The Role of Causal Representations in Moral Judgment

Dissertation

Zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades

"Doctor rerum naturalium"

der Georg-August-Universität Göttingen

im Promotionsprogramm Behavior and Cognition (BeCog)

der Georg-August University School of Science (GAUSS)

vorgelegt von

Neele Engelmann

aus Cuxhaven

Göttingen, 2022

Betreuungsausschuss

Prof. Dr. Michael R. Waldmann, Kognitionswissenschaft und Entscheidungspsychologie, Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Prof. Dr. Hannes Rakoczy, Kognitive Entwicklungspsychologie, Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Prof. Dr. Julia Fischer, Kognitive Ethologie, Deutsches Primatenzentrum, Universität Göttingen

Mitglieder der Prüfungskommission

Referent: Prof. Dr. Michael R. Waldmann, Kognitionswissenschaft und Entscheidungspsychologie, Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Korreferent: Prof. Dr. Hannes Rakoczy, Kognitive Entwicklungspsychologie, Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Julia Fischer, Kognitive Ethologie, Deutsches Primatenzentrum, Universität Göttingen

Prof. Dr. Annekathrin Schacht, Affektive Neurowissenschaft und Psychophysiologie, Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Prof. Dr. Nivedita Mani, Psychologie der Sprache, Georg-Elias-Müller-Institut für Psychologie, Universität Göttingen

Prof. Dr. Margarete Boos, Sozial- und Kommunikationspsychologie, Georg-Elias-Müller-

Institut für Psychologie, Universität Göttingen

Tag der mündlichen Prüfung: 29.08.2022

Acknowledgements

Thanks to my advisor, Michael Waldmann, for the excellent support during the realization of this dissertation project. It was often fun, but more importantly, I learned a lot from working with you (hopefully some of it shows). Thanks also to the other members of my Thesis Committee, Julia Fischer and Hannes Rakoczy, for your thoughtful comments during our meetings.

Thanks to Alex Wiegmann for countless helpful discussions, our collaborative projects, and for introducing me to some of the greatest philosophical work of our time (the "Rocky"-series).

Thanks to York Hagmayer, Simon Stephan and all "Quanti-Team"-members along the way for a great co-teaching experience, and thanks to the students in our courses for putting up with so many stats examples about cats. Thanks to Simon for many helpful comments on my work as well.

Thanks to all current and former colleagues at the Department of Cognitive and Decision Sciences and beyond for making this a great place to work: Birgit Bergmann-Bryant, Ronja Demel, Juan Carlos Marulanda Hernández, Jana Samland, Ralf Mayrhofer, and Niels Skovgaard-Olsen. Thanks to all (former) students whose B.Sc. and M.Sc. projects I could supervise.

Thanks to Lara Kirfel for your friendship and support, for hanging out at so many conferences and other places together, and for letting me practice all my talks on you.

Thanks to my fellow Dorothea Schlözer-mentees Karina Meyer, Maja Marcus, and Silke Möbius for accompanying me through the final year of this Ph.D., and thanks to Anne Burkard for being my mentor.

Finally, thanks to my family for your support and for always providing space to recover from work: Frauke Engelmann, Martin Scheschonka, Maruth Ordemann, and Simon Engelmann.

Preliminary Note

The present thesis is a publication-based (cumulative) dissertation. It is based on the following two original research articles that have been published in international peer-reviewed journals:

Engelmann, N., & Waldmann, M. R. (2022a). How causal structure, causal strength, and foreseeability affect moral judgments. *Cognition*, *226*, Article 105167. https://doi.org/10.1016/j.cognition.2022.105167

Engelmann, N., & Waldmann, M. R. (2022b). How to weigh lives. A computational model of moral judgment in multiple-outcome structures. *Cognition*, *218*, Article 104910. https://doi.org/10.1016/j.cognition.2021.104910

In this thesis, I will summarize the main empirical findings from the two articles, along with the relevant theoretical background, and provide an extended discussion. All parts of the dissertation were written by me and assistance of third parties was only accepted if it was scientifically justifiable and acceptable in regards to the examination regulations. All sources have been quoted.

The original articles are reprinted in Appendices A and B. I served as first author in both articles. In particular, I was responsible for (a) developing the theory, (b) designing and conducting the experiments, (c) analyzing and interpreting the data, and (d) writing up and publishing the manuscripts. Both articles have precursors in the form of peer-reviewed proceedings papers that I wrote during my PhD and which have been published by the Cognitive Science Society (Engelmann & Waldmann, 2019, 2021). These proceedings papers can be found in Appendices C and D. Apart from this main project, I also worked on additional projects during my PhD. These include projects on the concept of lying (Viebahn, Wiegmann, Engelmann, & Willemsen, 2021), the moral status of lying vs. misleading (Wiegmann & Engelmann, 2022), the relationship

between moral psychology and practical ethics (Wiegmann & Engelmann, 2020), and causal reasoning (Stephan, Engelmann, & Waldmann, 2021; Hagmayer & Engelmann, 2020). The abstracts of the articles that resulted from these projects are included in Appendices E - I.

Abstract

Morality and causation are deeply intertwined. For instance, the value of anticipated consequences is a crucial input for an action's moral permissibility, and assigning blame or responsibility for outcomes generally requires that a causal link connect the outcome with a potentially blameworthy agent's action. Psychological theories of moral judgment acknowledge this, but an explicit connection to theories of causal reasoning, and to theories of reasoning about outcomes, is missing. In this thesis, I present the results of two research projects that investigated, respectively, how (a) features of the causal relations connecting actions and outcomes, and (b) observers' subjective value of consequences affect moral judgments. In the first project, we found that chain structures connecting actions and harmful outcomes, compared to direct causal relations, can lead to a lower perceived strength of the relation, and thereby to attributions of diminished outcome foreseeability to agents. This explains why moral judgments about actions and agents can be more lenient in chains compared to direct relations. In the second project, we proposed and evaluated a computational model of reasoning about outcome trade-offs in moral scenarios. The model predicts permissibility judgments about actions from observers' subjective utilities of the action's consequences, and it accounted well for participants' judgments in two experiments. I argue that an improved understanding of how features of causal relations and the value of outcomes affect moral judgments would advance any contemporary theory of moral reasoning. The findings presented in this thesis aim to contribute to such an improved understanding. I conclude by discussing how features of causal relations and utilities might be formally integrated in causal representations, and lay out directions for future research.

Zusammenfassung

Moral und Kausalität sind eng verwoben. So ist z.B. der Wert der erwarteten Folgen ein entscheidender Faktor für die moralische Zulässigkeit von Handlungen, und die Zuweisung von Schuld oder Verantwortung für Ereignisse setzt im Allgemeinen voraus, dass eine kausale Verbindung zwischen dem Ereignis und der Handlung einer Person besteht. Psychologische Theorien des moralischen Urteilens erkennen das zwar an, jedoch fehlen explizite Verbindungen zu Theorien des kausalen Denkens und zu Theorien der Bewertung von Folgen. In dieser Arbeit stelle ich die Ergebnisse von zwei Forschungsprojekten vor, in denen untersucht wurde, wie (a) Merkmale der kausalen Beziehungen zwischen Handlungen und Ereignissen und (b) die subjektive Bewertung von Konsequenzen moralische Urteile beeinflussen. Das erste Projekt zeigte, dass die Repräsentation von kausalen Kettenstrukturen zwischen Handlungen und Ereignissen im Vergleich zu direkten Kausalbeziehungen zu einer als geringer wahrgenommenen Stärke der Beziehung und damit zu Zuschreibungen einer geringeren Vorhersehbarkeit der Folgen an die Handelnden führen kann. Das erklärt, warum moralische Urteile bei Ketten nachsichtiger ausfallen können. Im zweiten Projekt wurde ein komputationales Modell des Vergleichs der Werte von Konsequenzen in moralischen Szenarien vorgeschlagen und evaluiert. Das Modell sagt Zulässigkeitsurteile über Handlungen auf der Grundlage des subjektiven Nutzens der Konsequenzen voraus. Die Vorhersagen entsprachen in zwei Experimenten gut den Urteilen der Versuchspersonen. Ich argumentiere, dass ein verbessertes Verständnis davon, wie Merkmale von Kausalbeziehungen und der Wert von Folgen moralische Urteile beeinflussen, jede aktuelle Theorie des moralischen Denkens voranbringen würde. Die hier vorgestellten Ergebnisse sollen zu einem solchen besseren Verständnis beitragen. Abschließend erörtere ich, wie Merkmale von Kausalbeziehungen und subjektiver Nutzen formal in Kausalrepräsentationen integriert werden könnten, und zeige Richtungen für weitere Forschung auf.

Contents

List of Figures

1. Introduction

We do not merely observe the world around us, but we act on it every day. Some actions serve to satisfy our immediate and mundane needs, such as getting a sandwich at lunchtime. Other actions are part of long-term and sophisticated plans, such as securing a well-paid job in order to pay rent, support a family, or further other goals in life. We often have a large number of potential actions at our disposition, but we are not free to act in any way we want: Among other things, we must consider how our actions affect other people. As such, we may not steal our lunchtime sandwich from a street vendor, and we also should not accept a position as a professional hitman, even if it pays well. Clearly, these actions are also illegal, and we would expose ourselves to persecution and various other inconveniences if we undertook them. Other actions are legal, but we might not be allowed to do them due to *moral* considerations. For example, it might be seen as wrong to order a sandwich with meat in it, or to accept a well-paid position as a lobbyist for a fossil fuel company. While performing these actions would not expose us to *legal* persecution, they may elicit the *moral* disapproval of our peers, and plague our own conscience, too.

One important reason why these actions can be regarded as morally wrong is that they cause undesirable outcomes. The causal connection can be stronger or weaker, more or less direct, and rely on the presence of fewer or more additional causes and enabling conditions. The undesirable outcomes, in turn, can vary in their intensity. Stealing a sandwich from a street vendor directly causes the vendor to lose some income, or to become upset. Eating meat, on the one hand, is connected to outcomes that are much worse than one person's loss of a limited amount of money. Besides inflicting significant harm and suffering on animals, meat consumption is also a leading cause of greenhouse gas emissions and, thereby, global warming (see e.g. Stehfest et al., 2009). On the other hand, one person's action of buying a meat sandwich for lunch on a specific day

contributes only very weakly and indirectly to these highly negative effects. In everyday conversation, the crucial importance of the causal relations between our actions and negative outcomes often becomes apparent when potentially harmful actions are justified in the face of moral disapproval or criticism. Avid meat eaters can point to the negligible influence of one person's behaviour on global outcomes. Our friend at the fossil fuel company may argue that if not him, another applicant with potentially even less ethical qualms would have accepted the position. A careless flatmate may defend themselves against accusations by claiming that the pizza they forgot overnight would "just have burned up inside the oven" instead of posing a fire hazard. Conversely, an undeniable causal link between one's actions and negative consequences can give rise to feelings of guilt and blame even when the harm was completely accidental and unforeseeable. For example, people who caused the death of another person through no fault of their own can usually not avoid blaming themselves and sometimes suffer immensely as a result, leading to the formation of dedicated online support groups (see Anderson, Kamtekar, Nichols, & Pizarro, 2021).

The aim of my dissertation project was to elucidate how thinking about causal relationships and outcomes influences our thinking about morality. One project focused on moral judgments about causal chains (Engelmann & Waldmann, 2022a), while the second project investigated the role of outcome trade-offs in common-cause structures (Engelmann & Waldmann, 2022b).

In this chapter, I will first of all introduce *Causal Bayes Nets Theory*, one of the most successful accounts of human causal reasoning (Section 1.1), which served as the main theoretical foundation of our investigation in Engelmann and Waldmann (2022a). I will highlight the intuitive importance of the components of causal cognitive representations according to Causal Bayes Nets Theory, causal structure and causal strength, for the formation of moral judgments.

Subsequently, I will discuss the crucial role of outcomes for moral judgments (Section 1.2). Here, I will describe a model that has been proposed by Cohen and Ahn (2016) and that was the main point of departure for Engelmann and Waldmann (2022b). The model aims to predict moral judgments about actions from observers' subjective utility of consequences. I point out strengths and shortcomings of the model, and suggest that a generalized subjective-utilitarian model might be well-suited to explain the influence of outcomes on moral judgments.

Before presenting our own experiments, I briefly turn to contemporary theories of moral reasoning (Section 1.3). I show that almost all of them place reasoning about causal relations and outcomes at the heart of moral judgment, yet a specific account of the relationship is missing. Thus, I conclude that an improved understanding of how causal strength, causal structure, and outcomes factor into moral judgments would advance any contemporary global theory of moral reasoning.

I conclude the introduction by briefly pointing out several exemplary phenomena in moral reasoning that have already been successfully explained by drawing on aspects of people's causal model representations of situations (judgments about trolley dilemmas, responsibility attribution in groups, and causal selection, Section 1.4). These findings demonstrate the fruitfulness of analyzing moral scenarios in terms of causal relations and outcomes.

In contrast to these "top-down" lines of investigation (drawing on causal models to find an explanation of established phenomena in moral reasoning), our approach in the subsequent empirical chapters is more "bottom-up". In Chapter 2 (summarizing Engelmann & Waldmann, 2022a), I report the results of experiments that varied the causal structure and causal strength by which human actions and harmful outcomes are connected, keeping outcomes constant. We found that in causal chains, people tend to perceive relations as weaker, leading to a more positive moral evaluation of action and agent

compared to direct causal relations. In Chapter 3, I describe a computational model of reasoning about outcome trade-offs in common-cause structures that we developed and evaluated in Engelmann and Waldmann (2022b). Like its predecessor (Cohen & Ahn, 2016), the model predicts people's judgments about different outcome trade-offs based on subjective utilities. However, our model is applicable to a wider range of scenarios (all common-cause structures, whereas the previous model was restricted to dilemmas), and our experiments remedy some methodological problems of earlier studies (e.g., we collect judgments that are actually about morality, rather than about what one would personally do in a described situation). The model fit participants' judgments well. In the future, it could be used as one component of a more complete computational account of moral judgment that also considers causal relations and other important factors such as agents' beliefs and desires.

In the General Discussion (Chapter 4), I tentatively discuss some ways in which causal structure, causal strength, and outcome utilities could be formally integrated (Section 4.1), and lay out directions for future research.

1.1. How causal models support moral reasoning

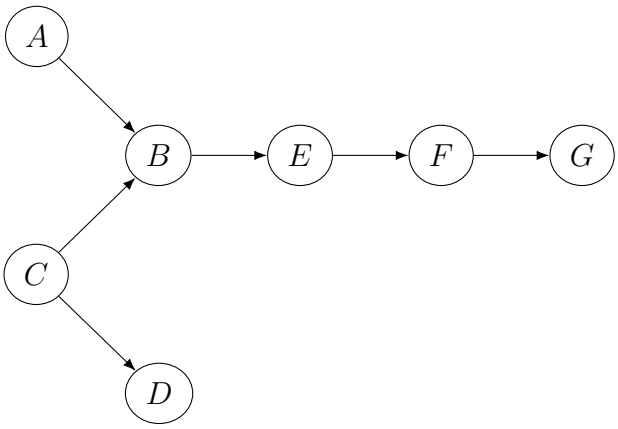
Thinking about the world in terms of cause and effect comes very naturally to us, perhaps so naturally that psychologists have not even identified and studied causal reasoning as a distinct and fundamental cognitive ability for a long time (see Waldmann, 2017). Around the turn of the last century, developments in computer science, philosophy, and statistics converged on the study of what is now called *Causal Graphical Models*, *Causal Bayes Nets*, or simply *Causal Models* (Pearl, 1988, 2000; Spirtes, Glymour, & Scheines, 1993, see also Pearl & Mackenzie, 2019, for an overview). These disciplines (and others, such as economics) use causal models to represent aspects of the external world, and aim to draw normative conclusions from them. Psychologists and cognitive scientists, on

the other hand, explored the potential of these models as the representational medium for causality in the human mind. Here, a main goal is to investigate to what extent people's causal inferences follow the normative models, and where and why they might systematically deviate from them (for overviews, see Waldmann, 2017; Sloman, 2005; Rottman & Hastie, 2014; Waldmann & Hagmayer, 2013). This section will provide a brief overview of the central properties of causal models, and illustrate their importance for the formation of moral judgments.

1.1.1. Causal structure

Constructing a causal model of some aspect of the world, be it as a deliberative scientific exercise or spontaneously in everyday life, first of all simply means to think about what causes what. We have to identify the relata of concern, and the causal relations between them. This is called causal structure. In the Causal Bayes Nets framework, the relata are conceptualised as variables that can take on different values (e.g., present vs. absent in the case of binary variables, or more values when a variable is continuous). The relations between them are represented by arrows. Figure 1 shows a simple example. Let's assume variable B is a person's action. For example, it could be a farmer's action of applying the potentially harmful herbicide glyphosate to some crops. For simplicity, we will treat all variables as binary in this exposition. Thus, if B takes on a value of 1, it would mean that the action was carried out. B = 0 would mean that it was not carried out (identically for all other variables). This action does not come out of nowhere, it also has causes. Let's assume that variable A represents the situation that the farmer is in (e.g., economic pressure), and variable C represents his character and attitudes (for instance, a certain carelessness with regards to environmental damage, or a naive trust that nothing bad will happen). Put simply, we could say that character and situation jointly cause the farmer's action (see e.g., Heider, 1958). Just like events can have multiple causes, they can also have multiple effects. The farmer's character,

Figure 1: A causal model.



Note: The nodes represent variables that can take on different states, and the arrows represent the causal relations between them.

for example, will also influence how well liked they are by their friends and colleagues (this could be variable D in the diagram).

Now let's assume that we do not care so much about the causes of the farmer's action as we care about it's consequences (variables E, F, and G in the diagram). As a result of the herbicide treatment, the farmer's crops could become poisoned (variable E). In turn, bees that are feeding on the crops may produce honey that is contaminated by glyphosate, rendering it inedible (variable F). Finally, a beekeeper who was planning on selling this honey may incur a substantial financial loss (variable G).[1]

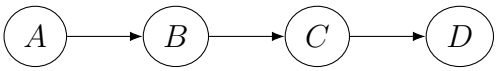
A crucial property of causal models is that once we know that a variable is present (e.g., the farmer's action has actually occurred), we do not need to know anything about its causes in order to make inferences about its consequences. Imagine we didn't know whether a particular farmer has used glyphosate or not, but we are aware of the generic causal model shown in Figure 1, and we also know that a beekeeper produces honey nearby. In that case, knowing that the farmer holds carefree attitudes about the use

[1]This example is inspired by a real-world case, see https://taz.de/Gericht-entscheidet-ueber-Glyphosat/!5788299/ (in German, last checked June 13, 2022).

of herbicides (variable A) would allow us to make a prediction about the beekeeper's financial loss: It should increase our confidence that such a loss could occur. In contrast, once we know for a fact that the herbicide has been applied, knowing about the farmer's underlying attitudes adds no useful information to our prediction of whether the outcome will occur in this particular case (although it might be informative for other inferences, such as about outcomes in other situations). This property of causal models is called the Markov condition (see, e.g., Rottman & Hastie, 2014): given its "parents" (its direct causes), any node in a causal model is statistically independent of all other nodes except its "descendants" (its direct and indirect effects). Applied to our model, it means that if we take the farmer's action to be the given "parent" node, we can ignore everything that precedes it to make inferences about plant poisoning, honey contamination, and the beekeeper's financial loss. This is also sometimes called "screening off" (see, e.g., Sloman, 2005): the farmer's action screens its causes (character, situation) off from its effects (plant poisoning), they become statistically independent given that the action has been carried out. This property of causal models is the reason why we are able to inspect and use causal models of isolated aspects or situations of the world at all, such as the farmer's action and its consequences. If we had to consider all causes of all variables of interest, no matter how far removed, we would be facing an impossible task when thinking about even the simplest causal queries. Thanks to the Markov condition, we can simplify the causal model of the glyphosate example, and focus on the action and its consequences only (although research has shown that people don't always make full use of this advantage, see Rottman & Hastie, 2014). This results in the simple causal chain depicted in Figure 2 (with A now being the farmer's action, B the plant poisoning, C the honey contamination, and D the beekeeper's financial loss).

Knowing if and how events are causally related, and, equally important, which events are *not* related, both enables and constrains our *moral* evaluations when human actions

Figure 2: A simple causal chain.



Note: The nodes represent variables that can take on different states, and the arrows represent the causal relations between them.

are part of a causal network (as they are in our example). For instance, the beekeeper in our example case might blame or even sue the farmer, based on the fact that the farmer's action caused the beekeeper to lose money. The farmer, on the other hand, might object that the specific causal chain of events that unfolded was rather long and maybe not entirely knowable in advance. Clearly, more than causality is under debate here (such as what each party could and should have foreseen). Causal relations are usually not *sufficient* to arrive at a complete moral, let alone legal evaluation of a situation. However, they are *necessary*. Without a causal pathway between the farmer's action and the beekeeper's financial loss, no discussion would be had between these two people at all.

1.1.2. Causal strength

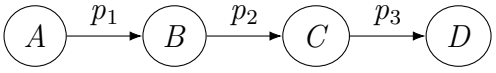
While the structure is the backbone of a causal model, its full potential can only be exploited when the structure is combined with a probabilistic theory, usually a theory of Bayesian inference (Pearl, 1988, 2000; Spirtes et al., 1993; Pearl & Mackenzie, 2019; Waldmann, 2017; Waldmann & Hagmayer, 2013; Sloman, 2005). Say we know that an action we plan to carry out has several effects, some wanted and some unwanted. For example, we may contemplate jaywalking across a busy street as a shortcut on our way to work. This will save us time, but it also bears some risk of causing an accident if cars have to break abruptly, potentially causing us or others to get hurt. We will intuitively assign probabilities to each thing happening, and most of us will only act if they take the

probability of accidents to be sufficiently low, and the probability of saving time to be sufficiently high. In other situations, we might ponder which of several potential causes underlies an observed effect. For instance, we might wonder why a friend hasn't texted us back. Are they mad at us? Did something happen to them? Or are they just busy? The answer we settle on will be informed by how likely each cause seems *per se* (its base rate), and by how likely it would produce the effect of our friend not texting back, given that the cause occurred (a conditional probability). Neither in the jaywalking case nor in this example would we be able to make a satisfying inference if all we knew about were qualitative causal relations.

In our honey example, equipping the causal chain with quantitative parameters results in the model displayed in Figure 3. Each arrow in the network has been assigned a probability, which represents the strength of the causal relation. For instance, given that the farmer applied the glyphosate (A = 1), there is a certain chance that his plants will be poisoned (B = 1). The strength of this relation is expressed by p_1. If plant poisoning indeed has no other causes than the application of glyphosate, p_1 is simply $p(\mathrm{B} = 1|\mathrm{A} = 1)$. Often, however, one might want to allow for an influence of alternative causes, even if they are unknown. For this reason, more refined measures of causal strength have been proposed (Cheng, 1997; Cheng & Novick, 1990, 1992; Novick & Cheng, 2004, see also Griffiths & Tenenbaum, 2005). Since we cut off everything before the farmer's action (A), it looks like the action has no causes. However, this is not the case, as we know. The influence of all causes of the farmer's action can be expressed by a base rate, p_A (not shown in the graph).

The combination of structural and quantitative causal knowledge enables powerful predictive and diagnostic inferences, and their relevance for moral evaluations is easy to see. In our example case, the farmer could have relied on such knowledge to think about the likelihood of someone being harmed by his action. In a causal chain that honors the

Figure 3: A parameterized causal chain.



Note: The nodes represent variables that can take on different states, and the arrows represent the causal relations between them. The parameters attached to the arrows indicate the strength of these relations.

Markov condition, the strengths of the individual links have to be multiplied to arrive at the probability of the final outcome given the root cause, here the farmer's action (see Stephan, Tentori, Pighin, & Waldmann, 2021). Unless all links are deterministic, it follows that earlier events are more likely to occur than events further down the chain, given the root cause. The farmer could use this fact to argue that even if he was aware of all qualitative causal relations, the occurrence of the undesired final outcome (the beekeeper's financial loss) was so unlikely that he couldn't have reasonably foreseen it to actually happen. Inversely, causal models also allow to ascertain the likely causes of events (diagnostic reasoning, see Meder, Mayrhofer, & Waldmann, 2014; Meder & Mayrhofer, 2017), or to determine whether a specific event actually produced another event in a particular situation (singular causation, see Stephan & Waldmann, 2018; Stephan, Mayrhofer, & Waldmann, 2020). Inferences like these play a crucial role for the beekeeper in our example. If he wasn't reasonably sure that his honey was contaminated because the bees fed on poisoned plants, and if he also wasn't reasonably sure that these plants were poisoned because the farmer applied glyphosate to them, he would have no reason to blame the farmer for his loss (and to subsequently sue for the damage).

In actuality, preceding the dispute between farmer and beekeeper was a chain of forward and backward causal inferences on the side of the beekeeper.[2] He observed that some plants close to his apiaries had shrivelled in a particular way, and suspected that

[2]see https://taz.de/Gericht-entscheidet-ueber-Glyphosat/!5788299/ (in German, last checked June 13, 2022).

they might have been treated with glyphosate. If that was the case, he considered it possible that his honey might be contaminated, as his bees would certainly have fed on the plants. Having the honey tested confirmed the suspicion.

1.1.3. Summary

Both the structure and strength of causal relations between actions and harmful outcomes have intuitive relevance for moral judgments about actions and agents. Chapter 2 will discuss in more detail how knowledge about causal structure, causal strength, and inferences about foreseeability may interact when people make moral judgments about causal chains (such as the chain between the farmer's action and the beekeeper's financial loss). I will contrast two potential models of the relationship between these components, and explain how we differentiated between the models with the help of three experiments. First, however, I will introduce another crucial input into moral judgments – the value of outcomes.

1.2. The role of outcomes in moral judgment

Moral judgments, in the honey example as well as in many other cases, are often concerned with situations in which bad things happened or are predicted to happen. When we judge someone's action as morally impermissible, this is often because we predict that it will have unfavourable consequences. Blame or punishment are assigned to agents for their contribution to outcomes that we perceive as negative (such as the beekeeper's financial loss). Moral dilemmas are particularly hard because no matter what one does, some kind of harm will occur in either case. To illustrate this point, think about a classical moral dilemma such as the *bystander* trolley case (Foot, 1967): An out-of control train is speeding towards five people and will run them over unless the train is redirected to a side-track on which just one person is standing. Here, the target action to be evalu-

ated, redirecting the trolley, has two effects. First and foremost, it saves the lives of five people who would otherwise be run over. Second, it will cause the death of one person on the sidetrack, who would otherwise survive. While the decision may seem difficult to many, it would not be difficult at all if the consequences of acting would not outweigh the consequences of doing nothing. It is a trivial point to most people that all else being equal, saving the lives of five people is better than saving the life of just one.

Obviously, and luckily, most people will never find themselves in a situation where they actively have to make a decision about the lives of other people. On such grounds, some researchers have criticized trolley dilemmas as unrealistic, or as poor predictors for people's actual moral behaviour (Bauman, McGraw, Bartels, & Warren, 2014; Bostyn, Sevenhant, & Roets, 2018; Schein, 2020). However, such criticism misses an important point. While studying moral behaviour in everyday situations is an important and interesting line of research in its own right, sometimes, we actually want to know how people *judge* the behaviour of others based on a description of their actions (Białek, Turpin, & Fugelsang, 2019). For instance, this is how the evaluation of most political decisions works. Decisions about the allocation of healthcare resources, for instance, usually have consequences that can be quantified quite precisely in terms of lives saved versus lives lost. The Covid-19 pandemic, unfortunately, has provided us with many further examples (e.g., political decisions about lockdowns, travel restrictions, potentially mandatory vaccines, triage protocols, and many others). The evaluation of such decisions, in everyday conversation or on social media, tends to put their consequences front and center (even though other factors, such as deciders' perceived motives, obviously play an important role as well). Thus, while we are rarely, if ever, the *actor* in a trolley dilemma, we are regular observers and judges. And our evaluations matter, as they might ultimately be reflected in voting behaviour, for example.

Artificial moral dilemmas then offer us a method of isolating and systematically in-

vestigating the influence of the many factors that affect our evaluations in everyday life (such as consequences, causal structure, mental states, etc.). For such an endeavour, the artificiality of moral dilemmas can be a feature rather than a bug. After all, we also wouldn't criticize the stimuli that are used in optical illusions (such as the two lines in the famous Müller-Lyer-Illusion) for their artificiality (Plunkett & Greene, 2019). They expose the fundamental mechanisms of certain abilities (vision, moral judgment) in a way that would not be possible in "messy" real-world scenarios. In Engelmann and Waldmann (2022b), we used moral dilemmas to investigate the influence of outcome-tradeoffs on judgments about the moral permissibility of actions (see Chapter 3). However, we also used more regular common-cause structures, that is, situations in which a potential action has several effects, but no changes to the status quo occur when the action is *not* executed.

Theories of moral reasoning generally acknowledge the importance of outcome tradeoffs for moral evaluations (and Section 1.3 will provide a more detailed overview of those theories). For instance, according to the *Doctrine of Double Effect*, which plays a central role in Mikhail (2007, 2011)'s Universal Moral Grammar Theory, a central condition for the moral permissibility of an action that has both positive and negative consequences (among other important requirements) is that the good outcomes outweigh the bad ones. In *Dual-Process Theories* (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Cushman, 2013; Crockett, 2013), the slow and deliberative mode of evaluation is taken to be concerned with the evaluation of the consequences of action versus inaction in a particular situation. However, figuring out how exactly reasoners compare the consequences of different paths of action has generally not been a focus of these theories. Instead, the paradigms employed in moral dilemma research (for overviews, see Waldmann, Nagel, & Wiegmann, 2012; Wiegmann & Engelmann, 2020) have generally mirrored the philosophical debate about

trolley dilemmas (Kamm, 2008; Unger, 1996; Foot, 1967; Thomson, 1976) insofar as they most often kept outcomes constant and varied other factors of interest, such as causal structure.

As a result, it is currently not clear what the functional form of an outcome integration mechanism in the context of moral judgment might look like. Initially, one might think that reasoners make a simple ordinal comparison. In a classical trolley dilemma, for instance, they realize that five lives are more than one life, which factors into the overall moral evaluation of the action as a "pro" reason. However, such a simple mechanism would fail to be applicable as soon as outcomes differ on other dimensions than the number of lives saved and lost. For example, most people probably value human lives higher than the lives of animals, and would therefore likely opt to save one human rather than, say, five cats. Comparisons can become even more complex when more abstract values such as liberty or security are invoked. Nevertheless, people readily evaluate trade-offs between all kinds of consequences. A recent striking demonstration comes from the "moral machine" experiment (Awad et al., 2018). Here, participants from all over the world were invited to evaluate dilemmas about out-of-control self-driving cars. Possible victims differed along many dimensions, such as number, age, species, or their role in society. Some stable preferences emerged, such as a greater readiness to save younger rather than older people (at least in some countries), and a greater readiness to save humans rather than animals. However, a theoretical account of how people arrive at these decisions is not provided.

1.2.1. Cohen & Ahn's (2016) *Subjective-Utilitarian Theory of Moral Judgment*

Cohen and Ahn (2016) recently proposed a novel theory of moral judgment that puts the consequences of actions front and center, and they also suggest a psychological mechanism for outcome comparisons (see also Cohen, Cromley, Freda, & White, 2021). According to their *Subjective-Utilitarian Theory of Moral Judgment* (later generalised

to *Psychological Value Theory*, see Cohen et al., 2021), outcomes are ultimately the only determinant of an action's moral evaluation. Precisely, it is an observer's subjective utility of the consequences that would arise with versus without acting that determines an action's moral status. This subjective utility is taken to be a composite of all cognitive and emotional reactions to a stimulus. For example, when evaluating a classical trolley dilemma in which five lives are pitted against one, the theory assumes that reasoners ascribe a certain value to one human life, and another, likely higher value to five lives. Importantly, the value of five lives does not need to be five times as large as the value of one life. In fact, Cohen and Ahn (2016) explicitly predicted that it wouldn't be, and that people's value functions would instead show patterns of diminishing marginal utility, as for example assumed in *Prospect Theory* (Kahneman & Tversky, 1979). Furthermore, more lives need not always be valued higher than fewer lives. If, for example, the one person on the sidetrack in a trolley dilemma was a close friend or relative of the observer, it is possible that the observer would value this person's life higher than the lives of five strangers.

Cohen and Ahn (2016) propose that people's representation of subjective utilities can be approximately described by Gaussian distributions. That is, there is a value that is most often assigned to each stimulus by a reasoner (the mean of the distribution), but depending on context or chance, the value of the stimulus can sometimes also be perceived as somewhat higher or lower (the variance). Crucially for the theory, the value distributions for many items are assumed to have substantial overlap. For example, the mean value of "five lives" is probably somewhat higher than the mean value of "one life" for most people. However, if these mean values are not drastically different, and conceptualised as the peaks of two normal distributions with non-negligible variance, there will be a wide range of values that could belong to either distribution.

According to the subjective-utilitarian theory, making a decision about acting or not

acting in a moral dilemma simply amounts to determining which set of consequences has the higher value. Cohen and Ahn (2016) suggest that people solve this task by repeatedly sampling from the value distributions of the items under consideration. On each trial, a single value from the distribution of "one life" is drawn, and compared to a single value drawn from the distribution of "five lives". It is noted which of these single values is higher. This process is repeated until enough evidence has been accumulated to consciously conclude that "five lives" actually has the higher value. With more overlap between any two distributions, this process will take longer and result in the experience of a harder decision. Furthermore, more overlap makes it more likely that "errors" occur, that is, that the item with the lower mean value is mistakenly thought to be more valuable.

Cohen and Ahn (2016) tested their theory by first asking a group of participants for their subjective values of a wide range of stimuli (people, animals, and inanimate objects). These values were elicited by providing a standard with a fixed value (they used "one chimpanzee", which was assigned a value of 1000), and instructing participants to compare each item against this standard (i.e., assigning a value of 2000 if they think that the item is twice as valuable as a chimpanzee, and so on). From these estimates, a value distribution could be constructed for each item, and the overlap between any two distributions could be determined. A different group of participants then completed a set of moral dilemmas consisting of a forced choice between any two of the previously valued items. Only one of them could be saved, the other one would be killed or destroyed. The involved items were randomly drawn on each trial. It turned out that, as predicted, the value estimates that the first group of participants provided predicted the choices of the second group very well. Besides the mean values predicting which item would be saved on a given trial, the overlap between the value distributions of any two items also predicted reaction times and error rates. More overlap led to longer reaction times

and more errors, that is, trials on which the item with the lower mean value was saved. Similar results were obtained in Cohen et al. (2021).

1.2.2. Strengths and shortcomings of the model

The form of Cohen and Ahn (2016)'s model was inspired by models that are well-established in other domains such as visual perception (so-called random walk or drift diffusion models, see e.g. Ratcliff & Rouder, 1998). The mechanism is formalized and yields quantitative predictions that could be confirmed in several experiments. The model is also parsimonious, relying just on the subjective utility of outcomes to explain people's choices in moral dilemmas with a large variety of different outcomes.

Nevertheless, there are also some shortcomings. Most importantly, the model is currently limited to decisions in two-option, forced choice dilemmas (such as classic trolley cases). Simply comparing which of two items has the higher subjective value will only suffice as an outcome analysis in such simple cases. However, we also readily make moral judgments about more complex situations. For instance, there are many cases in everyday life in which a gain to one group has to be traded off against a loss or harm to another group. Examples are tax alleviations for top incomes at the expense of others, or animal testing for research. Such cases are readily analyzed in terms of costs and benefits as well, which is one important input into their moral evaluation. Nevertheless, they do not necessarily have the structure of a moral dilemma. In the case of tax alleviations, one group would benefit if the measure was introduced, and others would lose some support or goods. If the measure is not introduced, however, the status quo would be maintained, and none of the groups would lose everything, or die (as in a moral dilemma). Such more "regular" multiple-outcome structures are way more frequent than moral dilemmas, but cannot be captured by the current form of Cohen and Ahn (2016)'s model.

Second, the test question that Cohen and Ahn (2016) used in all of their experiments

was "Would you save [Item A], causing [Item B] to be killed/destroyed?". Thus, even though the aim was to assess people's *moral* evaluations of dilemmas, a question was used that is more likely to elicit a self-assessment of the likelihood that one would act in the described situation (Royzman & Hagan, 2017). It is not surprising that people would be guided by their subjective values of consequences when thinking about what they would, personally, do. However, what one would do and what one takes to be morally right do not always have to coincide (Kahane & Shackel, 2010; Soter, Berg, Gelman, & Kross, 2021; Tassy, Oullier, Mancini, & Wicker, 2013). Whether subjective utilities also predict the *moral* evaluation of other people's actions thus remains an open question.

Finally, the subjective-utilitarian theory fails as a general theory of moral judgment (even if applicability was restricted to dilemmas) because it does not allow for a role of causal structure, intentionality, knowledge, action versus omission, and other factors that have been repeatedly demonstrated to affect moral judgments beyond the influence of consequences (Greene et al., 2001, 2004; Cushman, Young, & Hauser, 2006; Cushman, 2008; Hauser, Young, & Cushman, 2008, for overviews see also Waldmann et al., 2012; Wiegmann & Engelmann, 2020; May, 2018, see also Chapter 2). Cohen and Ahn (2016) attempted to manipulate "personal force" (Greene et al., 2009) in one of their experiments and found no effect, but due to their design, only few of all of their trials can actually have been appropriate tests of the effect.[3] Despite this criticism, a sufficiently generalized subjective-utilitarian model might be a very useful building block of a more complete computational account of moral reasoning, if it turns out that subjec-

[3]Effects of "personal force" are only expected between pairs of otherwise matched scenarios that are a) about human lives, and b) in which better consequences arise for acting than for not acting. In these cases, causing the greater good using "personal force" (e.g., by pushing someone) should be less permissible than doing so without personal force (e.g., by redirecting a threat). Whenever consequences are better without intervening at all, or when trade-offs are between animals or inanimate objects, no effects of personal force are expected. Cohen and Ahn (2016) randomly drew the items that would be involved in each scenario (humans, animals, or objects in different numbers and states). This way, it is hard to see how an overall effect of personal force could have become visible.

tive utilities actually predict people's *moral* evaluation of actions (all else being equal). We developed such a generalized subjective-utilitarian model and tested its predictions in Engelmann and Waldmann (2022b), and I am going to explain the new model and summarize the experiments in Chapter 3.

1.2.3. Summary

Both the valence of the outcomes that we bring about and the ways in which our actions are connected to those outcomes are important inputs for moral reasoning. Causal structure describes which events are causally related, and how. It thus encodes important *qualitative* causal knowledge (such as the claim that smoking causes lung cancer). Causal strength, on the other hand, encodes equally important *quantitative* causal knowledge (such as how likely someone's smoking habit is to give them lung cancer). Finally, moral reasoning is often concerned with outcomes that are valued negatively. For example, someone may feel contempt for companies that advertise addictive harmful substances like tobacco to people, and this person might therefore applaud efforts to restrict such advertisement. These reactions are based on both the value assessment that people getting lung cancer is very undesirable, and on the belief that advertisement encourages smoking, which in turn causes cancer in a non-negligible proportion of those who do smoke. If either lung cancer wasn't taken to be a serious harm, or if the observer didn't believe in a sufficiently strong causal relationship between advertisement and smoking behaviour, or between smoking and lung cancer, they would have no good reason to hold any of these positions.

The following section will review how contemporary theories of moral judgment incorporate reasoning about causal relationships and outcomes. It will demonstrate that all theories reserve some role for both in the generation of moral judgments, and that most theories even assign them a central role. However, none of the frameworks so far include (a) a detailed account of how different dimensions of causal networks (structure,

strength) affect moral inferences, or (b) a mechanism for evaluating and comparing the outcomes that result from different paths of action.

1.3. The role of causal representations in contemporary theories of moral reasoning

While the roots of moral psychology lie in the developmental theories of Piaget (1954) and Kohlberg (1974), there has been a particularly strong increase of interest by researchers in the past twenty years (see, e.g., Greene, 2015). This section will focus on this period, covering the most influential global theories of moral reasoning that have been proposed since the turn of the century. The presentation partly follows the discussion by Wiegmann and Engelmann (2020), but focuses on the role of causal representations (comprising outcomes and the ways in which they are brought about) in each theory.

1.3.1. Haidt's *Social-Intuitionist Model*

According to the *Social-Intuitionist Model* (from here on: SIM) (Haidt, 2001), moral judgments are primarily driven by intuitive emotional reactions to certain norm violations. Consciously available reasons for a moral condemnation are mere *post-hoc* justifications of these initial affective reactions, and play no causal role in their generation. The textbook example is a vignette about two siblings who engage in consensual incest, with many precautions guaranteeing that no harmful consequences will occur. Despite these guarantees, participants generally condemn the act, but are said to have trouble articulating reasons for their position (Haidt, Björklund, & Murphy, 2000). This effect, dubbed "moral dumbfounding", is the cornerstone of the theory. It shows, so the argument, that an action can be seen as deeply morally wrong in the absence of any rational justification in terms of harm or damage in the actual situation - thereby supposedly un-

dermining the famous and previously dominant rationalist accounts of moral reasoning by Piaget (1954) and Kohlberg (1974).

Under the SIM, explicit and consciously available causal representations of situations such as the incest case play no major role in the formation of moral judgments. After all, an act can be seen as wrong even though there is no causal connection between the act and any kind of harm. The intuitive condemnation of incest is explained by evolutionary adaptation, and analogous arguments are made for other kinds of violations (Graham et al., 2018, 2011, 2013). While the theory would not deny that reasoners *have* cognitive causal representations of cases, it disputes that they actually *use* them to make moral judgments. Instead, they may normally be used in the construction of the *post-hoc* justification of intuitive emotional reactions.

However, more recent findings indicate that the SIM may have been overlooking aspects of people's reasoning. Royzman, Kim, and Leeman (2015) demonstrated that in the incest case, many people do not actually believe the stipulation that no harm will occur. Stanley, Yin, and Sinnott-Armstrong (2019) furthermore show that reasoners increasingly condemn actions the more strongly they believe that harm *could have* occurred, even if it didn't in a particular case. Such observations point to a greater relevance of causal representations in moral judgment than the SIM acknowledges.

1.3.2. Greene's *Dual-Process-Theory*

Greene's *Dual-Process-Theory* (Greene et al., 2001, 2004) assigns a crucial role to emotions in moral judgment as well, but an equally important role to reasoning. The theory was inspired by people's different reactions to two versions of the classical philosophical thought experiment of the trolley case (Foot, 1967; Thomson, 1976). In both cases, an out of control trolley is headed towards five people who are tied to a rail track, and who will be killed by the trolley without intervention. In the *bystander* version of the case, the five can be saved by redirecting the trolley to another rail track on which just

one person is tied up. This person will die as a result of redirecting the trolley. In the *footbridge* version of the case, the trolley can only be stopped by pushing a heavy person in front of the train from a bridge above the tracks. Again, this person would die as a result. Even though the number of lives saved and lost is identical in both cases, it is well established that people take redirecting the train in *bystander* to be permissible, but object to pushing the man in *footbridge* (see Waldmann et al., 2012, for an overview). Greene's theory explains this difference by positing that the two dilemmas elicit different modes of processing. The *footbridge* dilemma, so the argument, has features that favour a primarily emotional processing. It involves serious and up-close harm to a person (generated by "personal force", see Greene et al., 2009) and, crucially, the harm does not result from deflecting a pre-existing danger from one person to another. In such cases, it is claimed, emotional processing outcompetes the rational, controlled processes that compare the consequences of acting and not acting in a scenario (and that would deem acting permissible when more lives are saved than lost). Evidence for a differential involvement of brain areas associated with emotional vs. rational processing in the *footbridge* vs. the *bystander* dilemma comes from fMRI studies (Greene et al., 2001), and the notion of competing processes is backed up by the observation that those people who think that pushing the man in *footbridge* is permissible only indicate this after longer reaction times.

Causal representations clearly matter in Greene's dual-process account, as the theory explicitly posits a cognitive process that is dedicated to computing and comparing the consequences of acting and not acting in a moral scenario (even if the results of this process can be outweighed by features that elicit a primarily emotional response). However, working out the details of this mechanism has not been a focus. Some evidence for the importance of outcomes comes from Shenhav and Greene (2010), who show that moral judgments are sensitive to the expected value of consequences.

1.3.3. Crockett and Cushman's *Dual-Process-Theory*

Greene's dual-process account has been criticized, among other things, for its simplified distinction between emotions and reasoning (see, e.g., Moll, De Oliveira-Souza, & Zahn, 2008; Pessoa, 2008), and for mapping "emotional" responses (which are seen as reacting to morally irrelevant features of moral dilemmas) to deontological ethics, and "rational" or "cognitive" responses to consequentialism (see, e.g., Berker, 2009; Kamm, 2009). Both Crockett (2013) and Cushman (2013) therefore proposed versions of dual-process accounts that avoid these assumptions. On their view, the two competing processes are a so-called "model-free" and a "model-based" algorithm. Put simply, a model-free algorithm evaluates an action based on past experience with that action. As such, pushing a person (as in *footbridge*) evokes predominantly negative associations, while pushing a button or pulling a lever (as in the *bystander* dilemma) is free of such negative associations. The model-based algorithm, on the other hand, analyses an action's consequences in the present situation. This process would determine that acting saves five lives and leads to the loss of one in both *bystander* and *footbridge*. Support for the theory comes from studies showing that people are averse to actions that usually have negative consequences, even though these consequences do not arise in a particular situation, for example shooting at someone with a fake gun (Cushman, Gray, Gaffey, & Mendes, 2012).

