-Review & Editing.

Manuscript 2 - Chapter 3

AK: Conceptualization, Methodology, Formal Analysis, Investigation, Data Collection, Visualization, Writing-Original Draft.

BB: Conceptualization, Methodology, Writing-Review & Editing, Supervision, Funding.

OD: Conceptualization, Methodology, Writing-Review, Supervision, Funding.

Manuscript 3 - Chapter 4

AK: Generation of the Study Concept, Methodology, Material Preparation, Investigation, Data Collection, Conceptualization, Formal Analysis, Visualization, Writing-Original Draft.

BH: Material Preparation, Data Collection, Writing-Review.

BB: Methodology, Writing-Review, Supervision.

OD: Material Preparation, Methodology, Writing- Review, Supervision, Funding

FC: Methodology, Writing-Review & Editing, Supervision, Funding.

- - - - - - - - - - - - -

Anita Keshmirian
Munich, 05 May 2021