Causal representations of situations are a crucial input for the model-based algorithm, as they alone enable an assessment of the effects of different hypothetical interventions on a situation. Generic causal knowledge can also play a role in the model-free algorithm, when an action's *usual* consequences affect its evaluation in a situation where these consequences are blocked. However, the model-free algorithm would also react to mere correlates of an action.

1.3.4. Mikhail's *Universal Moral Grammar*

Universal Moral Grammar Theory (Mikhail, 2007, 2011) is inspired by Chomsky's (1965) *Universal Grammar* and posits that moral judgment is governed by a set of universal and innate rules. Causality is at the heart of these rules. When processing a trolley dilemma, for example, the theory claims that the first step is to translate temporally ordered events such as the pushing of a button, the trolley being redirected, or a person dying, into a causal model of the situation. Further processing steps involve determining the valence of outcomes, and inferring agents' intentions (see also Levine, Mikhail, & Leslie, 2018). Once such a complete representation has been constructed, unconscious principles such as the *Doctrine of Double Effect* (DDE, see McIntyre, 2019) can be applied to determine an act's moral status. The DDE specifies several conditions under which otherwise forbidden actions, such as harming a person, become permissible. The action's positive effects have to outweigh its negative effects, only the positive but not the negative effects may be intended, and there is no way to bring about the positive effects without risking the negative ones. The DDE captures the distinction between the *bystander* and *footbridge* versions of the trolley case by positing that in *footbridge*, the death of the one person is a causal means of saving the lives of the five people (as their body is needed to stop the train, which inevitably kills them), and therefore intended (as means for bringing about an intended outcome are taken to be necessarily intended as well). Thus, *footbridge* does not meet the criteria of the DDE, while *bystander* does. When the trolley is redirected in *bystander*, the one person's death is merely a foreseen side-effect of saving the five. It would be possible to save the five even if the one person was removed from the scene, which is not the case in *footbridge*. Thus, causal relations and outcomes are of crucial importance in at least two steps of processing and evaluating a moral scenario, according to Universal Moral Grammar Theory. First, the basic causal representation of a situation, augmented with the valence of outcomes, serves as the basis

for further inferences, such as about agents' intentions. Second, higher-order principles such as the DDE can explicitly refer to causation in the conditions that they specify for an act's moral permissibility.

1.3.5. Gray, Waytz, & Young's *Theory of Dyadic Morality*

Human moral psychology is diverse. It is occupied with a wide range of topics and situations (Graham et al., 2018, 2011, 2013), and there are different kinds judgments that we readily make, such as judgments about permissibility, responsibility, blame, or punishment (Cushman, 2008). K. Gray, Young, and Waytz (2012) argue that this diversity is merely superficial. According to their *Theory of Dyadic Morality*, all moral reasoning is organised and unified by a cognitive template in which a moral agent harms a moral patient. This prototypical scheme, so the theory, is automatically activated in response to certain cues, such as potentially harmful actions being carried out, agents having bad intentions, or negative outcomes occurring. Crucially, K. Gray et al. (2012) posit that the template is automatically "completed" even in cases where dangerous actions are conducted, but no harm occurs (such as unsuccessful attempts to harm someone), or in cases where harm occurs, but there is no immediate intentional agent who can be held responsible (such as blaming fate, karma, or God for negative life events). The theory thereby explains, for example, people's aversion to Haidt et al. (2000)'s incest case. Even though the vignette explicitly stipulates that no harm will result from the siblings' decision to have sex, people can't help but rely on their intuitive dyadic template when evaluating the situation. People seem to automatically link moral wrongness and suffering in a variety of behavioural tasks (K. Gray, Schein, & Ward, 2014), and when asked to "list a morally wrong act", more than half of participants listed cases that are close to the prototypical dyadic template (such as murder, rape, and other forms of intentional harm) (K. Gray & Ward, 2011, as cited in K. Gray et al., 2012).

Causation is a crucial component of the posited dyadic moral template, and so are negatively valued outcomes. After all, the agent has to *cause harm* to the patient. The theory predicts that moral judgments should be less severe the less obvious or direct the causal relation between agent and patient is (K. Gray et al., 2012). While a more or less direct relation can easily be modeled by structure and strength parameters of causal models, the notion of agents and patients is more closely tied to a different class of theories of causation, namely force dynamics (White, 2006, 2007, 2009; Wolff, 2007, 2012; Talmy, 1988). According to such theories, causation is best described as the interplay between objects with certain dispositions (see also Mumford & Anjum, 2011). In the moral dyad, the objects are minds, and the main disposition of the agent is the ability to cause harm, while the main disposition of the patient is the ability to experience harm, or to suffer (see also H. M. Gray, Gray, & Wegner, 2007). Causal Bayes Nets and force dynamics are generally seen as competing theories of causation, but it has been shown that at least as psychological accounts of causal reasoning, they might be fruitfully integrated (Mayrhofer & Waldmann, 2015; Waldmann & Mayrhofer, 2016, see also Stephan, Engelmann, & Waldmann, 2021). In any case, the theory of dyadic morality reserves a central role for reasoning about causal relations and harmful outcomes, and might even require richer causal representations than typically afforded by Causal Bayes Nets.

1.3.6. Summary

All global theories of moral judgment that have been proposed in the past twenty years acknowledge the importance of causal relations and outcomes with positive or, more often, negative valence (except the SIM, which mostly reduces the function of causal reasoning to the construction of *post-hoc* justifications). If and how actions are causally connected to good and bad outcomes clearly matters for our evaluation of actions and agents. At the same time, the details are not worked out. None of the global frameworks

of moral reasoning makes a connection to a psychological theory of causal reasoning (except the Moral Dyad Theory, which was inspired by force dynamics), and we don't know how exactly the different dimensions of causal models are integrated in the formation of moral judgments. For instance, what are the relative contributions of causal strength and causal structure, and how do they interact with other important factors, such as inferences about agents' mental states? How are outcomes compared that result from the actual, hypothetical, or counterfactual instantiation of different paths in a causal network? The main contribution of this thesis consists in making some progress towards answering these questions (see Chapters 2 and 3).

Next, I will review some work that has already explicitly connected causal and moral reasoning. There are a number of compelling explanations for specific phenomena in moral cognition that draw heavily on causal reasoning. People use information about causal relations and outcomes in a variety of ways in different moral judgment tasks. This and the sophisticated accounts that have been developed to explain people's reasoning in particular tasks suggest that more detail in global theories of moral reasoning is needed when it comes to explaining the role of causal representations. After reviewing this evidence, I will proceed to the presentation of our own experiments.

1.4. Previous work on the relationship between causal and moral reasoning

In the following, we will highlight three specific phenomena in moral reasoning: patterns of judgments about trolley dilemmas, responsibility attribution in groups, and causal selection. We will explain how an analysis in terms of causal models has facilitated the understanding of these phenomena, and briefly review the evidence.

1.4.1. Explaining intuitions about trolley dilemmas

As the previous section has shown, the construction of theories of moral reasoning has often been guided by the goal to explain intuitions in moral dilemma situations, such the *bystander* and *footbridge* versions of trolley dilemmas. For instance, Mikhail (2007, 2011) argued that people implicitly apply a principle like the Doctrine of Double Effect (DDE), and that acting in *footbridge* is impermissible because one person is harmed as a causal means of saving the others in the scenario. However, the two dilemmas differ in many other aspects as well. Targeting one of those aspects, Waldmann and Dieterich (2007) point out that *bystander* involves an intervention on an agent of harm (the trolley), whereas *footbridge* involves an intervention on a patient of harm (the heavy person). In a series of experiments, they demonstrated that intervening on a patient of harm is seen as worse than intervening on an agent, and that the distinction between harming as a means and harming as a side effect ceases to alter people's intuitions once this so-called *locus of intervention* is held constant. The psychological explanation that Waldmann and Dieterich (2007) put forward is that people spontaneously compute two contrasts when evaluating a dilemma (see also Waldmann & Wiegmann, 2010, for an extension of the *Double Causal Contrast Theory* and further evidence). The global contrast is the overall value of consequences with and without intervention (e.g., five people alive and one person dead in the case of intervention vs. five people dead and one person alive in the case of no intervention). This global contrast is identical in *bystander* and *footbridge*. The local contrast, on the other hand, focuses on the direct consequences of intervening on the target of intervention (e.g., the trolley or the heavy person). In the case of intervening on the trolley in *bystander*, the local contrast is the same as the global contrast (if redirected, the trolley spares five people, but kills one). In case of intervening on a victim, such as the heavy person in *footbridge*, the local contrast highlights that this person will die when the proposed intervention is executed, but would remain alive

without it. The local contrast does not completely supersede or replace the favourable global contrast, but, so the argument, backgrounds it to some extent, thereby making patient interventions seem less morally acceptable.

A further puzzle that arose from research on these two most well-known trolley variants is that people tend to evaluate *bystander* positively when they judge it either in isolation, or before they judge *footbridge*, but more negatively when they have previously seen *footbridge* (Liao, Wiegmann, Alexander, & Vong, 2012). Wiegmann and Waldmann (2014) argue and demonstrated that the causal structure of *bystander* allows for selective attentional highlighting of either the causal path between intervention and saving, or of the causal path between intervention and harming (as saving and harming are separate effects of the same cause, redirecting the trolley). In *footbridge*, however, the path between intervention and saving the five cannot be highlighted independently, because harming the one person (by pushing them onto the rails to stop the trolley) lies on that same causal path and thus cannot be bypassed or backgrounded. When *footbridge* is presented first, harming the one person is highlighted by default. When *bystander* is presented afterwards, people transfer this attentional focus and now highlight the path between intervening and harming in *bystander* as well (rather than the path between intervening and saving, which they would normally highlight by default). Thus, *bystander* appears worse when it is presented after *footbridge*, compared to when people evaluate it before *footbridge* or in isolation.

Taken together, the findings and theories listed above show that an analysis of the causal structure of moral scenarios, in combination with hypotheses about how people direct their attention, can elucidate patterns in people's moral reasoning that are otherwise hard to explain. It should also be noted that both theories make and confirmed predictions for people's evaluation of completely novel versions of trolley dilemmas. Interventions seem to play a particularly important role in people's reasoning about these

dilemmas, as the locus of intervention determines the local contrast in the *Double Causal Contrast Theory*, and the fact that harming lies on the path between intervention and saving in *footbridge* prevents the selective highlighting of the intervention-saving relation. A formalization of interventions on causal networks is provided by Causal Bayes Nets Theory (Waldmann, 2017; Sloman, 2005; Pearl, 2000; Pearl & Mackenzie, 2019; Rottman & Hastie, 2014).

1.4.2. Responsibility attribution in groups

In moral dilemmas, one action has several effects, some good and some bad, some intended and some unintended. While dilemmas, and common-cause structures in general, are an important class of situations that trigger moral reasoning, other causal structures are common as well. One example are situations in which an effect has multiple causes (common-effect structures), such as a basketball team winning a match due to multiple players scoring points, or a group of friends working out the correct answer to a question in a pub quiz. Individual members of such teams are readily blamed or praised for their contribution to the overall outcome in such situations. A common reflex is to assume a simple "diffusion of responsibility" model for all cases in which credit or blame have to be allocated to multiple agents (see, e.g., Darley & Latané, 1968). However, research has shown that people are highly sensitive to the ways in which individual contributions combine to cause or prevent outcomes. An individuals' responsibility for a team outcome crucially depends on their own actual contribution, but also on those of the other team members, and even on sophisticated considerations about the number of possible counterfactual situations in which the agent could have made a difference to a global outcome, and how different these counterfactual scenarios are to what actually happened (Gerstenberg & Lagnado, 2010; Zultan, Gerstenberg, & Lagnado, 2012; Lagnado, Gerstenberg, & Zultan, 2013). Recent extensions of these frameworks combine such considerations about causal contribution with inferences about agents' character,

skills, or dispositions (Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021; Gerstenberg et al., 2018).

These models highlight the importance of counterfactual inferences for judgments of blame, praise, credit or responsibility. Counterfactual inferences, in turn, cannot get off the ground without a causal model. In order to gauge what would have happened if things had been slightly different, reasoners need to have some representation of how a situation's relevant variables generally affect each other. Causal Bayes Nets model counterfactual reasoning as simulated interventions on causal models, i.e., setting a target variable to a particular value and then reading off the resulting states of subsequent variables (Waldmann, 2017; Sloman, 2005; Pearl, 2000; Pearl & Mackenzie, 2019; Rottman & Hastie, 2014). While the responsibility attribution models described here emphasize the role of causal structure (i.e., whose actions affect the overall outcome) and of integration rules (i.e., how do multiple acts combine to bring about the outcome), the role of causal strength and of the value of outcomes is generally not investigated.

1.4.3. Causal selection

While multiple agents' actions (and other causes) contribute to outcomes in many cases, we often have the tendency to single out one factor and label it as "the" main or most important cause. This phenomenon is called causal selection (see, e.g., Hesslow, 1988). When two agents perform an otherwise identical action, and both of these actions are necessary for some harmful outcome to be produced, people often select the action that was norm-violating in some sense, for example by being unusual or forbidden. A famous example is the so-called *pen vignette* (Knobe & Fraser, 2008; Hitchcock & Knobe, 2009), in which two employees of a university department are described, both of which take pens from a receptionist's desk. One of the agents, an administrative assistant, is allowed to take pens, while the other one, a professor, is not. After each agent has taken a pen, the receptionist is out of pens, which leads to a problem later on when they have to take

an important note. When asked to what extent the two agents "caused the problem", people generally agree that the professor caused it, and generally disagree that the administrative assistant caused it. This general pattern has been replicated many times and in multiple scenarios (Icard, Kominsky, & Knobe, 2017, see also Willemsen & Kirfel, 2019, for an overview). Such findings have sometimes been interpreted as showing that causation is not actually the foundation of moral reasoning, but that the relationship may in fact be the other way round (Knobe & Fraser, 2008). After all, the *moral status* of an act seems to determine whether it is perceived to have *caused* an outcome here. If that was the case, an investigation into the role of causal representations in moral reasoning, as envisaged in this thesis, would be largely futile.

More recently, however, it has been recognized that setting up a descriptive causal model of a situation and selecting some cause as most important are two separate stages of causal reasoning, and that, in fact, the latter presupposes the former (see, e.g., Knobe & Shapiro, 2021). After all, a causal *selection* task can only arise once reasoners have understood that there are multiple causes of an outcome in the first place. More importantly, it has been demonstrated that considerations about the normative status of causes do not alter people's representations of causal structure or causal strength (Samland & Waldmann, 2016). However, when people use the verb "to cause something" in everyday life, it seems that they usually express much more than their descriptive world model. Instead, they use the term in an evaluative and selective way, indicating who should be blamed or praised (Livengood & Sytsma, 2020; Schwenkler & Sievers, in press; Samland & Waldmann, 2016), or what should be changed in future situations to produce good outcomes and prevent bad ones (Morris et al., 2018; Hitchcock & Knobe, 2009). Nevertheless, a descriptive causal model consisting of information about structure and strength is the necessary prerequisite for any such inferences and judgments.

1.4.4. Summary

In this section, we have seen that specific phenomena in moral judgment, such as the different evaluation of threat vs. victim interventions, or transfer effects between moral dilemmas, are best explained by analyses that draw on people's causal models of the situations in question (specifying which outcomes are brought about by particular actions, and how). These findings demonstrate that analyzing moral scenarios in terms of causal representations can be fruitful.

The aim of this thesis is to contribute to an improved understanding of how causal strength, causal structure, and outcomes generally affect moral judgments. In contrast to previous "top-down" approaches, in which particular phenomena in moral reasoning were singled out and explained with the help of a causal model analysis (such as transfer effects between moral dilemmas, responsibility attribution in groups, or patterns of causal selection), the approach of the work presented here is more "bottom-up". That is, we systematically varied different features of causal representations, namely causal structure, causal strength (Chapter 2), and outcomes (Chapter 3), and investigated how moral judgments are affected as a result. The aim is to deliver useful building blocks for further theorizing about the relationship between causal and moral reasoning.

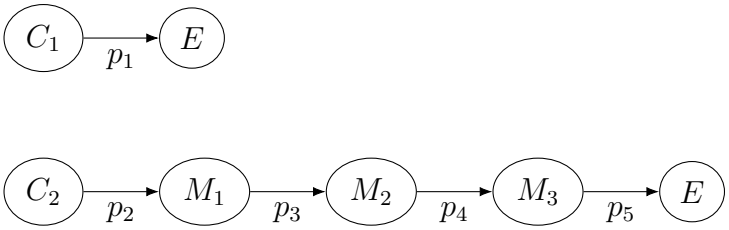
The following chapter presents experiments in which we manipulated the structure by which someone's action and accidental harm to another person are related (directly or via a longer chain of intermediate events), and the strength of the relation (high vs. low). We show that a chain structure can serve as a proxy for a weaker causal relation between action and harm, which leads to a more positive moral evaluation of action and agent. We also demonstrate that the more lenient moral evaluation is rooted in inferences about agents' mental states, based on the causal model. The harm that agents cause is kept constant in these experiments. The subsequent Chapter 3 pursues the opposite strategy, that is, varying outcomes while keeping strength and structure

fixed.

2. How causal structure, causal strength, and foreseeability affect moral judgments

Imagine a doctor in an emergency situation has to quickly decide which of two medications to administer to an unconscious patient. Both drugs would fulfill their main purpose, saving the patient's live, equally well. However, both drugs also carry a risk of an unwanted side effect, say, the development of a chronic feeling of dizziness. The doctor remembers the following: Drug A causes dizziness directly. Drug B, on the other hand, sometimes causes patients to feel more energized than usual after their recovery. This increase in energy makes patients sometimes more physically active in their day-to-day lives, which can lead to a depletion of iron levels. As a result of low iron, patients sometimes experience a chronic feeling of dizziness. Would it be better to administer drug B rather than drug A? Would the doctor be less blameworthy or less morally responsible for a patients' dizziness if they administered drug B? If so, why?

Figure 4: A direct causal relation and a longer causal chain.



Note: C = cause, E = effect, M = mechanism.

This example features two identical actions (administering a medication), and two identical, negatively valued outcomes (a patient suffering from a chronic feeling of dizziness). The action can also cause the outcome in both cases. However, the nature of the relation differs. Drug A is construed as causing dizziness directly, while in the case of drug B, several intermediate steps are described. Figure 4 provides an illustration.

Causal chains, in comparison to their direct counterparts, are an interesting point of departure for an investigation into the effects of causal structure and causal strength on moral judgments. It has repeatedly been demonstrated in different contexts and paradigms that agents who cause harm indirectly are evaluated more positively than agents who do so directly (Royzman & Baron, 2002; Ziano et al., 2021; Cushman et al., 2006; Greene et al., 2001; Hauser et al., 2008, for an overview see Sloman, Fernbach, & Ewing, 2009). As such, actions can be seen as more permissible when their potential harms are construed as more distant. And in retrospect, agents can be seen as less morally responsible for harms when their action is described as a less direct cause of the harm in question. These intuitions are also echoed in legal contexts, where someone's action is sometimes not considered as sufficiently "proximate" to some damage or harm for them to be held liable for it (Knobe & Shapiro, 2021; J. T. Johnson & Drobny, 1985).

However, a shortcoming of these accounts is that they do not clarify why and how indirectness can lead to a more lenient moral evaluation of agents and actions. Particularly, they do not distinguish between the respective influence of causal structure and causal strength on moral judgments. The most salient difference between direct and indirect relations is certainly their causal structure. As a consequence of this structural difference, however, they can also differ in their perceived causal strength, with the overall relation between action and outcome being seen as weaker in chains than in direct relations (Stephan, Tentori, et al., 2021). Either difference could produce the more favourable moral evaluation of actions and agents who cause harm indirectly. Below, I will explain and contrast two competing models of the cognitive mechanism by which causal chains might soften moral judgments. On one view (which we called the probabilistic model), the effect is ultimately driven by inferences about lower causal strength in chains. Another possibility is that there is a genuine effect of causal structure on moral judgments that is independent of causal strength. We called this the indirect-

ness model. The two models make different predictions about the respective effects of variations in causal structure and causal strength on moral judgments. We tested these predictions in three experiments.

Furthermore, we hypothesized that differences in people's causal model representations of moral scenarios, be they on the level of causal structure or on the level of causal strength, do not impact moral judgments *directly*. It is well established that moral judgments depend not only on the objective relations between actions and their effects in the world, but also crucially on agents' presumed or actual mental states (Cushman, 2008; Samland & Waldmann, 2016; Cushman, 2013; J. T. Johnson & Drobny, 1985; Alicke, 2000; Paharia, Kassam, Greene, & Bazerman, 2009; Fincham & Jaspars, 1983; Lagnado & Channon, 2008; Kirfel & Lagnado, 2021a; Kneer & Skoczeń, 2021; Kneer & Machery, 2019; Nobes & Martin, 2021). In the context of chains, outcome foreseeability seems particularly relevant. We propose that chains lead to more positive moral evaluations of actions and agents because reasoners take the agents to be less able to *foresee* harm being caused by their action, compared to when the relation is direct. We tested this hypothesis as well. Both the indirectness model and the probabilistic model can accommodate a mediating role of foreseeability attributions. Under the probabilistic model, outcomes are seen as less foreseeable because they are seen as less likely to occur. On the indirectness model, the indirect nature of the causal relation itself creates an impression of lower foreseeability.

2.1. The probabilistic model

Formally, causal structure and causal strength are independent in Causal Bayes Nets Theory. Looking back at the two structures in Figure 4, this means that the mere fact that one of the relations is indirect gives us no indication that it should be weaker than its direct counterpart. It all depends on the strength of the individual links p_1 to

p_5. Assuming the validity of the Markov condition (that is, parent nodes screen their children off from more distant parts of the network), the overall strength of the relation between an initial and final node in a causal chain is given by multiplying the strengths of all individual links that are part of the chain. For Figure 4, this is $p_2 \times p_3 \times p_4 \times p_5$. Depending on how the product of the strengths p_2 to p_5 compares to the strength of the single causal link in the direct relation, p_1, and unless all causal links are deterministic ($p = 1$, which is rarely the case), the indirect relation could be weaker, stronger, or equally strong as the direct one. Then how should effects of indirectness on moral judgments be rooted in inferences about weaker causal relations, as the probabilistic model suggests? It has in fact been demonstrated that people tend to assume roughly constant priors for the strength of causal links in the absence of exact knowledge about them (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Stephan, Tentori, et al., 2021). Such assumptions might also be aided by the use of verbal labels that convey probabilistic information, such as "sometimes causes" (Meder & Mayrhofer, 2017). If reasoners assign a roughly constant value to all individual causal links in Figure 4, no matter if they are part of the direct relation or the causal chain, it follows that the overall causal relation is weaker in the indirect case. For instance, if the strength of all single links p_1 to p_5 was 0.8, the strength of the relationship between action and harmful outcome would be 0.8 in the direct relation, but $0.8^4 = 0.41$ in the indirect relation.

As such, causing harm indirectly may be seen as morally better simply because the harm is perceived as less likely to be produced by an action, compared to when the causal relation is construed as direct. The probabilistic model thus proposes that effects of causal structure on moral judgments are ultimately due to people's inferences about causal strength. When overall relations are perceived as equally strong in a chain as in a direct relation, the probabilistic model predicts that effects of causal structure on moral judgments should disappear.

2.2. The indirectness model

The indirectness model makes the opposite prediction. According to this hypothesis, causal structure has a genuine effect on moral judgments, and causing harm indirectly rather than directly may be seen as morally better even when the probabilistic relation between action and harm is equally strong in both cases. A theoretical justification for such effects could come from Construal Level Theory (Trope & Liberman, 2010), according to which different forms of psychological distance (e.g., spatial, temporal, social, and others) can have similar downstream effects on a range of judgments and decisions (but see Calderon, Mac Giolla, Ask, & Granhag, 2020; Maier et al., 2022; Žeželj & Jokić, 2014). Possibly, a chain representation thus creates an impression of "causal distance" that can have a softening effect on moral judgments.

Empirically, effects of indirectness on moral judgments are well established (Royzman & Baron, 2002; Ziano et al., 2021; Cushman et al., 2006; Greene et al., 2001; Hauser et al., 2008, for an overview see Sloman et al., 2009). Some studies also show that indirect harms are perceived as less likely than direct harms (Royzman & Baron, 2002; Ziano et al., 2021). However, it is unclear whether people evaluate indirect harms as less negative because of the indirect relation itself (as the indirectness model predicts), or because of the lower likelihood of harm (as the probabilistic model predicts). Deciding between the two accounts requires fully crossing causal structure and causal strength in an experiment.

2.3. The mediating role of outcome foreseeability

According to normative ethical theories, the fact that one's action causally contributed to some outcome is generally regarded as necessary, but not sufficient for the ascription of moral responsibility for that outcome (see Rudy-Hiller, 2018). It is also required that the agent could reasonably *foresee* that the outcome might be brought about by their

actions. Similar requirements are found in legal definitions of negligence or recklessness (see Dubber, 2015, pp. 42-46, see also Kneer & Skoczeń, 2021; Nobes & Martin, 2021). Descriptively, people are also clearly sensitive to agents' mental states when it comes to making judgments about moral permissibility (Cushman, 2008, 2013; Paharia et al., 2009), blame (Fincham & Jaspars, 1983; Lagnado & Channon, 2008; Alicke, 2000; Samland & Waldmann, 2016), punishment (Cushman, 2008, 2013), liability (J. T. Johnson & Drobny, 1985), or agent causation (Fincham & Jaspars, 1983; Kirfel & Lagnado, 2021a; Lagnado & Channon, 2008; Alicke, 2000).

We thus expected that any effects of causal model representations, be it effects of causal structure or causal strength, would only alter moral judgments through attributions of outcome foreseeability. After all, when someone has no way of knowing that their action can cause harm, it should not matter for their moral evaluation whether this hidden causal relation is direct or indirect, weak or strong. When relations are known, on the other hand, knowing about an indirect or weak relation likely affords less foreseeability than knowing about a direct or strong relation. We proposed that this difference in attributed foreseeability drives the more favourable moral evaluation of actions and agents in causal chains compared to direct relations. That is, we think that when reasoners learn about an agent whose action can cause harm indirectly, they take this person to be less able to foresee that the harmful outcome will actually be brought about by their action, compared to an agent whose action can cause harm directly. According to our hypothesis, this is why the action is typically seen as more permissible, and the agent as less responsible, when the relation between their action and harm is indirect rather than direct. In cases where outcomes are made equally foreseeable or unforeseeable in a chain and a direct relation, we predicted no difference in moral judgments about actions and agents.

This hypothesis is compatible with both the probabilistic model and the indirectness

model. The two models only differ in what they take to be the cause of lower foreseeability in chains. According to the probabilistic model, harm becomes less foreseeable in chains because it is perceived as less likely. According to the indirectness model, harm appears less foreseeable simply because the relation is indirect.

2.4. Prospective and retrospective moral judgments

In our experiments, we investigated whether causal structure (chains vs. direct relations) would affect prospective judgments of the moral permissibility of actions, and retrospective judgments about the moral responsibility of agents. We opted for these two types of moral judgments because each of them requires the use of a distinct key function of causal reasoning. Permissibility judgments were prospective because participants had to make them *before* they knew whether harm would actually occur or not, based on information about a generic causal relation between action and harm. This requires *prediction* (see, e.g., Lagnado & Shanks, 2002). Judgments of moral responsibility were retrospective because participants had to make them *after* they knew both that the action had been performed, and that a harmful outcome had actually occurred. This, among other things, requires thinking about *singular causation*, that is, whether the action really caused the outcome in this case, or whether their co-occurrence is a mere coincidence (Stephan & Waldmann, 2018; Cheng & Novick, 2005). We were interested to see how both types of causal judgments would be affected by our causal structure manipulation, and how they, in turn, would affect foreseeability attributions and moral judgments.

2.5. Previous research on causal chains

Past research on causal chains in moral or legal contexts has largely focused on causal *selection*. For cases in which a causal sequence of events produced a harmful outcome,

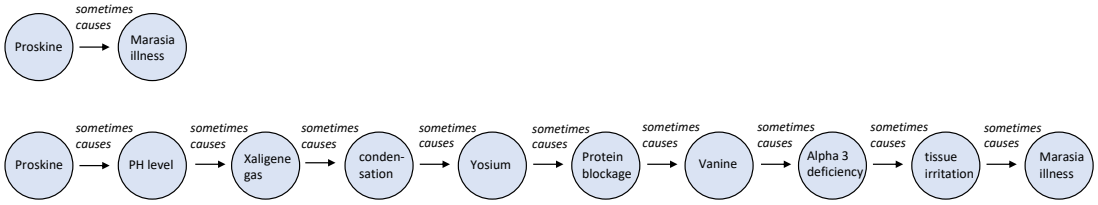
researchers were interested in how people determine which of these events should be seen as the main or most important cause of harm (Hart & Honoré, 1985; Hilton, McClure, & Sutton, 2010; McClure, Hilton, & Sutton, 2007; Livengood & Sytsma, 2020; Spellman, 1997; Lagnado & Channon, 2008; Knobe & Shapiro, 2021; J. T. Johnson & Drobny, 1985; Pearl & Mackenzie, 2019). For example, if one person shoots a gun at another, and this second person, while fleeing from the gunshots, accidentally pushes a bystander into oncoming traffic, should the shooter be liable for the bystander's death (for a similar case, see Pearl & Mackenzie, 2019, p. 288)? Or does the pushing by the fleeing man supersede the initial action, and should therefore be regarded as the cause of death? A number of factors have been identified that may influence the likelihood of a cause being selected as most important. These include the causes' position in the chain of events, its probabilistic relationship to the final outcome, its status as a physical event versus an intentional action, and its normality (Hilton et al., 2010; McClure et al., 2007; Lagnado & Channon, 2008; Spellman, 1997; Knobe & Shapiro, 2021). In any case, the task in these paradigms is to compare several causes within a causal chain, and classify one of them as most important. In contrast, we were interested in comparisons between different causal chains. Our chains featured the same initial action and the same final outcome, but differed in the number of intermediate variables. In the cases we used, all intermediate variables were physical events. We expected that the initial action would always be seen as the main cause of the final outcome in these cases (see also the initial example of the doctor prescribing two different medications), but we asked whether, and why, this action's causal proximity to harm might affect its moral evaluation. This is a different research question, but the results may also inform the causal selection debate. After all, one recent proposal is that the extent to which causes in chains are seen as "abnormal" (which includes "morally bad") affects their likelihood of being selected as most important (Knobe & Shapiro, 2021).

To our knowledge, only one study has directly compared cases like the ones we were interested in. J. T. Johnson and Drobny (1985) presented their participants with the case of a truck driver who forgets to replace some safety pin in his truck after an inspection. As a result, the steering later fails, resulting in an accident. In one condition ("short chain"), the accident causes a fire, which causes a nearby house to burn down. In another condition ("long chain"), some burning gasoline floats down a hill and across a river, setting fire to some grass on the other side, in turn also causing a house to burn down. Participants took the driver to be equally negligent in both conditions, but they rated him as less liable for the damage to the house in the "long chain" condition. They also took him to be less able to foresee the damage to the house in that condition, and indicated that it was less likely. These findings are consistent with the effects that we were expecting. However, J. T. Johnson and Drobny (1985) did not investigate the cognitive mechanism behind these effects. For instance, were liability judgments reduced because harm was seen as increasingly unlikely in the "long chain" condition, and therefore less foreseeable? Or did the instruction of an indirect relation suffice?

2.6. Summary of the empirical findings (Engelmann & Waldmann, 2022a)

In the following, I will summarize three experiments that investigated whether, and why, the presentation of a longer causal chain between an action and a harmful outcome leads to a more favourable moral evaluation of action and agent, compared to a direct causal relation. For more details, see Engelmann and Waldmann (2022a). For all materials, data, and code, see `https://osf.io/5bmgc/`. All studies were implemented in Unipark Questback and run as online experiments. Participants were recruited via `www.prolific.co`.

Figure 5: Materials of Experiment 1 in Engelmann & Waldmann (2022a).



Note: A direct causal relation and a longer causal chain with the same start- and endpoint. This figure is reproduced from Engelmann & Waldmann (2022a, Fig. 3).

2.6.1. Experiment 1

The aim of this experiment was to set the stage for our project by establishing that chains lead to more positive moral judgments about actions and agents than direct relations. We also wanted to find out whether this effect was mediated by attributions of outcome foreseeability to agents. We thus manipulated the causal structure connecting action and harm (chain vs. direct relation, within-subject), and whether agents were aware of the relation (knowledge vs. no knowledge, between subjects). We asked participants for judgments about the moral permissibility of actions, and the moral responsibility of agents. Both the indirectness model and the probabilistic model predict that chains should lead to more positive moral judgments, albeit for different reasons. We furthermore predicted that this effect should only occur when agents are aware of the causal relations between their action and harm. Without such knowledge, harms are equally unforeseeable in chains as in direct relations, which should result in identical moral judgments.

We constructed three vignettes about actions that could lead to harmful outcomes, either directly or via a chain of intermediate events. We kept the material rather artificial in all scenarios, to avoid any effects of knowledge or assumptions about link strength. For instance, we told participants that producing a certain (fictitious) chemical in a lab

could lead to a certain (likewise fictitious) disease in people who are exposed to the chemical. This relation was either described as direct, that is, producing the chemical in a lab "sometimes causes" the disease in exposed people, or as indirect (here: mediated by a chain of chemical reactions and mechanisms in the human body). In the indirect conditions, all individual links were also labeled with "sometimes causes". Besides the verbal description, participants were shown an illustration of the relationship (see Figure 5).

All participants saw both a case involving a direct causal relation, and a case involving a causal chain, but each case was presented within a different cover story. Cover story and causal structure were combined using a Latin Square, resulting in six unique combinations.

After learning about the generic causal relationship between an action and harm, participants were confronted with the case of an agent who is about to perform the action in question (e.g., producing the chemical). In the knowledge conditions, we informed participants that the agent was aware of all the information that they had previously learned as well. In the no knowledge conditions, we said that there was no way that the agent could have known about the relation, since the relevant research was not available to them.

Participants were asked to assess how morally permissible the action was. On a subsequent screen, they were informed that the harmful outcome had now actually occurred (e.g., a colleague of the agent's had contracted the disease). We asked them to what extent they considered the agent to be morally responsible for this outcome. After providing moral judgments about both cases (the chain and the direct relation), the cases were presented to participants again, and we asked them for a predictive causal judgment (given that the action has been performed, how likely is it that the harmful outcome will be produced?), and for a foreseeability attribution (to what extent could

the agent foresee that someone would be harmed by their action?), for each case.

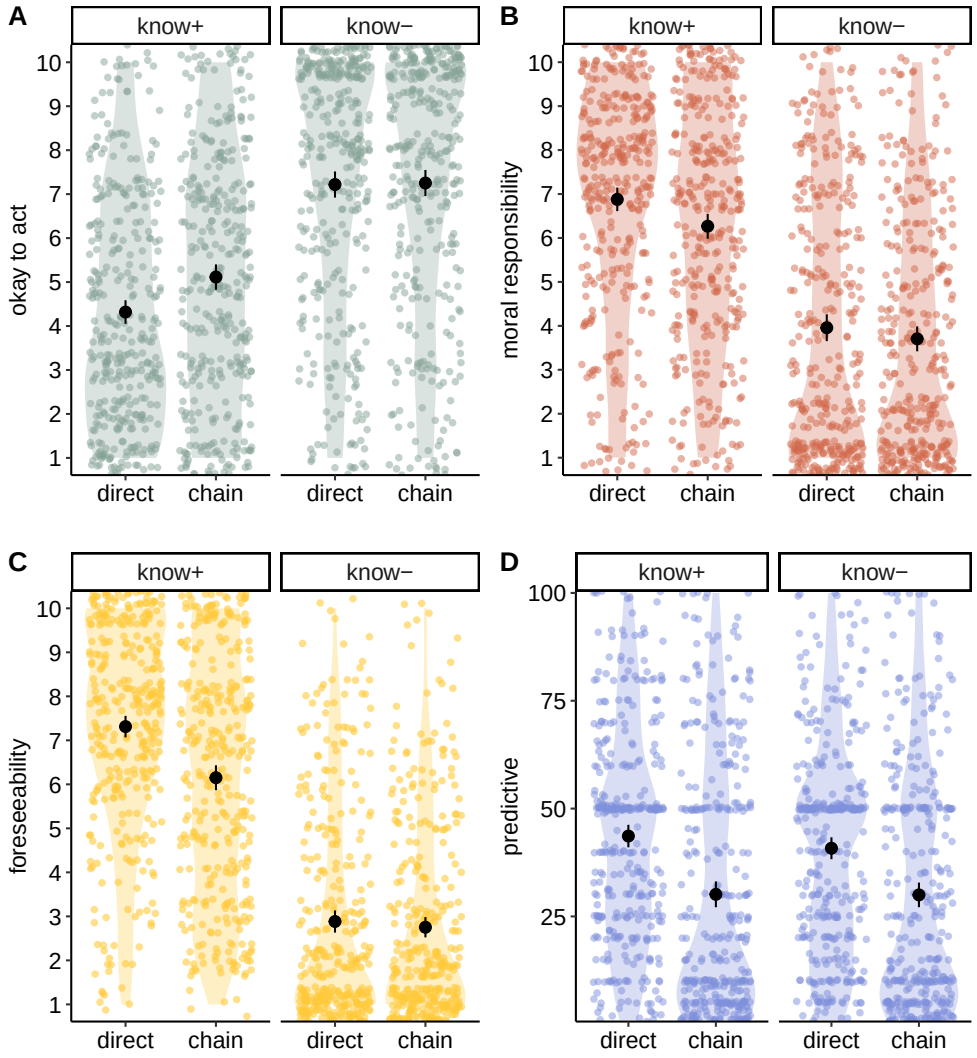
The results showed that actions and agents were indeed evaluated more positively when the relation between action and harm was a causal chain rather than a direct causal relation (see Figure 6). Actions were seen as more permissible, and agents as less responsible, when the relation was described as indirect rather than as direct. However, this effect depended on agents' knowledge, as predicted. When agents were unaware of the causal relations between their actions and harm, actions were generally seen as permissible, and agents as not responsible. For permissibility ratings, no detectable effect of the causal structure manipulation persisted without knowledge. For responsibility, there was still a significant effect without knowledge (thus, unlike for permissibility judgments, we did not find the predicted interaction between causal structure and knowledge for responsibility judgments). However, this effect was very small (see Figure 6B).

The foreseeability ratings confirmed that a chain representation indeed led to less outcome foreseeability than a direct relation, but only when agents were aware of the relations. Without such knowledge, outcomes were taken to be equally unforeseeable in chains as in direct relations.

Finally, participants predictive causal judgments showed that they took outcomes to be less likely to be produced by actions in chains, irrespective of agents' knowledge.

All in all, the results of this experiment confirmed that actions and agents are evaluated more favourably, in terms of morality, when the relation between their action and harm is construed as indirect rather than direct. It also confirmed the mediating role of attributions of outcome foreseeability. While we saw that participants also judged harms to be less likely to occur in chains, the results were still consistent with both the probabilistic model and the indirectness model. Deciding between the two models requires fully crossing causal structure and causal strength.

Figure 6: Results of Experiment 1 in Engelmann & Waldmann (2022a).



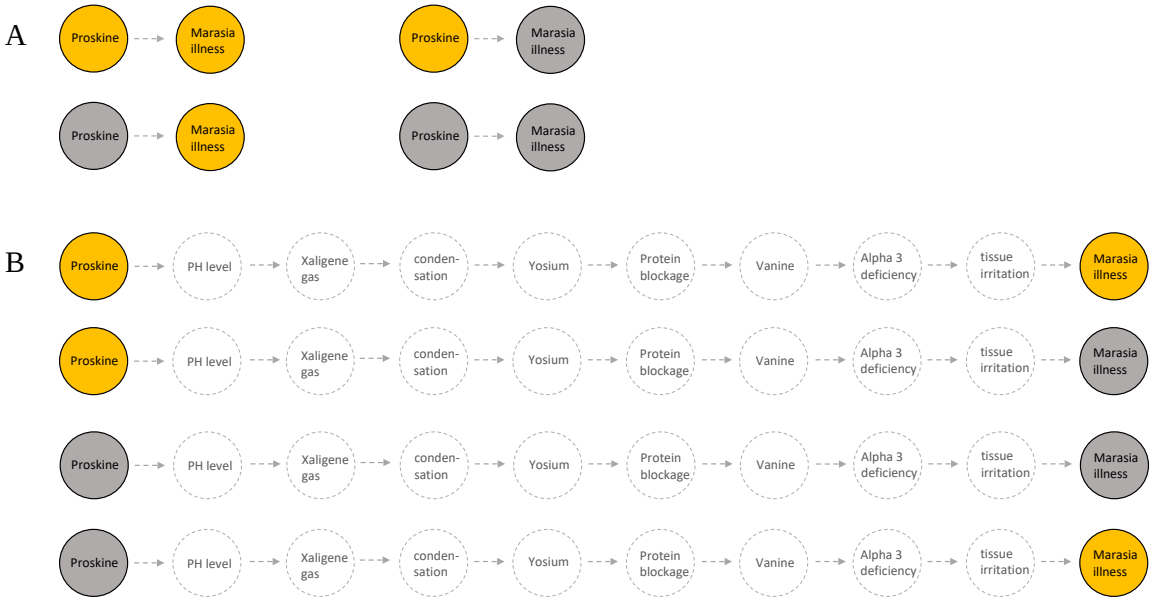
Note: Means and 95% confidence intervals of moral permissibility ratings (A), moral responsibility ratings (B), foreseeability ratings (C), and predictive causal judgments (D) per condition: Causal chains vs. direct relations, knowledge (*know+*) vs. no knowledge (*know-*) of agents about the causal relations. This figure is reproduced from Engelmann & Waldmann (2022a, Fig. 4).

2.6.2. Experiment 2

The aim of this experiment was to clarify whether agents and their actions are evaluated more favourably in indirect relations because the causal relation between their action and harm is perceived to be weaker (the probabilistic model), or whether indirectness itself drives the effect (the indirectness model). To this end, we now manipulated both causal structure (direct vs. chain, between subjects) and causal strength (high vs. low, between subjects). The probabilistic model predicts that only causal strength affects moral judgments: When strength is low, the permissibility of actions should be higher, and moral responsibility for harms should be lower, compared to when strength is high. Given equal strength, causal structure (direct vs. chain) should not matter for moral evaluations. The indirectness model, on the other hand, predicts that chains should still lead to more positive moral judgments of actions and agents than direct relations when the strength of the relations is equal.

The initial instruction phase of this experiment was identical to Experiment 1. Participants learned about a generic causal relation between an action and a harmful outcome, using the same three cover stories as in the previous experiment (the different cover stories were now manipulated between subjects). After the structure learning phase was completed, we informed participants that researchers had now also collected data about the strength of the causal relations. In the chemical example, we said that this data had been obtained by inspecting health records of lab workers who were exposed vs. not exposed to the chemical in question, and had or had not developed the disease. Participants then learned the contingency between the chemical and the disease in a trial-by-trial observational learning task with 40 observations (see Figure 7 for example stimuli). The true contingencies were $p(\text{outcome} \mid \text{action}) = .80$ in the high strength condition, and $p(\text{outcome} \mid \text{action}) = .20$ in the low strength condition. The harmful outcomes were never observed without the actions; $p(\text{outcome} \mid \text{no action}) = 0$.

Figure 7: Materials of Experiment 2 in Engelmann & Waldmann (2022a).



Note: Participants saw 40 learning trials involving these stimuli (either chains or direct relations). Grey nodes meant that an event was absent in a single case (e.g., a person had not been exposed to the chemical, or a person had not developed the disease). Yellow nodes meant that the event was present (e.g., a person had been exposed to the chemical, or a person had developed the disease). Light-grey, dashed nodes meant that the variable was not measured, and it was therefore unknown whether it was present or absent. This figure is reproduced from Engelmann & Waldmann (2022a, Fig. 5).

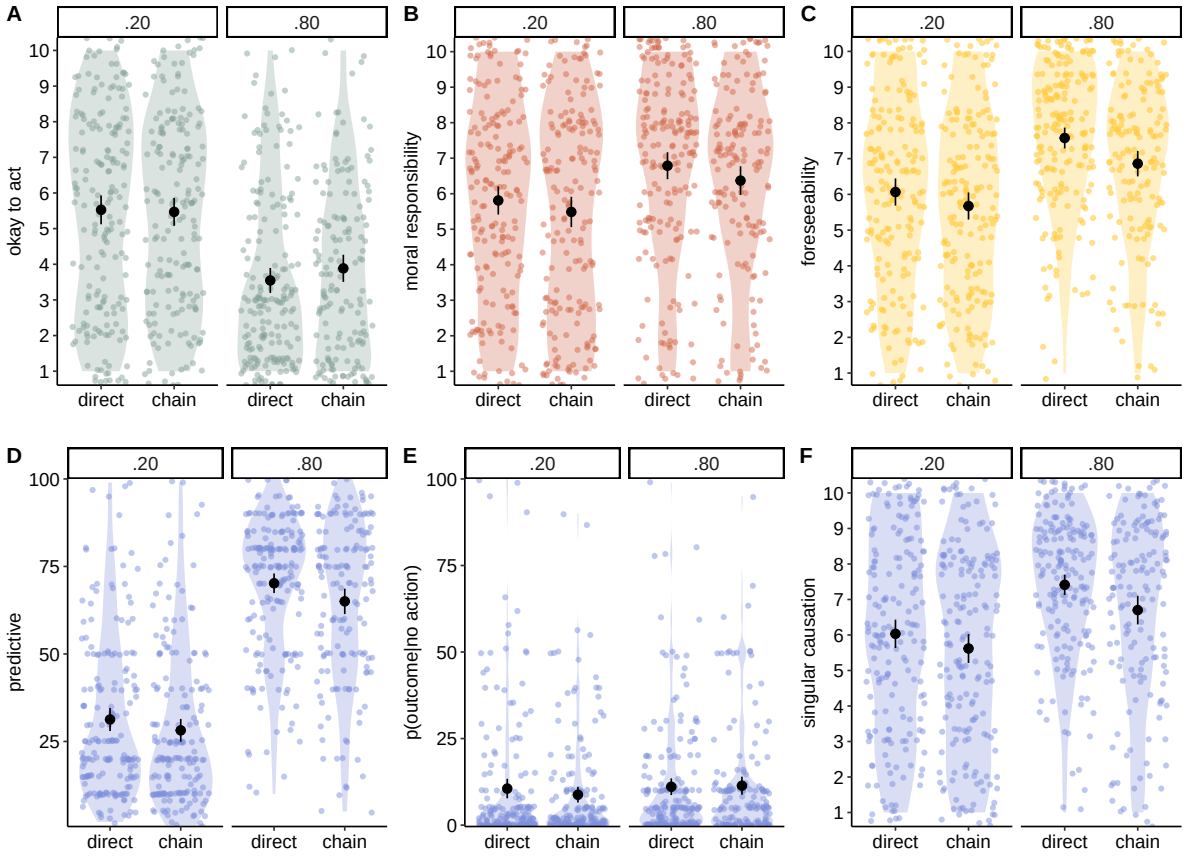
Once the strength learning phase was completed, participants were presented with the case of an agent who was about to perform the action in question (e.g., producing the chemical in a lab), as in the previous experiment. We asked for moral permissibility judgments, and for judgments of moral responsibility after the harmful outcome had actually occurred.

On the final page, we asked participants for some additional judgments about the case they had seen. As in Experiment 1, we assessed predictive causal judgments, which would allow us to see whether causal strengths were correctly learned. We also asked for foreseeability ratings, which we expected to increase with higher causal strength. A new measure in this experiment was singular causation. That is, we asked participants how confident they were that the agents' action had actually caused the harmful outcome in the case that they had evaluated. Next to foreseeability, the fact that actions are actually causally connected to outcomes is generally seen as a requirement for moral responsibility (Rudy-Hiller, 2018; Driver, 2008). All else being equal, and unless the existence of alternative causes of outcomes is categorically ruled out, confidence in singular causation should increase with higher causal strength (Stephan & Waldmann, 2018; Cheng & Novick, 2005). Finally, we asked participants how often they take the harmful outcome to occur *without* the action in question. Such cases were never observed in our learning task, but we wanted to capture people's assumptions about them.

While participants learned the contingencies well overall, the results (see Figure 8) showed that they still perceived chains to be somewhat weaker than direct relations, even though the true contingencies were identical for both structures (Figure 8D). Correspondingly, effects of causal structure were also observed for foreseeability attributions, confidence in singular causation, and moral responsibility (all lower in chains than in direct relations). Permissibility judgments were not affected by causal structure.

Given that chains are seen as weaker than direct relations, these effects are pre-

Figure 8: Results of Experiment 2 in Engelmann & Waldmann (2022a).



Note: Means and 95% confidence intervals of moral permissibility ratings (A), moral responsibility ratings (B), foreseeability ratings (C), predictive causal judgments (D), ratings of p(outcome|no action) (E), and singular causation ratings (F) per condition: Causal chains vs. direct relations. This figure is reproduced from Engelmann & Waldmann (2022a, Fig. 6).

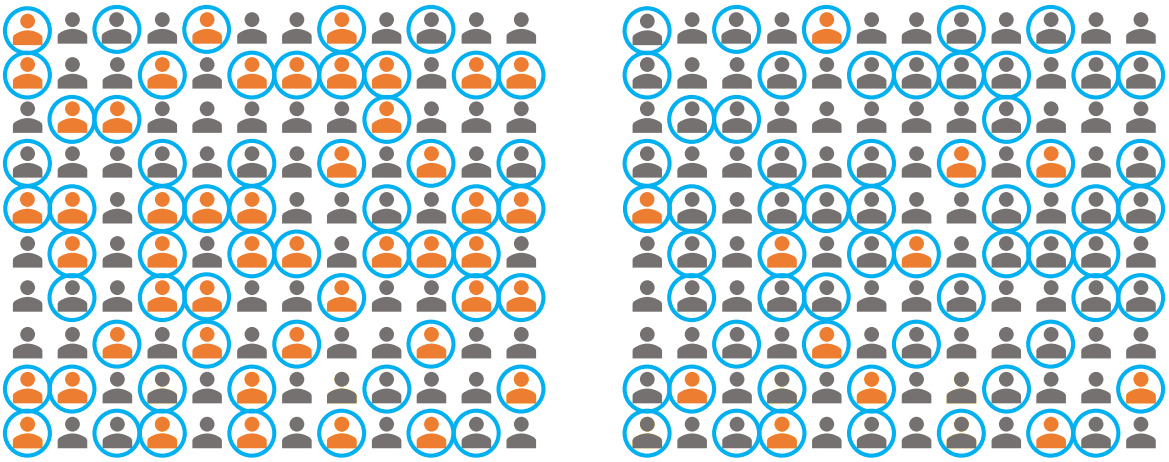
dicted by the probabilistic model. Furthermore, all relevant measures were substantially affected by causal strength in the expected directions (lower strength led to less foreseeability, lower confidence in singular causation, and a more lenient moral evaluation). These observations, and the fact that these effects were larger than the remaining effects of causal structure, also pointed towards the probabilistic model. The size of the remaining effect of structure on moral responsibility ratings was also substantially reduced, compared to a meta-analytic estimate based on Experiment 1 and pilot studies. Thus, it seemed that effects of causal structure on moral judgments were at least to a large extent mediated by inferences about causal strength. The crucial question remained whether they would disappear entirely when strength is not only objectively equal between chains and direct relations, but also *perceived* as equal by reasoners.

2.6.3. Experiment 3

In this experiment, we again crossed causal structure (direct vs. chain, within-subject) and causal strength (high vs. low, between subjects), but presented a higher number of observations in the strength learning task. We assumed that in Experiment 2, effects of structure on predictive causal judgments remained because 40 learning trials were not sufficient to overwrite participants' lower strength priors in chains, which they presumably formed when initially learning about the causal structure (Stephan, Tentori, et al., 2021). If this explanation is correct, presenting a larger number of cases should eventually lead to the perception of equally strong relations, creating the appropriate conditions for a strict test of the probabilistic model against the indirectness model.

A secondary aim was to further explore foreseeability requirements. In our previous experiments, agents in the scenarios were either aware of the full causal relation connecting action and outcome (causal structure and causal strength), or not aware of any aspect of it. However, it is also possible to be aware of the existence of a causal relation without knowing its strength (structure knowledge), or to be aware of a statistical

Figure 9: Materials of Experiment 3 in Engelmann & Waldmann (2022a): Learning data.



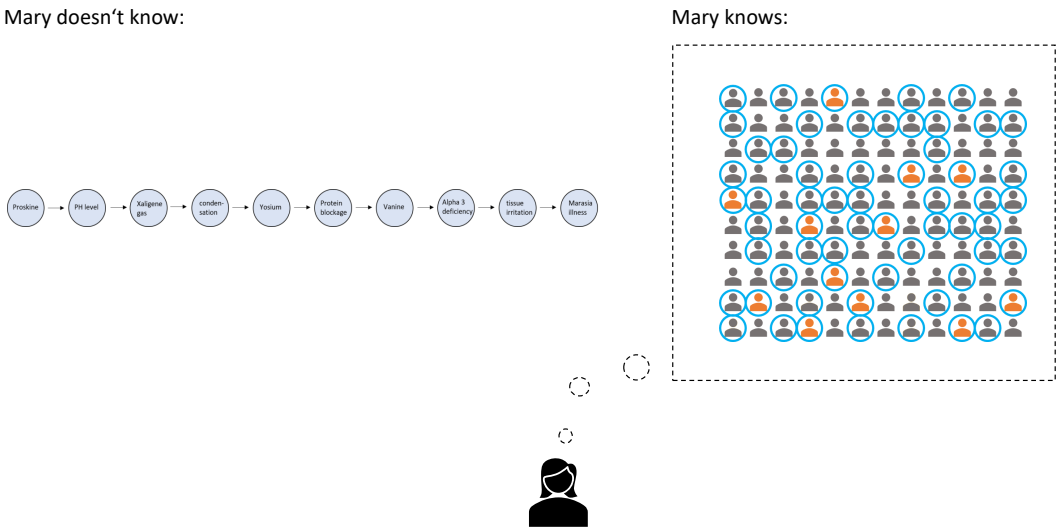
Note: Blue circles meant that the action had been performed (e.g., the chemical was produced), red person icons meant that the harmful outcome had occurred (e.g., the person had developed the disease). The left panel shows data for a high contingency (.80), the right panel shows data for a low contingency (.20). This figure is reproduced from Engelmann & Waldmann (2022a, Fig. 7).

association without knowing the underlying causal structure (strength knowledge). We were interested to see which implications such partial knowledge about causal models would have for moral judgments, and thus also varied knowledge in four levels (full vs. none vs. structure vs. strength, between subjects).

The experiment began in the same way as the previous one, but we used a summary format for the strength learning task instead of trial-by-trial learning (see Figure 9). This way, we could present participants with three times as many cases as before (120 instead of 40), while keeping the experiment at a reasonable length. The true contingencies were identical to Experiment 2. When initially instructing the causal structures, we also removed the "sometimes causes"-labels from the individual causal links, and instead said that "there is a causal relation between [A] and [B]". By removing the probabilistic labels, we hoped to further facilitate the learning of equally strong relations.

When describing agents' knowledge, we either said that they were aware of all the

Figure 10: Materials of Experiment 3 in Engelmann & Waldmann (2022a): Knowledge manipulation.



Note:. In this case, the agent is aware of the statistical association between action and harmful outcome, but does not know the underlying causal structure (strength knowledge). This figure is reproduced from Engelmann & Waldmann (2022a, Fig. 8).

information that participants had learned (full knowledge), none of this information (no knowledge), or just some of it. In the structure knowledge condition, we said that agents were aware of the lab study in which the causal mechanism connecting action and outcome was uncovered. However, the research in which strength was determined was not available to them. In the strength knowledge condition, we gave the opposite instruction. We also provided illustrations of agents' knowledge (see Figure 10 for an example). Only participants who correctly answered manipulation check questions about agents' knowledge were included in the final analyses. All other measures were identical to Experiment 2.

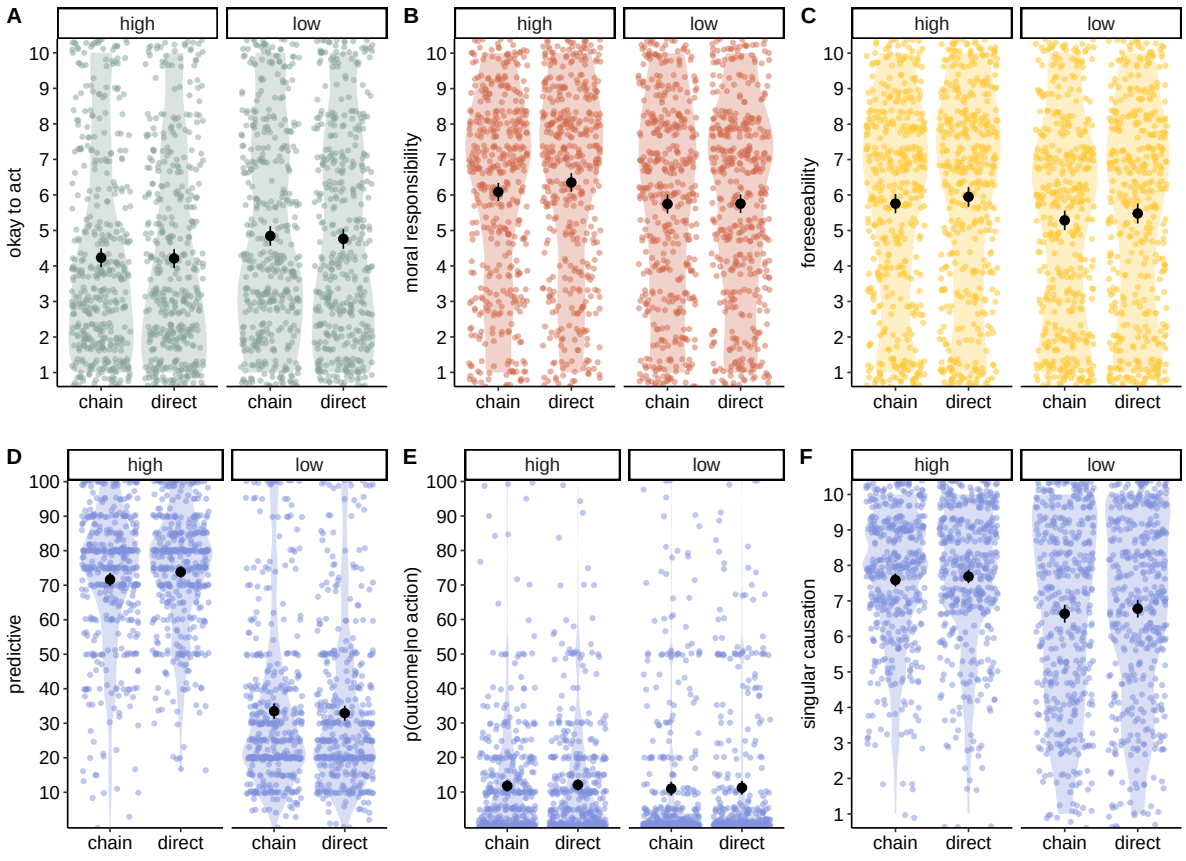
As we had hoped, participants now considered the overall causal relations between action and harm to be equally strong in chains as in direct relations, and they were also equally confident that actions caused outcomes (singular causation) in chains as in direct relations. In line with the predictions of the probabilistic model (and contrary to

the predictions of the indirectness model), neither judgments of moral permissibility nor judgments of moral responsibility were substantially affected by causal structure under these circumstances.[4] Actions only became more permissible, and agents were only seen as less responsible when harmful outcomes became less likely (see Figure 11 for the results of this experiment, and see Fig. 12 for a comparison of the effect sizes for structure and strength between all three experiments presented in this chapter). The only finding that was predicted by the indirectness model and not the probabilistic model was the observation that outcomes were still considered to be slightly less foreseeable in chains than in direct relations. However, this effect was very small. The fact that this small difference in foreseeability did not alter moral judgments is in line with the observation of previous experiments (here and also in Engelmann & Waldmann, 2021) that more considerable differences in foreseeability seem to be required for moral judgments to be affected (that is, effects of manipulations on foreseeability were usually larger than effects on moral judgments).

The new knowledge manipulation revealed that agents with partial knowledge were judged as roughly equally responsible, and their actions as roughly as impermissible as those of agents with full knowledge. Participants might have expected agents to derive strength knowledge from structure knowledge, and vice versa. Another possibility is that it was difficult for participants to represent such states of partial knowledge. Agents without any knowledge about causal relations, on the other hand, were generally seen as not responsible, and their actions as permissible. While we found no significant interactions of the strength manipulation with knowledge, descriptively, effects of strength on moral judgments were largest when agents had full knowledge, and absent without knowledge. Strength and structure knowledge were in between.

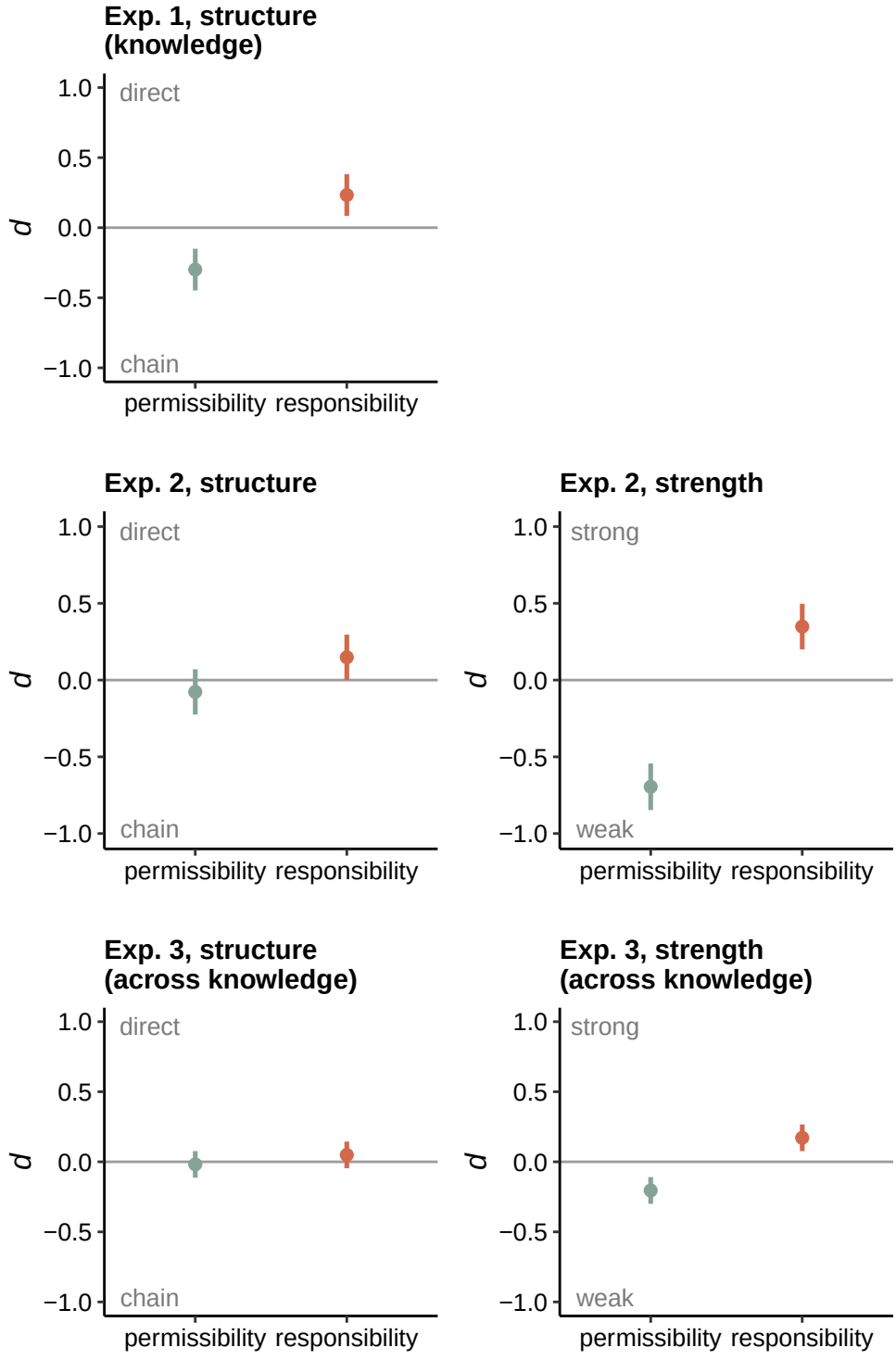
[4]There was a small effect of structure for responsibility that was only significant in a one-tailed t-test. Including causal structure in the regression model for responsibility ratings that also accounted for the other manipulations did not improve the model fit.

Figure 11: Results of Experiment 3 in Engelmann & Waldmann (2022a).



Note: Means and 95% confidence intervals for moral permissibility ratings (A), moral responsibility ratings (B), foreseeability ratings (C), predictive causal judgments (D), ratings of p(outcome|action) (E), and singular causation ratings (F) per condition: Causal chains vs. direct relations, and high strength vs. low strength. Data are collapsed across knowledge conditions. Descriptively, effects of causal strength on moral judgments and foreseeability ratings were larger than the overall effects that are visible here when agents had full or partial knowledge about the causal relations, and they were largely absent without such knowledge. However, the interaction between knowledge and causal strength was not statistically significant. See Appendix of Engelmann & Waldmann (2022a) for plots of these measures per knowledge condition. This figure is an extended version of Fig. 9 in Engelmann & Waldmann (2022a).

Figure 12: Comparison of effect sizes for causal structure and causal strength between all experiments presented in Engelmann & Waldmann (2022a).



Note: Cohen's d and 95% CIs. When causal structure and causal strength were manipulated independently (Exp. 2 and Exp. 3), the influence of structure on judgments of moral permissibility and moral responsibility became smaller and finally disappeared.

2.7. Summary and Discussion

This project started with the premise that descriptive causal models of situations underlie moral judgments about the agents and actions involved in that situation (Sloman et al., 2009; Waldmann, Wiegmann, & Nagel, 2017; Lagnado & Gerstenberg, 2017). Variations in causal model representations (causal structure and causal strength) should therefore be able to produce variations in moral judgments. We explored this general prediction using the case of moral judgments about causal chains between actions and accidental harms, comparing them to direct causal relations. Causal chains are a particularly interesting starting point for such an investigation for two reasons. First, previous research has revealed that knowledge about causal structure can affect inferences about causal strength in chains (Stephan, Tentori, et al., 2021; Bés, Sloman, Lucas, & Raufaste, 2012). Second, the fact that causing harm indirectly rather than directly is seen as morally better is a well-established phenomenon in moral psychology (Royzman & Baron, 2002; Ziano et al., 2021; Cushman et al., 2006; Greene et al., 2001; Hauser et al., 2008; J. T. Johnson & Drobny, 1985, for an overview see Sloman et al., 2009), but whether such effects are grounded in differences in causal structure or causal strength is unclear.

We derived two competing models that could account for a more favourable moral evaluation of agents and actions in indirect relations. The probabilistic model posits that effects of causal structure manipulations on moral judgments are ultimately due to inferences about causal strength. In a causal chain, the overall relationship between action and outcome simply appears weaker than in a direct relation, which is why actions are seen as more permissible, and agents as less responsible. According to the indirectness model, on the other hand, causal structure itself is the driving force behind the effect. On this view, actions and agents should receive more favourable moral evaluations in chains than in direct relations even if the probabilistic relationship between action and outcome

is equally strong in both cases. Both models assume a mediating role of attributions of outcome foreseeability. That is, both propose that the reason for the more positive moral evaluations in chains is that people take agents to be less able to foresee the harmful outcome, compared to direct relations. However, the probabilistic model holds that outcomes are seen as less foreseeable because they are perceived as less likely, while the indirectness model posits that the indirect nature of the relationship per se creates an impression of lower foreseeability.

We explored the competing predictions of the two models in three experiments. Experiment 1 confirmed that actions are seen as more permissible, and agents as less responsible, when the relation between their action and a harmful outcome is indirect rather than direct. It also confirmed that this effect is mediated by attributions of outcome foreseeability to agents. While these results are compatible with both models, Experiments 2 and 3 were designed to test the two models against each other. In the end, the results we found were more compatible with the probabilistic model than with the indirectness model. Once the overall relations between actions and outcomes were perceived as equally strong in chains and direct relations, moral judgments ceased to be affected by causal structure. However, they were reliably modulated by causal strength. In weaker relations, actions were rated as more permissible and agents as less responsible, compared to stronger causal relations. The only finding that was predicted by the indirectness model and not the probabilistic model was the fact that outcomes were still seen as slightly less foreseeable in chains than in direct relations in Experiment 3, even though the relations were now perceived to be equally strong. This suggests that indirectness itself is in principle capable of affecting foreseeability, but this influence seems to be very weak and consequently it had no implications for moral judgments here.

2.7.1. Relationship to theories of moral judgment

Generally, the results we found are compatible with all global theories of moral judgment that reserve a role for causal reasoning (see Section 1.3). Some researches have also developed more concrete accounts of single types of moral judgments. For instance, Cushman (2008) proposes that prospective moral judgments such as permissibility are primarily influenced by agents' mental states, whereas retrospective moral judgments such as responsibility or blame are affected by mental states, but also by the actual causal connection between action and harmful outcome. This is consistent with our results. For prospective judgments of moral permissibility, actions became more permissible the less foreseeability participants attributed to agents in the scenario regarding the occurrence of the harmful outcome. When changes in causal models did not alter foreseeability (no knowledge conditions in Experiments 1 and 3), they did not lead to changes in permissibility judgments either.

Likewise, agents were judged as less morally responsible when participants attributed less a priori foreseeability of harm. Without foreseeability, they were overall not taken to be morally responsible, even though participants still agreed that their action caused the harm in question (singular causation). We did not independently manipulate singular causation here. That is, we did not include cases in which agents acted, knowing they could cause harm by doing so, but the harm either didn't occur or was produced by alternative causes (but see Cushman, 2008). However, we found that in Experiments 2 and 3, the effect of causal strength on judgments of moral responsibility was fully mediated by foreseeability attributions *and* confidence in singular causation (see Supplementary Materials of Engelmann & Waldmann, 2022a). Thus, our findings are at least consistent with the view that moral responsibility requires both foreseeability and singular causation (see also Rudy-Hiller, 2018).

An interesting avenue for further research could be to confirm that a *causal* notion of

foreseeability is at play here. That is, we expect that agents need not only foresee that some harm could occur, given their action. This requirement would also be fulfilled if action and harm were both caused by a third variable, and only spuriously associated. Instead, agents presumably need to foresee that the harmful outcome might be *caused* by their action. Future studies could design scenarios that differentiate between these two senses of foreseeability. Our prediction is that moral judgments would only depend on the causal sense.

2.7.2. Content effects, transitivity, and the granularity of causal relations

Given the findings we obtained in our experiments, one may wonder whether *any* harmful action would appear morally better when more of the variables that mediate between action and harm are made explicit. After all, the chains we contrasted could potentially represent the same causal mechanism, just at different levels of granularity. What if the harmful action was hitting someone's head with a hammer, and the negative outcome that second person's injury? Intuitively, it does not seem that explaining the exact way in which the attacker's brain sent signals to their muscles, which led them to contract, causing their arm to move in such a way that the hammer in their hand collided with the victim's head would make a moral difference. But why? Several factors could be at play here. First, hitting someone's head with a hammer is likely intentional. We focused on cases of *accidental* harms in our experiments, showing how variations in foreseeability led to more lenient evaluations. It is possible that intending harm (which normally entails maximal foreseeability) and successfully causing it is rather immune to variations in the causal structure and strength connecting action and outcome (although the literature on so-called deviant causal chains suggests that not just the outcome, but also the causal mechanism may need to be foreseen for full responsibility, see Alicke & Rose, 2012; Pizarro, Uhlmann, & Bloom, 2003). Another possibility is that in the hammer example, our prior knowledge about the high strength of the relation would

not allow it to be sufficiently weakened by the introduction of additional intermediate variables.

Finally, research has shown that how connected certain variables in a causal network appear to be is sometimes not only determined by the strength of the probabilistic relations between them. On the one hand, assumptions about transitivity (if A causes B, and B causes C, A also causes C) can make people infer a causal relation between two variables in the absence of a statistical association between them (Von Sydow, Hagmayer, & Meder, 2016). On the other hand, even strong statistical associations can sometimes be disregarded when the variables in question do not belong to the same semantic schema or "chunk" (S. G. Johnson & Ahn, 2015). S. G. Johnson and Ahn (2015) for instance presented participants with a case in which a person steps on a dog's tail, causing the dog to growl. The dog's growling, in turn, scared a kid. Even though participants agreed that both causal links were strong, they tended not to agree that the person stepping on the dog's tail caused the kid to be scared. Thus, they took the relation to be intransitive. In other cases in which individual links were seen as equally strong, such as the case of a person exercising, getting thirsty, and hence drinking water, relations were seen as transitive. The hammer example seems strongly "chunked" as well, maybe to an extent that would make it difficult to plausibly construe the relationship as indirect in any way.

How semantic chunking is best explained is currently unclear. To test whether a given three-element chain (A $\rightarrow$ B $\rightarrow$ C) was perceived as chunked or not, S. G. Johnson and Ahn (2015) asked participants whether mentioning the middle element, B, in conversation would be necessary to explain why A led to C. If not, the relationship between A and C was classified as chunked. This test question points towards a possible relationship between semantic chunking and preferences about the granularity (or resolution) of causal relations. In principle, we could zoom into each "direct" causal relation down to the level of atoms and their interaction (see also Stephan, Tentori, et al., 2021). Yet,

we do not usually care about that much detail in everyday life. How we determine the appropriate granularity of causal relations is subject to ongoing debate (see, e.g., Woodward, 2021). In our experiments, chains might have been more detailed and "low-level" than participants would have naturally construed such cases (see also Kinney & Lombrozo, 2022). It is possible that cases which are more naturally construed as longer chains would produce stronger effects of causal structure on moral judgments, given equal overall strength. In any case, investigating the relationship between granularity preferences, semantic chunking, (in-)transitivity of chains and moral judgments seems like a promising avenue for further research.

2.7.3. Accidental harms – Influences beyond causation and foreseeability

Our experiments showed that variations in causal model representations can change the level of outcome foreseeability that is attributed to agents when they cause harm accidentally. These modulations of foreseeability, in turn, can lead to a more or less severe moral evaluation of agent and action. In the case of moral responsibility judgments, another relevant influence seems to be confidence in singular causation, that is, in the claim that the agents' action actually caused the harmful outcome in this case.

Legally, agents can be liable for accidental harms that were caused by their actions if they are found to have acted negligently or recklessly. Put briefly, acting negligently means that while the agent may not have been aware that their action could be harmful when they performed the action, they could and should have been aware. An agent acts recklessly when they were in fact aware of the risk that their action would pose at the time of acting, but proceeded regardless (see e.g. Dubber, 2015, pp. 42 - 46). The conditions for liability that are spelled out here, causation plus a "guilty mind" (*mens rea*), match our experimental manipulations. A number of other empirical studies have also recently shown that people's ascription of what agents could and should have foreseen play a crucial role in the assessment of culpability for accidental harms (Kneer &

Machery, 2019; Kneer & Skoczeń, 2021; Nobes & Martin, 2021). But are causation and foreseeability always sufficient for liability for accidental harms? In legal contexts, this is not necessarily the case. In *United States v. Carroll Towing Co.* (1947)[5], a formula is famously suggested according to which an actor might *not* be liable for accidental harm caused by their actions when the burden of taking adequate precautions against it would have outweighed the severity of the harm caused, multiplied by its probability. We expect that such considerations will also be at play in laypeople's moral judgments. Furthermore, it is relevant to consider the utility and probability of the primary goal that an agent was pursuing with their action (Engelmann & Waldmann, 2022b). Formulating a principle from all of these requirements and considerations might converge into something similar to the *Doctrine of Double Effect* (see McIntyre, 2019). In any case, we predict that people will not morally condemn any agent who caused harm foreseeably, but that their judgments will instead be some product of the interplay of the cited factors.

2.7.4. Conclusion and outlook

This chapter presented the results of an investigation into the effects of causal structure and causal strength on moral judgments. Structure and strength are the two dimensions of causal representations according to Causal Bayes Nets Theory, one of the most successful accounts of human causal reasoning. As argued in Section 1.3, nearly all global theories of moral reasoning reserve a central role for causal reasoning, yet none of them specifies how exactly thinking about causes and effects guides and constrains thinking about morality. There are also a number of accounts of specific phenomena in moral reasoning (e.g., patterns of judgment about trolley cases, causal selection, responsibility attribution in groups, see Section 1.4) that are well explained by analyzing the

[5]see https://www.lexisnexis.com/community/casebrief/p/casebrief-united-states-v-carroll-towing-co-1383630741, last accessed May 05, 2022

causal structure of scenarios. But a systematic, bottom-up investigation of the effects of structure and strength on moral judgments is still missing.

We here aimed to lay the groundwork for such an investigation, starting with the case of moral judgments about causal chains. Past research has shown that structure and strength can interact in interesting ways in causal chains (Stephan, Tentori, et al., 2021; Bés et al., 2012). Independently, the fact that indirect harm is often seen as morally better than direct harm is well established in moral psychology (Royzman & Baron, 2002; Ziano et al., 2021; Cushman et al., 2006; Greene et al., 2001; Hauser et al., 2008, for an overview see Sloman et al., 2009), but the cognitive foundations of this effect remained unclear. There are of course several ways in which harm can be indirect. We focused on one of them here, which is the case where several physical events mediate the relationship between an initial action and harm to another person. We established that agents and actions are evaluated more favourably in such causal chains, compared to direct relations with the same start- and endpoints. Subsequently, we contrasted two models that might explain this effect, of which one was based primarily on causal structure and one was based primarily on causal strength. The data were more in line with the probabilistic model, which posits that in causal chains like the ones we investigated, effects of structure are ultimately mediated by inferences about lower strength. We also found that this effect is mediated by attributions of outcome foreseeability.

The cognitive mechanism that our results favoured is rational as long as that the assumption of equal link strengths in chains (leading to a lower strength of the overall relation) is justified (see, e.g., Stephan, Tentori, et al., 2021). Lower strength leading to lower foreseeability of harm is also plausible, and matches how foreseeability is understood in legal discussion (see, e.g., Moore, 2019). Finally, the fact that foreseeability and causation are crucial inputs into assessments of moral responsibility is in line with

normative philosophical theories (see Rudy-Hiller, 2018). Thus, our data show that people integrate information about causal structure, strength, and foreseeability in a largely rational manner when it comes to making moral judgments about causal chains.

The fact that effects of structure were driven by inferences about strength here of course does not mean that causal structure is generally unimportant. To the contrary, only the fact that a causal relation between action and outcome exists in the first place enables its strength to matter morally. Plus, as Section 1.4 has shown, there are many circumstances where the exact kind of causal structure, and/or how multiple causes combine to bring about effects, are crucial for explaining and predicting people's judgments.

Future research, besides exploring the directions that were pointed out in Sections 2.7.1 - 2.7.3, might also want to employ other structures than chains in the context of moral judgments. Relations can be indirect in many interesting ways, and many real-world scenarios with serious moral implications are much more causally complex than the scenarios we used in our experiments. One example is the climate crisis, where many actors' actions and decisions combine in complex ways in bringing about harms for others. For this reason, it is sometimes thought that seeking legal compensation for climate-related harms that already occur, or protection against harms that will occur unless action is taken (e.g., lawsuits against individual states or corporations) will be difficult due to the hurdle of establishing causation. Yet, climate science has devised a number of attribution methods to determine the causal contribution of individual actors to different kinds of harms (Stuart-Smith et al., 2021). An interesting avenue for further research could be to investigate how the results of different methods of gauging individual contribution in complex causal models are understood by laypeople and lawyers, and which implications they have for their judgments.

The following chapter, however, will first of all focus on another crucial component of causal representations. So far, we have varied the structure and strength by which an ac-

tion and a negative outcome are causally connected, keeping the outcome constant. The valence of outcomes that are caused by an action, and trade-offs between several outcomes of different valences, however, are another important input into moral judgments. Like the influence of causal structure and causal strength, the ways in which outcome trade-offs psychologically inform moral judgments is underspecified in current theories of moral reasoning, even though all theories agree that outcomes are important. Earlier, I introduced a model by Cohen and Ahn (2016), which predicts judgments about the moral permissibility of actions from observer's subjective utilities of consequences (see Section 1.2). I discussed strengths as well as shortcomings of the model in its current form. Based on this critique, we proposed and evaluated a more generalized subjective-utilitarian model in Engelmann and Waldmann (2022b), and tested the model against people's judgments in two experiments.

3. How to weigh lives. A computational model of moral judgment in multiple-outcome structures

Consequences clearly matter for our moral evaluation of actions, and of the agents performing them. According to Cohen and Ahn (2016)'s subjective-utilitarian theory of moral judgment, they are even all that matters. As we argued earlier (see Section 1.2), we think that Cohen and Ahn (2016)'s account fails as a general theory of moral judgment, because it does not consider the influence of factors like causal structure, agents' knowledge or intentionality, all of which have been demonstrated to matter for moral judgments when consequences are kept constant (see Waldmann et al., 2012; Wiegmann & Engelmann, 2020; May, 2018). Nevertheless, we believe that a subjective-utilitarian model could be very useful as one component of a more complex account of moral judgment. It could serve as the mechanism that compares the values of the outcomes that result from different courses of action. For this purpose, however, the model should be applicable to a broader range of situations than it is in its current form. Furthermore, more evidence is necessary to ensure that the model actually predicts *moral* judgments, not just judgments about what observers would personally do if they were in the described situation.

3.1. A generalized subjective-utilitarian model (GSUM)

In Engelmann and Waldmann (2022b), we proposed and evaluated a generalized subjective-utilitarian model of moral judgment in multiple-outcome structures (GSUM). GSUM is inspired by Cohen and Ahn (2016)'s model, but it is applicable to a much wider range of situations, not just simple two-option forced-choice dilemmas. GSUM achieves this by explicitly representing all relevant actual, hypothetical, or counterfactual states of entities that are affected by an action. Consider a case in which an action improves the

lives of five people (for example by increasing their life expectancy and/or giving them better health), but also leads to the death of one person. However, if the action was not performed, nothing at all would have changed for either the five people or for the one person (we will call cases with such a structure *improving* cases from now on). For a subjective-utilitarian analysis of this situation, we need to think about our valuation of four states: 1) The state of the five people without intervention (normal), 2) the state of the five people after intervention (improved), 3) the state of the one person without intervention (normal), and 4) the state of the one person after intervention (dead). Out of these four values, a subjective utility for acting in this scenario can be calculated (from here: *scenario utility*). For the sketched case, such a scenario utility would look like this (SU = subjective utility):

$$\begin{aligned} Scenario\ Utility\ (improving) = (SU_{five\,improved} \ - \ SU_{five\,normal}) + \\ (SU_{one\,dead} \ - \ SU_{one\,normal}) \end{aligned} \tag{1}$$

The first part of Equation 1 captures the gain that results from making the lives of the five people somewhat better, and the second part represents the loss that is incurred by the death of the one person. If the improvement to the lives of the five people is so large that it outweighs the death of the one person, the scenario utility becomes positive. While this example is about a case in which someone's action has two effects, further effects could be added, and would factor into the scenario utility in the same way as the first two.

Note that the same analysis is also applicable to cases that have a typical dilemma structure (from here on: *saving* cases). Here's an example in which either the lives of five people or the life of one person can be saved:

$$\begin{aligned} Scenario\ Utility\ (saving) = (SU_{five\,normal} \ - \ SU_{five\,dead}) + \\ (SU_{one\,dead} \ - \ SU_{one\,normal}) \end{aligned} \tag{2}$$

The first part of Equation 2 captures the gain that results from saving five people, and the second part, again, represents the loss that is incurred by the death of the one person. If the valuation of dead states is zero, or at least constant (as Cohen & Ahn, 2016's model presupposes), GSUM reduces to a comparison of the alive/intact values of the two items under comparison, and will therefore make identical predictions to Cohen and Ahn (2016)'s model for saving cases.

In the following discussion, and in our experiments, we will focus on moral judgments about actions with two effects that are either saving or improving cases (with varying numbers of lives saved or improved). To generate predictions for people's moral evaluation of such cases, we will first assess their subjective utilities of different numbers of lives, in different states (normal, dead, or improved). Our model will then proceed in the following way (see also Fig. 13, for a depiction of the computational steps): On a given trial, four values will be randomly drawn from the utility estimates of participants, corresponding to the values that are relevant for a scenario (e.g., values for normal and dead states in a saving scenario, values for normal, dead, and improved states in an improving scenario). From these estimates, a scenario utility will be calculated, as described in Equations 1 and 2. This procedure will be repeated many times for each scenario (we are going to use 10,000 iterations). In the end, the proportion of positive values among all scenario utilities will be calculated, and used as the main predictor for moral evaluations. We predicted that the higher the proportion of positive scenario utilities for a case becomes, the more positively it will be evaluated by participants. While we focused on the retrospective moral evaluation of actions that have already been executed, the model can also be applied to the prospective evaluation of planned or considered actions. In the first case, some states of the involved entities are actual (what happened because of the action) and some counterfactual (what would have happened without it), and in the latter case, some states are actual (the status quo of all affected entities) and some

are hypothetical (what would happen because of the action).

3.2. Summary of the empirical findings (Engelmann & Waldmann, 2022b)

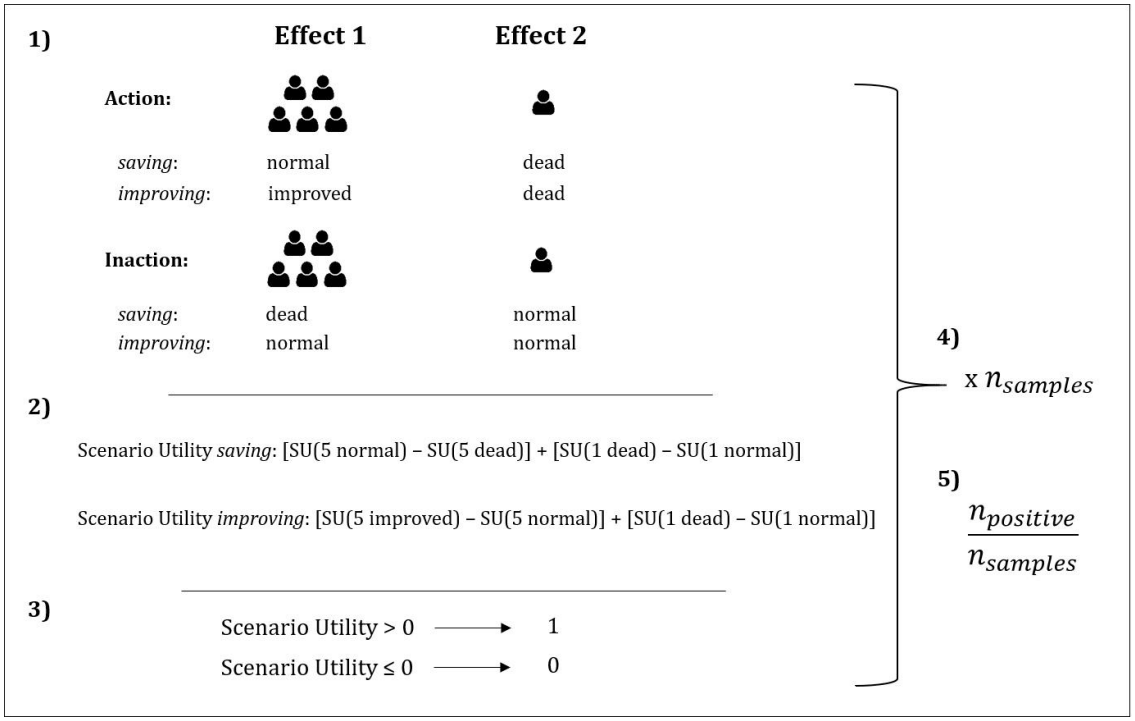
In the following, I will provide a summary of the experiments that have been conducted to evaluate GSUM. For more details, see Engelmann and Waldmann (2022b). For all materials, data, and code, see `https://osf.io/682uc/`. All studies were conducted as online experiments, the survey implementation was in Unipark Questback, and participants were recruited via `www.prolific.co`.

3.2.1. Utility estimation study

First, we aimed to elicit the input data for our model with an utility estimation task (following the method used in Cohen & Ahn, 2016, but with some modifications). We asked participants to indicate their subjective value of a range of entities (people, animals, plants), in a range of group sizes (one, five, ten, 20, 100) and states (normal, dead, improved). These stimuli were the same ones that would later be used in moral judgment tasks in Experiment 1 and 2. We chose to include animals and plants besides humans because we wanted to elicit a wide range of subjective utilities and, later, ensure sufficient variation in moral evaluations to be able to test our model.

We invited 125 participants to complete the experiment online (final sample size after attention checks = 123). Each participant saw all stimuli, that is: people, monkeys, fish, trees, and roses, described as either in a normal state, an improved state, or as dead. The improved state corresponded to the description in the later moral judgment tasks. Examples are “100 people, all of them with improved health and increased lifespan”, or “five trees, all of them bigger and more resistant to pests than usual”. Furthermore, the group size of entities varied (one, five, ten, 20, 100). Thus, each participant saw and

Figure 13: Illustration of GSUM's calculation steps for a saving and an improving case (Engelmann & Waldmann, 2022b).



Note: SU = subjective utility. Step 1 depicts the actual, hypothetical, or counterfactual states of the affected people that have to be considered for the saving and the improving case. In Step 2, a scenario utility is determined for each case (from four randomly sampled values of each relevant class). In Step 3, the model determines whether the scenario utility is positive or not. This process is repeated n times (Step 4, in our case n = 10,000). Finally (Step 5) the proportion of positive scenario utilities is determined (the predictor of moral judgments). This figure is reproduced from Engelmann & Waldmann (2022b, Fig. 1).

evaluated 5 (kind of entity) × 3 (state of entity) × 5 (number) = 75 stimuli.

In the instructions, we informed participants that their task would be to indicate how important or valuable certain items seem to them, or how good or bad they take it to be that these things exist. These assessments should be given in the form of numbers, with positive values indicating that items are valued positively (are "something good"), and negative values indicating that items are valued negatively (are "something bad"). The scale was constrained at -1000 and +1000, with a value of zero described as corresponding to the value of "pieces of a broken tea cup" (i.e., worthless). Participants were provided with practice trials to get used to the task format, and they were able to inspect all stimuli before assigning any values, allowing them to calibrate their valuation scale.

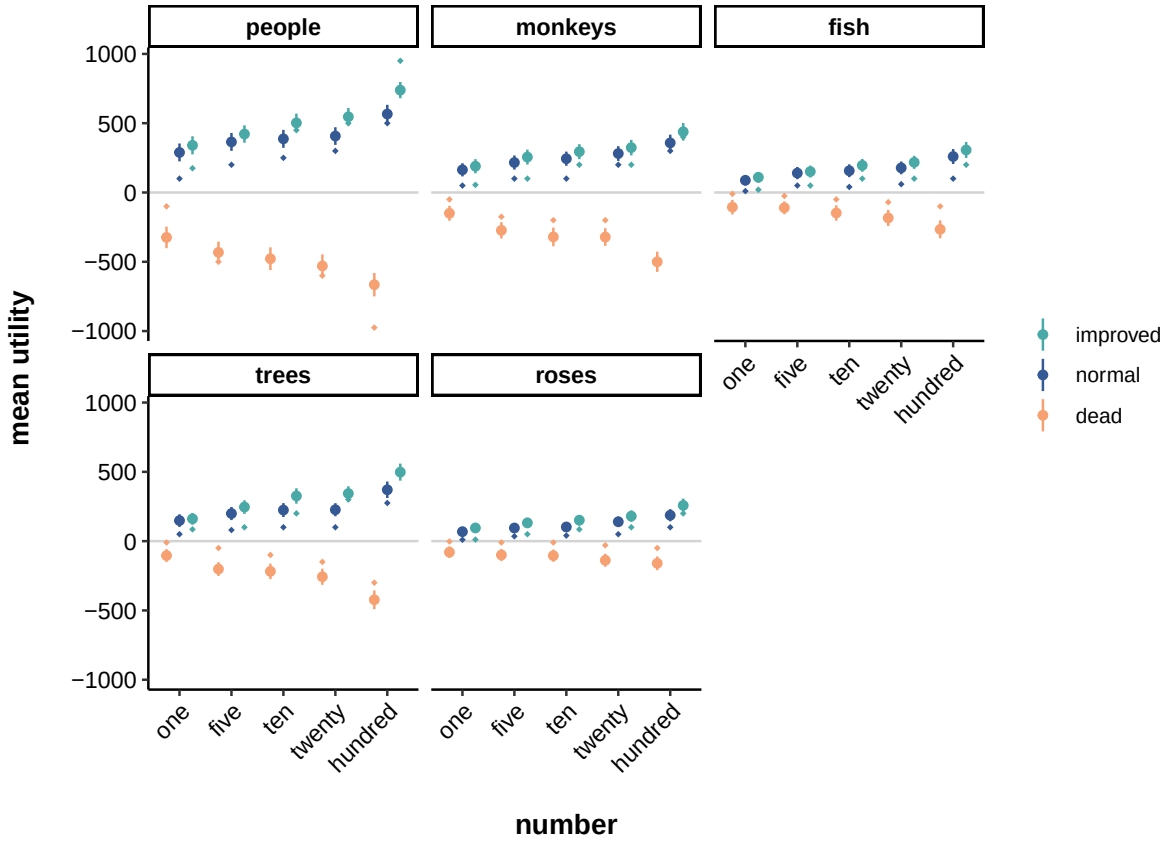
Figure 14 shows the results. Participants assigned the highest values to the lives of people, lower values to animals, and the lowest values to plants (although monkeys and trees did not differ). Normal and improved states were valued positively, and their values increased with larger group sizes. Improved states were valued somewhat more highly than normal states, but the difference was not large. Dead states were generally assigned negative values, with values becoming more negative the higher the number of dead entities became.

Thus, participants completed the utility estimation task in a sensible and expected manner. We therefore proceeded to generate predictions for moral judgments about a range of scenarios using GSUM, and compared them against different participants' moral judgments in Experiments 1 and 2.

3.2.2. Experiment 1

The aim of this experiment was to test whether subjective utilities of consequences actually predict people's *moral* judgments about scenarios, not just their assessments of what they would personally do. We therefore presented a new sample of participants with a range of moral dilemmas (that is, saving cases. Experiment 2 will compare

Figure 14: Results of the utility estimation task in Engelmann & Waldmann (2022b).



Note: Mean (large dots, error bars = 95% CIs) and median (small dots) utility estimates per type of entity (people, monkeys, fish, trees, roses), number (one, five, ten twenty, hundred) and state (normal, dead, improved). This figure is reproduced from Engelmann & Waldmann (2022b, Fig. 2).

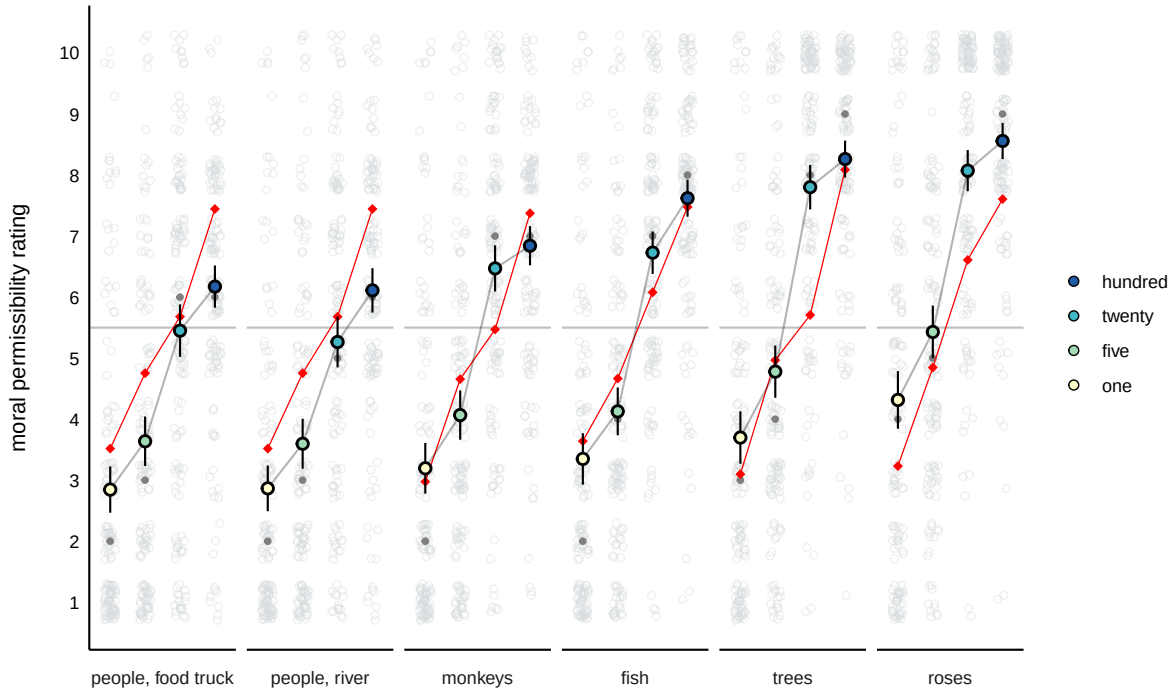
saving and improving cases). These dilemmas involved the stimuli whose utilities had previously been estimated by different participants. Each scenario described a trade-off between ten lives that would be lost on the one hand, and either one, five, twenty, or hundred lives that would be saved on the other hand. The agent in the scenario always opted to intervene, that is, they always caused the death of ten in order to save one, five, 20, or 100 others. All lives within a scenario were of the same kind, that is, all people, all monkeys, all trees, etc. We manipulated the number of saved lives between-subjects, and the kind of entities within-subject, resulting in a 4 (number saved, between) $\times$ 6 (kind of entity[6], within) design. We invited 615 participants to complete the experiment online (final sample size after attention checks: 594).

In each scenario, we described that one group was currently threatened by death, for example a remote village whose 20 inhabitants were close to starvation. A person in charge (e.g., a government official) had the opportunity to save the threatened group, for instance by redirecting a food truck towards the village. However, this action would cause the death of another group, for example of another village's ten inhabitants, who would otherwise have received the delivery. In each scenario, the agent decided to act, and both outcomes occurred, that is, one group was saved and the other group died. In all scenarios, simple illustrations were provided, depicting the group sizes, and their respective states before and after intervention. We asked participants: "To what extent was [agent]'s action morally permissible?", with answers given on a scale ranging from 1 ("not at all") to 10 ("fully").

The results are shown in Fig. 15. People rated actions to be more permissible the more lives were saved, and they also took acting to be more permissible when the involved entities were plants rather than animals, and animals rather than people. GSUM's predictions (shown in red in Fig. 15) captured these patterns well (R^2 between .76 and

[6]We used two scenarios about human lives, because we also used two about animals and two about plants. Therefore the "kind of entity" factor has six rather than five levels here.

Figure 15: Results of Experiment 1 in Engelmann & Waldmann (2022b).



Note: Mean moral permissibility ratings (large coloured dots) and 95% confidence intervals per scenario (people/food truck, people/river, monkeys, fish, trees, roses) and number of lives saved (100, 20, five, one). Dark grey dots are medians, light-grey jittered circles are individual data points. GSUM's predictions are depicted in red. This figure is reproduced from Engelmann & Waldmann (2022b, Fig. 5).

.79, depending on specifications of analyses).

To sum up, Experiment 1 confirmed that observers' subjective utilities of the consequences of other people's actions in fact predict their moral evaluation of these actions. The higher the perceived gain of acting (scenario utility), the better the moral evaluation. Importantly, the utility estimates on which GSUM's predictions are based were provided by a different group of participants, who were not engaged in a moral judgment task at all. Thus, it seems that GSUM has potential to serve as the outcome integration mechanism in a computational account of moral judgment. Experiment 2 will further explore this potential by testing the model on scenarios that are not dilemmas (improving cases).

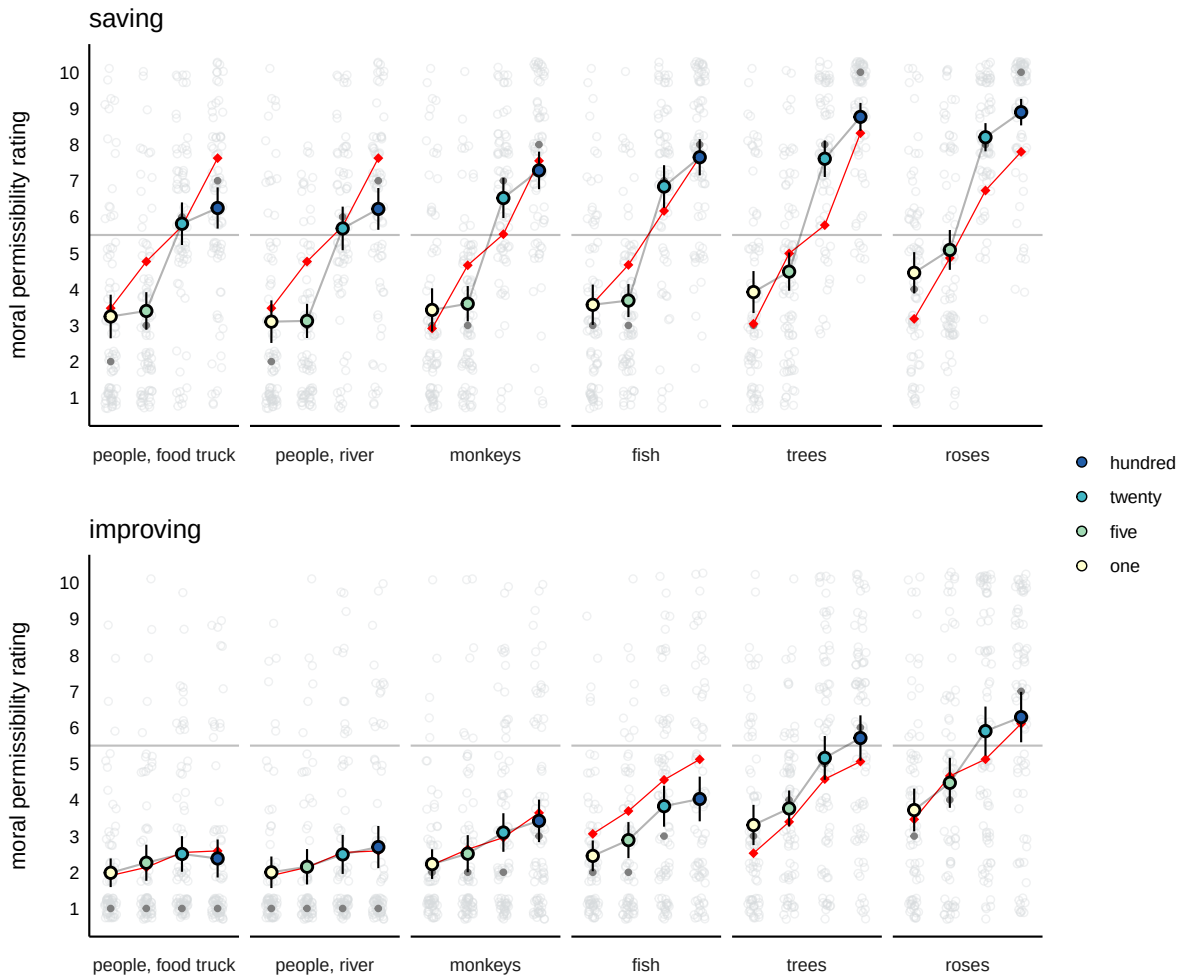
3.2.3. Experiment 2

A main advantage of GSUM over the previous model (Cohen & Ahn, 2016; Cohen et al., 2021) is that it is applicable to scenarios beyond simple trolley-style dilemmas. In this experiment, we therefore added such cases to our design for a further evaluation of our model. In our previous scenarios (saving cases), it was always a re-allocation of some resource (e.g., food, water) that saved a threatened group, while another group died from lack of the same resource as a consequence. This setup is easily modified for improving cases, allowing for closely matched scenarios. In the food truck case described earlier, we simply stated that the government official learned that the health of a number of people in a particular village could be vastly improved and their lifespan extended if more food was available to them. The only option to provide them with additional food is to redirect a truck that was originally headed towards a different village. The inhabitants of this village would die without the delivery. As before, the person in charge decides to intervene, and both anticipated effects occur (e.g., the health of the one group is greatly improved and their lifespan extended, and the other group dies). Saving vs. improving was added to the design as a new between-subjects manipulation. Everything

else (material, procedure, test question) remained identical to Experiment 1, resulting in a 2 (saving vs. improving, between) × 4 (number saved, between) × 6 (kind of entity, within) design. We invited 621 participants to complete the experiment online (final sample size after attention checks = 607).

See Fig. 16 for the results. Acting in saving cases was generally seen as much more permissible than in improving cases. Within the class of saving scenarios, we generally observed the same pattern as in Experiment 1. Acting became more permissible with higher numbers of lives saved, and it was again more permissible for plants than for animals, and more permissible for animals than for people. While improving cases were generally regarded as less permissible, they still showed an influence of the number of improved lives, and of the kind of involved entities. Within the lower half of the rating scale, acting still become more permissible when more lives were improved, and it was more permissible for plants than for animals or people. We created predictions for all cases using GSUM, and again, the model fit the data well (R^2 between .76 and .82 for saving cases, and R^2 between .85 and .90 for improving cases, depending on specifications of analyses). GSUM captured the difference between saving and improving (resulting from the fact that the utility gain due to improving was generally perceived as small, while the loss due to death was generally perceived to be large). It also captured the effect of numbers, and, somewhat surprisingly, tracked the differences in permissibility ratings between improving scenarios about people, animals, and plants. GSUM is able to capture these differences because people's utilities of improved and normal states showed roughly the same distance for all groups of entities (improved states were valued somewhat higher than normal states, see Fig. 14), resulting in comparable gains due to improving across species. Losses, however, would be more pronounced for the higher-valued entities (e.g., people), because dead states were valued much more negatively for them than for others (e.g., plants).

Figure 16: Results of Experiment 2 in Engelmann & Waldmann (2022b).



Note: Mean moral permissibility ratings (large coloured dots) and 95% confidence intervals per structure condition (saving vs. improving), scenario (people/food truck, people/river, monkeys, fish, trees, roses), and number of lives saved or improved (100, 20, five, one). Dark grey dots are medians, light-grey jittered circles are individual data points. GSUM's predictions are depicted in red. This figure is reproduced from Engelmann & Waldmann (2022b, Fig. 6).

3.3. Summary and Discussion

We have demonstrated that observer's subjective utilities of consequences predict their *moral* judgments about the permissibility of other people's actions. As such, a generalized subjective-utilitarian model (GSUM) might be well-suited to serve as the outcome integration mechanism of a full computational account of moral judgment. As we have argued earlier, all contemporary theories of moral reasoning require such a mechanism, but none of them have spelled it out in detail so far. The only exception is Cohen and Ahn (2016)'s subjective-utilitarian theory of moral judgment, but their model is limited to two-option, forced choice dilemmas. Furthermore, it had only been tested against people's assessments of their own likelihood of intervening in such dilemmas ("Would you (...)?"). Our generalized model is applicable to any kind of multiple-outcome structure, and we demonstrated that it predicts judgments about dilemmas as well as about more regular cases in which actions have more than one effect.

3.3.1. Beyond consequences

Unlike Cohen and Ahn (2016), we do not claim that outcomes are all that matters for people's moral evaluation of actions. It is well documented that moral judgments are also affected by factors such as intentionality or knowledge (Sloman, Fernbach, & Ewing, 2012; Kirfel & Lagnado, 2021a; Cushman et al., 2006; Cushman, 2008; Samland & Waldmann, 2016; Engelmann & Waldmann, 2021; Lagnado & Channon, 2008, see also Chapter 2 of this thesis and Engelmann & Waldmann, 2022a), causal structure (Mikhail, 2007, 2011; Wiegmann & Waldmann, 2014; Cushman et al., 2006), action versus omission (Cushman et al., 2006; Sloman et al., 2009, but see Willemsen & Reuter, 2016), "personal force" (Greene et al., 2001, 2004), and possibly others (for overviews, see Waldmann et al., 2012; Wiegmann & Engelmann, 2020; May, 2018; Sloman et al., 2009). May (2018) recently suggested that all of these dimensions could be subsumed

under the term "agential involvement". The more involved an agent is taken to be in bringing about a harmful outcome, the more severe our moral evaluation. In any case, a complete computational account of moral judgment needs to take all of the cited factors into account, in addition to outcomes. Nevertheless, outcomes are crucial. Within such an account, GSUM could serve as the mechanism that compares the consequences that are obtained when different causal paths in a network are instantiated.

3.3.2. The valuation of different states and species

GSUM's predictions rest fundamentally on the input values that participants generated in the utility estimation task. Utilities and how to measure them is a much-debated topic in philosophy and economics (for an overview, see Narens & Skyrms, 2020). A classical method is to infer people's utilities from their choices (revealed preferences). However, Cohen and Ahn (2016) and Cohen et al. (2021) aimed to assess subjective utilities independently of choices, in order to be able to test, and potentially also falsify, whether utilities *predict* choices (see Cohen et al., 2021, for further discussion). Our method was inspired by Cohen and Ahn (2016), but we also made some modifications.

Most importantly, we allowed for negative utilities. We did this primarily based on considerations about the value of dead states. In Cohen and Ahn (2016)'s model, only dilemmas are considered, in which the values of dead states cancel out (assuming they are equal for the two items in a dilemma). Thus, it doesn't matter which specific values are assigned. However, this is not true for other cases. Our improving scenarios, for example, require a specification of the value of the dead state. One obvious possibility is of course to simply assign a value of zero to all dead entities. Since some entities are assigned much higher values than others in their normal state (e.g., people are valued higher than roses), assigning a value of zero to all dead states would still accurately capture that a person dying is worse than a rose dying. In the case of a person, a high value would be reduced to zero (a large loss), whereas in the case of a rose, a much lower

value would be reduced to zero (a smaller loss). However, people might additionally wish to express that the fact that a person is dead is worse than the fact that a rose is dead. Such a difference could not be expressed by assigning a constant value of zero to all entities when dead, and requires using negative utilities. We ultimately left the choice to participants by providing a valuation scale that allowed both. Participants readily assigned negative values, thereby expressing both that a person dying is a greater loss than a rose dying (because of the greater difference between the positive normal state and the negative dead state), and also that the fact that a person is dead is worse than the fact that a rose is dead (because of the more negative value for a dead person than for a dead rose).

We realize that these issues touch upon more complex questions that have been discussed extensively in philosophy (Kagan, 2012; Kamm, 2020; Nagel, 2012), like whether and in what way death can be bad for a person (or an animal, a plant), and thus if and on what basis death should be regretted on behalf of those who die. Our simple valuation study cannot answer how exactly people think about these questions. It can only tell us that people take others being dead to be something negative, and more so when more are dead, or when those who are dead are people rather than animals or plants. However, given the centrality of matters of life and death in moral psychology, this seems like an important avenue for future research.

The results of our utility estimation task furthermore show that people generally value human lives highest, and assign lower values to animals and plants. However, they still differentiated between different forms of non-human life, likely reflecting the increased recognition of a moral status of at least animals (Korsgaard, 2018; Singer, 1975). Plants may be valued instrumentally (trees), or for aesthetic reasons (roses), or maybe they are also valued in themselves by some participants.

Interestingly, the different valuation patterns between species explained some patterns

in moral judgments that might otherwise be attributed to deontological constraints (see Caviola et al., 2020). Especially in improving cases, we observed that improving the lives of some people at the expense of others was seen as clearly less permissible than doing the same with animals or plants. One way of explaining how these different responses come about is the following: In all cases (people, animals, plants), reasoners realize that performing the action would yield a roughly equal gain, and also a roughly equal loss. However, they then decide that humans have special rights that protect them against trade-offs outside of emergency situations (a deontological constraint). These rights are not, or only to a lesser extent, granted to other species, resulting in the differential moral evaluation of improving scenarios that we observed. However, we have seen in our experiments that the differences manifest much earlier, namely already in the valuation of people, animals and plants. In improving cases, the loss of human lives is simply not considered as equally compensated by the gain due to improving as it is for other species. Thus, differences in subjective utilities here seem to explain some patterns in people's judgments that might otherwise be attributed to an influence of deontological thinking. However, we also observed that acting was less permissible for people than animals and plants in saving cases (although the effect was somewhat smaller than for improving cases), which is not currently captured by our model. Thus, deontological constraints might play a role as well.

Generally, the different valuation of people, animals, and plants may reflect speciesist attitudes on the side of participants. The psychological foundations of speciesism have received some attention in recent years (Caviola et al., 2020; Caviola, Everett, & Faber, 2019; Caviola, Schubert, Kahane, & Faber, 2022; Crimston, Bain, Hornsey, & Bastian, 2016; Goodwin & Benforado, 2015; Horta, 2010). Caviola et al. (2020) could show that even when frequently cited factors such as intellectual ability or suffering capacity are held constant, people still extend more consideration to humans than to other species. A

further exploration of the cognitive mechanisms that underlie such attitudes is another important avenue for further research (see Caviola et al., 2022, for steps in this direction).

3.3.3. Conclusion and outlook

We have proposed and evaluated a generalized subjective-utilitarian model of moral judgment (GSUM). The model predicted participants' moral judgments about moral dilemmas and other multiple-outcome structures well in two experiments. We thus suggest that GSUM can be used as one building block of a more complete and ideally formalized account of moral judgment. Specifically, it can serve as the mechanism that compares the consequences that are obtained when different paths in a causal network are instantiated (actually, hypothetically, or counterfactually). The model is compatible with any contemporary theory of moral reasoning. Central challenges for future research include an integration of GSUM with other important factors such as agents' mental states or dispositions, causal structure, and others. Moreover, it would be interesting to further explore the sources of people's differential valuation of different kinds of lives, and of different states (e.g., how death is represented).

4. General Discussion

Causation and morality are deeply intertwined. This fact has been recognized in all major ethical frameworks, despite their differences. Roughly speaking, consequentialism dictates that the effects of our actions determine their moral status (see Sinnott-Armstrong, 2021), deontology holds that other factors than consequences matter as well (such as whether the way in which consequences are brought about is in line with certain rights and duties, see Alexander & Moore, 2021; Kamm, 2008), and virtue ethics focuses on moral character, which might be seen as a root cause of many of our actions (see Hursthouse & Pettigrove, 2018). Defining and establishing causation is also a central task in legal discussions and proceedings (see Hart & Honoré, 1985; Moore, 2019; Lagnado, 2021; Lagnado & Gerstenberg, 2017). In everyday life, the importance of causation for moral judgments is reflected in many ways. For instance, we blame others for the negative consequences of their actions and praise them for positive ones (see, e.g., Malle, Guglielmo, & Monroe, 2014), we infer and evaluate others' intentions (Knobe, 2003) and character (Uhlmann, Pizarro, & Diermeier, 2015; Siegel, Crockett, & Dolan, 2017; Montealegre, Bush, Moss, Pizarro, & Jimenez-Leal, 2020; Hartman, Blakey, & Gray, 2022) based on their behaviour, and we use such knowledge to decide who to associate or cooperate with in the future (see, e.g., Tomasello & Vaish, 2013).

Psychological theories of moral reasoning all include a commitment to a central role of causal thinking, more or less explicitly. Dual-process theories posit a cognitive mechanism that assesses the consequences of performing an action in a particular situation (Greene et al., 2001, 2004; Cushman, 2013; Crockett, 2013), universal moral grammar theory places a strong emphasis on people's causal representation of a situation (Mikhail, 2007, 2011), and the theory of dyadic morality's central construct is a causal template of an agent harming a patient (H. M. Gray et al., 2007; K. Gray & Ward, 2011; K. Gray et al., 2012, 2014). The only theory that mostly reduces the role of causal reasoning

to post-hoc justification is Haidt's social-intuitionist model (Haidt et al., 2000; Haidt, 2001). Despite the wide-spread acknowledgement of a crucial role of causal reasoning for moral judgments, the details are hardly spelled out. Specifically, a connection to psychological theories of reasoning about causal relationships and to theories of reasoning about outcomes is missing. This is what we attempted to begin in the projects presented in this thesis.

The first project (Chapter 2, Engelmann & Waldmann, 2022a) was inspired by Causal Bayes Nets Theory (see Pearl, 2000; Pearl & Mackenzie, 2019; Rottman & Hastie, 2014; Sloman, 2005; Waldmann, 2017) and focused on the roles of causal structure, causal strength, and foreseeability in moral judgments about causal chains. We discovered that when reasoners represent a causal relation between an action and an accidental harmful outcome as indirect rather than direct (a difference in causal structure), this representation can lead to the impression that the causal relation is weaker (a difference in perceived causal strength). Causal relations that are perceived as weaker, in turn, can lead to the inference that agents are less able to foresee that harm will actually occur. Attribution of diminished outcome foreseeability then give rise to the more positive moral evaluation of actions and agents in chains compared to direct relations. The harmful outcomes that agents caused were held constant in these experiments.

The second project (Chapter 3, Engelmann & Waldmann, 2022b) was concerned with the influence of outcome trade-offs on moral judgments in common-cause structures. The theoretical background of this project lies in theories about decision-making based on subjective utilities, which Cohen and Ahn (2016) recently applied to moral judgment. Based on a critique of their model, we developed a generalized subjective-utilitarian model of moral judgment (GSUM), and evaluated the model's predictions in two experiments. GSUM predicted people's permissibility judgments well, both in classic moral dilemmas and in common-cause structures that are not dilemmas. We concluded that

the model could be used as one component of a more complete computational account of moral judgment. Within such an account, it could serve as the mechanism that that compares the values of the consequences that arise from different courses of action (actually, hypothetically, or counterfactually).

4.1. Combining structure, strength and utilities in a causal network

Causal Bayes Nets can connect actions and outcomes, representing both as variables that can take on several values (e.g., present vs. absent, or a greater range of values in the case of continuous variables). The states of variables refer to events (see, e.g., Samland & Waldmann, 2016; Waldmann & Mayrhofer, 2016). The *utility* of events, however, is not normally encoded. Formally, the event of a traffic light turning green is no different from the event of a person dying. In classical expected utility theory (which inspired the subjective-utilitarian model presented by Cohen & Ahn, 2016 and thereby Engelmann & Waldmann, 2022b), on the other hand, the role of *causality* is usually at best implicit (but see Weirich, 2020).

As the projects presented in this thesis have shown, both the causal relations connecting actions and outcomes and the observer's subjective value of these outcomes are important for moral judgments. A formal account that accurately represents morally salient situations therefore needs to capture both. This is recognized by Mikhail (2011, 2007), who suggests that people construct a series of increasingly rich representations when confronted with a moral scenario. A merely temporal order is followed by a causal structure, which is then augmented to a "moral structure", in which the valence of events is encoded as well. Sloman et al. (2009) propose to use Causal Bayes Nets as the "representational infrastructure for moral judgment", arguing that they can represent actions, outcomes, as well as other important factors such as agents' mental states or intentions (see also Sloman et al., 2012). Causal Bayes Nets might thus be able to em-

body Mikhail's "moral structure". However, Sloman et al. (2009) do not explicate how the values of events can become part of a causal network.

Jern and Kemp (2015) suggest to use so-called Decision Networks to model how people think about other's actions and plans. Decision Network, like Causal Bayes Nets, encode events and the causal relations between them, but they also include dedicated "utility nodes", and information about agents' choice strategies (such as the assumption that an agent will maximize their utility). Utility nodes are represented like further effect nodes of the events from whose occurrence an agent receives positive or negative utility. A utility function specifies how the value of the utility variable changes depending on the states of the variables that represent events in the world. Determining the aggregate utility of an action would require to sum across all utility nodes in a decision network that instantiates the action. Such utility nodes could be added to classical Causal Bayes Nets as well, but Jern and Kemp (2015) argue that Causal Bayes Nets with utility nodes are not able to capture goal-directed behaviour, e.g., the fact that an agent's behaviour can change when the utility function specified for those nodes changes. Kleiman-Weiner, Gerstenberg, Levine, and Tenenbaum (2015) use similar networks to model intentions in moral dilemmas as goal-directed plans (see Bratman, 1987), rather than as nodes that are causes of behaviour in a causal network, as Sloman et al. (2009) and Sloman et al. (2012) propose.

No matter which type of network ultimately turns out to be better at capturing people's inferences in moral scenarios, it seems that including (subjective) utilities in a causal representation is feasible. If the events from which agents receive utility or disutility are conceptualised as changes between states of the world (see, e.g., Casati & Varzi, 2020), the utility of an event (five people dying) might be comprised of the difference in utility between the two states that make up the event (the utility of five people being dead minus the utility of five people being alive, as we modeled it in

Engelmann & Waldmann, 2022b).

In modelling such situations, it is important to distinguish between the utilities of the person who judges a moral scenario (e.g., the participants in an experiment) and the utilities of the agent in the scenario, known or inferred. In Engelmann and Waldmann (2022b), we predicted judgments about agents from the *observer's* subjective utilities of outcomes. More theoretical work is clearly needed to spell out how people take their own vs. other's (potentially diverging) utilities of events into account when making moral judgments, and to formalize such inferences appropriately. Here, we merely wanted to point out that the utilities of events can in principle be included in causal representations, along with other important factors such as causal structure and causal strength.

4.2. Making moral judgments based on causal networks

We have seen that causal representations may likely be able to, in one form or another, capture information about many relevant ingredients for moral judgments: actions, outcomes with utilities, intentions, perhaps also knowledge and character (see Kirfel & Lagnado, 2021b; Sloman et al., 2009), and possibly others. Presuming that this is possible, it would still not tell us what we should *do* with such a model when we want to make a moral judgment, i.e., which combinations of values of the different variables in the network should give rise to which moral verdicts. Causal models alone are descriptive, they need to be combined with a normative theory to make predictions about moral scenarios (see also Sloman et al., 2009). Theories of moral reasoning (see Section 1.3) allow to derive different answers to the question how causal models might be used to arrive at moral judgments.

Within dual process theories (Greene et al., 2001, 2004; Cushman, 2013; Crockett, 2013), a causal model could be used to trace the consequences that would result from actions in particular situations (the second, "rational" or "model-based" process). The

input from this process would then factor into the global moral evaluation, which also considers input from the first, "emotional" or "model-free" process. Beyond comparisons between the classic *bystander* vs. *footbridge* versions of trolley dilemmas, it is not always easy to derive concrete predictions about the input that the two processes would deliver, and how the global moral evaluation would be affected.

Mikhail (2007, 2011) suggests that internalized deontic rules are applied to the final representation of a scenario. The theory focuses on the prohibition of "battery", defined as "purposefully or knowingly causing harmful or offensive contact with another individual or otherwise invading another individual's physical integrity without his or her consent" (Mikhail, 2007), and the Doctrine of Double Effect (DDE). The DDE specifies under which conditions there can be exceptions from the prohibition of battery. While the DDE is consistent with people's typical judgments about the classic variations of the trolley dilemma, Waldmann and Dieterich (2007) have shown that the distinction between harming as a means and harming as a side effect (which is crucial to the DDE) ceases to affect people's judgments when the point of intervention in a moral dilemma is held constant (intervention on a victim vs. on a threat, which is confounded in the contrast between the classic versions of *bystander* and *footbridge*). Waldmann and Dieterich (2007) and Waldmann and Wiegmann (2010) proposed an alternative theory, which captured known patterns in reasoning about moral dilemmas, and made correct predictions for novel cases (Double Causal Contrast Theory, see Section 1.4). Another possibility is that Mikhail (2007, 2011) is right in positing that unconscious deontic rules guide moral reasoning, but the correct rules have simply not been discovered yet. For instance, Kamm (2008) criticizes the DDE and proposes alternative deontic principles that draw heavily on causal structure as well.

Finally, it is conceivable that a "morally bad action" is represented as a prototype category, and that people judge actions as worse the more similar they are to the pro-

totype. In fact, several researchers have suggested that the seemingly disparate list of factors that affect moral judgments (causal relations, outcomes, intentionality, knowledge, omission versus commission, "personal force" or "battery", locus of intervention, and possibly others, for overviews see Waldmann et al., 2012; Wiegmann & Engelmann, 2020; Sloman et al., 2009; May, 2018) might be unified in such a way. Greene (2014) argues that we have evolved an (emotional) aversion against "violence", by which he means intentional and direct physical harm being inflicted on another person. K. Gray et al. (2012)'s "moral dyad" works in a similar way, but emphasizes the importance of a perceived mind in both victim and perpetrator of harm. May (2018) subsumes the listed factors under the term "agential involvement". The more involved an agent is perceived to be in bringing about harm, so the prediction, the more severe our moral evaluation. Sloman et al. (2009) argue for an "idealized causal model" against which actual situations are compared. Their proposal for such a model only consists of an intention node and an outcome node, with a causal link connecting the two. Making a situation less similar to this model in any way (e.g., unintentional, a causal link that is weak, atypical or nonexistent, no bad outcome or a less severe one) should lead to a more favourable moral evaluation, on this view. While the details may be debated, the proposal that the prototype of a morally bad action could be represented in the form of a causal network seems plausible. When making a moral judgment, people would then compare their causal representation of the actual situation against the causal representation of the prototype, and more similarity would lead to a harsher moral evaluation. The prototype idea, however, clearly needs specification and systematic testing before it should be considered as a serious competitor of existing theories of moral judgment.

The findings presented in this thesis add the following building blocks for further theorizing about the role of causal representations in moral judgment, no matter which overarching theoretical framework of moral reasoning is adopted: First, causal structure

can affect inferences about causal strength, and thereby about agent's level of outcome foreseeability and moral judgments, when no clear information about strength is provided (Engelmann & Waldmann, 2022a). Second, the value of outcomes can be represented in terms of observer's subjective utilities (Engelmann & Waldmann, 2022b).

4.3. Conclusion and directions for future research

The causal relations by which actions are connected to good or bad outcomes, and these outcomes themselves, play an important role in moral reasoning. In this thesis, I presented the results of two projects that empirically investigated the respective influence of features of causal relations (structure, strength) and of outcomes on judgments about the moral permissibility of actions and the moral responsibility of agents. We found that in causal chains, structure can serve as a cue for strength, which can in turn affect attributions of outcome foreseeability, and thereby moral judgments. In another project, we proposed and evaluated a computational model of reasoning about outcome trade-offs in common-cause structures. Hopefully, these findings will serve as useful building blocks for further theorizing about the relationship between causal and moral reasoning, and for the empirical investigation of this relationship. Central challenges for future research include the formal integration of the many factors that affect moral judgments into complete descriptive models, and determining how people use these descriptive models to arrive at normative evaluations.

References

Alexander, L., & Moore, M. (2021). Deontological ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2021 ed.). Metaphysics Research Lab, Stanford University.

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

Alicke, M., & Rose, D. (2012). Culpable control and deviant causal chains. *Social and Personality Psychology Compass*, *6*(10), 723–735. doi: https://doi.org/10.1111/j.1751-9004.2012.00459.x

Anderson, R. A., Kamtekar, R., Nichols, S., & Pizarro, D. A. (2021). "False positive" emotions, responsibility, and moral character. *Cognition*, *214*, Article 104770. doi: https://doi.org/10.1016/j.cognition.2021.104770

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64. doi: https://doi.org/10.1038/s41586-018-0637-6

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. doi: https://doi.org/10.1111/spc3.12131

Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy & Public Affairs*, *37*(4), 293–329. doi: https://doi.org/10.1111/j.1088-4963.2009.01164.x

Bés, B., Sloman, S., Lucas, C. G., & Raufaste, E. (2012). Non-bayesian inference: Causal structure trumps correlation. *Cognitive Science*, *36*(7), 1178–1203. doi: https://doi.org/10.1111/j.1551-6709.2012.01262.x

Białek, M., Turpin, M. H., & Fugelsang, J. A. (2019). What is the right question for moral psychology to answer? Commentary on Bostyn, Sevenhant, and Roets

(2018). *Psychological Science*, *30*(9), 1383–1385. doi: https://doi.org/10.1177/0956797618815171

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. doi: https://doi.org/10.1177/0956797617752640

Bratman, M. (1987). *Intention, plans, and practical reason*. CSLI Publications.

Calderon, S., Mac Giolla, E., Ask, K., & Granhag, P. A. (2020). Subjective likelihood and the construal level of future events: A replication study of Wakslak, Trope, Liberman, and Alony (2006). *Journal of Personality and Social Psychology*, *119*(5), e27–e37. doi: https://doi.org/10.1037/pspa0000214

Casati, R., & Varzi, A. (2020). Events. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2020 ed.). Metaphysics Research Lab, Stanford University.

Caviola, L., Everett, J. A., & Faber, N. S. (2019). The moral standing of animals: Towards a psychology of speciesism. *Journal of Personality and Social Psychology*, *116*(6), 1011–1029. doi: https://doi.org/10.1037/pspp0000182

Caviola, L., Kahane, G., Everett, J. A., Teperman, E., Savulescu, J., & Faber, N. S. (2020). Utilitarianism for animals, kantianism for people? Harming animals and humans for the greater good. *Journal of Experimental Psychology: General*, *150*(5), 1008–1039. doi: https://doi.org/10.31234/osf.io/j3rgm

Caviola, L., Schubert, S., Kahane, G., & Faber, N. S. (2022). Humans first: Why people value animals less than humans. *Cognition*, *225*, Article 105139. doi: https://doi.org/10.1016/j.cognition.2022.105139

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405. doi: https://doi.org/10.1037/0033-295X.104.2.367

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*(4), 545–567. doi: https://doi

.org/10.1037/0022-3514.58.4.545

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365–382. doi: https://doi.org/10.1037/0033-295x.99.2.365

Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review*, *112*(3), 694–706. doi: https://doi.org/10.1037/0033-295x.112.3.694

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General*, *145*(10), 1359–1381. doi: https://doi.org/10.1037/xge0000210

Cohen, D. J., Cromley, A. R., Freda, K. E., & White, M. (2021). Psychological value theory: The psychological value of human lives and economic goods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. doi: https://doi.org/10.1037/xlm0001047

Crimston, C. R., Bain, P. G., Hornsey, M. J., & Bastian, B. (2016). Moral expansiveness: Examining variability in the extension of the moral world. *Journal of Personality and Social Psychology*, *111*(4), 636–653. doi: https://doi.org/10.1037/pspp0000086

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366. doi: https://doi.org/10.1016/j.tics.2013.06.005

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

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–292. doi: https://doi.org/10.1177/1088868313495594

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion*, *12*(1), 2–7. doi: https://doi.org/10.1037/a0025071

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089. doi: https://doi.org/10.1111/j.1467-9280.2006.01834.x

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4, pt.1), 377–383. doi: https://doi.org/10.1037/h0025589

Driver, J. (2008). Attributions of causation and moral responsibility. In W. Sinnott-Armstrong (Ed.), *Moral psychology. The cognitive science of morality: Intuition and diversity* (Vol. 2, pp. 423–440). MIT Press.

Dubber, M. D. (2015). *An introduction to the model penal code.* Oxford University Press.

Engelmann, N., & Waldmann, M. R. (2019). Moral reasoning with multiple effects: Justification and moral responsibility for side effects. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 41th meeting of the cognitive science society* (pp. 1703–1709).

Engelmann, N., & Waldmann, M. R. (2021). A causal proximity effect in moral judgment. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd meeting of the cognitive science society* (pp. 2330–2336).

Engelmann, N., & Waldmann, M. R. (2022a). How causal structure, causal strength, and foreseeability affect moral judgments. *Cognition*, *226*, Article 105167. doi: https://doi.org/10.1016/j.cognition.2022.105167

Engelmann, N., & Waldmann, M. R. (2022b). How to weigh lives. A computational model of moral judgment in multiple-outcome structures. *Cognition*, *218*, Article

104910. doi: https://doi.org/10.1016/j.cognition.2021.104910

Fincham, F., & Jaspars, J. (1983). A subjective probability approach to responsibility attribution. *British Journal of Social Psychology*, *22*(2), 145–161. doi: https://doi.org/10.1111/j.2044-8309.1983.tb00575.x

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

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*(1), 166–171. doi: https://doi.org/10.1016/j.cognition.2009.12.011

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? From expectations to responsibility judgments. *Cognition*, *177*, 122–141. doi: https://doi.org/10.1016/j.cognition.2018.03.019

Goodwin, G. P., & Benforado, A. (2015). Judging the goring ox: Retribution directed toward animals. *Cognitive Science*, *39*(3), 619–646. doi: https://doi.org/10.1111/cogs.12175

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology* (Vol. 47, pp. 55–130). Elsevier.

Graham, J., Haidt, J., Motyl, M., Meindl, P., Iskiwitch, C., & Mooijman, M. (2018). Moral foundations theory. In K. Gray & J. Graham (Eds.), *Atlas of moral psychology* (pp. 211–222). Guilford Publications.

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. doi: https://doi.org/10.1037/a0021847

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619. doi: https://doi.org/10.1126/science.1134475

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. doi: https://doi.org/10.1037/a0036149

Gray, K., & Ward, A. (2011). The harm hypothesis: Perceived harm unifies morality (unpublished manuscript).

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124. doi: https://doi.org/10.1080/1047840X.2012.651387

Greene, J. (2014). *Moral tribes: Emotion, reason, and the gap between us and them.* Penguin Books.

Greene, J. (2015). The rise of moral cognition. *Cognition*, *135*, 39–42. doi: https://doi.org/10.1016/j.cognition.2014.11.018

Greene, J., Cushman, F., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371. doi: https://doi.org/10.1016/j.cognition.2009.02.001

Greene, J., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400. doi: https://doi.org/10.1016/j.neuron.2004.09.027

Greene, J., 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. doi: https://doi.org/10.1126/science.1062872

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384. doi: https://doi.org/10.1016/j.cogpsych

.2005.05.004

Hagmayer, Y., & Engelmann, N. (2020). Asking questions to provide a causal explanation – Do people search for the information required by cognitive psychological theories? In E. A. Bar-Asher Siegal & N. Boneh (Eds.), *Perspectives on causation: Selected Papers from the Jerusalem 2017 workshop* (pp. 121–147). Springer International Publishing.

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

Haidt, J., Björklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason (unpublished manuscript).

Hart, H. L. A., & Honoré, T. (1985). *Causation in the law*. Oxford University Press.

Hartman, R., Blakey, W., & Gray, K. (2022). Deconstructing moral character judgments. *Current Opinion in Psychology*, *43*, 205–212. doi: https://doi.org/10.1016/j.copsyc.2021.07.008

Hauser, M., Young, L., & Cushman, F. (2008). Reviving Rawls' linguistic analogy: Operative principles and the causal structure of moral action. In W. Sinnott-Armstrong (Ed.), *Moral psychology. The cognitive science of morality: Intuition and diversity* (pp. 107–143). MIT Press.

Heider, F. (1958). *The psychology of interpersonal relations*. Lawrence Earlbaum Associates.

Hesslow, G. (1988). The problem of causal selection. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 11–32). New York University Press.

Hilton, D. J., McClure, J., & Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European*

Journal of Social Psychology, *40*(3), 383–400. doi: https://doi.org/10.1002/ejsp.623

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, *106*(11), 587–612. doi: https://doi.org/10.5840/jphil20091061128

Horta, O. (2010). What is speciesism? *Journal of Agricultural and Environmental Ethics*, *23*(3), 243–266. doi: https://doi.org/10.1007/s10806-009-9205-2

Hursthouse, R., & Pettigrove, G. (2018). Virtue ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2018 ed.). Metaphysics Research Lab, Stanford University.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93. doi: https://doi.org/10.1016/j.cognition.2017.01.010

Jern, A., & Kemp, C. (2015). A decision network account of reasoning about other people's choices. *Cognition*, *142*, 12–38. doi: https://doi.org/10.1016/j.cognition.2015.05.006

Johnson, J. T., & Drobny, J. (1985). Proximity biases in the attribution of civil liability. *Journal of Personality and Social Psychology*, *48*(2), 283–296. doi: https://doi.org/10.1037/0022-3514.48.2.283

Johnson, S. G., & Ahn, W.-K. (2015). Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. *Cognitive Science*, *39*(7), 1468–1503. doi: https://doi.org/10.1111/cogs.12213

Kagan, S. (2012). *Death*. Yale University Press.

Kahane, G., & Shackel, N. (2010). Methodological issues in the neuroscience of moral judgement. *Mind & Language*, *25*(5), 561–582. doi: https://doi.org/10.1111/j.1468-0017.2010.01401.x

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, *47*(2), 278–291.

Kamm, F. M. (2008). *Intricate ethics: Rights, responsibilities, and permissible harm.* Oxford University Press.

Kamm, F. M. (2009). Neuroscience and moral reasoning: A note on recent research. *Philosophy & Public Affairs*, *37*(4), 330–345. doi: https://doi.org/10.1111/j.1088-4963.2009.01165.x

Kamm, F. M. (2020). *Almost over: Aging, dying, dead.* Oxford University Press.

Kinney, D., & Lombrozo, T. (2022). Evaluations of causal claims reflect a trade-off between informativeness and compression. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th meeting of the cognitive science society.*

Kirfel, L., & Lagnado, D. (2021a). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, *212*, Article 104721. doi: https://doi.org/10.1016/j.cognition.2021.104721

Kirfel, L., & Lagnado, D. (2021b). Changing minds — Epistemic interventions in causal reasoning. *PsyArXiv*. doi: https://doi.org/10.31234/osf.io/db6ms

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th meeting of the cognitive science society* (pp. 1123–1128).

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition*, *182*, 331–348. doi: https://doi.org/10.1016/j.cognition.2018.09.003

Kneer, M., & Skoczeń, I. (2021). Outcome effects, moral luck, and the hindsight bias. *SSRN*. doi: http://dx.doi.org/10.2139/ssrn.3810220

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194. doi: https://doi.org/10.1093/analys/63.3.190

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experi-

ments. In W. Sinnott-Armstrong (Ed.), *Moral psychology. The cognitive science of morality: Intuition and diversity* (Vol. 2, pp. 441–448). MIT Press.

Knobe, J., & Shapiro, S. (2021). Proximate cause explained. *The University of Chicago Law Review*, *88*(1), 165–236. doi: https://doi.org/10.2139/ssrn.3544982

Kohlberg, L. (1974). *Zur kognitiven entwicklung des kindes [On the cognitive development of children]*. Suhrkamp.

Korsgaard, C. M. (2018). *Fellow creatures: Our obligations to the other animals*. Oxford University Press.

Lagnado, D. A. (2021). *Explaining the evidence: How the mind investigates the world*. Cambridge University Press.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770. doi: https://doi.org/10.1016/j.cognition.2008.06.009

Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 565–602). Oxford University Press.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *37*(6), 1036–1073. doi: https://doi.org/10.1111/cogs.12054

Lagnado, D. A., & Shanks, D. R. (2002). Probability judgment in hierarchical learning: A conflict between predictiveness and coherence. *Cognition*, *83*(1), 81–112. doi: https://doi.org/10.1016/S0010-0277(01)00168-8

Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology*, *129*, Article 101412. doi: https://doi.org/10.1016/j.cogpsych.2021.101412

Levine, S., Mikhail, J., & Leslie, A. M. (2018). Presumed innocent? How tacit assumptions of intentional structure shape moral judgment. *Journal of Experimental Psychology: General*, *147*(11), 1728–1747. doi: https://doi.org/10.1037/xge0000459

Liao, S. M., Wiegmann, A., Alexander, J., & Vong, G. (2012). Putting the trolley in order: Experimental philosophy and the loop case. *Philosophical Psychology*, *25*(5), 661–671. doi: https://doi.org/10.1080/09515089.2011.627536

Livengood, J., & Sytsma, J. (2020). Actual causation and compositionality. *Philosophy of Science*, *87*(1), 43–69. doi: https://doi.org/10.1086/706085

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955–984. doi: https://doi.org/10.1037/a0013256

Maier, M., Bartoš, F., Oh, M., Wagenmakers, E.-J., Shanks, D., & Harris, A. J. L. (2022). Publication bias in research on construal level theory. *PsyArXiv*. doi: https://doi.org/10.31234/osf.io/r8nyu

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186. doi: https://doi.org/10.1080/1047840X.2014.877340

May, J. (2018). *Regard for reason in the moral mind*. Oxford University Press.

Mayrhofer, R., & Waldmann, M. R. (2015). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science*, *39*(1), 65–95. doi: https://doi.org/10.1111/cogs.12132

McClure, J., Hilton, D. J., & Sutton, R. M. (2007). Judgments of voluntary and physical causes in causal chains: Probabilistic and social functionalist criteria for attributions. *European Journal of Social Psychology*, *37*(5), 879–901. doi: https://doi.org/10.1002/ejsp.394

McIntyre, A. (2019). Doctrine of double effect. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2019 ed.). Metaphysics Research Lab, Stanford

University.

Meder, B., & Mayrhofer, R. (2017). Diagnostic reasoning. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 433–458). Oxford University Press.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*(3), 277–301. doi: https://doi.org/10.1037/a0035944

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152. doi: https://doi.org/10.1016/j.tics.2006.12.007

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press.

Moll, J., De Oliveira-Souza, R., & Zahn, R. (2008). The neural basis of moral cognition: Sentiments, concepts, and values. *Annals of the New York academy of sciences*, *1124*(1), 161—180. doi: https://doi.org/10.1196/annals.1440.005

Montealegre, A., Bush, L., Moss, D., Pizarro, D., & Jimenez-Leal, W. (2020). Does maximizing good make people look bad? *PsyArXiv*. doi: https://doi.org/10.31234/osf.io/2zbax

Moore, M. (2019). Causation in the law. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University.

Morris, A., Phillips, J. S., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). Causal judgments approximate the effectiveness of future interventions. *PsyArXiv*. doi: https://doi.org/10.31234/osf.io/nq53z

Mumford, S., & Anjum, R. L. (2011). *Getting causes from powers*. Oxford University Press.

Nagel, T. (2012). *Mortal questions*. Cambridge University Press.

Narens, L., & Skyrms, B. (2020). *The pursuit of happiness: Philosophical and psycho-*

logical foundations of utility. Oxford University Press.

Nobes, G., & Martin, J. W. (2021). They should have known better: The roles of negligence and outcome in moral judgments of accidental actions. *British Journal of Psychology*, *113*(2), 370–395. doi: https://doi.org/10.1111/bjop.12536

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review*, *111*(2), 455–485. doi: https://doi.org/10.1037/0033-295X.111.2.455

Paharia, N., Kassam, K. S., Greene, J. D., & Bazerman, M. H. (2009). Dirty work, clean hands: The moral psychology of indirect agency. *Organizational Behavior and Human Decision Processes*, *109*(2), 134–141. doi: https://doi.org/10.1016/j.obhdp.2009.03.002

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.

Pearl, J., & Mackenzie, D. (2019). *The book of why: The new science of cause and effect*. Penguin Books.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, *9*(2), 148–158.

Piaget, J. (1954). *Das moralische urteil beim kinde [The moral judgment of the child]*. Rascher.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology*, *39*(6), 653–660. doi: https://doi.org/10.1016/S0022-1031(03)00041-6

Plunkett, D., & Greene, J. D. (2019). Overlooked evidence and a misunderstanding of what trolley dilemmas do best: Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science*, *30*(9), 1389–1391. doi: https://doi.org/10.1177/

0956797619827914

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356. doi: https://doi.org/10.1111/1467-9280.00067

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*, *140*(1), 109–139. doi: https://doi.org/10.1037/a0031903

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*(2), 165–184.

Royzman, E. B., & Hagan, J. P. (2017). The shadow and the tree. In J. F. Bonnefon & B. Trémolière (Eds.), *Moral inferences* (pp. 56–74). Psychology Press.

Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of Julie and Mark: Unraveling the moral dumbfounding effect. *Judgment & Decision Making*, *10*(4), 296–313.

Rudy-Hiller, F. (2018). The epistemic condition for moral responsibility. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2018 ed.). Metaphysics Research Lab, Stanford University.

Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition*, *156*, 164–176. doi: https://doi.org/10.1016/j.cognition.2016.07.007

Schein, C. (2020). The importance of context in moral judgments. *Perspectives on Psychological Science*, *15*(2), 207–215. doi: https://doi.org/10.1177/1745691620904083

Schwenkler, J., & Sievers, E. (in press). Cause, "cause", and norm. In P. Willemsen & A. Wiegmann (Eds.), *Advances in experimental philosophy of causation.* Bloomsbury Publishing.

Shenhav, A., & Greene, J. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667–677. doi: https://doi.org/10.1016/j.neuron.2010.07.020

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, *167*, 201–211. doi: https://doi.org/10.1016/j.cognition.2017.05.004

Singer, P. (1975). *Animal liberation: A new ethics for our treatment of animals.* Harper-Collins.

Sinnott-Armstrong, W. (2021). Consequentialism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2021 ed.). Metaphysics Research Lab, Stanford University.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives.* Oxford University Press.

Sloman, S., Fernbach, P. M., & Ewing, S. (2009). Causal models: The representational infrastructure for moral judgment. In B. Ross (Ed.), *Psychology of learning and motivation* (Vol. 50, pp. 1–26). doi: https://doi.org/10.1016/S0079-7421(08)00401-5

Sloman, S., Fernbach, P. M., & Ewing, S. (2012). A causal model of intentionality judgment. *Mind & Language*, *27*(2), 154–180. doi: https://doi.org/10.1111/j.1468-0017.2012.01439.x

Soter, L. K., Berg, M. K., Gelman, S. A., & Kross, E. (2021). What we would (but shouldn't) do for those we love: Universalism versus partiality in responding to others' moral transgressions. *Cognition*, *217*, Article 104886. doi: https://doi.org/10.1016/j.cognition.2021.104886

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, *126*(4), 323–348. doi: https://doi.org/10.1037/0096-3445.126.4.323

Spirtes, P., Glymour, C. N., & Scheines. (1993). *Causation, prediction, and search.* Springer.

Stanley, M. L., Yin, S., & Sinnott-Armstrong, W. (2019). A reason-based explanation for moral dumbfounding. *Judgment & Decision Making*, *14*(2), 120–129.

Stehfest, E., Bouwman, L., Van Vuuren, D. P., Den Elzen, M. G., Eickhout, B., & Kabat, P. (2009). Climate benefits of changing diet. *Climatic Change*, *95*(1), 83–102. doi: https://doi.org/0.1007/s10584-008-9534-6

Stephan, S., Engelmann, N., & Waldmann, M. R. (2021). The perceived dilution of causal strength [manuscript submitted for publication].

Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2020). Time and singular causation — A computational model. *Cognitive Science*, *44*(7), Article e12871. doi: https://doi.org/10.1111/cogs.12871

Stephan, S., Tentori, K., Pighin, S., & Waldmann, M. R. (2021). Interpolating causal mechanisms: The paradox of knowing more. *Journal of Experimental Psychology: General*, *150*(8), 1500–1527. doi: http://dx.doi.org/10.1037/xge0001016

Stephan, S., & Waldmann, M. R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science*, *10*(1), 242–257. doi: https://doi.org/10.1111/tops.12309

Stuart-Smith, R. F., Otto, F. E., Saad, A. I., Lisi, G., Minnerop, P., Lauta, K. C., ... Wetzer, T. (2021). Filling the evidentiary gap in climate litigation. *Nature Climate Change*, *11*(8), 651–655. doi: https://doi.org/10.1038/s41558-021-01086-7

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*(1), 49–100. doi: https://doi.org/10.1207/s15516709cog1201_2

Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology*, *4*, Article 250. doi: https://doi.org/10.3389/fpsyg.2013.00250

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

Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology*, *64*, 231–255. doi: https://doi.org/10.1146/annurev-psych-113011-143812

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*(2), 440–463. doi: https://doi.org/10.1037/a0018963

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81. doi: https://doi.org/10.1177/1745691614556679

Unger, P. K. (1996). *Living high and letting die: Our illusion of innocence*. Oxford University Press.

Viebahn, E., Wiegmann, A., Engelmann, N., & Willemsen, P. (2021). Can a question be a lie? An empirical investigation. *Ergo*, *8*, Article 7. doi: https://doi.org/10.3998/ergo.1144

Von Sydow, M., Hagmayer, Y., & Meder, B. (2016). Transitive reasoning distorts induction in causal chains. *Memory & Cognition*, *44*(3), 469–487. doi: https://doi.org/10.3758/s13421-015-0568-5

Waldmann, M. R. (Ed.). (2017). *The Oxford handbook of causal reasoning*. Oxford University Press.

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science*, *18*(3), 247–253. doi: https://doi.org/10.1111/j.1467-9280.2007.01884.x

Waldmann, M. R., & Hagmayer, Y. (2013). Causal reasoning. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 733–752). Oxford University Press.

Waldmann, M. R., & Mayrhofer, R. (2016). Hybrid causal representations. In B. Ross

(Ed.), *Psychology of learning and motivation* (Vol. 65, pp. 85–127). doi: https://doi.org/10.1016/bs.plm.2016.04.001

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. Holyoak & R. G. Morris (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 364–389). Oxford University Press.

Waldmann, M. R., & Wiegmann, A. (2010). A double causal contrast theory of moral intuitions in trolley dilemmas. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd meeting of the cognitive science society* (pp. 2589–2594).

Waldmann, M. R., Wiegmann, A., & Nagel, J. (2017). Causal models mediate moral inferences. In J. F. Bonnefon & B. Trémolière (Eds.), *Moral inferences* (pp. 37–55). Psychology Press.

Weirich, P. (2020). Causal decision theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2020 ed.). Metaphysics Research Lab, Stanford University.

White, P. A. (2006). The causal asymmetry. *Psychological Review*, *113*(1), 132–147. doi: https://doi.org/10.1037/0033-295X.113.1.132

White, P. A. (2007). Impressions of force in visual perception of collision events: A test of the causal asymmetry hypothesis. *Psychonomic Bulletin & Review*, *14*(4), 647–652. doi: https://doi.org/10.3758/bf03196815

White, P. A. (2009). Perception of forces exerted by objects in collision events. *Psychological Review*, *116*(3), 580–601. doi: https://doi.org/10.1037/a0016337

Wiegmann, A., & Engelmann, N. (2020). Entwicklungen und probleme der moralpsychologie zu beginn des 21. jahrhunderts [Developments and problems in moral psychology in the 21st century]. In N. Paulo & J. C. Bublitz (Eds.), *Empirische ethik - Grundlagentexte aus psychologie und philosophie* (pp. 139–175). Suhrkamp Verlag.

Wiegmann, A., & Engelmann, N. (2022). Is lying morally different from misleading? An empirical investigation. In L. Horn (Ed.), *From lying to perjury: Linguistic and legal perspectives on lies and other falsehoods* (pp. 89–111). De Gruyter. doi: https://doi.org/10.1515/9783110733730-005

Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, *131*(1), 28–43. doi: https://doi.org/10.1016/j.cognition.2013.12.004

Willemsen, P., & Kirfel, L. (2019). Recent empirical work on the relationship between causal judgements and norms. *Philosophy Compass*, *14*(1), Article e12562. doi: https://doi.org/10.1111/phc3.12562

Willemsen, P., & Reuter, K. (2016). Is there really an omission effect? *Philosophical Psychology*, *29*(8), 1142–1159. doi: https://doi.org/10.1080/09515089.2016.1225194

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111. doi: https://doi.org/10.1037/0096-3445.136.1.82

Wolff, P. (2012). Representing verbs with force vectors. *Theoretical Linguistics*, *38*(3-4), 237–248. doi: https://doi.org/10.1515/tl-2012-0015

Woodward, J. (2021). *Causation with a human face: Normative theory and descriptive psychology*. Oxford University Press.

Žeželj, I. L., & Jokić, B. R. (2014). Replication of experiments evaluating impact of psychological distance on moral judgment. *Social Psychology*, *43*(3), 223–231. doi: https://doi.org/10.1027/1864-9335/a000188

Ziano, I., Wang, Y. J., Sany, S. S., Ho, N. L., Lau, Y. K., Bhattal, I. K., ... others (2021). Perceived morality of direct versus indirect harm: Replications of the preference for indirect harm effect. *Meta-Psychology*, *5*, Article MP.2019.2134. doi: https://doi.org/10.15626/MP.2019.2134

Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Causality and

counterfactuals in group attributions. *Cognition*, *125*(3), 429–440. doi: 10.1016/j.cognition.2012.07.014

Appendix A Engelmann & Waldmann (2022a)

Cognition 226 (2022) 105167

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

How causal structure, causal strength, and foreseeability affect moral judgments

Neele Engelmann *, Michael R. Waldmann

Department of Psychology, University of Göttingen, Gosslerstr 14, 37073 Göttingen, Germany

ARTICLE INFO

Keywords:
Causal reasoning
Moral judgment
Causal chains
Foreseeability
Causal proximity
Indirect harm

ABSTRACT

Causal analysis lies at the heart of moral judgment. For instance, a general assumption of most ethical theories is that people are only morally responsible for an outcome when their action causally contributed to it. Considering the causal relations between our acts and potential good and bad outcomes is also of crucial importance when we plan our future actions. Here, we investigate which aspects of causal relations are particularly influential when the moral permissibility of actions and the moral responsibility of agents for accidental harms are assessed. Causal strength and causal structure are two independent properties of causal models that may affect moral judgments. We investigated whether the length of a causal chain between acts and accidental harms, a structural feature of causal relations, affects people's moral evaluation of action and agent. In three studies ($N = 2285$), using a combination of vignettes and causal learning paradigms, we found that longer chains lead to more lenient moral evaluations of actions and agents. Moreover, we show that the reason for this finding is that harms are perceived to be less likely, and therefore less foreseeable for agents, when the relation is indirect rather than direct. When harms are considered equally likely and equally foreseeable, causal structure largely ceases to affect moral judgments. The findings demonstrate a tight coupling between causal representations, mental state inferences, and moral judgments, and show that reasoners process and integrate these components in a largely rational manner.

1. Introduction

Morality and causation are deeply intertwined. For instance, we typically only hold agents accountable for outcomes they have presumably caused. Considering the potential consequences of actions is also of crucial importance *before* these actions are performed. As such, we may criticise our friends' decision to smoke, eat meat, or not get vaccinated by pointing to potential harms that their actions may cause. For many practical purposes, saying that someone "caused" some unwanted outcome is equivalent to saying that they are to blame for it (Livengood & Sytsma, 2020; Samland & Waldmann, 2016; Schwenkler & Sievers, 2022). And even when we clearly are not at fault (e.g., for lack of relevant knowledge or control), we usually cannot help but feel guilt when we harm others. An extreme example are cases of people who caused severe injury or even death to others accidentally and blamelessly, but nevertheless feel guilty, sometimes for the rest of their lives (Anderson et al., 2021).

Considering causal relations in moral judgments feels spontaneous, casual, and natural to us. However, this type of reasoning is a product of sophisticated cognitive operations. Condemning an action because of its negative consequences requires an understanding of the fact that the action *causes* these consequences, and is not merely statistically associated with them. Moreover, we would naturally consider how *likely* these negative consequences actually are, whether there are alternative causes of the outcomes, and whether there are any additional positive or negative effects. In the meat example, a friend might argue that they, too, regret the environmental impact of meat consumption, and don't deny that our food habits are causes. However, they may additionally emphasize that the link between their particular choice of meal today and global outcomes is weak, and that there are substantial alternative causes of global warming that should be addressed instead, such as the use of fossil fuels. Finally, they may argue that in light of these considerations, the positive effect of enjoying their meal outweighs the negligible impact on the planet. If we argue back, we will probably cite different causal relations to change their minds, such as the claim that one person going vegan inspires others, and that the behavior of a large number of people may have an impact on global outcomes after all. How we acquire and confidently use such causal knowledge is by no means a

* Corresponding author.
E-mail address: neele.engelmann@uni-goettingen.de (N. Engelmann).

https://doi.org/10.1016/j.cognition.2022.105167
Received 4 January 2022; Received in revised form 17 April 2022; Accepted 10 May 2022
Available online 31 May 2022

trivial question (see, e.g., Gopnik & Schulz, 2007; Waldmann, 2017).

1.1. How causal models support moral reasoning

Causal Bayes Nets provide us with a structured method of describing and studying such and similar inferences within causal models. They have their origins in computer science and Artificial Intelligence (Pearl, 2000; Pearl & Mackenzie, 2019; Spirtes et al., 1993), but are currently also one of the most successful theories of human causal reasoning (Waldmann, 2017; Sloman, 2005; Rottman and Hastie, 2014). They have two crucial dimensions, *causal structure* and *causal strength*. Causal structure describes, simply put, what causes what. The causal relata are conceptualised as variables, which can either be binary or continuous. A binary variable takes on a value of one when a particular event occurs (e.g., our friend is mad at us), and a value of zero otherwise (our friend is not mad at us). Causal models are often depicted as graphs, with nodes representing cause and effect variables, and arrows representing the directed causal relationships between them. Fig. 1 shows three simple examples of causal structures. One cause can have several effects (common-cause structure), one effect can have several causes (common-effect structure), or several causes can be lined up in a row, each being its predecessors effect (causal chains). Causal structure thus represents important *qualitative* causal knowledge. We can learn about causal structures by observing statistical regularities in our environment (Cheng, 1997; Griffiths & Tenenbaum, 2005; Meder et al., 2014), or by explicit instruction (e.g., reading a textbook or a study, or by asking questions, see Hagmayer & Engelmann, 2020).

The second crucial dimension of causal models is *causal strength*. Causal strength parameters express how strong the causal relationship between two directly linked variables within a network is. In the Causal Bayes Nets framework, this strength is conceptualised as probabilistic dependence. That is, saying that A directly causes B expresses that the presence of A increases or reduces the probability of B. How exactly causal strength is best measured is debated (Perales et al., 2017). A simple way to estimate causal strength is to observe how often B occurs when A has occurred, and subtract how often B occurs when A is absent. This measure is known as a contingency, or as Δ-p, a difference between two conditional probabilities (Cheng & Novick, 1990, 1991). Cheng (1997) proposed a modified measure in her Power PC theory, which expresses causal strength as the probability of an effect in the presence of the cause in the hypothetical absence of all alternative causes. Whenever an effect has just one cause, the strength of the link between them can simply be estimated by p(effect|cause) (a simple conditional probability), in both the Δp and the Power PC accounts.

A further feature of causal models is that they encode assumptions about conditional independence (Markov condition). In a three-variable chain, for example, the default assumption is that the initial cause covaries with the final effect but becomes statistically independent once the mediating variable is held constant (see for example Mayrhofer & Waldmann, 2015). Assuming the validity of the Markov condition allows us to make inferences about indirect probabilistic relations. For example, in a causal chain with three variables, the two link strengths can be used to calculate the indirect strength between the initial and the terminal variable.

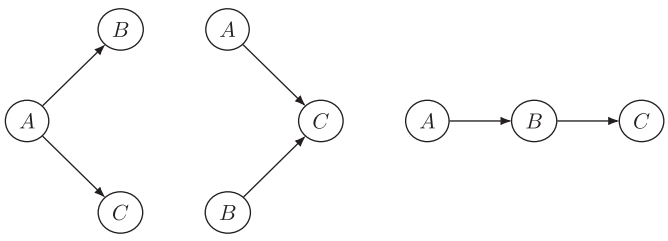


Fig. 1. Examples for different causal structures: a common-cause structure (left), a common-effect structure (middle), and a causal chain (right).

There are several ways in which we can learn about probabilities, and thereby about causal strength. Probabilities can be explicitly stated, like when we learn about a medication's possible side effects by reading the packaging. In other cases we learn via direct interaction with our environment, for instance when we track how our friends tend to react to different behaviours of ours. The former is known as "learning by description" and the latter as "learning by experience" (Hertwig et al., 2004, 2018). Description and experience formats can also be mixed. Both formats can be used to convey quantitative information.

Equipped with knowledge about causal structure and causal strength, we can make a variety of inferences that are crucial for moral judgments (Sloman et al., 2009; Waldmann et al., 2017; Waldmann & Wiegmann, 2012). Generic causal knowledge, that is, knowledge about which causal relations generally exist in the world (structure knowledge) allows us to tentatively predict outcomes that matter to us, both on a generic level (e.g., "does smoking cause cancer?") as well as in singular cases ("will Peter's smoking cause him to get cancer?"). However, additional knowledge about strength allows us to make more specific quantitative estimates. Both structure and strength are relevant for prospective moral judgments, that is, when we gauge the morality of an action prior to knowing which of its possible consequences will actually occur. Structure knowledge is important because it tells us which effects we, and the person contemplating the act, can and should anticipate. An agent who knowingly puts others at risk will be evaluated negatively in most cases. Strength information may be an additional factor affecting the degree of an action's permissibility. If we learn that the causal relationship between action and harm is only weak, we will likely see the action as more permissible than when the relationship is strong.

Causal models not only enable predictive judgments, they also allow us to think backwards, and consider the causes of events that have already happened (Meder & Mayrhofer, 2017). For instance, we may conclude that someone who caused an accident may have been intoxicated. In other cases, we might observe a potential cause and an effect in a specific case, and wonder whether the potential cause actually produced the effect in this case, or whether it was only a co-incidence that both were present. For example, someone who took a drug that can affect alertness may have been involved in an accident, but in this particular case the drug may not have causally contributed to the accident. Such questions, which are about singular causation (Cheng & Novick, 2005; Stephan & Waldmann, 2018), are highly relevant because holding someone responsible for some outcome generally requires that their action caused it (Alicke, 2000; Driver, 2008; Rudy-Hiller, 2018).

It is generally acknowledged that causal reasoning matters for moral judgments (Lagnado & Gerstenberg, 2017; Sloman, Fernbach, & Ewing, 2009; Waldmann, Wiegmann, & Nagel, 2017). For instance, people's moral judgments are sensitive to whether or not there was a causal connection between someone's action and harm to another person (Cushman, 2008), analyzing the causal structures of moral dilemmas can explain moral inferences (Waldmann et al., 2017; Wiegmann & Waldmann, 2014), and counterfactual inferences on causal models determine how we allocate responsibility in group settings (Zultan et al., 2012). Nevertheless, the interplay between the two dimensions of causal models, causal structure and causal strength, has not yet been systematically investigated in the context of moral reasoning. The aim of this article is to start filling this gap, focusing on the case of causal chains.

1.2. Causal chains and moral judgment

In a causal chain consisting of the variables A, B, and C, A may directly cause B, and B may cause C (see Fig. 1). While three variables are minimally required, causal chains can of course also be longer. Chains are a particularly interesting starting point for our project because the interaction between causal structure and causal strength seems very likely in moral judgments about chains, and has in fact been

stipulated, but not yet investigated (see Sloman et al., 2009). To illustrate, consider a case in which one person's action has a risk of causing harm to another. For instance, a doctor may consider which of two medications to administer to an unconscious patient in an emergency. Assume that both drugs would fulfill their main purpose, saving the patient's life, equally well. However, both can also cause chronic dizziness as an unwanted side effect. Say that for drug A, the doctor remembers that the drug sometimes causes dizziness directly. For drug B, on the other hand, the doctor recalls that the drug sometimes causes patients to feel more energized than usual after their recovery, which in turn sometimes leads to increased levels of physical activity. Because of the higher physical activity, patients' iron levels can sometimes become depleted, which, finally, sometimes leads to a chronic feeling of dizziness. In sum, whereas drug A leads to the negative side effect directly, drug B leads to the same side effect through a relatively long chain of events. Fig. 2 illustrates the two different causal structures that are involved in this example. Does it seem better to prescribe drug B rather than drug A? And if the patient actually suffers from chronic dizziness after their recovery, is the doctor who prescribed drug A (direct relation) more blameworthy or morally responsible than the doctor who prescribed drug B (indirect relation)? If so, why?

While the difference between the causal structures mediating the side effect (direct vs. chain) may be the most salient difference in the example, the case also touches upon intuitions about causal strength, by using probabilistic expressions such as "sometimes causes". By featuring multiple probabilistic links, the overall relationship between the initial action (prescribing the medicine) and the final outcome (chronic dizziness) may appear to be weaker in the causal chain compared to the direct relation (which includes just one probabilistic link). If all links have equal strengths, and the chain honors the Markov condition, a lower probability of the dizziness given the medication is indeed formally entailed (see Stephan et al., 2021). However, the overall relation between medication and dizziness can also be equally strong in the chain as in the direct relation if the individuals links in the chain are sufficiently strong.

In our experiments, we will investigate whether actions that can cause harm to another person are seen as more morally permissible in chains compared to direct relations. We will also investigate whether agents are viewed as less morally responsible for harms that were brought about via a chain of events, as compared to a direct relation. If we actually find more lenient moral judgments in the chain conditions, there are two prominent candidate explanations for them, and both have their roots in the underlying causal representation of the situation.

1.2.1. The probabilistic model

One possibility is that effects of causal structure (chains vs. direct relations) on moral judgments are ultimately driven by inferences about causal strength. On this view, actions and agents are evaluated more favourably in chains because people take harms to be less likely, compared to direct relations. If harm is less likely to result from an action, it makes intuitive sense that the action should be seen as more permissible in a prospective moral evaluation. But why would the a priori likelihood of harm matter in a *retrospective* assessment of moral responsibility? After all, it is already known that harm has occurred at this stage. One reason could be that retrospective judgments may be influenced by singular causation judgments. These judgments assess whether a present potential cause and a present effect are indeed causally related in a particular situation, or whether their co-occurrence is merely a coincidence. Research on singular causation judgments has demonstrated that causal strength influences these inferences (Stephan & Waldmann, 2018). Everything else equal, stronger causes are more likely to have caused an outcome than weaker causes. If people are less confident that an action actually caused the harm in question, it makes sense that they would hold the agent less morally responsible. Another important reason might be reduced a priori foreseeability of the outcome on the part of the agent (see 1.3).

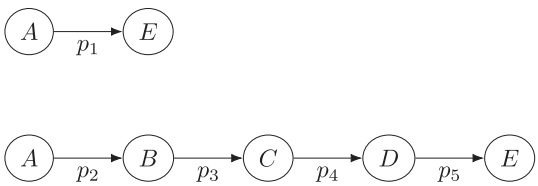


Fig. 2. Illustration of a direct causal relation (top) and a longer causal chain (bottom). *p*'s stand for the strengths of causal links.

There is evidence that people actually tend to infer a weaker overall relation in causal chains compared to direct relations (Bes et al., 2012; Stephan et al., 2021). As pointed out earlier, whether this inference is normatively justified or not depends on the assumed strengths of the causal links in the different conditions. In Fig. 2, if all individual causal links p_1 - p_5 have roughly the same probabilistic strength, and the chain honors the Markov condition, it is analytically true that the direct effect has a higher probability than the indirect effect. This is true because in Markov chains the strength of the relation between the initial cause and the final effect can be calculated by multiplying the strengths (measured as Δp) of all mediating links (Stephan et al., 2021). Thus, if all links had a strength of 0.7, the overall strength of the relation between action and harm in the direct relation in Fig. 2 would also be 0.7, but it would be 0.7^4 in the chain. Verbal cues such as "sometimes causes" probably convey to participants that all links are probabilistic and roughly equally strong (Meder & Mayrhofer, 2017). When all links are deterministic (strength = 1, which is almost never the case in real-world scenarios), or when no information about link strength is available at all (which means that the overall relation could be weaker, stronger, or equally strong in chains compared to direct relations), weakening is not normative.

The probabilistic model thus predicts that more positive moral evaluations of action and agent in chains compared to direct relations should be observed because participants often infer that harm is less likely in chains. When reasoners know that the overall relation between action and harm is just as strong in chains as in direct relations, no effect on moral judgments is expected.

1.2.2. The indirectness model

Another possibility is that causal structure itself drives effects of chain length on moral judgments because chains make the harm appear more indirect, independent of its likelihood. Royzman & Baron (2002) demonstrated that people regard indirect harm as morally better than direct harm across a range of scenarios featuring different kinds of indirectness (for a replication, see Ziano et al., 2021). People also considered harm to be less likely to be caused when the relation was indirect rather than direct. However, these studies did not investigate whether the better moral evaluation was due to the lower probability of harm being caused, or whether both are separate effects of indirectness. Psychologically, effects of indirectness itself might be explained by Construal Level Theory (Trope & Liberman, 2010). According to this theory, different kinds of distance (spatial, temporal, social, hypothetical or counterfactual worlds, etc.) are represented as one unified dimension of "psychological distance" in our minds (but see Calderon, Mac Giolla, Ask, & Granhag, 2020; Žeželj & Jokić, 2014; Maier et al., 2022). Increasing psychological distance in any way is predicted to have similar downstream effects on a range of judgments and decisions.

The indirectness model predicts that actions and agents should be evaluated more positively in chains than in direct relations, even when the probabilistic relationship between action and harm is known to be equally strong in both cases. In other words, the indirectness model assumes that there is a genuine effect of causal structure on moral judgments, independent of causal strength.

1.3. The role of foreseeability

To be held morally responsible for a harmful outcome, agents are generally not just required to have caused the outcome. It also has to be reasonably foreseeable to them that the harmful outcome might be produced by their actions. This requirement is reflected in the so-called *epistemic condition* in philosophical theories of moral responsibility (Rudy-Hiller, 2018), and in definitions of negligent or reckless behaviour in the law (see, e.g., Dubber, 2015, *pp*. 42–46). There is robust evidence that foreseeability plays a crucial role in the judgments of laypeople as well, be they about permissibility (Cushman, 2008, 2013; Paharia et al., 2009), blame (Alicke, 2000; Fincham & Jaspars, 1983; Lagnado & Channon, 2008; Samland & Waldmann, 2016), punishment (Cushman, 2008, 2013), liability (Johnson & Drobny, 1985), or agent causation (Alicke, 2000; Fincham & Jaspars, 1983; Kirfel & Lagnado, 2021; Lagnado & Channon, 2008).

We therefore expect that inferences about foreseeability are a crucial mediator between causal models and moral judgments. Generally, we predict that people are held less accountable for harmful outcomes the less they were able to foresee the outcome. Prospectively, actions should become more permissible the less an agent can foresee it to produce harmful consequences. For the comparison between direct relations and chains, we predict that if agents and actions are evaluated more positively in chains, this will be because people take agents to be less able to foresee the actual occurrence of harm, compared to direct relations. This hypothesis is compatible with both the probabilistic model and the indirectness model. Under the probabilistic model, harm may seem less foreseeable because it becomes less likely. Under the indirectness model, harm may seem less foreseeable because of the indirect relation between action and outcome. Different levels of foreseeability between chains and direct relations only arise when agents in both cases are *aware* of the respective causal relation between their action and some harmful outcome. We thus predict that chain length will only affect moral judgments when agents know that and how their action is connected to a harmful outcome. Conversely, when harm is taken to be equally foreseeable or unforeseeable in chains and direct relations, we predict no difference in moral evaluations.

1.4. Previous work on causal chains and moral judgment

Causal chains have long been of interest to psychologists, philosophers, and legal scholars alike (Hart & Honoré, 1985; Hilton et al., 2010; Knobe & Shapiro, 2021; Livengood & Sytsma, 2020; Spellman, 1997). However, a popular research question has been how people *select* "the" (main, most important, proximate, or legal) cause of an outcome from a sequence of events that led to the outcome in question. For example, if one person shoots a gun at another, and this second person, while fleeing from the gunshots, accidentally pushes a bystander into oncoming traffic, should the shooter be liable for the bystander's death (*cf*. Pearl & Mackenzie, 2019, *p*. 288)? Or does the pushing by the fleeing man supersede the initial action, and should therefore be regarded as the cause of death? The answers that people give about such or similar cases have been shown to depend on the cause's position in the chain, its probabilistic relation to subsequent events and to the final outcome, or whether the cause is an intentional action or merely a physical event (Hilton et al., 2010; Lagnado & Channon, 2008; McClure et al., 2007; Spellman, 1997). In any case, determining the main, proximate, or legal cause of an outcome comes down to comparing different events *within* a chain, and then designating the most important one as the cause.

In contrast to this line of research, the cases that we are focusing on here are not selection tasks. In the doctor example, prescribing the medication is arguably the main cause of patients experiencing dizziness, in the chain version as well as in the direct version of the case. However, we are going to *vary* the actions' proximity to outcomes within the causal model representation, construing the relation as direct or as indirect. This amounts to a comparison *between* two possible ways in which the same action and outcome could be related. We are interested in a level of causal reasoning that is more fundamental than causal selection (see also Samland & Waldmann, 2016). Before a cause can be selected from a chain, people have to arrive at a cognitive representation of the cause's relation to other events of interest.

To our knowledge, only one study has so far compared chains of different length while holding the initial action and the final outcome constant. Johnson & Drobny, 1985 presented their participants with two versions of a case in which a truck driver forgets to replace a safety pin in his truck after an inspection. In either case, the steering fails as a result, and an accident occurs. In the "short chain" condition, the accident causes a fire, and the fire burns down a nearby house. In the "long chain" conditions, the fire ignites some gasoline, which flows down a hill and across a river, sets fire to grass on the other side, which finally also causes a house to burn down. Participants indicated that the driver was equally negligent in both conditions, but that he was less able to foresee the damage to the house in the long compared to the short chain condition. They also judged the driver to be less liable for the damage in the long chain condition. These findings thus provide initial evidence for an effect of causal structure. However, Johnson & Drobny, 1985 did not investigate the cognitive mechanism behind their findings. Is the driver less liable because he was less able to foresee harm? Was he less able to foresee harm because it became increasingly unlikely to be caused by his negligent omission in the longer chain? Moreover, their experiment confounds causal structure with the type of chain events. The short chain consists of saliently different events than the long chain. We control for this confound by comparing chains and direct relations that could, in principle, be underwritten by the same causal mechanism at different levels of granularity.

2. Overview of experiments

Experiment 1 set the stage for the project by establishing that actions are seen as more permissible, and agents as less responsible, when actions and harmful outcomes are connected via a chain rather than directly. The experiment also confirmed that this effect depends on the attribution of different levels of outcome foreseeability to agents.

Experiment 2 aimed to test the probabilistic model and the indirectness model against each other by crossing causal strength and causal structure. Participants learned about the strength of the overall relations between initial actions and final negative outcomes. The same contingency data linking these two types of events were presented for chains and direct relations. The results confirmed predictions of the probabilistic model, but relations were not yet perceived as equally strong in chains and direct relations. Thus, there were still some findings that are compatible with both models.

In Experiment 3, the number of observations for the contingency learning task was increased to improve learning. The results now showed that chains and direct relations were perceived as equally strong. The results for moral judgments generally supported the probabilistic model. There was also some evidence that structure itself matters, which is predicted by the indirectness model. However, effects of structure alone were small and not consistently detected. Experiment 3 also explored foreseeability in a more fine-grained manner.

The materials, data, and analysis code for all experiments (as well as additional analyses and figures) are available at https://osf.io/5bmgc/(from here on: Supplementary Materials). For all analyses and figures, we used R (Core Team, 2020) and RStudio (RStudio Team, 2020) in combination with the following packages (in alphabetical order): *effsize* (Torchiano, 2020), *ez* (Lawrence, 2016), *ggpubr* (Kassambara, 2020), *Hmisc* (Harrell, 2020), *lavaan* (Rosseel, 2012), *lmtest* (Zeileis & Hothorn, 2002), *MASS* (Venables & Ripley, 2002), *MBESS* (Kelley, 2020), *mediation* (Tingley et al., 2014), *meta* (Balduzzi et al., 2019), *nlme* (Pinheiro et al., 2020), *rcompanion* (Mangiafico, 2021), *reshape2* (Wickham, 2007), *tidyverse* (Wickham et al., 2019), *xtable* (Dahl et al., 2019).

3. Experiment 1: Causal structure and outcome foreseeability

The aim of this experiment was to test whether the representation of a causal chain between an action and a harmful final outcome would lead to a more positive moral evaluation of the action or the agent than the representation of action and outcome as being directly linked. Furthermore, we tested whether these effects are mediated by attributions of outcome foreseeability to agents. Both the probabilistic model and the indirectness model predict that agents in chains are evaluated more leniently because people take them to be less able to foresee harm, compared to agents confronted with direct relations. Chains only lead to less foreseeability than direct relations when agents know about the respective causal relation between their action and harm. We thus vary agents' knowledge, and predict that there will only be a moral difference between chains and direct relations when agents know about the relations. When they are unaware, there should be no difference between these two types of causal structure.

The results of this experiment will not distinguish between the probabilistic model and the indirectness model. Both models predict an effect of causal structure and the described interaction with foreseeability, albeit for different reasons. Experiments 2 and 3 will implement more focused tests. However, we will assess whether overall relations are perceived as weaker in chains than in direct relations. Such an effect would be a necessary prerequisite for the validity of the probabilistic model.

We manipulated causal structure (direct relations vs. chains) and the awareness of agents of the relation between their action and the potential harmful outcome (knowledge vs. no knowledge). Participants were asked for a prospective moral evaluation of permissibility ("is it okay to act?"), and for a retrospective evaluation of agents' moral responsibility for the harms caused.

3.1. Methods

3.1.1. Design and participants

We used a 2 (structure: *direct* vs. *chain*, within-subject) x 2 (knowledge about the causal relation: *knowledge* vs. *no knowledge*, between-subjects) design. We created three cover stories, which were combined with the two levels of the structure manipulation in a Latin Square design, such that each participant saw the *direct* case in a different cover story than the *chain* case. This design resulted in six unique combinations of cover story and relation. Participants were randomly assigned to one of those combinations. The sample size was determined by simulation (see Supplementary Material for the code). Based on pilot studies we assumed an effect of $d = 0.22$ for moral permissibility ratings and an effect of $d = 0.30$ for moral responsibility ratings in the *knowledge* conditions. We predicted null effects on both measures in the *no knowledge* conditions. With a sample size of at least 700 participants, we were able to detect a) the predicted effects on moral judgments in separate, one-sided *t*-tests as well as b) the predicted interaction between causal structure and knowledge in a 2×2 ANOVA with a power of $>90\%$ for each measure (combined power for both measures and all predicted effects: 87%). 726 participants completed the survey. After applying the exclusion criteria, data of 704 participants remained for all analyses ($M_{age} = 34.81$, $SD_{age} = 13.14$, 353 women, 343 men, 7 non-binary, 1 no answer).

3.1.2. Materials and procedure

The experiment began with information about the generic causal relationship between an action and a harmful outcome. One of the three cover stories, here shown in the *chain* version, for example, read (see Supplementary Material for the other stories):

> A group of scientists is investigating the effects of exposure to a certain chemical called Proskine. In their studies, they found the following results:
>
> - When Proskine is produced and stored, this sometimes causes changes in the PH level within a storage container.
> - When these changes occur, it sometimes causes Xaligene gas to develop in the container.
> - When Xaligene gas is present in a container, this sometimes causes condensation.
> - When this condensation occurs, it sometimes causes another chemical called Yosium to form in the container.
> - When Yosium is present, it sometimes causes certain proteins to be blocked in the human body (for people in the vicinity).
> - When these proteins are blocked, it sometimes causes Vanine, a transmitter substance, to build up.
> - When Vanine has built up, this sometimes causes a deficiency of a molecule called Alpha 3.
> - When there is an Alpha 3 deficiency, it sometimes causes some tissue irritation in the lung.
> - When this tissue irritation occurs, it sometimes causes Marasia illness, a severe respiratory condition.
>
> Please take a moment to study and understand the illustration.

In the *direct* condition, just one link was described ("When Proskine is produced and stored, this sometimes causes people in its vicinity to get Marasia illness, a severe respiratory condition"). Fig. 3 shows an example of the illustrations that were used in each condition. We decided to use relatively long chains because previous experiments have shown that a substantial reduction in predictive causal judgments is needed to affect moral judgments (Engelmann & Waldmann, 2021). The nodes connecting action and outcome were physical or biochemical events in all stories. We kept the material artificial to preclude knowledge or any strong assumptions about the strength of the causal links between them. After participants had learned about the generic causal relation, they were presented with the case of an agent who was about to perform the harmful action. In the *knowledge* conditions, we pointed out that the agent was aware of the causal relation that participants had just learned about:

> In a pharmaceutical lab, the chemist Mary produces and stores some Proskine, which she needs for her research. The lab is shared with several colleagues.
> Since the scientists studying Proskine have published their results, she is aware of the previously described findings.

In the *no knowledge* conditions, we stated:

> Since the scientists studying Proskine have not published their results so far, Mary cannot be aware of them. To the best of her knowledge, there are no special risks associated with producing and storing Proskine.

We asked participants: "From a moral point of view, is it okay for Mary to produce and store Proskine?" Ratings were provided on a 10-point scale ranging from "not at all" to "fully". On the next page, we informed participants that the harmful outcome had actually occurred ("It later turns out that Mary's colleague Andrew contracted Marasia illness") and asked them to assess the agent's moral responsibility ("To what extent is Mary morally responsible for Andrew contracting Marasia illness?", using the same scale). After participants had morally evaluated two scenarios in this way (one with a direct relation and one with a chain), both scenarios were presented anew and we asked participants for a predictive causal judgment ("Given that Proskine is produced and stored, how likely is it for a person close by to develop Marasia illness?", from 0-100%). We also asked them to rate the agents' ability to foresee the harm ("To what extent could Mary foresee that someone would be harmed by her action?", 10-point scale from "not at all" to "fully"). The experiment ended with a debriefing and the assessment of demographic variables.

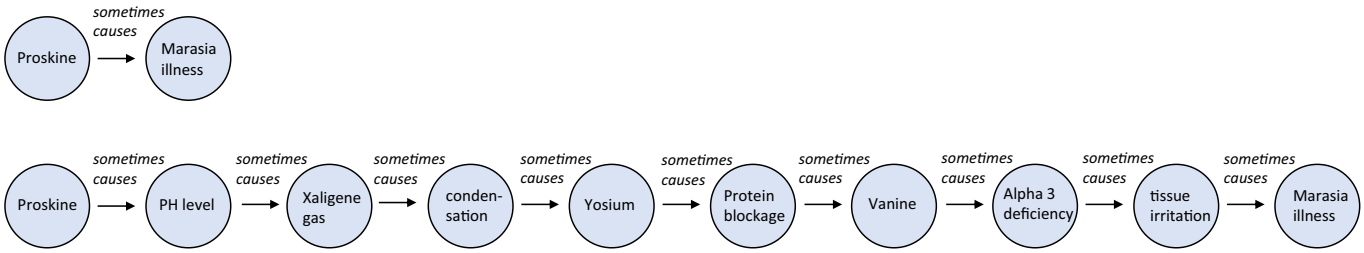


Fig. 3. Example illustrations of a direct relation (top) and the corresponding chain (bottom) in Experiment 1.

3.2. Results and discussion

Fig. 4 provides an overview of the results, and Table 1 shows the descriptive statistics for all conditions. As predicted, actions were seen as more permissible in the chain conditions than in the direct relation conditions, but only when agents were aware of the relation between their action and the harmful outcome (see Table 2 for the results of the planned *t*-tests for both moral judgment questions, and Table 3 for the results of ANOVAs for all measures). Participants also judged agents to be less morally responsible for the harm caused by their actions in the chain conditions. Against our predictions, the effect of structure on responsibility judgments did not entirely disappear in the no knowledge conditions, but the remaining effect was very small there (see Table 2). Thus, we showed that causal chains lead to a more lenient moral evaluation of actions and agents than direct causal relations, and that this effect is largely mediated by attributions of outcome foreseeability.

While participants always perceived outcomes as less likely to be caused in chains than in direct relations, this weaker causal relationship only led to lowered attributions of outcome foreseeability when agents were aware of the respective relations (see Fig. 4, Table 3). Without such knowledge, harm was generally taken to be unforeseeable, actions as permissible, and agents as not morally responsible, no matter the causal structure.

Table 10 in the Appendix shows the correlations between all measures in this experiment (put briefly, permissibility ratings are negatively correlated with all other measures, while all other measures are positively correlated). See Supplementary Materials for additional mediation analyses and for a figure showing participants' response trajectories across the different measures.

The results of the experiment are in line with the probabilistic model, but they cannot rule out the indirectness model either. After all, reasoners might always perceive harms as less foreseeable in chains, even when causal strength is equally high as in a direct relation. The probabilistic model, on the other hand, assumes that outcomes only become less foreseeable because they are perceived as less likely. For a more thorough investigation of the mediating role of predictive causal judgments, we are going to vary both causal structure and causal strength in the subsequent experiments.

4. Experiment 2: Causal structure and causal strength

In this experiment we crossed the structure by which actions and outcomes are related (chains vs. direct relations) with the strength of the overall relation between action and outcome (weak vs. strong). The probabilistic model predicts that causal structure should cease to affect foreseeability, and thereby moral judgments, when the overall relation between action and outcome is equally strong in chains and in direct relations. Attributions of outcome foreseeability and moral judgments should only be affected by strength on this view. Low strength should lead to less foreseeability, and a more positive moral evaluation of action and agent, compared to high strength. The indirectness model, on the other hand, predicts that causal structure still affects outcome foreseeability and moral judgments (less foreseeability and more positive moral evaluation in chains) when relations are perceived to be equally strong.

We also added a singular causation question to our procedure in this experiment. That is, we retrospectively asked participants how confident they were that the action actually caused the harmful outcome in this situation. Singular causation is generally seen as a prerequisite for assigning moral responsibility (see, e.g., Driver, 2008; Rudy-Hiller, 2018). Thus, an additional reason why moral responsibility is lower in chains than in direct relations (apart from lower outcome foreseeability) could be that participants are less confident that agents actually have caused the harmful outcomes in chains.

Just as causal strength, confidence in singular causation depends on how often reasoners observe an effect in the *presence* of its putative cause, as well as in its *absence* (Cheng & Novick, 2005; Stephan & Waldmann, 2018). We aimed to control both of these probabilities in our experiment (manipulating p(outcome|action), and keeping p(outcome|¬action) fixed at zero) to better control what subjects assume about the observed causal relations. We decided to present these conditional probabilities in a trial-by-trial observational learning task (see Stephan et al., 2021, for a similar paradigm). That is, participants repeatedly observed whether cause and effect were present or absent in particular cases.

4.1. Methods

4.1.1. Design and participants

We employed a 2 (strength: p(outcome|action) =.80 vs. p(outcome|action) =.20) x 2 (structure: *direct* vs. *chain*) x 3 (cover story: *apples* vs. *chemical* vs. *computer*) design. In all conditions p(outcome|no action) was zero. All factors were manipulated between subjects. We collected the same measures as in Experiment 1, but we added a control question about the probability of the effect in the *absence* of the target action (p(outcome|no action)) and the aforementioned question about singular causation. We conducted a power analysis based on meta-analytic effect size estimates (including data of pilot studies) for the effect of causal structure on ratings of moral permissibility (d = 0.25 [0.17;0.32]) and moral responsibility (d = 0.29 [0.21;0.36], see Supplementary Material for details on the meta analysis and the data of the pilot studies). Based on this analysis we planned to collect data of at least 694 participants in this experiment. With this sample size we were able to detect both effects with 94% power in one-sided paired *t*-tests (Faul et al., 2007). [1] The power analysis focused on the effect of causal structure on the two moral measures because we expected these effects to be the smallest in the design. 727 participants completed the survey. 16 were excluded, leaving data of 711 participants for the analyses (M_{age} = 36.38, SD_{age} = 13.52, 416 women, 287 men, 6 non-binary, 2 no answer).

4.1.2. Materials and procedure

In each condition participants first learned about the generic causal relation between an action and a harmful outcome (*direct* vs. *chain*) in

[1] For the lower bounds of the meta-analytic 95% confidence intervals (d = 0.17 for permissibility and d = 0.21 for responsibility), the planned sample size yields a power of 75% and 89%, respectively.

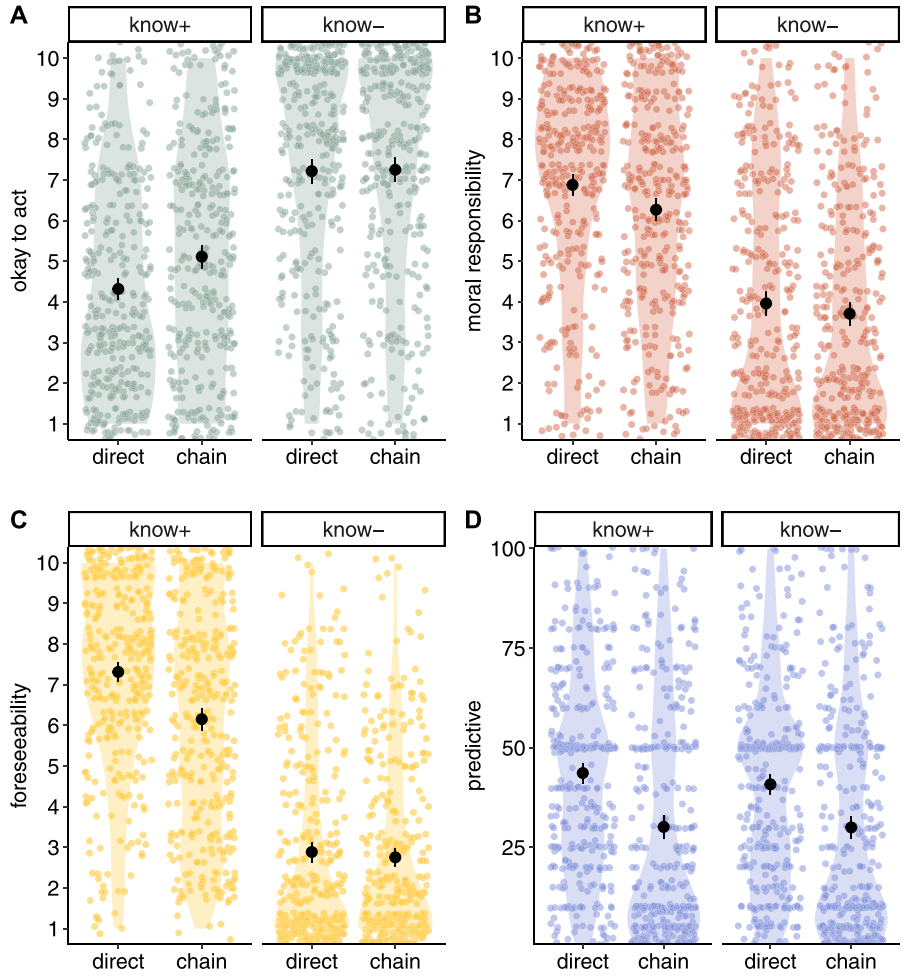


Fig. 4. Means and 95% confidence intervals for ratings of moral permissibility (A), moral responsibility (B), attributions of outcome foreseeability (C), and predictive causal judgments (D) in Experiment 1.

Table 1
Means and standard deviations per condition in Experiment 1.

Condition	Query	Relation	*M*	*SD*	*n*
knowledge	permissibility	direct	4.32	2.56	352
		chain	5.11	2.75	
	responsibility	direct	6.88	2.55	
		chain	6.27	2.69	
	foreseeability	direct	7.31	2.34	
		chain	6.15	2.69	
	predictive	direct	43.63	24.88	
		chain	30.12	28.57	
no knowledge	permissibility	direct	7.22	2.83	352
		chain	7.25	2.83	
	responsibility	direct	3.96	2.91	
		chain	3.70	2.71	
	foreseeability	direct	2.88	2.40	
		chain	2.75	2.20	
	predictive	direct	40.80	24.24	
		chain	29.99	27.20	

Table 2
Results of the planned *t*-tests for effects of causal structure (chains vs. direct relation) on judgments of moral permissibility and moral responsibility in Experiment 1. *P*-values are one-tailed for the knowledge conditions, and two-tailed for the no knowledge conditions (in line with the hypotheses).

Condition	Query	*t(df)*	*p*	*d (95 % CI)*
knowledge	permissibility	−5.08 (351)	<.001	−0.30 (−0.42; −0.18)
	responsibility	3.80 (351)	<.001	0.23 (0.11; 0.35)
no knowledge	permissibility	−0.26 (351)	0.79	−0.01 (−0.10; 0.08)
	responsibility	2.03 (351)	0.04	0.09 (0; 0.18)

the same way as in the previous experiment. Subsequently they were informed that scientists were also reviewing health records of people who were or were not exposed to the substance or item in question. We informed participants that they were going to see 40 of these health records in the form of illustrations. These illustrations were representations of the causal structure, very similar to the depictions of the relation between action and outcome in the structure instruction phase (see Fig. 5). The presence versus absence of causes and effects were indicated by the colors of nodes (yellow: present, grey: absent). In the *chain* conditions we instructed participants that the variables connecting action and outcome were not measured in the study of health records. Thus, it was unknown whether they were present or absent in any single case. Visually, this was represented by depicting them as light-grey nodes with dashed outlines. Arrows were also dashed and light-grey in all conditions. After the illustrations were explained, participants were requested to answer a set of instruction check questions to confirm that they understood the meaning of all elements of the illustrations (see Supplementary Materials). Proceeding to the next part of the experiment was only possible after all check questions had been answered correctly.

The next phase was an observational learning task with 40 trials (see Table 4 for an overview of trials). In both strength conditions

Table 3
Anova results for all measures of Experiment 1. We report 90% CIs for η_p^2 (see Steiger, 2004).

Query	Factor	*F (df)*	*p*	η_p^2 (90% CI)
permissibility	knowledge	194,96 (1,702)	<.001	0.22 (0.17; 0,26)
	structure	16,63 (1,702)	<.001	0.02 (0.01; 0,04)
	knowledge × structure	14,01 (1,702)	<.001	0.02 (0.01; 0,04)
responsibility	knowledge	237.60 (1,702)	<.001	0.25 (0.21; 0.30)
	structure	18.08 (1,702)	<.001	0.03 (0.01; 0.05)
	knowledge × structure	3.11 (1,702)	0.08	0 (NA;0.02)
foreseeability	knowledge	590.96 (1,702)	<.001	0.46 (0.41; 0.49)
	structure	57.96 (1,702)	<.001	0.08 (0.05; 0.11)
	knowledge × structure	36.89 (1,702)	<.001	0.05 (0.03; 0.08)
predictive	knowledge	0.68 (1,702)	0.41	0 (NA;0.01)
	structure	202.56 (1,702)	<.001	0.22 (0.18; 0.27)
	knowledge × structure	2.49 (1,702)	0.1	(NA; 0.01)

participants saw 20 cases in which the cause (the action) was present and 20 cases in which it was absent. In the "high strength" conditions the effect was present in 16 out of the 20 cases in which the cause was present (p(outcome|action) = 0.80). In the "low strength" conditions the effect was only present in 4 out of the 16 cases in which the cause was present (p(outcome|action) =.20). The effect was never present without the cause in either condition (p(outcome|no action) = 0). The order of trials was randomized, and each trial was visible on screen for 4 seconds, followed by a white mask lasting 0.5 seconds. The animation was created in Adobe Animate 2015 and lasted around 02:30 minutes in total (see Supplementary Materials for an example video). Once the learning phase was over, the experiment proceeded exactly as in the previous studies. We presented participants with an agent who was about to perform the action in question, and asked for judgments of moral permissibility and moral responsibility. On the final page, participants were asked to provide a predictive causal judgment; they were asked to estimate p(outcome|no action), to rate the agent's level of outcome foreseeability, and to indicate their confidence in the claim that the action actually caused the harmful outcome in this particular case (singular causation). An example question for p(outcome|no action) is: "How likely is it for a person to get Marasia illness if they have not been exposed to Proskine in any way?" Ratings were provided on a slider ranging from 0 to 100%. The singular causation question asked for instance: "How confident are you that Mary's action of producing and storing Proskine was the cause of Andrew's Marasia illness?" Here, ratings were provided on a 10-point scale ranging from "not at all" to "completely". All other questions and scales were identical to the ones in Experiment 1. Since each participant saw only one cover story in this experiment, no information was repeated on the last page, it only contained the final questions.

4.2. Results and discussion

Fig. 6 provides an overview of the results and Table 5 shows the

Table 4
Learning trials in Experiment 2 per strength condition (p(outcome|action) =.80 vs. p(outcome|action) =.20).

Strength	Action	Harm	Observations
.80	yes	yes	16
	yes	no	4
	no	yes	0
	no	no	20
.20	yes	yes	4
	yes	no	16
	no	yes	0
	no	no	20

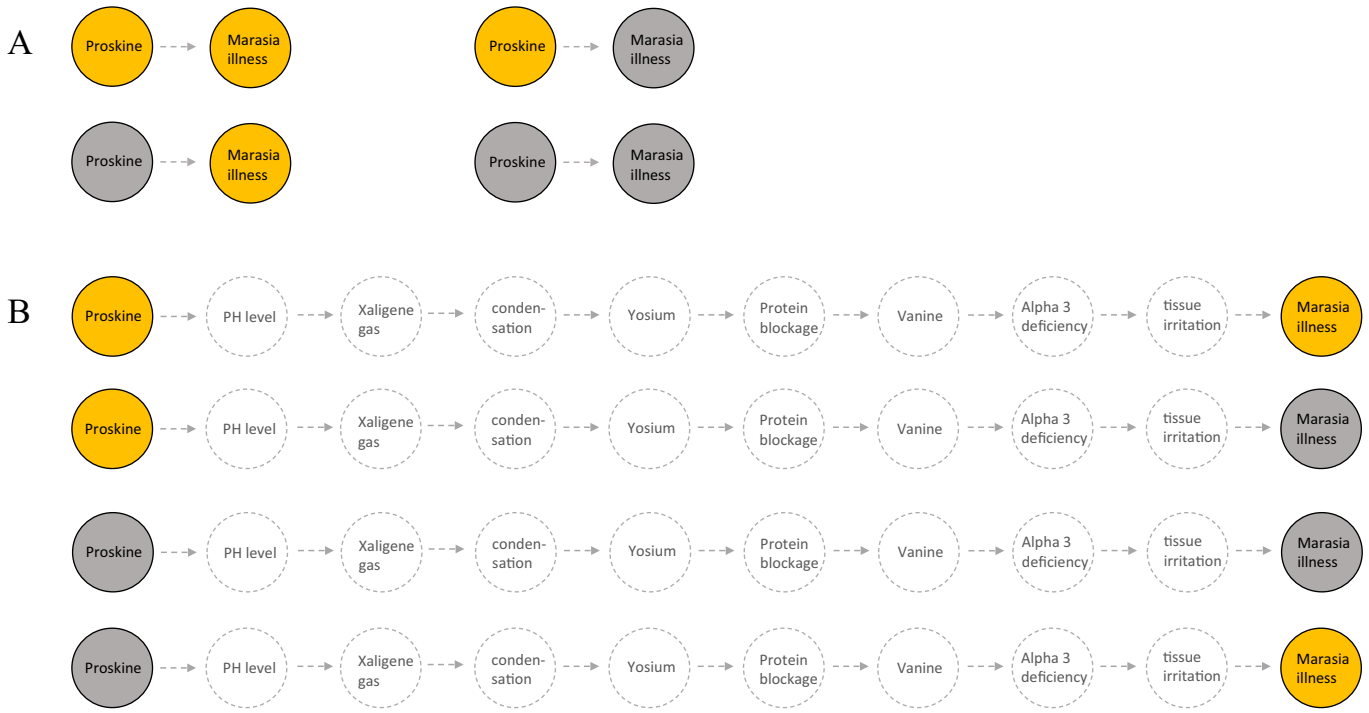


Fig. 5. Example illustrations for direct relations (A) and chains (B) in Experiment 2. Yellow nodes indicated that actions or outcomes were present, grey nodes indicated that they were absent. Dashed circles meant that the status of a variable was unknown.

descriptive statistics for all conditions. Table 6 lists the results of the planned *t*-tests for the effects of causal structure on moral judgments, and, for comparison, also the results of *t*-tests for the effects of causal strength. As expected under the probabilistic model, the causal structure connecting action and harm ceased to affect prospective judgments of moral permissibility in this experiment. Permissibility was low when causal strength was high, no matter whether the causal structure connecting action and harm was a chain or a direct relation. Permissibility was higher when causal strength was low, again independent of causal structure.

Contrary to the predictions of the probabilistic model, a small effect of causal structure persisted for judgments of moral responsibility. Here, agents were still seen as somewhat less responsible in the chain conditions compared to the direct relation conditions[2], even though equal contingencies were presented for both structures. However, this effect was very small. The causal strength manipulation, on the other hand, had clear and distinct effects on judgments of moral permissibility and of moral responsibility in the expected directions (see Table 6), as predicted by the probabilistic model. Participants rated actions as more permissible when they had observed a lower contingency between action and harm, and agents as more responsible when participants had observed a higher contingency.

Table 7 shows the results of exploratory ANOVAs for all remaining measures. As predicted by the probabilistic model, participants considered harm to be less foreseeable by agents when they had learned about a weaker association between action and outcome, but also when action and outcome were connected via a chain rather than directly. The remaining effect of causal structure is not predicted by the probabilistic model. It is consistent with the indirectness model, but there is also an alternative explanation. Looking at participants' predictive causal judgments, we find that despite observing identical contingencies between actions and harmful outcomes, participants still considered the causal relations to be slightly, but significantly weaker in chains than in direct relations (see Fig. 6D, Table 7). This difference might explain the remaining effects of causal structure on foreseeability attributions, and, thereby, on judgments of moral responsibility. When relations are perceived as weaker, the probabilistic model predicts lower outcome foreseeability and a more positive moral evaluation.

Thus, the results are overall consistent with the probabilistic model, although the indirectness model cannot be entirely ruled out yet (based on the effect of causal structure on judgments of moral responsibility). Nevertheless, the evidence for the probabilistic model is clearly stronger than for the indirectness model at this point. The remaining effect of causal structure on moral responsibility ratings was very small ($d = 0.15$ [0.00;0.30]) and only came out significant in some analyses. Plus, this effect lies outside the meta-analytic confidence interval that we determined for this measure based on three preceding studies, in which causal strength was not independently manipulated ($d = 0.29$ [0.21;0.36]). Thus, keeping causal strength constant substantially reduced the effect (and eliminated it entirely for judgments of moral permissibility). Variations in causal strength, on the other hand, had clear and distinct effects on foreseeability and moral judgments in the expected directions.

Why did participants still perceive chains to be weaker than direct relations, even though the objective contingencies were identical? We suspect that learning about a causal chain creates a prior belief about lower causal strength (see also Stephan et al., 2021). Our forty learning trials were apparently only able to partially overwrite this prior. If this explanation is correct, presenting a larger number of observations should lead to the impression of equally strong relations.

Confidence in the fact that actions actually caused the harmful outcomes in the described situations (singular causation) was generally high (see Fig. 6, Table 5), and increased with stronger causal relationships. It was also higher in direct relations than in chains. Participants' estimates of effects occurring in the *absence* of their instructed causes were generally low and not affected by any manipulation. However, they were significantly higher than zero ($M = 10.43$, $SD = 17.10$, $t_{710} = 16.27$, $p<.001$). Since participants thus assumed that alternative causes of our fictitious harmful outcomes existed, the fact that their singular causation ratings increased with the perceived strength of causal relations is in line with normative computational models (Stephan & Waldmann, 2018). Confidence in singular causation correlated with participants' judgments of moral responsibility ($r = 0.55$ [0.50; 0.60], $t_{709} = 17.61$, $p<.001$), as we hypothesized.[3] For further clarification, we predicted responsibility judgments from judgments about singular causation, foreseeability attributions, causal structure, and causal strength in a linear regression. Only singular causation ($\beta = 0.39$, $t_{706} = 10.03$, $p < .001$) and foreseeability ($\beta = 0.34$, $t_{706} = 8.24$, $p < .001$) emerged as significant predictors of moral responsibility judgments in this analysis. Thus, both factors explain unique variance in responsibility judgments. Moreover, causal strength and causal structure cease to predict responsibility when the influence of singular causation and foreseeability is accounted for, providing further evidence for the mediating roles of these latter factors. In the Supplementary Materials, we provide additional mediation analyses and a figure showing participants' response trajectories across the different measures in each condition.

In sum, we set out to test whether causal structure would cease to affect moral judgments when causal strength is kept constant. This is predicted by the probabilistic model, but not by the indirectness model. While we still observed a very small effect of causal structure on one of the moral judgments (moral responsibility) in this experiment, we view the data overall as more in line with the probabilistic model.

5. Experiment 3: Improving learning and exploring foreseeability

This experiment had two aims: First, we substantially increased the number of observations in the causal strength learning phase to facilitate the learning of different causal relations as being equally strong. If participants perceive chains and direct relations as equally strong, the probabilistic model predicts that causal structure should cease to affect moral judgments. Higher strength alone should lead to more severe moral judgments, and lower strength to more lenient judgments, independent of structure. The indirectness model, on the other hand, predicts that chains should still be evaluated more positively than direct relations, even when both are equally strong.

Second, we wanted to explore what exactly agents need to be aware of for their actions to become impermissible, or for them to be seen as morally responsible for harmful outcomes. Given that causal structure and causal strength are separate dimensions of causal models it is possible for agents to be aware *that* a certain causal relation exists between their action and some harmful outcome (i.e., knowledge about structure) but not how strong (or weak) the relation is. Likewise, it is

[2] In line with our *a priori* power analysis (and for a stricter test of our hypothesized null effects), we report the results of one-sided *t*-tests for the effects of causal structure on moral judgments. However, it should be noted that the effect of causal structure on judgments of moral responsibility was *not* significant in a 2 × 2 ANOVA of moral responsibility ratings (see Table 7). Thus, there is only very weak inconsistent evidence for an effect of structure on responsibility judgments independent of strength in this experiment. For judgments of moral permissibility, there was no difference in patterns of significance between ANOVA and *t*-test.

[3] Although there were significant correlations between almost all measures (see Table 11 in the Appendix), the only other measure that was associated with singular causation to a comparable extent as moral responsibility was foreseeability ($r = 0.58$ [0.53; 0.63], $t_{709} = 19.15$, $p<.001$). Fully clarifying the role of singular causation judgments for moral responsibility will require manipulating singular causation independently in future research.

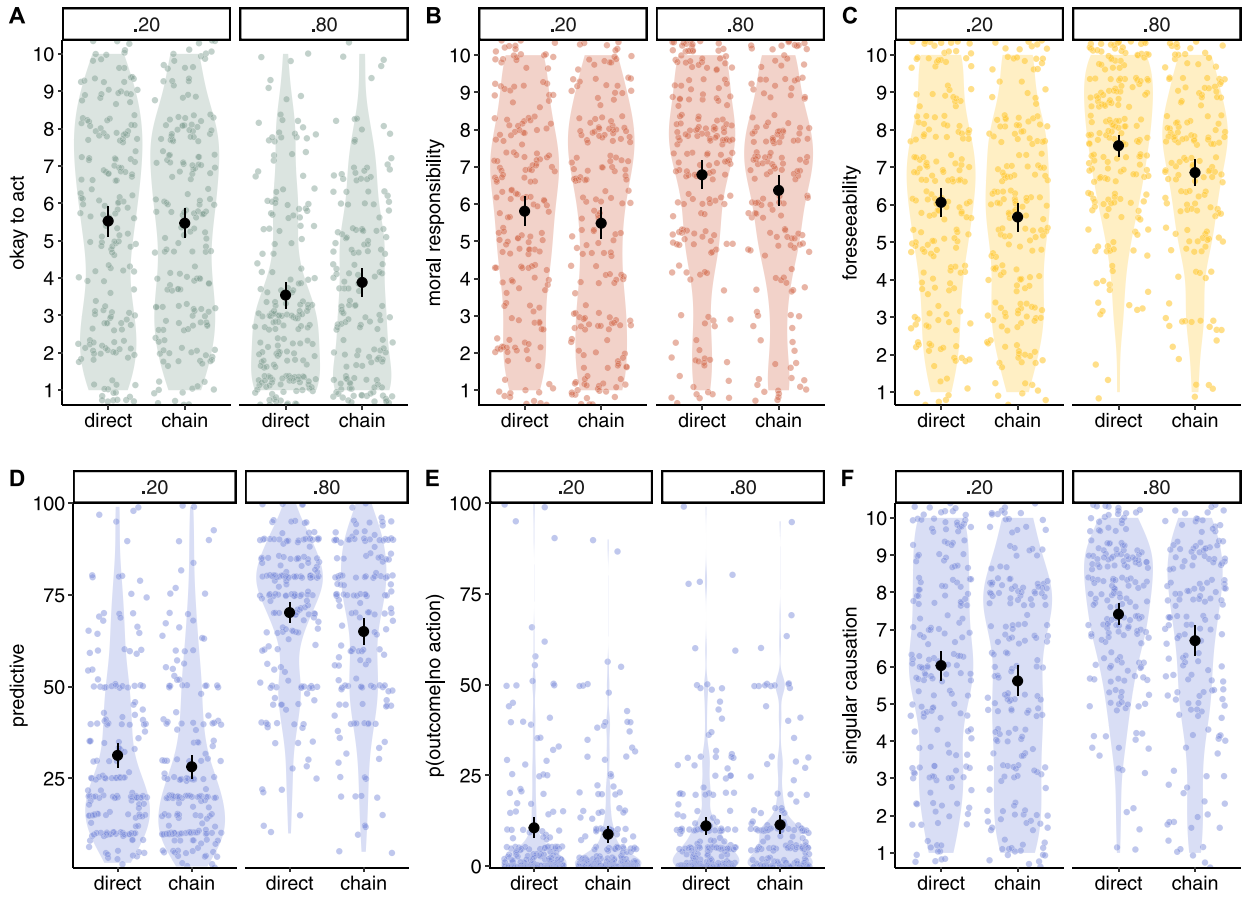


Fig. 6. Means and 95% confidence intervals for ratings of moral permissibility (A), moral responsibility (B), attributions of outcome foreseeability (C), predictive causal judgments (D), *p*(outcome|no action) (E), and singular causation judgments (F) in Experiment 2.

Table 5
Means and standard deviations per condition in Experiment 2.

Relation	Strength	Query	*M*	*SD*	*n*
direct	.20	permissibility	5.53	2.79	184
		responsibility	5.81	2.73	
		foreseeability	6.07	2.62	
		predictive	31.28	22.69	
		p(outcome|no action)	10.55	19.36	
		singular causation	6.03	2.75	
	.80	permissibility	3.54	2.45	189
		responsibility	6.79	2.64	
		foreseeability	7.58	2.03	
		predictive	70.19	19.53	
		p(outcome|no action)	11.09	16.79	
		singular causation	7.41	1.98	
chain	.20	permissibility	5.47	2.67	181
		responsibility	5.49	2.91	
		foreseeability	5.67	2.58	
		predictive	28.17	22.26	
		p(outcome|no action)	8.78	15.39	
		singular causation	5.62	2.79	
	.80	permissibility	3.89	2.42	157
		responsibility	6.37	2.57	
		foreseeability	6.86	2.26	
		predictive	65.03	23.05	
		p(outcome|no action)	11.39	16.50	
		singular causation	6.70	2.56	

possible to be aware of a probabilistic dependence between action and outcome (i.e., knowledge about strength) without knowing which causal structure underlies the association. Or the agent could be simultaneously

Table 6
Effects of causal structure (chain vs. direct) and causal strength (high vs. low) on moral judgments in Experiment 2. *P*-values are one-tailed and not adjusted for multiple comparisons.

	Query	*t (df)*	*p*	*d (95% CI)*
causal structure	permissibility	−1.03 (707.17)	0.15	−0.08 (−0.22; 0.07)
	responsibility	1.97 (698.71)	0.02	0.15 (0.0; 0.30)
causal strength	permissibility	−9.29 (706.61)	<.001	−0.70 (−0.85; −0.54)
	responsibility	4.65 (708.6)	<.001	0.35 (0.20; 0.50)

aware of both properties of the causal model. While Experiment 1 already established that attributions of outcome foreseeability mediate the effects of causal model representations on moral judgments, the experiment could not differentiate between the influences of structure knowledge and strength knowledge on foreseeability, and thereby on moral judgments.

5.1. Methods

5.1.1. Design and participants

We employed a 2 (structure: *direct* vs. *chain*, within-subject) x 2 (strength: *p*(outcome|action) =.80 vs. *p*(outcome|action) =.20, between-subjects) x 4 (knowledge: *full knowledge* vs. *no knowledge* vs. *structure knowledge* vs. *strength knowledge*, between-subjects) design. In all conditions *p*(outcome|no action) was zero. The cover stories were randomly selected for each participant from a set of four stories (*apples* vs. *chemical*

Table 7
ANOVA results per measure in Experiment 2. *P*-values are not adjusted for multiple comparisons.

Query	Factor	*F (df)*	*p*	η_p^2 (90% CI)
permissibility	structure	0.48 (1,707)	0.49	0 (NA;0.01)
	strength	85.11 (1,707)	<.001	0.11 (0.07; 0.14)
	strength × structure	1.04 (1,707)	0.31	0 (NA;0.01)
responsibility	structure	3.27 (1,707)	0.07	0 (NA;0.02)
	strength	20.87 (1,707)	<.001	0.03 (0.01; 0.05)
	strength × structure	0.05 (1,707)	0.82	0 (NA;0.0)
foreseeability	structure	9.36 (1,707)	0.002	0.01 (0; 0.03)
	strength	57.29 (1,707)	<.001	0.07 (0.05; 0.11)
	strength × structure	0.82 (1,707)	0.36	0 (NA;0.01)
predictive causal judgment	structure	6.23 (1,707)	0.013	0.01 (0; 0.02)
	strength	533.78 (1,707)	<.001	0.43 (0.39; 0.47)
	strength × structure	0.39 (1,707)	0.53	0 (NA;0.01)
singular causation	structure	8.59 (1,707)	0.003	0.01 (0; 0.03)
	strength	42.33 (1,707)	<.001	0.06 (0.03; 0.09)
	strength × structure	0.61 (1,707)	0.43	0 (NA;0.01)
p(outcome | no action)	structure	0.36 (1,707)	0.55	0 (NA;0.01)
	strength	1.39 (1,707)	0.24	0 (NA;0.01)
	strength × structure	0.65 (1,707)	0.42	0 (NA;0.01)

vs. *computer* vs. *varnish*) with the constraint that the *direct* and *chain* conditions would be presented in different cover stories for each participant. All measures were identical to the ones in Experiment 2. We decided to collect at least 200 valid responses for each level of the knowledge manipulation (and thus at least 800 valid responses in total). With this sample size we were able to detect an effect of the structure manipulation (*direct* vs. *chain*) on moral judgments of $d = 0.18$ with 80% power in a one-tailed *t*-test in each knowledge condition (Faul et al., 2007). The same test yields a power of 97% for an effect of $d = 0.25$, and we were thus able to detect such an effect in *all* of the eight one-tailed t-tests (four knowledge conditions, two moral judgment measures) with a power of $0.97^8 = 78\%$. Our best estimate for an effect of causal structure in the face of equal causal strength comes from Experiment 2 and is $d = 0.15$ with a 95% confidence interval ranging from 0 to 0.30. We expected causal structure to be the smallest effect of interest in the design, and focused our power analysis on the moral judgment measures. 1165 participants completed the experiment. Eight participants were excluded from the analyses for taking the survey twice (only their first participation was retained), 57 were excluded for either using a smartphone against instructions or for failing the simple attention check. Of the remaining 1100 participants, we excluded those who failed to give correct answers to a manipulation check question about the agents' knowledge in at least one of the scenarios (see next section for the questions). The accuracy of the answers to these test questions ranged from 76% in the *structure knowledge* condition to 96% in the *no knowledge* condition. The final sample consisted of 854 participants ($M_{age} = 32.21$, $SD_{age} = 11.1$, 422 women, 417 men, 14 non-binary, 1 no answer).

5.1.2. Materials and procedure

As in the previous experiments, this experiment also began with the instruction of a generic causal relation between an action and a harmful outcome. This relation could either be direct or a long chain (9 links). The same cover stories as in the previous experiments were used, plus one new story (*varnish*, see Supplementary Materials). In the present experiment, we removed the "sometimes causes" labels from the illustrations. In the verbal description of the causal relations, we also removed this expression and said instead "there is a causal relation between [cause] and [effect]". We decided to do this because we wanted to find out whether an effect of causal structure on moral judgments would persist even when the probabilistic nature of the causal relations and the similarity of the strengths of individual links in the direct and chain relations were not highlighted. In our previous experiments, it was reasonable for participants to form an initial belief that the overall strength in the chain conditions is lower than in the conditions that instructed direct relations.

After the instruction of the generic causal relation, we informed participants that they would now be presented with the results of a second study in which scientists had reviewed health records. Other than in the previous experiment, the data were presented in summary format (see Fig. 7). The presence of a blue circle meant that the target action was present (e.g., someone in a lab had produced and stored the potentially harmful chemical), and a red person icon meant that the harmful outcome had occurred (e.g., the person had developed the disease). The summary format allowed us to present participants with three times as many observations as in the previous experiment (120 instead of 40), while keeping the experiment at a reasonable length. Before participants were able to see the data, they had to correctly answer two instruction check questions to confirm that they understood the meaning of the symbols (see Supplementary Materials).

Once the learning phase was completed, we presented participants with the case of an agent who planned to perform the target action, as in the previous experiments. We manipulated the agent's knowledge in four levels. The agent either knew about everything the participant had just learned (*full knowledge*), the agent had none of this information (*no knowledge*), the agent only knew the underlying causal structure (*structure knowledge*), or the agent was only aware of the study in which the strength of the statistical association between action and outcome was measured (*strength knowledge*).

Here is an example of the knowledge manipulation in the *strength knowledge* condition: "She knows about the strength of the association between producing Proskine and the occurrence of Marasia illness, that is, how often Marasia illness occurs when Proskine is produced and stored versus not produced and stored (the result of the study of health records). However, she does not know that the relation between producing and storing Proskine and the occurrence of Marasia illness is causal (the causal chain described earlier). To sum up, Mary knows how often Marasia illness occurs when Proskine is present in a lab, but she does not know what the causal relation behind this association is." We provided illustrations to remind participants of both the objective causal structure and strength as well as the agents' knowledge (see Fig. 8 for an example). As in the previous experiments, we requested judgments of moral permissibility before the act was initiated as well as judgments of moral responsibility after the harmful outcome had occurred. Once participants had completed the moral judgment tasks for a scenario, the main information about the case was presented again (summarised on a single page), and participants were asked to provide predictive causal judgments of the probability of the outcome in the presence and absence of the action. Moreover, they were asked to assess agents' foreseeability, and for singular causation judgments. We also asked participants about the agents' knowledge at the time of acting as a manipulation check. An

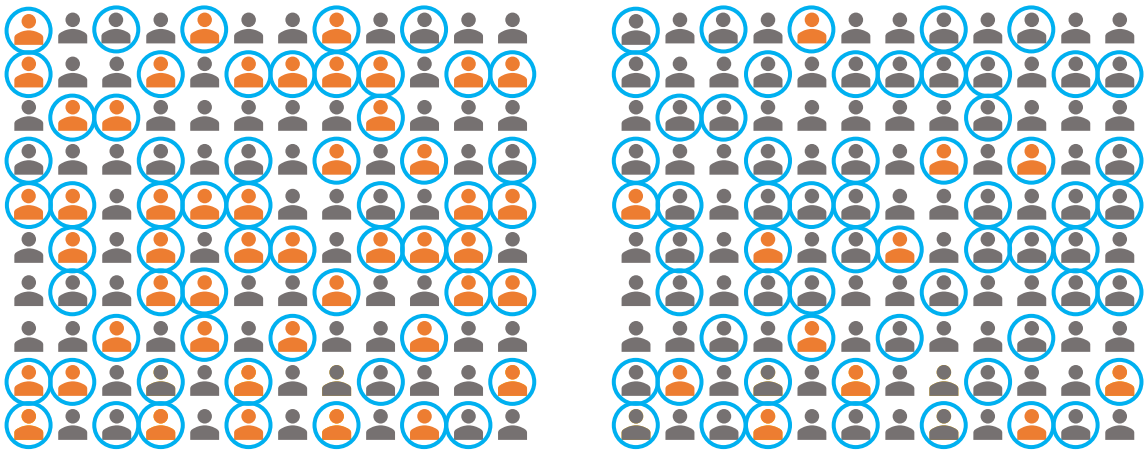


Fig. 7. Learning data used in Experiment 3. Participants were instructed that the presence of a blue circle meant that the target cause (an action) was present, and a red person icon meant that the target effect (a harmful outcome) had occurred. The left panel shows a contingency of .80, the right panel shows a contingency of.20. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

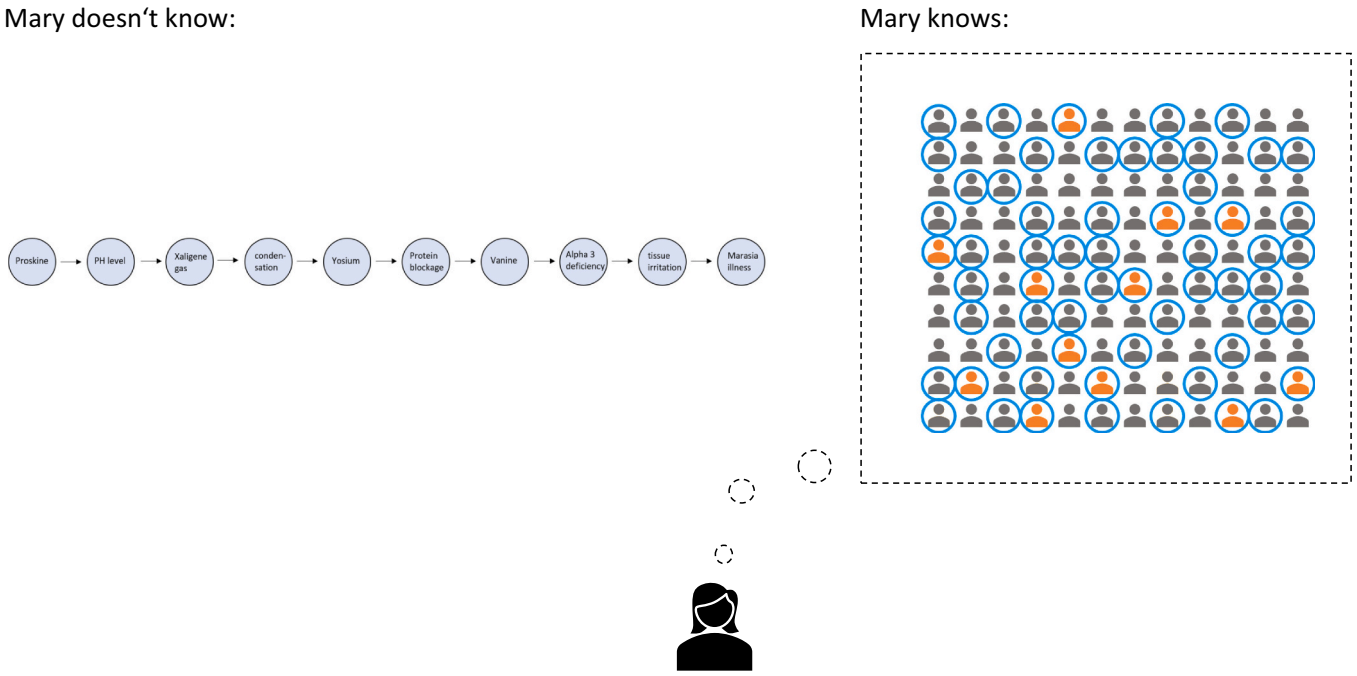


Fig. 8. Example illustration of an agent's knowledge in Experiment 3. Here the agent is aware of the strength of the association between an action and a harmful outcome (right), but unaware of the causal relation behind the association (left).

example for this question is: “When producing and storing Proskine, Mary knew about…”, with the four options “the lab study (that is, the fact that Proskine can cause Marasia illness via a chain of other intermediate events)”, “the study of health records (that is, how often Marasia illness occurs when Proskine is produced and stored)”, “both studies”, or “neither study”.

5.2. Results and discussion

See Fig. 9 for the results for permissibility, responsibility, and foreseeability judgments. Table 13 in the Appendix shows the descriptive statistics for all measures and conditions. Table 8 contains the results of the planned *t*-tests for moral judgments. We fit linear mixed models to the data and determined which combination of predictors best described participants’ responses using model comparisons. For the final models, see Table 9.

As we had hoped to achieve with the larger number of observations in the learning task, participants now perceived the overall causal relations between actions and outcomes to be equally strong in chains and in direct relations. As predicted by the probabilistic model (but not the indirectness model), neither judgments of moral permissibility nor judgments of moral responsibility were robustly affected by causal structure under these circumstances (see Fig. 9, Table 9, Table 8). Judgments of moral permissibility were only affected by causal strength (with lower causal strength leading to higher permissibility ratings, as the probabilistic model predicts), and by the agents’ knowledge. For judgments of moral responsibility, ratings increased with higher causal strength, and were also affected by knowledge. There was a small effect of causal structure on responsibility judgments in a one-sided t-test (see Table 8). However, adding structure to the model did not improve the fit when the other factors were accounted for (see Table 9 and Supplementary Materials). Thus, overall, only causal strength and agents’

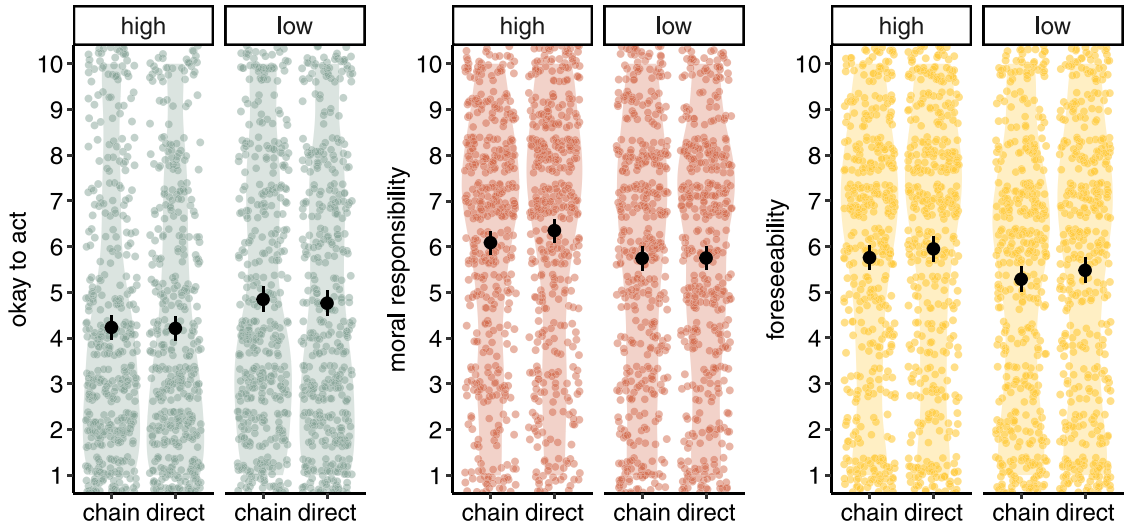


Fig. 9. Means and 95% confidence intervals for ratings of moral permissibility (left), moral responsibility (middle), and foreseeability (right) in Experiment 3. Data are collapsed across knowledge conditions, see Appendix for graphs per knowledge condition.

Table 8
Effects of causal structure (chain vs. direct) and causal strength (high vs. low) on moral judgments in Experiment 3, collapsed across knowledge conditions (see Appendix for tests per knowledge condition). *P*-values are one-tailed.

	Query	*t (df)*	*p*	*d (0.95 CI)*
causal structure	permissibility	0.63 (853)	.26	0.02 (−0.04, 0.08)
	responsibility	−1.75 (853)	0.04	−0.05 (−0.1, 0.01)
causal strength	permissibility	−4.23 (1703.52)	<.001	−0.2 (−0.3,−0.11)
	responsibility	3.54 (1704.91)	<.001	0.17 (0.08, 0.27)

knowledge proved to exert substantial and robust influences on moral judgments.

As for the knowledge manipulation, we found that participants only considered complete ignorance about causal relations to be exculpating. When agents knew about either the structure by which action and harm were related or about the strength of the relation, they were judged to be roughly equally morally responsible and their actions roughly equally impermissible as agents with complete knowledge about the causal relations (see Table 9, for additional graphs per knowledge condition see Fig. 10 – Fig. 13 in the Appendix). The only exception was that agents with only strength knowledge were seen as slightly less morally responsible than agents with full knowledge, $d = 0.16$ [0.03,0.30].

Adding a two-way interaction term between knowledge and strength did not improve the model fit for either moral judgment measure, possibly due to a lack of power for the reliable detection of interaction effects. Nevertheless, descriptively, the largest effects of causal strength on moral judgments were observed when agents were fully aware of causal relations, and no effects of strength could be detected in the no knowledge conditions. Results for partial knowledge were in between (see Fig. 10 – Fig. 13 and Table 14, all in the Appendix). These observations again support the mediating role of foreseeability. Unexpectedly, strength still affected moral judgments when agents were only aware of the causal structure. It may have been difficult for participants to screen off strength information in their moral judgments after the lengthy learning task, or they may have had difficulties to differentiate between states of partial knowledge.

When explicitly asked about outcome foreseeability, participants indicated that agents were less able to foresee harm when causal strength was low rather than high, and also when the relation was direct rather than a chain (see Fig. 9, Table 9). The effect of causal structure is predicted by the indirectness model, but not by the probabilistic model. However, the influence of strength ($d = 0.16$ [0.07,0.26]) was stronger than the influence of structure ($d = 0.07$ [0.02,0.11]). The outcomes were judged as unforeseeable in the no knowledge conditions, and as highly foreseeable in the full knowledge conditions. Partial knowledge (only causal structure, only causal strength) also somewhat reduced perceived foreseeability compared to full knowledge. One might wonder how these results for foreseeability square with the fact that moral judgments were not affected by causal structure or partial knowledge relative to full knowledge, given that we claim that foreseeability attributions mediate effects of the causal model representations on moral judgments. However, we have seen in all of our experiments (here as well as in Engelmann & Waldmann, 2021) that the direct effects of our causal model manipulations on foreseeability are larger than their indirect effects on moral judgments. A relatively large difference in outcome foreseeability seems to be required for moral judgments to be affected. The small effects of structure and foreseeability that we observed here may not have been strong enough to push through to the moral measures.

Singular causation judgments, just as predictive causal judgments, were only affected by causal strength in this experiment, no detectable effects of causal structure was observed (see Table 9). As in Experiment 2, confidence in singular causation correlated with moral responsibility ratings ($r = .41$ [.37,.45], $t_{1706} = 18.82$, $p < .001$). As before, we analyzed whether responsibility judgments were affected by judgments about singular causation, foreseeability attributions, causal structure, and causal strength in a linear regression. In this analysis, only singular causation ($\beta = 0.27$, $t_{1703} = 12.04$, $p < .001$) and foreseeability ($\beta = 0.55$, $t_{1703} = 31.28$, $p < .001$) were significant predictors of moral responsibility judgments. Table 12 in the Appendix shows the correlations between all measures. We provide the results of additional mediation analyses and figures showing participants' response trajectories across the different measures in the Supplementary Materials.

In sum, the results that we obtained here supported the probabilistic model more strongly than the indirectness model. Both moral judgment measures were clearly and consistently affected by the strength of the causal relations between action and harmful outcomes as well as by the agents' knowledge. The dominant pattern was that actions became more permissible and agents less responsible when harm was less likely to be caused by the action. Partial knowledge about causal relations (only strength, only structure) sufficed for moral condemnation of agents and actions. In the conditions in which agents were completely ignorant of the causal relations, no significant effects of strength on judgments of

Table 9
Final regression models for all measures in Exp. 3.

Permissibility				
Random: participant ID	Intercept	Residual		
SD	1.68	1.73		
Fixed: strength, knowledge	Estimate	*SE*	*t (df)*	*p*
low strength	0.58	0.14	4.11 (849)	<.001
no knowledge	3.64	0.20	18.04 (849)	<.001
strength knowledge	0.17	0.20	0.81 (849)	0.42
structure knowledge	0.35	0.20	1.76 (849)	0.08
responsibility				
Random: participant ID	Intercept	Residual		
SD	1.77	1.61		
Fixed: strength, knowledge	Estimate	*SE*	*t (df)*	*p*
low strength	−0.48	0.14	−3.36 (849)	<.001
no knowledge	−3.31	0.20	−16.22 (849)	<.001
strength knowledge	−0.42	0.21	−2.03 (849)	.04
structure knowledge	−0.24	0.20	−1.20 (849)	.23
foreseeability				
Random: participant ID	Intercept	Residual		
SD	1.43	1.41		
Fixed: structure, strength, knowledge	Estimate	*SE*	*t (df)*	*p*
direct	0.19	0.07	2.84 (853	.005
low strength	−0.50	0.12	−4.22 (849)	<.001
no knowledge	−5.51	0.17	−32.52 (849)	<.001
strength knowledge	−1.17	0.17	−6.83 (849)	<.001
structure knowledge	−1.27	0.17	−7.51 (849)	<.001
predictive				
Random: participant ID	Intercept	Residual		
SD	16.38	13.0		
Fixed: strength	Estimate	*SE*	*t (df)*	*p*
low strength	−39.52	1.29	−30.72 (852)	<.001
singular causation				
Random: participant ID	Intercept	Residual		
SD	1.78	1.47		
Fixed: strength	Estimate	*SE*	*t (df)*	*p*
low strength	−0.93	0.14	−6.59 (852)	<.001
p(outcome|no action)				
Random: participant ID	Intercept	Residual		
SD	14.13	12.86		
Fixed: none				

moral permissibility or moral responsibility could be detected. This is in line with the hypothesis that attributions of outcome foreseeability mediate the effect of the causal features on moral judgments (which was also confirmed in Experiment 1, see also mediation analyses in the Supplementary Materials).

The only finding that is more in line with the indirectness model than with the probabilistic model was the observation that causal structure to a small extent affected attributions of outcome foreseeability when *p* (outcome|action) was not only objectively kept constant, but also perceived as constant by participants. Thus, even though participants estimated outcomes as equally likely in the chain and direct relation conditions, they still considered agents to be somewhat less able to foresee the harmful outcome when action and outcome were related via a chain rather than directly. This finding is not predicted by the probabilistic model. The effect was very small, but nevertheless it seems that indirectness can have an effect on attributions of outcome foreseeability that is independent of causal strength. The stronger and more consistently observed effect was, however, that foreseeability is dependent on causal strength, leading to downstream effects on moral judgments.

6. General discussion

Making moral judgments requires causal analysis. We judge people's actions by their anticipated and actual consequences, and we hold them responsible for the good or bad outcomes that they caused. Causal analysis focuses on the causal relations that exist in the world and the strengths of these relations. These two dimensions of causal models, causal structure and causal strength, are the foundation of a number of inferences that are crucial for moral reasoning (see, e.g. Cushman, 2008, 2013; Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021; Sloman, Fernbach, & Ewing, 2009; Waldmann, Wiegmann, & Nagel, 2017). Even though the general claim that causation lies at the heart of moral reasoning is largely undisputed, the specific interplay of causal structure and causal strength in the formation of moral judgments has not yet been investigated in a systematic way. In the present research we aimed to lay the groundwork for such an investigation, starting with the case of direct causal relations in comparison to causal chains.

We focused on chains whose initial element is an intentional action and whose final outcome is some form of harm to another person. The intermediate variables were (bio-)chemical events. In our first experiment we showed that people evaluate agents and actions more positively when agents cause harm via a chain of intermediate events, rather than directly. Specifically, participants in the chain condition considered actions to be more morally permissible *before* they were executed, and they considered agents to be less morally responsible *after* harmful outcomes had actually occurred. Experiment 1 also demonstrated that causal representations did not directly alter moral judgments, but that their effects were mediated by inferences about the agents' mental states. Causal chains tend to lead to the impression that outcomes are less foreseeable to agents, which in turn leads to a more positive moral evaluation.

Experiments 2 and 3 were dedicated to uncovering the cognitive mechanisms that underlie this effect. We contrasted two hypotheses: According to the probabilistic model outcomes are less foreseeable in chains because they are perceived as less likely to occur. On this view, effects of causal structure (chains vs. direct relations) on moral judgments are ultimately dependent on inferences about causal strength. Alternatively, according to what we called the indirectness model, causal structure itself might be driving the effect. On this view, outcomes are perceived as less foreseeable in chains merely because the relation is indirect. The two models make diverging predictions for situations in which the overall causal relationship between action and harm is equally strong in chains as in direct relations. When the overall causal strength is the same, the probabilistic model predicts that there should be no difference between the moral evaluation of a direct and an indirect relation. The indirectness model, on the other hand, predicts that agents and actions should be judged more positively in chains, even when the overall relation between the action and the final outcome is just as strong as in a direct relation.

Varying both causal structure and causal strength in the two experiments we found that the evidence more strongly favoured the probabilistic model. When chains and direct relations were equally strong, no effects of causal structure on judgments of moral permissibility were observed. Although we still observed effects on moral responsibility judgments in some conditions, they were very weak and inconsistent. Causal strength, on the other hand, had a clear effect on both types of moral judgments in both Experiments 2 and 3. Actions were judged as more permissible when harmful consequences were less likely, and agents were seen as less morally responsible when the a priori likelihood of harm was low.

Experiment 3 also shed further light on the kind of foreseeability that affects permissibility and responsibility judgments. It turned out that agents do not need to be aware of all aspects of a causal relation between their action and harm. Knowing that a certain causal relation exists suffices even when its strength is unknown. Likewise, knowing about a

statistical association between action and harm suffices, even when the causal structure that underlies the association is unknown. In our experiments, the learning data that participants saw made it very easy to infer the existence of a causal relation between action and harm. We thus suspect that reasoners expected others who are confronted with the same data (i.e., the agents in our scenarios) to arrive at the same conclusion that they themselves drew, namely that the relationship is actually causal. Future studies should investigate in more detail how reasoners' own judgments about relations affect the inferences that they expect others to draw.

Even though the probabilistic model emerged as the dominant path by which causal structure influences moral judgments, we also found some effects that are only predicted by the indirectness model. In Experiment 3, participants considered outcomes as somewhat less foreseeable in chains than in direct relations, even though they took the probabilistic relation to be equally strong in both cases. Thus, it seems that causal structure can have a direct effect on foreseeability. However, the effect was not large enough to robustly affect moral judgments as well. We repeatedly observed that moral evaluations are only altered when outcomes become substantially, not just slightly less foreseeable to agents (see also Engelmann & Waldmann, 2021).

We studied two types of moral judgments, prospective permissibility and retrospective responsibility judgments. Both types of judgments are affected by causal relations and mental states (foreseeability) but whereas prospective judgments rely on predictive causal judgments, our results indicate that retrospective judgments rely on singular causation judgments. We found that participants' confidence in the claim that actions actually caused harmful outcomes in the case at hand predicted their moral responsibility ratings beyond the predictive power of foreseeability attributions. Thus, it seems that both a sufficient level of outcome foreseeability and confidence in singular causation are required for the attribution of moral responsibility. Fully clarifying how singular causation is affected by strength and structure knowledge, and how it in turn interacts with foreseeability to shape judgments of moral responsibility will require systematically manipulating singular causation in future research (but see Cushman, 2008, for a demonstration of the general relevance of singular causation in moral judgments).

6.1. Beyond causal strength and causal structure

Structure and strength are the two crucial components of causal networks. As we have demonstrated in the present research, their interplay, along with the inferences that they license about agents' mental states, can explain patterns in people's moral reasoning. However, there may be additional factors that might affect moral judgments. Some of them may interact with our beliefs about causal structure and causal strength, while others might have other sources.

For example, domain knowledge will probably affect default assumptions about causal structure and causal strength. Strickland et al. (2017) showed that when people reason about events from the physical domain, they tend to represent them as the products of linear causal chains, whereas psychological events are assumed to be the product of many independent causes (common-effect structures, see Fig. 1). Furthermore, causal links in the physical domain were estimated to be stronger than links in the psychological domain. Thus, domain-specific assumptions seem to shape our causal representations in the absence of clear information about structure and strength. To the extent that they do, we would expect moral judgments to follow suit. For example, if all links were perceived to be completely deterministic in a scenario, chain length should no longer affect moral judgments, as multiplying the links would no longer reduce the probability of the outcome in a chain.

A further important property of causal chains that honor the Markov condition is transitivity. That is, if A causes B, and B causes C, people should usually also agree that A causes C. People made this assumption in our experiments, as wis suggested by their singular causation ratings, for example. These ratings were always above the scale midpoint, indicating that participants took actions to have actually caused the final outcomes. Sometimes, however, people do not assume chains to be transitive. (Johnson & Ahn, 2015) proposed that chains tend to be viewed as intransitive when the first and the last element are not part of the same semantic schema, or "chunk". For instance, participants in their experiments agreed that there was a strong causal link between a person stepping on a dog's tail and the dog growling. They also perceived a strong causal link between the dog growling and a child becoming scared. However, the link between stepping on a dog's tail and a child becoming scared was judged to be weak (and weaker than in other cases in which the individual links were perceived as equally strong). We predict that judgments about foreseeability and moral judgments might also vary with perceived strength in such cases. That is, a child becoming scared might be judged as rather unforeseeable given that someone stepped on a dog's tail, and the agent would probably not be seen as very responsible for it. These results are in line with our predictions, though, since in these cases the relation between initial action and final outcome is viewed as weak. Thus, when chains are seen as intransitive because of semantic chunking, even shorter chains than the ones we used in our experiments might lead to a considerably more positive moral evaluation.

Finally, semantic chunking seems to be closely related to preferences about the granularity at which we represent causal relations. Any causal relationship can, in principle, be construed on many different levels of granularity (i.e., very abstract and high-level vs. down to the level of atoms and their interactions). How exactly reasoners choose the appropriate level of granularity for a given relationship is subject to an extensive debate (see, e.g., Woodward, 2021). One plausible factor is whether mediating events constitute suitable targets of intervention whose manipulation would allow some control over the causal relationship (Woodward, 2005). Based on Johnson & Ahn (2015)'s work, another factor may be whether an intermediate event B is required to explain why A leads to C. When instructing long chains in our experiments, we mostly described fairly low-level biochemical mechanisms as intermediate events, involving fictitious substances and devices. We did this in an effort to minimize any effects of knowledge of or assumptions about the strengths of individual links on participants' judgments. However, low-level biochemical mechanisms are not usually what we *care* about in everyday life, unless we have special reason to be interested in them. Thus, the chains we used could have seemed to participants as the results of "zooming in" on a causal relation in which the only interesting parts are the beginning and the end. If this was the case, effects of causal structure on inferences about overall strength, foreseeability, and moral judgments might be more pronounced when the intermediate events seem more relevant to participants.

6.2. Causation and foreseeability in the law

The extent to which agents could foresee or should have foreseen harm that resulted from their actions is not just of crucial importance for moral judgments, it also informs central legal notions such as negligence or recklessness. An agent is typically considered negligent when they cause harm that they should have reasonably foreseen, or that a reasonable person would have foreseen (even if in the actual situation, the agent did not foresee it). Recklessness requires that agents were aware of a substantial risk of harm at the time of acting, and acted despite this knowledge (see, e.g., Dubber, 2015, *pp*. 42–46). Kneer & Skoczeń (2021) and Kneer & Machery (2019) found that judgments that expressed that agents should have foreseen the harms they caused (i.e., attributions of negligence) explained effects that would otherwise be described as instances of moral luck or as direct effects of outcomes on moral judgments. Nobes & Martin (2021) experimentally manipulated negligence and recklessness in cases of accidental harms by instructing that agents forgot about risks (negligence) or ignored known risks (recklessness). They found that either suffices for a negative moral evaluation, and that participants often infer that agents who caused

harm were negligent when no information to the contrary was provided.

In our scenarios, agents could also be described as negligent or even reckless in the conditions in which they were aware of the harms they might cause. In the conditions in which they were unaware, attributions of negligence are likely blocked because we always pointed out that agents *could* not possibly have been aware of the risk, as the relevant scientific results were not available to them (i.e., a reasonable person could not have foreseen harm). Thus, the results we reported here are consistent with the view that attributions of negligence or recklessness might influence moral judgments about accidental harms.

In legal discussions of negligence, however, it is generally not only taken into account whether someone should have been able to foresee harm, but also how severe the harm is that they should have been able to foresee, and whether the burden of taking reasonable precautions against harm would have been acceptable. In *United States v. Carroll Towing Co.* (1947) a formula is famously suggested according to which an act should *not* count as negligent when the burden of taking adequate precautions would have outweighed the severity of harm, multiplied by its probability. Thus, it is possible for agents to reasonably foresee harm, and still not be deemed negligent. If harm is a side effect, one might furthermore add the probability and utility of the primary goal that an agent was pursuing with their action (see Engelmann & Waldmann, 2022; Waldmann & Wiegmann, 2012). Future studies need to investigate the interplay of these factors in laypeople's judgments. We suspect that people will not condemn just any foreseeable harm, but that their attributions might instead roughly reflect the conditions that are specified in the formula cited above. Anecdotally, we can report that many participants in our studies remarked that they wondered whether agents took adequate precautions to prevent harm.

6.3. Conclusion

We set out to test how causal structure and causal strength affect moral judgments about causal chains. When agents caused harms via a longer chain of intermediate events, rather than directly, participants saw actions as more permissible, and agents as less morally responsible. We demonstrated that this effect mainly arises because reasoners take harm to be less likely, and therefore less foreseeable to agents in chains. Thus, effects of causal structure were predominantly mediated by inferences about causal strength and about agents' mental states. The mere indirectness of a relation can also lower foreseeability, but these differences were mostly not strong enough to change moral evaluations. Future research should investigate the interplay between chain length and semantic chunking or granularity, as well as how foreseeability interacts with the utility of agents' primary goals and the perceived burden of taking adequate precautions against harm.

CRediT authorship contribution statement

Neele Engelmann: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Michael R. Waldmann**: Conceptualization, Writing – review & editing, Supervision.

Acknowledgments

We would like to thank Alex Wiegmann for helpful comments. We also thank Jonathan F. Kominsky, an anonymous reviewer, as well as audiences at the Georgetown Law and Language Lab, and the Chicago/Michigan Psychological and Law Studies Lab. Portions of this work have appeared in the *Proceedings of the Cognitive Science Society*: Engelmann, N., & Waldmann, M. R. (2021). A causal proximity effect in moral judgment. In *Proceedings of the 43th Annual Conference of the Cognitive Science Society* (pp. 2330 - 2336). Austin, TX: Cognitive Science Society.

Appendix A

Table 10
Correlations between all measures of Experiment 1. All $p < .001$ (unadjusted).

	Perm.	Resp.	Foresee	Pred.
perm.		−0.57	−0.53	−0.30
resp.	−0.57		0.64	0.33
foresee	−0.53	0.64		0.33
pred.	−0.30	0.33	0.33	

Table 11
Correlations between all measures of Experiment 2. All $p < .001$ (unadjusted) except between responsibility and p(E|noC) ($p = 0.50$).

	Perm.	Resp.	Foresee	Pred.	Singular	p(E|noC)
perm.		−0.46	−0.44	−0.41	−0.43	0.14
resp.	−0.46		0.53	0.31	0.55	−0.03
foresee	−0.44	0.53		0.38	0.58	−0.13
pred.	−0.41	0.31	0.38		0.40	0.14
singular	−0.43	0.55	0.58	0.40		−0.15
p(E|noC)	0.14	−0.03	−0.13	0.14	−0.15	

Table 12
Correlations between all measures in Experiment 3. All $p < .001$ (unadjusted) except between permissibility and p(E|noC) ($p = .002$), and between foreseeability and p(E|noC) ($p = .002$).

	Perm.	Resp.	Foresee	Pred.	Singular	p(E|nonC)
perm.		−0.60	−0.58	−0.15	−0.24	0.08
resp.	−0.60		0.65	0.16	0.41	−0.11
foresee	−0.58	0.65		0.12	0.32	−0.07
pred.	−0.15	0.16	0.12		0.31	0.08
singular	−0.24	0.41	0.32	0.31		−0.26
p(E|nonC)	0.08	−0.11	−0.07	0.08	−0.26	

Table 13
Descriptive statistics per condition in Exp. 3.

	Chain		Direct		
full knowledge, high strength	***M***	***SD***	***M***	***SD***	***n***
permissibility	3	1.89	3.05	2.14	98
responsibility	7.23	2.15	7.59	2.07	
foreseeability	7.81	1.84	8.17	1.74	
predictive	71.54	21.06	74.98	18.89	
singular	7.45	2.15	7.89	1.95	
p(outcome|no action)	15.72	20.6	12.98	16.61	
full knowledge, low strength	***M***	***SD***	***M***	***SD***	***n***
permissibility	3.89	2.47	3.86	2.53	110
responsibility	6.75	2.64	6.43	2.59	
foreseeability	7.27	2.3	7.27	2.34	
predictive	31.74	19.42	32.41	21.47	
singular	6.64	2.68	6.69	2.65	
p(outcome|no action)	11.91	20.86	11.71	21.78	
no knowledge, high strength	***M***	***SD***	***M***	***SD***	***n***
permissibility	6.98	2.94	6.94	2.78	108
responsibility	3.43	2.41	3.89	2.7	
foreseeability	2.34	2.07	2.26	1.84	
predictive	69.44	22.53	72.75	17.08	
singular	7.38	1.89	7.42	2.15	
p(outcome|no action)	12.48	17.98	12.16	16.44	
no knowledge, low strength	***M***	***SD***	***M***	***SD***	***n***
permissibility	7.41	2.84	7.08	3.01	108
responsibility	3.64	2.7	3.75	2.59	
foreseeability	1.87	1.3	1.99	1.54	
predictive	33.11	26.87	34.96	26.92	
singular	6.3	2.92	6.66	2.73	
p(outcome|no action)	13.31	23.89	13.53	22.85	
strength knowledge, high strength	***M***	***SD***	***M***	***SD***	***n***
permissibility	3.32	1.99	3.42	2.06	113
responsibility	6.88	2.01	6.72	2.33	
foreseeability	6.73	1.98	6.94	2.07	
predictive	72.47	17.84	73.23	16.76	
singular	7.57	1.86	7.53	2.09	
p(outcome|no action)	9.54	17.06	11.57	18.24	
strength knowledge, low strength	***M***	***SD***	***M***	***SD***	***n***
permissibility	4.02	2.3	3.72	2.23	97
responsibility	6.14	2.61	6.55	2.31	
foreseeability	5.88	2.32	6.25	2.22	
predictive	36.51	23.22	35.44	23.32	
singular	6.49	2.59	7.02	2.38	
p(outcome|no action)	11.21	20.12	12.25	22.3	
structure knowledge, high strength	***M***	***SD***	***M***	***SD***	***n***
permissibility	3.56	2.11	3.36	2.19	108
responsibility	6.88	2.21	7.31	2.06	
foreseeability	6.3	2.02	6.59	1.94	
predictive	73.06	18.02	74.47	16.06	
singular	7.94	1.75	7.95	1.71	
p(outcome|no action)	9.82	12.3	11.69	16.06	
structure knowledge, low strength	***M***	***SD***	***M***	***SD***	***n***
permissibility	4.03	2.28	4.3	2.44	112
responsibility	6.46	2.21	6.34	2.42	
foreseeability	6.12	2.09	6.42	2.22	
predictive	32.98	23.25	29.28	18.58	
singular	7.1	2.36	6.78	2.57	
p(outcome|no action)	7.62	18.54	7.71	16.58	

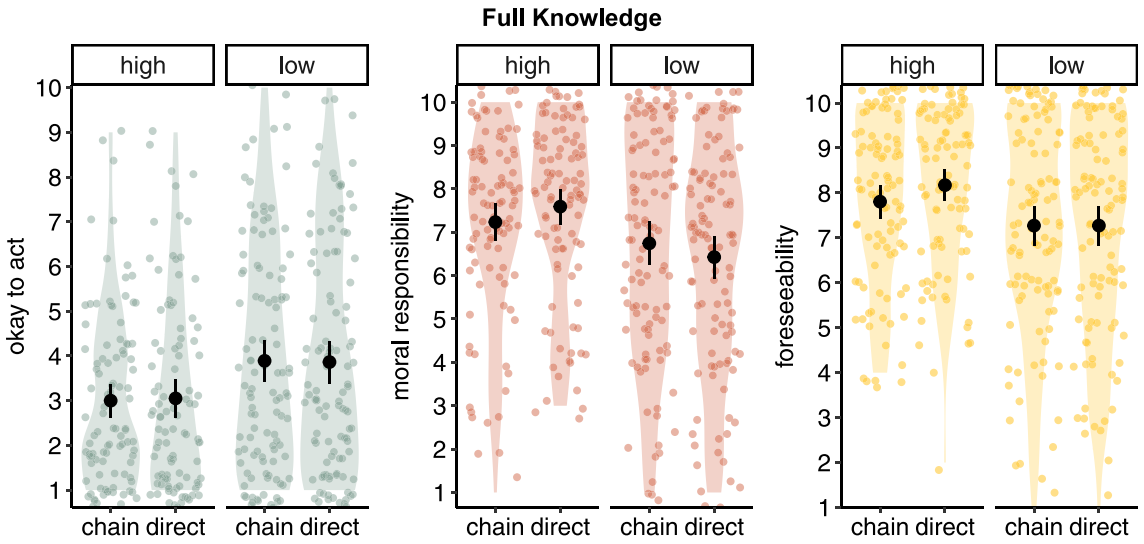


Fig. 10. Means and 95% CIs for ratings of moral permissibility (left), moral responsibility (middle), and foreseeability (right) in the full knowledge conditions of Experiment 3.

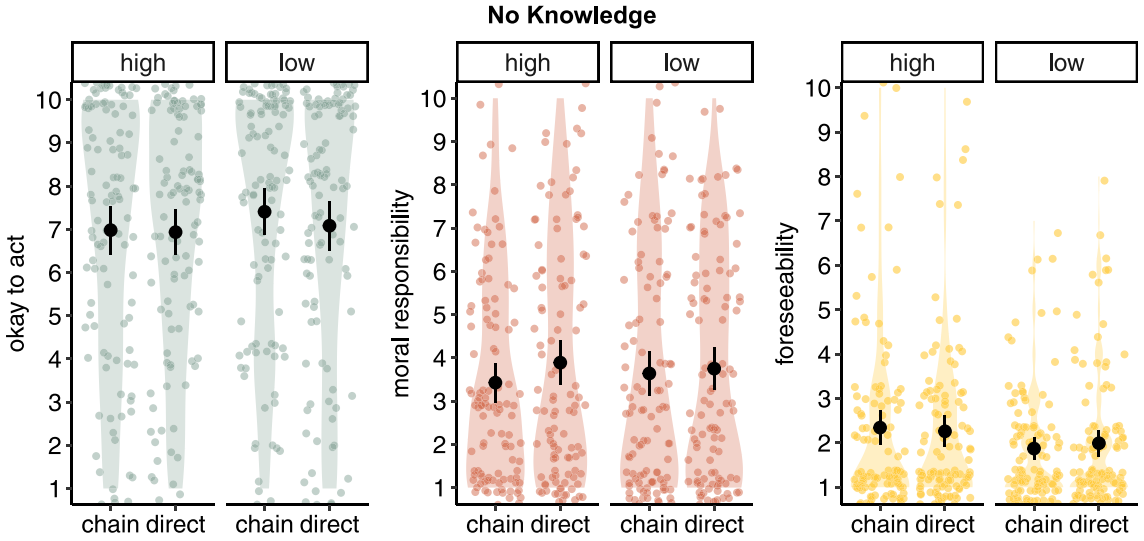


Fig. 11. Means and 95% CIs for ratings of moral permissibility (left), moral responsibility (middle), and foreseeability (right) in the no knowledge conditions of Experiment 3.

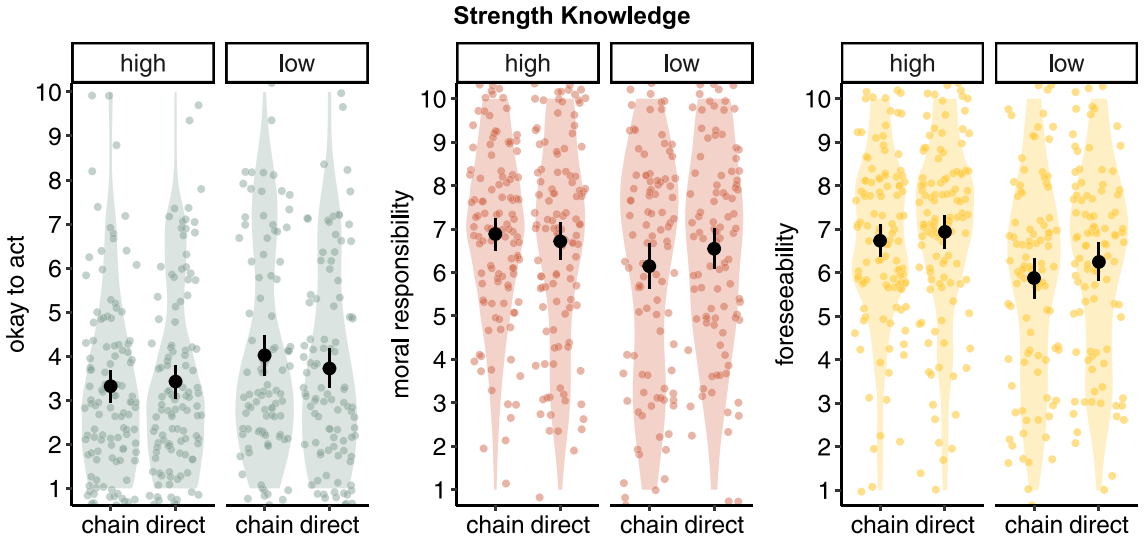


Fig. 12. Means and 95% CIs for ratings of moral permissibility (left), moral responsibility (middle), and foreseeability (right) in the strength knowledge conditions of Experiment 3.

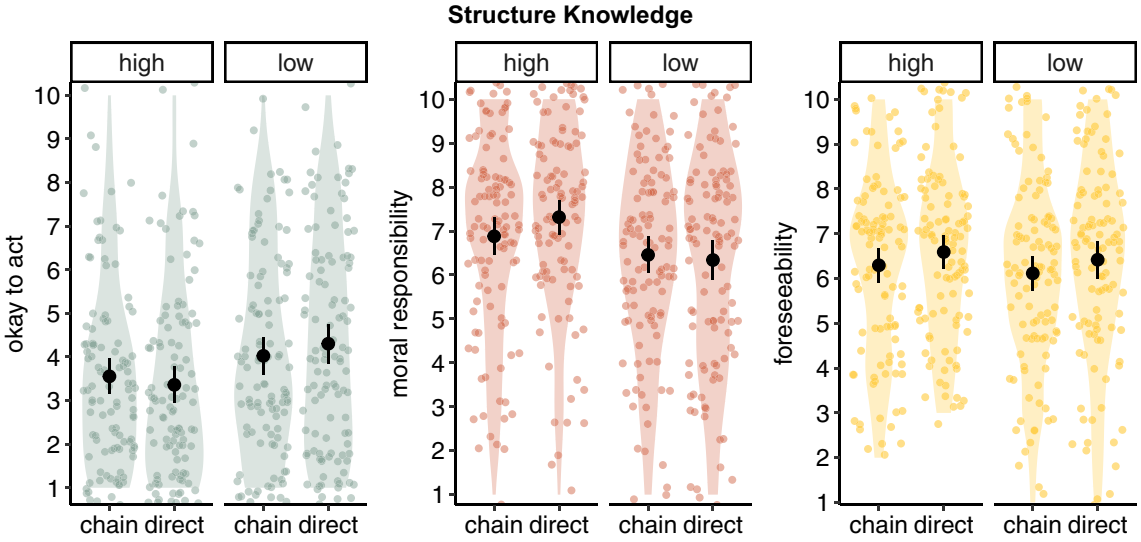


Fig. 13. Means and 95% CIs for ratings of moral permissibility (left), moral responsibility (middle), and foreseeability (right) in the structure knowledge conditions of Experiment 3.

Table 14
Effects of causal structure and causal strength on judgments of moral permissibility and moral responsibility per knowledge condition in Exp. 3.

	Knowledge	Query	*t (df)*	*p*	*d (0.95 CI)*
causal structure	full	permissibility	−0.06 (207)	.52	0.0 (−0.14, 0.13)
		responsibility	0 (207)	.50	0 (−0.14, 0.14)
	none	permissibility	1.11 (215)	.13	0.06 (−0.05, 0.18)
		responsibility	−1.95 (215)	.03	−0.11 (−0.22, 0)
	strength	permissibility	0.43 (209)	.34	0.04 (−0.14, 0.21)
		responsibility	−0.60 (209)	0.27	−0.04 (−0.17, 0.09)
	structure	permissibility	−0.29 (219)	.62	−0.02 (−0.15, 0.11)
		responsibility	−1.04 (219)	.15	−0.07 (−0.2, 0.06)
causal strength	full	permissibility	−3.84 (410.14)	<.001	−0.37 (−0.57, −0.18)
		responsibility	3.56 (410.10)	<.001	0.35 (0.15, 0.54)
	none	permissibility	−1.03 (429.79)	.15	−0.10 (−0.29, 0.09)
		responsibility	−0.15 (429.63)	.56	−0.01 (−0.20, 0.17)
	strength	permissibility	−2.36 (390.13)	.009	−0.23 (−0.43, −0.04)
		responsibility	1.99 (387.89)	.02	0.20 (0, 0.39)
	structure	permissibility	−3.29 (436.52)	<.001	−0.31 (−0.50, −0.12)
		responsibility	3.29 (437.29)	<.001	0.31 (0.13, 0.50)

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105167.

References

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

Anderson, R. A., Kamtekar, R., Nichols, S., & Pizarro, D. A. (2021). False positive emotions, responsibility, and moral character. *Cognition, 214*, 104770. https://doi.org/10.1016/j.cognition.2021.104770

Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: a practical tutorial. *Evidence-Based Mental Health, 22*, 153–160. https://doi.org/10.1136/ebmental-2019-300117

Bes, B., Sloman, S., Lucas, C. G., & Raufaste, E. (2012). Non-bayesian inference: Causal structure trumps correlation. *Cognitive Science, 36*(7), 1178–1203. https://doi.org/10.1111/j.1551-6709.2012.01262.x

Calderon, S., Mac Giolla, E., Ask, K., & Granhag, P. A. (2020). Subjective likelihood and the construal level of future events: A replication study of Wakslak, Trope, Liberman, and Alony (2006). *Journal of Personality and Social Psychology, 119*(5), e27–e37. https://doi.org/10.1037/pspa0000214

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*(2), 367. https://doi.org/10.1037/0033-295X.104.2.367

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58*(4), 545–567.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition, 40*(1–2), 83–120. https://doi.org/10.1016/0010-0277(91)90047-8

Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review, 112*, 694–706. https://doi.org/10.1037/0033-295X.112.3.694

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

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292. https://doi.org/10.1177/1088868313495594

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2019). xtable: Export Tables to LaTeX or HTML. *R package version 1.*, 4–8.

Driver, J. (2008). Attributions of causation and moral responsibility. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol. 2. The cognitive science of morality: Intuition and diversity* (pp. 423–439). Cambridge, MA: MIT Press.

Dubber, M. D. (2015). *An introduction to the model penal code*. New York: Oxford University Press.

Engelmann, N., & Waldmann, M. R. (2021). A causal proximity effect in moral judgment. In , *43. Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 2330–2336). Cognitive Science Society.

Engelmann, N., & Waldmann, M. R. (2022). How to weigh lives. a computational model of moral judgment in multiple-outcome structures. *Cognition, 218*, 104910. https://doi.org/10.1016/j.cognition.2021.104910

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.
Fincham, F., & Jaspars, J. (1983). A subjective probability approach to responsibility attribution. *British Journal of Social Psychology, 22*(2), 145–161. https://doi.org/10.1111/j.2044-8309.1983.tb00575.x
Gopnik, A., & Schulz, L. (2007). *Causal learning: Psychology, philosophy, and computation*. New York: Oxford University Press.
Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*(4), 334–384. https://doi.org/10.1016/j.cogpsych.2005.05.004
Hagmayer, Y., & Engelmann, N. (2020). Asking questions to provide a causal explanation - Do people search for the information required by cognitive psychological theories? In E. A. Bar-Asher Siegal, & N. Boneh (Eds.), *Perspectives on causation: Selected papers from the Jerusalem 2017 workshop, Jerusalem studies in philosophy and history of science* (pp. 121–147). Cham: Springer International Publishing.
Harrell, F. E. Jr. (2020). Hmisc: Harrell Miscellaneous. *R package version 4*, 2–4.
Hart, H. L. A., & Honoré, T. (1985). *Causation in the law*. New York: Oxford University Press.
Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*(8), 534–539. https://doi.org/10.1111/j.0956-7976.2004.00715.x
Hertwig, R., Hogarth, R. M., & Lejarraga, T. (2018). Experience and description: Exploring two paths to knowledge. *Current Directions in Psychological Science, 27*(2), 123–128. https://doi.org/10.1177/0963721417740645
Hilton, D. J., McClure, J., & Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology, 40*(3), 383–400. https://doi.org/10.1002/ejsp.623
Johnson, J. T., & Drobny, J. (1985). Proximity biases in the attribution of civil liability. *Journal of Personality and Social Psychology, 48*(2), 283. https://doi.org/10.1037/0022-3514.48.2.283
Johnson, S. G., & Ahn, W.-k. (2015). Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. *Cognitive Science, 39*(7), 1468–1503. https://doi.org/10.1111/cogs.12213
Kassambara, A. (2020). *ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0*.
Kelley, K. (2020). *MBESS: The MBESS R Package. R package version 4.8.0*.
Kirfel, L., & Lagnado, D. (2021). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition, 212*, 104721. https://doi.org/10.1016/j.cognition.2021.104721
Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition, 182*, 331–348. https://doi.org/10.1016/j.cognition.2018.09.003
Kneer, M., & Skoczeń, I. (2021). Outcome effects, moral luck and the hindsight. *Bias*. Available at SSRN https://ssrn.com/abstract=3810220.
Knobe, J., & Shapiro, S. (2021). Proximate cause explained. *The University of Chicago Law Review, 88*(1), 165–236.
Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition, 108*(3), 754–770. https://doi.org/10.1016/j.cognition.2008.06.009
Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 565–602). New York: Oxford University Press.
Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology, 129*, 101412. https://doi.org/10.1016/j.cogpsych.2021.101412
Lawrence, M. A. (2016). ez: Easy analysis and visualization of factorial experiments. *R package version 4.4-0*.
Livengood, J., & Sytsma, J. (2020). Actual causation and compositionality. *Philosophy of Science, 87*(1), 43–69. https://doi.org/10.1086/706085
Maier, M., Bartoš, F., Oh, M., Wagenmakers, E., Shanks, D., & Harris, A. J. L. (2022). *Publication bias in research on construal level theory*. Available at https://psyarxiv.com/r8nyu/.
Mangiafico, S. (2021). *rcompanion: Functions to Support Extension Education Program Evaluation. R package version 2.4.0*.
Mayrhofer, R., & Waldmann, M. R. (2015). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science, 39*(1), 65–95. https://doi.org/10.1111/cogs.12132
McClure, J., Hilton, D. J., & Sutton, R. M. (2007). Judgments of voluntary and physical causes in causal chains: Probabilistic and social functionalist criteria for attributions. *European Journal of Social Psychology, 37*(5), 879–901. https://doi.org/10.1002/ejsp.394
Meder, B., & Mayrhofer, R. (2017). Diagnostic causal reasoning with verbal information. *Cognitive Psychology, 96*, 54–84. https://doi.org/10.1016/j.cogpsych.2017.05.002
Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review, 121*(3), 277. https://doi.org/10.1037/a0035944
Nobes, G., & Martin, J. W. (2021). They should have known better: The roles of negligence and outcome in moral judgments of accidental actions. *British Journal of Psychology, 113*(2), 370–395. https://doi.org/10.1111/bjop.12536
Paharia, N., Kassam, K. S., Greene, J. D., & Bazerman, M. H. (2009). Dirty work, clean hands: The moral psychology of indirect agency. *Organizational Behavior and Human Decision Processes, 109*(2), 134–141. https://doi.org/10.1016/j.obhdp.2009.03.002
Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge, UK: Cambridge University Press.
Pearl, J., & Mackenzie, D. (2019). The book of why: The new science of cause and effect. *Penguin Books*.
Perales, J. C., Catena, A., Cándido, A., & Maldonado, A. (2017). Rules of causal judgment: Mapping statistical information onto causal beliefs. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 29–51). New York: Oxford University Press.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R. Core Team. (2020). nlme: Linear and nonlinear mixed effects models. *R package version 3*, 1–149.
R. Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02
Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin, 140*(1), 109–139. https://doi.org/10.1037/a0031903
Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research, 15*(2), 165–184.
RStudio Team. (2020). *RStudio: Integrated development environment for R. RStudio*. Boston, MA: PBC.
Rudy-Hiller, F. (2018). The epistemic condition for moral responsibility. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. Fall 2018 edition.
Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition, 156*, 164–176. https://doi.org/10.1016/j.cognition.2016.07.007
Schwenkler, J., & Sievers, E. (2022). Cause, "cause", and norm. In P. Willemsen, & A. Wiegmann (Eds.), *Advances in experimental philosophy of causation*. Bloomsbury.
Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. New York: Oxford University Press.
Sloman, S., Fernbach, P. M., & Ewing, S. (2009). Causal models: The representational infrastructure for moral judgment. *Psychology of Learning and Motivation, 50*, 1–26. https://doi.org/10.1016/S0079-7421(08)00401-5
Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General, 126*(4), 323–348. https://doi.org/10.1037/0096-3445.126.4.323
Spirtes, P., Glymour, C. N., & Scheines. (1993). *Causation, prediction, and search*. New York: Springer.
Steiger, J. H. (2004). Beyond the f test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9*(2), 164–182. https://doi.org/10.1037/1082-989X.9.2.164
Stephan, S., Tentori, K., Pighin, S., & Waldmann, M. R. (2021). Interpolating causal mechanisms: The paradox of knowing more. *Journal of Experimental Psychology: General, 150*(8), 1500–1527. https://doi.org/10.1037/xge0001016
Stephan, S., & Waldmann, M. R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science, 10*(1), 242–257. https://doi.org/10.1111/tops.12309
Strickland, B., Silver, I., & Keil, F. C. (2017). The texture of causal construals: Domain-specific biases shape causal inferences from discourse. *Memory & cognition, 45*(3), 442–455.
Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software, 59*(5), 1–38.
Torchiano, M. (2020). *effsize: Efficient effect size computation. R package version 0.8.1*.
Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*(2), 440–463.
Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. New York: Springer. Fourth edition.
Waldmann, M. R. (2017). *The Oxford handbook of causal reasoning*. New York: Oxford University Press.
Waldmann, M. R., & Wiegmann, A. (2012). The role of the primary effect in the assessment of intentionality and morality. In , *34*. *Proceedings of the annual meeting of the cognitive science society*. Cognitive Science Society.
Waldmann, M. R., Wiegmann, A., & Nagel, J. (2017). Causal models mediate moral inferences. In J. F. Bonnefon, & B Trémolière (Eds.), *Moral inferences* (pp. 45–63). Psychology Press.
Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*(12), 1–20. https://doi.org/10.18637/jss.v021.i12
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition, 131*(1), 28–43. https://doi.org/10.1016/j.cognition.2013.12.004

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

Woodward, J. (2021). *Causation with a human face: Normative theory and descriptive psychology*. New York: Oxford University Press.

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News, 2*(3), 7–10.

Žeželj, I. L., & Jokić, B. R. (2014). Replication of experiments evaluating impact of psychological distance on moral judgment. *Social Psychology, 45*(3), 223–231. https://doi.org/10.1027/1864-9335/a000188

Ziano, I., Wang, Y. J., Sany, S. S., Ho, N. L., Lau, Y. K., Bhattal, I. K., , … Chan, H. Y. C., et al. (2021). Perceived morality of direct versus indirect harm: Replications of the preference for indirect harm effect. *Meta-Psychology, 5*. https://doi.org/10.15626/MP.2019.2134

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

Appendix B Engelmann & Waldmann (2022b)

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

How to weigh lives. A computational model of moral judgment in multiple-outcome structures

Neele Engelmann *, Michael R. Waldmann

Department of Psychology, University of Göttingen, Germany

ARTICLE INFO

Keywords:
Moral judgment
Moral dilemmas
Moral reasoning
Deontology
Consequentialism
Utilitarianism

ABSTRACT

When is it allowed to carry out an action that saves lives, but leads to the loss of others? While a minority of people may deny the permissibility of such actions categorically, most will probably say that the answer depends, among other factors, on the number of lives saved versus lives lost. Theories of moral reasoning acknowledge the importance of outcome trade-offs for moral judgments, but remain silent on the precise functional form of the psychological mechanism that determines their moral permissibility. An exception is Cohen and Ahn's (2016) subjective-utilitarian theory of moral judgment, but their model is currently limited to decisions in two-option life-and-death dilemmas. Our goal is to study other types of moral judgments in a larger set of cases. We propose a computational model based on sampling and integrating subjective utilities. Our model captures moral permissibility judgments about actions with multiple effects across a range of scenarios involving humans, animals, and plants, and is able to account for some response patterns that might otherwise be associated with deontological ethics. While our model can be embedded in a number of competing contemporary theories of moral reasoning, we argue that it would most fruitfully be combined with a causal model theory.

1. Introduction

Most of us will never be in the unlucky position of the agent in a trolley dilemma (Foot, 1967). Our moral concerns are usually much more mundane than the question of whether or not we should let one person get run over by a train in order to save five others from the same fate, for example. Some people, however, routinely make life-and-death decisions. Many political actions, take the allocation of healthcare resources as just one example, have outcomes that can be quantified in terms of lives saved versus lives lost. While most of us do not actively get a say in these large-scale matters, we judge those who do. Everyday moral discourse, be it in person or on social media, is rife with both condemnation and justification of actions which, more or less directly, trade off lives or other goods. Examples of such trade-offs are policies implementing speed limits in traffic, the introduction of social distancing measures during the Covid-19 pandemic, or the European Union closing its borders to refugees.

Trolley dilemmas have, in recent years, often been criticized for lacking such real-life context or for poorly predicting actual moral behaviour (see, for example, Bauman, McGraw, Bartels, & Warren, 2014; Bostyn, Sevenhant, & Roets, 2018; Schein, 2020). Against this criticism, others have argued that moral psychology does not only address the question of how people behave in real-world situations, but also what they judge to be right and wrong. Moral judgment, so the argument, is an interesting psychological phenomenon in its own right (Białek, Turpin, & Fugelsang, 2019). Furthermore, moral dilemmas are not always meant to be representative of actual situations. As Plunkett and Greene (2019) argue, contrasts between different artificial moral dilemmas can serve the same purpose as contrasts between visual stimuli in artificial optical illusions. They can expose the core mechanics that are untraceable in more content-laden "realistic" situations.

Inspired by an initially exclusively philosophical debate ignited by Foot (1967) and Thomson (1985), moral psychologists have now spent at least two decades empirically investigating people's intuitions about moral dilemmas. Mirroring the philosophical debate about trolley dilemmas, the dominant research strategy in psychology has been to keep the outcomes of an action constant and vary other factors of interest. This strategy has revealed some relatively stable patterns (see May, 2018; Waldmann, Nagel, & Wiegmann, 2012, for detailed overviews). Everything else being equal, people find it morally worse if a negative outcome is brought about intentionally rather than by accident (Cushman, 2008; Cushman, Young, & Hauser, 2006; Young & Saxe, 2011),

* Corresponding author at: Department of Psychology, University of Göttingen, Gosslerstraße 14, 37073 Göttingen, Germany.
E-mail address: neele.engelmann@uni-goettingen.de (N. Engelmann).

https://doi.org/10.1016/j.cognition.2021.104910
Received 15 December 2020; Received in revised form 15 September 2021; Accepted 16 September 2021

through an action rather than an omission (Cushman et al., 2006; Cushman & Young, 2011; Spranca, Minsk, & Baron, 1991, but see Willemsen & Reuter, 2016), as a causal means for a positive primary outcome rather than a side-effect (Cushman et al., 2006; Cushman & Young, 2011; Feltz & May, 2017; Hauser, Cushman, Young, Kang-Xing Jin, & Mikhail, 2007), and by so-called "personal force" or "battery" rather than indirectly (Greene et al., 2009; Hauser et al., 2007; Mikhail, 2007, 2011). Overall, all of these features taken together may constitute the prototype of a harmful, morally bad action (see Greene, 2013, p. 247).

In contrast to these studies, the focus of the present research is on the role of outcomes in moral judgments. A common response is to associate outcomes with consequentialist and acts with deontological ethical theories. However, outcomes play a role in all ethical frameworks, including deontological theories. For example, the deontological Doctrine of Double Effect (see Alexander & Moore, 2016, for an overview) holds that an action which causes serious harm (such as a person's death) can be morally permissible given that, among other things, the harm is outweighed by the action's positive effects. But can one death be considered as outweighed when two other lives are saved? Are there degrees of permissibility when a larger or smaller number of lives are saved? Further complications arise when the lives involved in a trade-off belong to different categories (e.g., people vs. animals) or lives are traded off against other goods, such as inanimate objects or abstract values. Any rule based on a simple numerical comparison will fail to be applicable as soon as trade-offs involve more than one kind of entity (while causing the death of one person to save five others may be permissible, it may not be permissible to cause one person's death in order to save five fish, for example). Normative philosophical theories cover a wide range of positions on both the kind of trade-offs that are allowed and the circumstances under which they are allowed (see Alexander & Moore, 2016). Psychologically, judging trade-offs between different kinds of entities can certainly be requested from subjects, as has recently been strikingly demonstrated by the "moral machine" experiment (Awad et al., 2018). Here, participants made choices in dilemmas pitting a wide range of possible victims against each other (differing in number, age, role in society, and other features). Some stable patterns emerged, for example a preference to save more rather than fewer lives, or to save humans rather than animals. However, this study does not answer the question of how different entities are compared.

1.1. The role of outcomes in psychological theories of moral judgment

While most of the general psychological theories of moral judgment do not spell out an outcome integration mechanism in detail, all of them assume such a mechanism. Dual-process accounts posit that there are two competing modes of moral reasoning, with the first one reacting to situational features, such as personal force, intentionality, or the distinction between action and omission. In the theory of Greene and colleagues (Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001) a slow and deliberative second process follows and rationally determines whether the outcome trade-off is favourable or not. In Cushman's (2013) and Crockett's (2013) versions of dual-process theories, this second process is described in more detail and characterized as a model-based algorithm which evaluates an action based on all immediate outcomes in a specific situation. Still, the focus of these theories is on the "big picture" of moral judgment, for example on explaining to what extent it is driven by affective and cognitive processes. Figuring out the details of outcome trade-offs is not the main aim (but see Shenhav & Greene, 2010).

A competitor of dual-process theories is Mikhail's Universal Moral Grammar theory (Mikhail, 2007, 2011), which is inspired by deontological ethics. In this theory, the *Doctrine of Double Effect* (DDE) plays a central role with its focus on the distinction between intended and foreseen harm. As mentioned above, the DDE addresses outcome trade-offs in its *proportionality condition*: for an action that causes serious harm to be morally permissible, the harm in question must, among other specified conditions, not be "out of proportion" to the action's positive effects. Mikhail has proposed a formalism for comparing outcomes, the so-called *Moral Calculus of Risk* (Mikhail, 2011, pp. 140–142). It consists of the values of the positive and negative outcomes of a candidate action and considers their respective probabilities. Furthermore, the "Necessity of the Risk" is included, which is the probability that the agent's purpose (default: bringing about the positive effect) would not be achieved without risking the negative effect. Briefly put, an action should become more permissible with a better expected value, but less permissible when it is more likely that the positive effect could also have been produced without risking the negative effect at all. Dilemmas are defined by a limited set of options: the agent cannot bring about a positive effect without also causing a negative one. Therefore, the Moral Risk Calculus will, in most dilemma scenarios, come down to a simple expected value calculation: the actual numbers of lives saved versus lost, weighted by the respective probabilities of them being saved versus lost given the action. To our knowledge, the Moral Risk Calculus has not been subjected to a systematic empirical investigation.

Cohen and Ahn (2016) recently defended a novel one-system approach to reasoning about outcomes in moral scenarios, inspired by philosophical utilitarianism and decision theory (see Steele & Stefánsson, 2020, for an overview). According to their *Subjective-Utilitarian Theory of Moral Judgment* (henceforth: STMJ), *only* outcomes matter for evaluating a moral dilemma, which is contrary to all psychological theories of moral judgments discussed above. More specifically, the value that an observer attaches to the outcomes of each available course of action determines, according to STMJ, the probability that this course of action is selected as morally preferable. The proposed mechanism is formalized, and yields quantitative predictions about judgments in moral dilemmas. Applied to the standard trolley case with five lives saved and one life lost, the typical majority opinion that acting is permissible is explained by the fact that, all else being equal, people think that five lives are more important or valuable than just one. However, STMJ does not claim that people simply count and compare lives saved and lost. Instead, it is possible that one particular life (e.g., of a close friend) has a higher subjective value to someone evaluating the dilemma than the lives of five others combined. In this case, the theory predicts that the action that saves this one person is favoured.

The underlying decision process is described as a cumulative sampling of values from internal distributions until a decision criterion is reached. The form of the mechanism is inspired by a random-walk decision process, a model that has been confirmed in other domains, such as visual perception (e.g., Ratcliff & Rouder, 1998). Spelled out for the standard trolley case, STMJ would claim that an observer has some internal representation of the value of one life, and also of the value of five lives. These representations take the form of Gaussian distributions. The mean of the value distribution for five lives is likely to be higher than the mean of the value distribution of one life, but the two distributions might also overlap to some extent. When an observer is faced with the task of identifying the higher-valued stimulus of a pair, they repeatedly sample and compare value pairs from both distributions. At some point, enough evidence will have been accumulated to consciously conclude that five lives have the higher value. Crucially, the more two distributions overlap, the longer this process will take, resulting in the experience of a harder decision and in longer reaction times. More overlap between two distributions also creates noise, sometimes leading to prediction errors in which the option with a lower mean value dominates.

Cohen and Ahn (2016) had participants explicitly indicate the subjective values of a variety of stimuli: people, animals, and inanimate objects. Values were elicited by asking participants to compare each item against a standard with a fixed, arbitrary value (a chimpanzee with a value of 1000). From the values participants generated, a distribution for each item and the overlap between any two distributions was

determined. Different participants then completed a series of moral dilemma tasks using the pretested set of stimuli. In each trial, two stimuli were randomly drawn and presented together in a situation in which only one of them could be saved, and the other one would be killed or destroyed. Participants had to answer the question "*Would you save [Item A], causing [Item B] to be killed/destroyed?*" Their choices as well as response times were recorded for each trial. The overlap between value distributions of any two items turned out to predict both measures very well. Based on these results, Cohen and Ahn (2016) conclude that people are subjective utilitarians when it comes to moral judgments – that is, that they base their moral judgment only on the subjective values of an action's outcomes. Predictions of STMJ converge with findings from different lines of research. For example, when weighing different numbers of lives against each other, STMJ would not predict that the mean of the value distribution for "five lives" is five times higher than the mean for "one life". Instead, a concave relationship between the number of lives and values is assumed (see also Cromley & Cohen, 2019). And indeed, the distributions of some items, such as "one adult" and "five adults", showed a near complete overlap in Cohen and Ahn's (2016) studies, indicating that five lives were only valued marginally higher than one life. In brief, STMJ is parsimonious, firmly grounds moral judgment in well-established domain-general mechanisms, and makes quantitative predictions about choices in moral dilemmas that could be confirmed in several experiments.

Nonetheless, there are also some shortcomings and open questions. In its current form, the model is only applicable to classic moral dilemmas in which the action under consideration leads to a trade-off between saving and killing (or destroying). While such dilemmas are important, there are many decisions with multiple outcomes that are not life-and-death dilemmas. For example, a political action may benefit some groups at the expense of others (such as tax alleviations for top incomes), while nothing at all would have changed if the action had not been performed. Moreover, killing versus saving does not exhaust the realm of moral actions. An agent may also consider improving people's lives and compare the outcomes with an act that simply retains the status quo (e.g., health-related policy interventions). In these situations, the value of people's lives is not the only relevant quantity, but their status in the presence versus absence of a potential action needs to be compared. It is therefore desirable to generalize the model, and make it applicable to these other kinds of multiple-outcome situations as well.

Next, it is questionable whether participants in Cohen and Ahn's (2016) experiments actually provided *moral* judgments. After all, the test question in all experiments was "Would *you* save [Item A], causing [Item B] to be killed/destroyed?" (emphasis added). What people say they would do can be very different from what they think is morally right. For example, people might say that they would save their best friend rather than five strangers, while at the same time denying that this is the correct thing to do from a moral point of view (Kahane & Shackel, 2010, see also Tassy, Oullier, Mancini, & Wicker, 2013, Soter, Berg, Gelman, & Kross, 2021). Royzman and Hagan (2017) demonstrated that the "would you…" question used in many experiments may actually not track moral judgment, but a self-assessment of the likelihood that one would act in the described situation. Matters are further complicated by the fact that many dilemma studies, including those conducted by Cohen and Ahn (2016), frame the participant as the actor in a dilemma. However, people can give different judgments about a case when they are mere observers and thus morally evaluate someone else's action (Nadelhoffer & Feltz, 2008). Arguably, a large proportion of day-to-day moral judgments, and certainly the examples cited in the introduction, concern the actual behaviour of other people rather than hypothetical scenarios about oneself. Whether subjective utilities of outcomes predict judgments that (1) are actually about morality, and (2) concern the behaviour of other people thus remains an open question.

1.2. A Generalized Subjective-Utilitarian Model (GSUM)

To address these concerns, we propose and evaluate a *Generalized Subjective-Utilitarian Model* (GSUM). GSUM is based on the sampling of values, like the model proposed by Cohen and Ahn (2016). As described above, their model compares values of relevant entities in their alive or intact state, for example the value of five lives against the value of one life. An underlying assumption seems to be that when killed or destroyed, the value of entities reduces to zero (or another constant), and is therefore cancelled out when comparing the action alternatives. This may be a plausible simplification in life-and-death dilemmas, but it limits the range of applicability of the model. To generalize the model, all relevant actual, hypothetical, or counterfactual states of entities need to be explicitly represented. Imagine that an action improves the lives of five people, but also leads to the death of one person (henceforth: *improving* cases). Here, the gain of the first group needs to be traded off against the death of one person. In the case of a retrospective moral evaluation of an already executed action, the relevant comparison is between the actual state of affairs after the intervention, and the counterfactual state that would have obtained in the absence of the intervention. However, the same comparison can be made for a prospective evaluation of moral permissibility, in which case the predicted states in the presence and absence of an intervention are both hypothetical.

In a case in which two groups of people (or animals, plants) are affected by an action, our model therefore considers four subjective utilities[1]: (1) the state of Group 1 without intervention, (2) the state of Group 1 after intervention, (3) the state of Group 2 without intervention, and (4) the state of Group 2 after intervention. From these four values, the subjective utility of acting in this particular scenario (henceforth *scenario utility*) can be calculated. In the case of a classic moral dilemma, and assuming that subjective utilities of dead entities cancel out, the model reduces to the comparison of alive or intact values, as described by Cohen and Ahn (2016). But other cases require it to explicitly represent the values of entities in the contrasted states. Here is an example with a classic life-and-death dilemma case in which five lives are saved at the expense of one (SU = subjective utility):

$$\text{Scenario Utility } (\textit{saving}) = [\text{SU (5 normal)}–\text{SU (5 dead)}]+ [\text{SU (1 dead)}–\text{SU (1 normal)}]$$

And for an *improving* case with the same numbers:

$$\text{Scenario Utility } (\textit{improving}) = [\text{SU (5 improved)}–\text{SU (5 normal)}]+ [\text{SU (1 dead)}–\text{SU (1 normal)}]$$

If the action has more favourable outcomes than inaction, the scenario utility becomes positive in both cases.

GSUM takes as its input subjective utility assessments for items in different numbers and states. To make predictions for a particular scenario in which two items are traded off, four values are randomly sampled from the relevant pool of utility estimates (for example: one value for "five people in normal condition", one value for "five dead people", and so on), and the scenario utility is calculated. If a scenario utility is positive, a value of 1 is stored, otherwise it is represented as 0. To arrive at a robust prediction for each scenario, a large number of sampling iterations and scenario utility calculations are performed for each scenario (we are going to use 10,000 iterations). The proportion of positive scenario utilities among this large number of iterations is used as the predictor for a scenario's moral evaluation. The higher the proportion of positive scenario utilities, the higher are the predicted moral permissibility ratings for acting. Fig. 1 illustrates the procedure for a moral dilemma (*saving*) and for a case in which the action leads to an improvement of otherwise unchanged entities (*improving*).

GSUM thus embodies straightforward intuitions about the functional

[1] Any number of outcomes can be added to this equation.

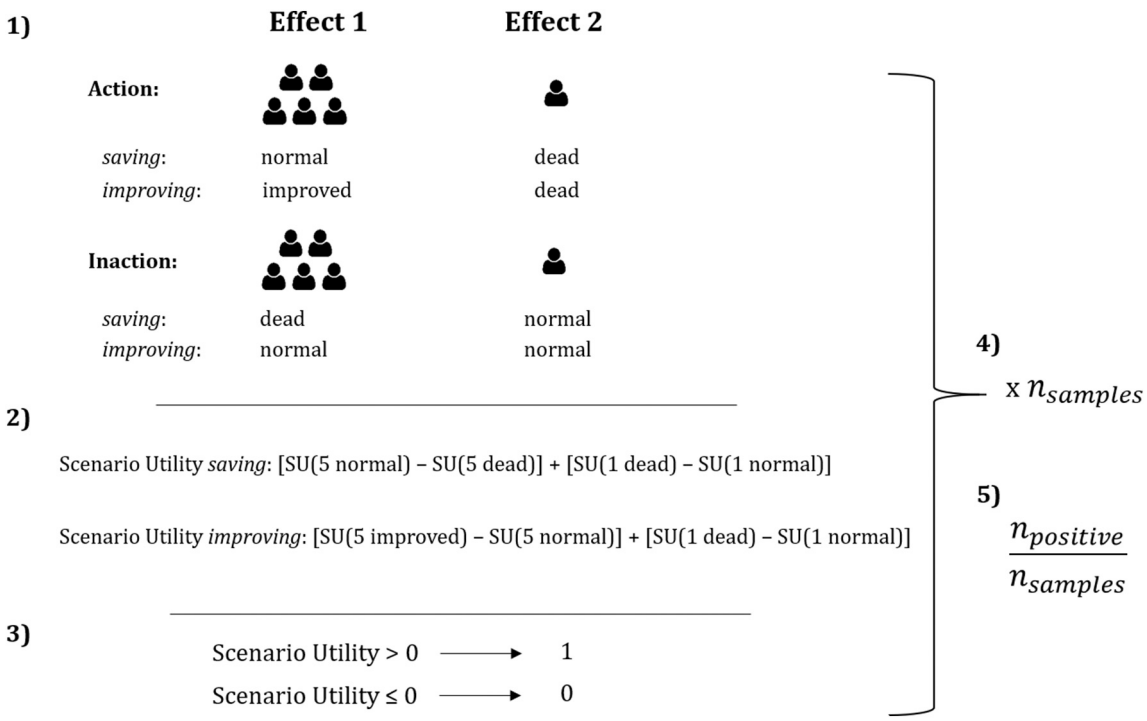


Fig. 1. Illustration of GSUM (with example values for subjective utilities) for a moral life-and-death dilemma (saving) and a case with two effects that is not a life-and-death dilemma (improving). SU = subjective utility.

form of a psychological outcome integration mechanism in the context of moral judgment. In a single sampling iteration, the model considers the aggregated value of all changes that are brought about by an action, and compares it to the aggregated value of an inaction. The crucial question for a moral evaluation of the action is whether the outcomes of acting outweigh the outcomes of inaction (or of an alternative action). GSUM represents this as a binary as well. As more and more samples are drawn, uncertainty caused by similar values of action and inaction or by large variations of the estimates becomes represented.

While our model is inspired by the model of Cohen and Ahn (2016), there are some key differences. The most obvious difference is the explicit modelling of state changes, resulting in a consideration of four rather than two values in each sample. Other differences arise due to the focus on moral instead of action preference judgments. Cohen and Ahn (2016) focus on binary choices. A choice counts as correctly predicted when the item with the higher mean utility (as identified in their independent utility estimation task) is saved. By contrast, we are interested in the extent to which people regard another person's action as morally permissible. We take moral permissibility to be a continuous evaluative reaction ranging from stark opposition to strong approval, rather than a binary choice. To predict moral judgments from participants' subjective utilities, we thus do not need to define a correct choice against which responses are compared. Our hypothesis is that the size of moral permissibility ratings will be proportional to the difference between the valuations of acting versus not acting. Formally, this is reflected in our model in the following way: we count the proportion of samples in which the outcomes of simulated actions outweigh the outcomes of simulated inactions. We use this proportion as a direct predictor of continuous moral permissibility judgments for actions.

2. Utility estimation study

In this study, we aimed to elicit the input data for our model, that is, subjective utility estimates for different entities in a range of numbers and states. The stimuli whose values we asked participants to estimate are the same ones that were used in the subsequent moral judgment tasks of Experiment 1 (life-and-death dilemmas) and Experiment 2 (life-and-death dilemmas vs. improving cases). Different kinds of entities (people, animals, plants) were compared in order to elicit a wide range of values, which allowed us to model a wide range of permissibility judgments in the subsequent moral judgment tasks.

2.1. Methods

2.1.1. Participants and design

We varied the number (one, five, ten, twenty, hundred), state (normal, dead, improved), and kind (people, monkeys, fish, trees, roses) of entities, all within-subject. We aimed for a sample size of 120 valid responses. Sample size was determined via simulation based on effects observed in a pilot study (small effect of numbers, $\eta_g^2 = 0.02$, large effects of state, $\eta_g^2 = 0.39$, and entity, $\eta_g^2 = 0.24$, two-way interactions between number and state, $\eta_g^2 = 0.01$, state and entity, $\eta_g^2 = 0.04$, and a three-way interaction, $\eta_g^2 = 0.004$). With 120 participants in a fully within-subject design with a conservative estimate for the correlation between repeated measures ($r = 0.1$), we achieve a power of at least 80% to detect each of these effects. Note, however, that the principal aim of this experiment was to collect input data for our model, not to test any specific hypotheses. All analyses should be therefore regarded as exploratory.

We invited 125 participants on *prolific* (www.prolific.co). Inclusion criteria were being at least 18 years old and a native English speaker, having an acceptance rate of previous studies on the platform of at least 90%, and not having participated in any previous studies using similar materials. Participants were paid £1.50 for an estimated 15 min of their time.

2.1.2. Materials and procedure

Participants were presented with the following instructions (see also Cohen & Ahn, 2016):

In the following study, your task will be to provide numerical value estimates for certain stimuli that will be presented to you. These stimuli can be people, animals, plants, or objects. You can understand the values that we will

ask you to estimate as an indication of how important, valuable or meaningful something /someone is, or how good or bad it is that something/ someone exists or does not exist, in your opinion. These values do not need to correspond to monetary value. For example, the first teddy bear you had as a child might have a high value to you, but only a very low monetary value. Likewise, something expensive could mean very little to you personally.

For example, an item in the experiment could be "a new bicycle". If you think that this is something good, then you should assign a positive value to this item. If you think that this is something bad, then you should assign a negative value. You could also assign a value of 0, to indicate that you are indifferent about the item. Moreover, the size of the value that you assign should reflect how positive or negative an item is, in your opinion. For example, assume you assigned a positive value of 10 to some item. If you value a second, different item ten times as much as this first item, you should assign a value of roughly 100 to the second item. The same is true for the negative direction. If you assign a negative value of − 10 to some item, and there is another item that is ten times worse than the first, in your opinion, you should assign a value of − 100 to the second item.

To help you come up with the numerical estimates, the task will be structured as follows:

You will see all items whose value we will ask you to estimate at once, on the same page. We encourage you to read through the whole list of items before assigning any values. When assigning the values, please use the following benchmarks as a reference:

- *Assume that "pieces of a broken tea cup" would be assigned a value of zero*
- *The highest possible value is + 1000*
- *The lowest possible value is − 1000*

Note that you can, but do not have to make use of the full range of the scale.

We chose "pieces of a broken tea cup" as a representative example for the scale value zero because we expected this item to be both familiar and naturally associated with a value of zero (worthless). Before the main task began, we presented participants with some practice trials ("a dead penguin", "two diamonds", "your best friend", "three healthy elephants", "a house that is burned down") and four instruction check questions (see Supplementary Materials). Participants were able to proceed to the main task once they had answered all instruction check questions correctly. Before entering any value estimates, participants had to scroll through the list of all 75 items (to help them calibrate their value estimates to the provided scale). On the next page, all items were presented again, and participants entered their value estimate for each item into a text field. The entries into text fields were not restricted, but participants were reminded to stick to the instructed scale (from −1000 to +1000).

2.2. Supplementary Materials

Data, materials, and code for this and all following experiments are available at https://osf.io/682uc/ (from here on: Supplementary Materials). For all statistical analyses and figures, we used R (R Core Team, 2019) and RStudio (RStudio Team, 2016) in combination with the following packages (in alphabetical order): *car* (Fox & Weisberg, 2019), *effsize* (Torchiano, 2020), *ez* (Lawrence, 2016), *faux* (DeBruine, 2020), *ggpubr* (Kassambara, 2019), *lmtest* (Zeileis & Hothorn, 2002), *MASS* (Venables & Ripley, 2002), *MBESS* (Kelley, 2019), *nlme* (Pinheiro et al., 2020), *nls2* (Grothendieck, 2013), *rcompanion* (Mangiafico, 2019), *reshape2* (Wickham, 2007), and the *tidyverse* (Wickham et al., 2019).

2.3. Results and discussion

Two participants were excluded because they failed a simple attention check,[2] resulting in a final sample size of 123 participants (mean age = 34.35, $SD = 13.16$, 56% women, 43% mean, 1% non-binary or no answer). Prior to the analyses we checked whether participants' entries conformed to the instructed response format (only numbers between −1000 and 1000, no text) and excluded those entries that did not. This resulted in the exclusion of 26 entries (0.3% of all entries). Fig. 2 shows the results. For all species, dead entities were predominantly assigned negative utilities (i.e., disutilities), and these values became more negative with higher numbers of dead entities. Normal and improved entities were assigned positive values that increased with larger numbers. Moreover, normal and improved states were valued very similarly overall. The highest values were assigned to people and the lowest to roses. Stepwise model comparisons revealed that the data were best described by a model containing main effects of number, entity, and state, the two-way interactions number × state and entity × state, plus the three-way interaction (see Table 1 for the output of the final model). The model explained 56% of the variance of the responses (Cragg & Uhler Pseudo-R^2). The number x state interaction reflects the fact that estimates became more positive with higher numbers for the improved and normal states, but more negative with higher numbers for the dead states (post-hoc tests[3] revealed that the effect was roughly medium-sized for all states, $\varepsilon^2 = 0.07$ for dead states, 0.08 for normal states, and 0.1 for improved states, all $ps < .001$, just the direction changed; see Mangiafico, 2016, for benchmarks of ε^2). Likewise, the entity × state interaction reflects that when in a normal or improved state, the highest values were provided for people, then monkeys and trees, then fish, and then roses (all $p < .001$, with Bonferroni-adjustment). When entities were dead, however, this order was reversed, with the most negative values assigned to people, then monkeys and trees, then fish, then roses (all $p < .001$, with Bonferroni-adjustment). Again, the size of the effect was medium for all three states ($\varepsilon^2 = 0.14$ for dead states, 0.10 for normal states, and 0.13 for improved states, all $p < .001$). The three-way interaction indicates that the difference in slopes for the manipulation of numbers of normal, improved, and dead states differed slightly between entities.

We also compared the fit of linear and nonlinear (exponential) models to the utility estimates, separately for each entity for alive (combining normal and improved) vs. dead states (see Supplementary Materials for the models and plots). We found that exponential models described the trajectory of utilities better than linear models for all entities and states. Utility estimates rise (or, for the dead states, fall) more quickly in the lower compared to the higher numerical ranges, thus showing patterns of diminishing marginal (dis-)utility or numbing (Slovic, 2007).

The main purpose of collecting this dataset was to use it as input for GSUM. We now turn to collecting moral permissibility judgments for a range of scenarios involving the stimuli whose values we have assessed in the Utility Estimation Study. We will also generate predictions for these cases using GSUM and compare them to participants' responses.

3. Experiment 1: Life-and-death dilemmas

The purpose of this experiment was to collect data from a new sample of participants for an initial evaluation of our model in an actual moral judgment task. We examined dilemmas in which ten entities

[2] "If Peter is taller than Alex, and Alex is taller than Max, who is the shortest among them?" This attention check was used in all subsequent experiments (presented on the final page).

[3] Friedman rank sum tests based on the data of all participants who provided no invalid entries ($N = 114$). *P*-values are Bonferroni-adjusted for the number of Friedman tests conducted (6 tests).

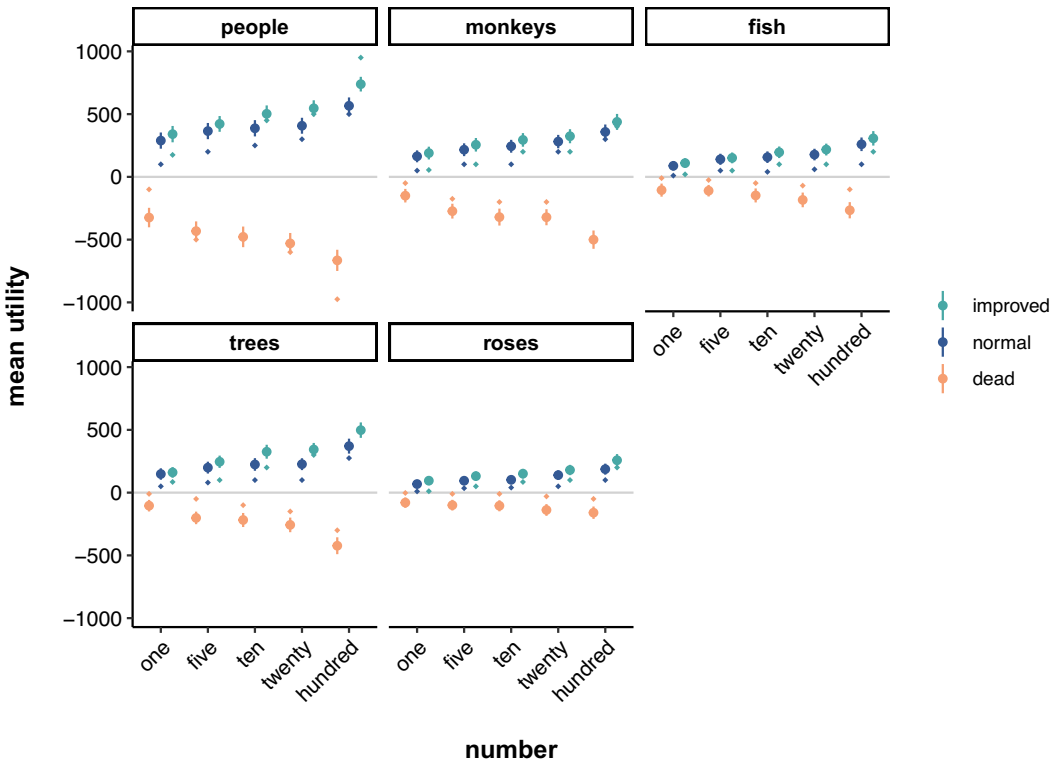


Fig. 2. Mean and median utilities assessed in the Utility Estimation Study. The large dots are means, the error bars are 95% confidence intervals. The small dots are medians. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(people, animals, or plants – two scenarios for each category) were killed in order to save either one hundred, twenty, or five others, or just one. Previous work (Cohen & Ahn, 2016) has only tested questions about personal action preferences, judged from the actor's perspective ("Would you…"). We systematically varied both the number and kind of entities (people, animals, plants) involved in a trade-off. This design allowed us to investigate whether the numerical ratio of lives saved versus lost influences moral judgments about trade-offs between human lives similarly as trade-offs between lives of animals and plants. The main goal was to explore whether potential value differences between humans, animals, and plants in different states explain differences in permissibility judgments.

3.1. Methods

3.1.1. Participants and design

We employed a 4 (number saved: hundred vs. twenty vs. five vs. one, between-subjects) × 6 (scenario: people 1 (*foodtruck* case) vs. people 2 (*river* case)[4] vs. monkeys vs. fish vs. trees vs. roses, within subject) design. We expected that the between-subjects effect of the number of affected entities will be the smallest effect in the design. We invited 615 participants to participate in our survey on the platform *prolific* (www.prolific.co). To be included in the experiment, participants had to be native speakers of English, not have participated in any previous studies using similar materials, and have a 90% acceptance rate of previous tasks on the platform at least. Participants were paid 0.50 GBP for an estimated five minutes of their time (6 GBP/h). 21 participants were excluded for failing a simple attention check, leaving data of 594 participants for the analyses (mean age = 36.3, $SD = 12$, 60% female, 39% male, 1% another identity/no answer). This sample size yielded a power of approximately 80% to detect a between-subjects effect of numbers at Cohen's $f = 0.14$ ($\eta_p^2 = 0.019$), and a power of approximately 90% to detect the effect at Cohen's $f = 0.16$ ($\eta_p^2 = 0.025$; determined with G*Power 3.1.9.2, Faul, Erdfelder, Lang, & Buchner, 2007, and Superpower, Lakens & Caldwell, 2021).

3.1.2. Materials and procedure

In each of the six vignettes an agent is facing a dilemma. By performing a certain action, they can save a number of lives (hundred, twenty, five, or one), but will inevitably also cause ten deaths (this number was kept constant across all scenarios and conditions). The threat to one group was described as resulting from external circumstances such as natural disasters or illness. The sole means of saving was a re-allocation of limited resources (e.g., food, water), where receiving extra resources would save the threatened group. Given that these resources are limited, re-allocating more to the threatened group would lead to the death of the other, formerly unthreatened group (by lack of food or water, for example). Thus, harming was a side-effect of helping, never a means. We described agents as authorized to make the decision in question (via roles in government or management) in order to preclude participants from making judgments about legal rather than moral permissibility. Personal force or physical contact were not part of the scenarios. Moreover, the consequences of acting were never self-beneficial to agents. In each vignette, all entities are of the same kind (all human, all animals, or all plants). The agent is aware of all the outcomes and is motivated by the positive, but not the negative outcomes. In all cases, the agent decides to act, and both outcomes occur. Scenarios were presented in random order. After reading each scenario, participants were asked to provide a rating of the moral permissibility of

[4] Since we included two scenarios about animals (monkeys, fish) and plants (trees, roses), we also included two scenarios about people. These only differed in terms of the cover story: In the *food truck* case, lives could be saved by redirecting a food truck from one village to another; in the *river* case, people could be saved by redirecting a river.

Table 1
Summary of the selected regression model of the data collected in the Utility Estimation Study.

Random effects: participant ID					
	Intercept				Residual
SD	136.04				275.14
Fixed effects:					
	Estimate	SE	*df*	*t*	*p*
(Intercept)	−323.87	27.88	9002	−11.62	<0.001
Five	−108.02	35.37	9002	−3.05	0.002
Ten	−153.89	35.37	9002	−4.35	<0.001
Twenty	−205.94	35.37	9002	−5.82	<0.001
Hundred	−341.02	35.37	9002	−9.64	<0.001
Monkeys	173.93	35.3	9002	4.93	<0.001
Fish	218.58	35.37	9002	6.18	<0.001
Trees	220.17	35.37	9002	6.22	<0.001
Roses	244.44	35.37	9002	6.91	<0.001
Normal	612.88	35.3	9002	17.36	<0.001
Improved	664.36	35.3	9002	18.82	<0.001
Five, normal	183.81	49.92	9002	3.68	<0.001
Ten, normal	252.01	49.92	9002	5.05	<0.001
Twenty, normal	329.13	49.97	9002	6.59	<0.001
Hundred, normal	618.27	49.92	9002	12.38	<0.001
Five, improved	189.37	49.92	9002	3.79	<0.001
Ten, improved	315.77	49.92	9002	6.33	<0.001
Twenty, improved	411.92	49.92	9002	8.25	<0.001
Hundred, improved	739.07	49.92	9002	14.8	<0.001
Monkeys, normal	−299.41	49.87	9002	−6	<0.001
Fish, normal	−419.6	49.92	9002	−8.4	<0.001
Trees, normal	−361.01	49.97	9002	−7.22	<0.001
Roses, normal	−465.14	49.92	9002	−9.32	<0.001
Monkeys, improved	−324.86	49.87	9002	−6.51	<0.001
Fish, improved	−446.12	49.97	9002	−8.93	<0.001
Trees, improved	−399.61	49.92	9002	−8	<0.001
Roses, improved	−489.84	49.92	9002	−9.81	<0.001
Five monkeys	−14.89	49.97	9002	−0.3	0.766
Ten monkeys	−16.33	49.97	9002	−0.33	0.744
Twenty monkeys	34.25	49.92	9002	0.69	0.493
Hundred monkeys	−8.93	49.97	9002	−0.18	0.858
Five fish	102.98	49.97	9002	2.06	0.039
Ten fish	111.56	50.02	9002	2.23	0.026
Twenty fish	127.97	50.02	9002	2.56	0.011
Hundred fish	180.35	49.97	9002	3.61	<0.001
Five trees	10.59	50.02	9002	0.21	0.832
Ten trees	39.88	50.02	9002	0.8	0.425
Twenty trees	52.79	50.02	9002	1.06	0.291
Hundred trees	21.55	49.97	9002	0.43	0.666
Five roses	87.29	50.03	9002	1.74	0.081
Ten roses	128.86	50.02	9002	2.58	0.01
Twenty roses	147.69	50.08	9002	2.95	0.003
Hundred roses	260.43	49.97	9002	5.21	<0.001
Five monkeys, normal	−8.09	70.6	9002	−0.11	0.909
Ten monkeys, normal	−1.52	70.57	9002	−0.02	0.983
Twenty monkeys, normal	−39.77	70.57	9002	−0.56	0.573
Hundred monkeys, normal	−73.78	70.57	9002	−1.05	0.296
Five fish, normal	−126.69	70.57	9002	−1.8	0.073
Ten fish, normal	−140.86	70.6	9002	−2	0.046
Twenty fish, normal	−162.41	70.67	9002	−2.3	0.022
Hundred fish, normal	−286.04	70.57	9002	−4.05	<0.001
Five trees, normal	−35.31	70.64	9002	−0.5	0.617
Ten trees, normal	−62.72	70.67	9002	−0.89	0.375
Twenty trees, normal	−97.57	70.67	9002	−1.38	0.167
Hundred trees, normal	−76.55	70.6	9002	−1.08	0.278
Five roses, normal	−136.65	70.6	9002	−1.94	0.053
Ten roses, normal	−193.18	70.6	9002	−2.74	0.006
Twenty roses, normal	−199.63	70.68	9002	−2.82	0.005
Hundred roses, normal	−419.01	70.57	9002	−5.94	<0.001
Five monkeys, improved	−0.36	70.57	9002	−0.01	0.996
Ten monkeys, improved	−40.66	70.57	9002	−0.58	0.564
Twenty monkeys, improved	−105.98	70.53	9002	−1.5	0.133
Hundred monkeys, improved	−140.66	70.57	9002	−1.99	0.046
Five fish, improved	−146.22	70.6	9002	−2.07	0.038
Ten fish, improved	−190.62	70.64	9002	−2.7	0.007
Twenty fish, improved	−229.28	70.64	9002	−3.25	0.001
Hundred fish, improved	−384.56	70.6	9002	−5.45	<0.001
Five trees, improved	−7.08	70.6	9002	−0.1	0.92
Ten trees, improved	−36.99	70.6	9002	−0.52	0.6
Twenty trees, improved	−76.18	70.6	9002	−1.08	0.281
Hundred trees, improved	−82.76	70.57	9002	−1.17	0.241
Five roses, improved	−132.09	70.6	9002	−1.87	0.061
Ten roses, improved	−234.94	70.6	9002	−3.33	<0.001
Twenty roses, improved	−268.6	70.64	9002	−3.8	<0.001
Hundred roses, improved	−495.69	70.57	9002	−7.02	<0.001
AIC	129,970.4				
Pseudo-R^2 (Cragg & Uhler)	0.56				

the action ("To what extent was [agent]'s action morally permissible?") on a scale ranging from 1 ("not at all") to 10 ("fully"). For each scenario, illustrations were shown indicating the numbers of entities as well as their states before and after the agent's action. Here is an example of a scenario in which 100 people are saved and ten are killed (see Supplementary Materials for all other scenarios):

Olivia is the prime minister of Tolosia, a mountainous country with many distant and small valleys. She is authorised to make all decisions about the inhabitants' welfare.

One day, she learns that one valley, Morhall, is suffering from a drought that left its inhabitants in poor health due to lack of water. Exactly 100 people live in Morhall, all of whom are in critical condition and will die if nothing is done.

Olivia could order to open a dam that would redirect a mountain river towards Morhall. With a quick water supply, the 100 inhabitants would recover. However, the redirection of the river would also cause a lack of water in another mountain village, Lorness, causing its 10 inhabitants to die of thirst within a few days. All of the 10 inhabitants of Lorness are fine at the moment.

Since both valleys are inaccessible to any means of transport, redirecting the river is currently the only available measure to influence the wellbeing of the inhabitants.

Here is an illustration of the two valleys and the current state of their inhabitants (Fig. 3).

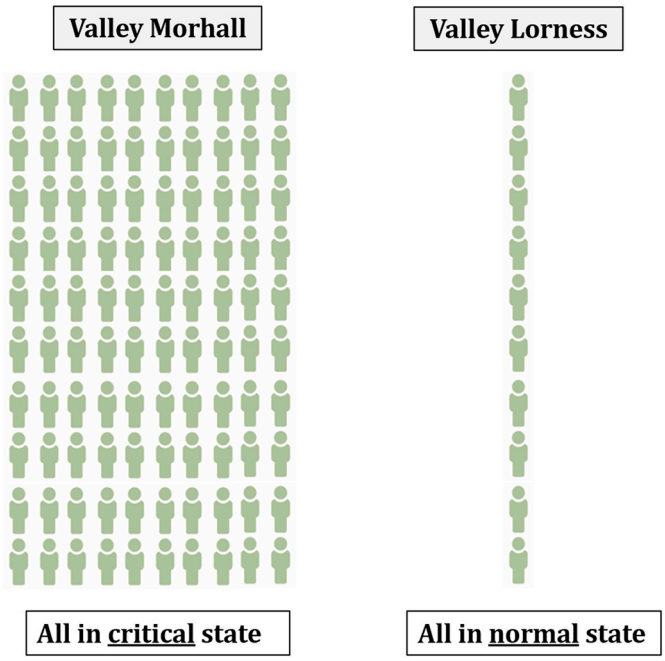


Fig. 3. Example of illustrations used in Experiment 1: States of affected groups before intervention.

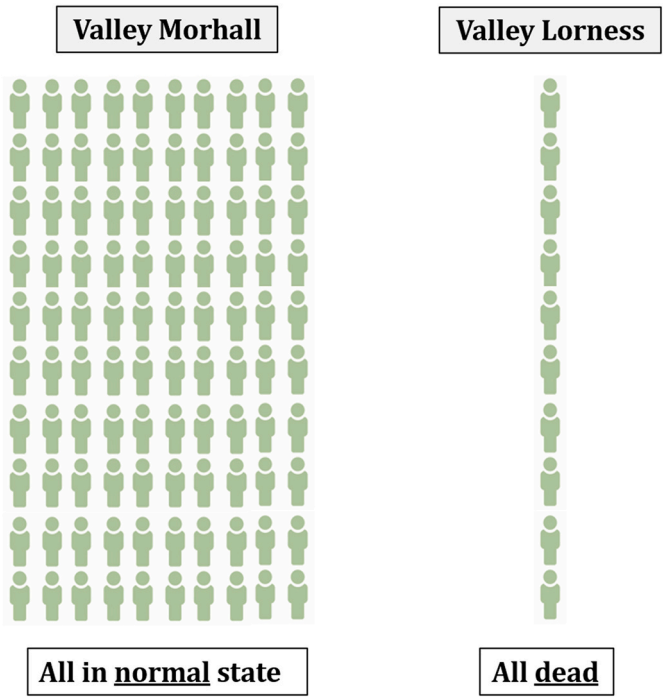


Fig. 4. Example of illustrations used in Experiment 1: States of affected groups after intervention.

Olivia is aware of all the facts. She wants the 100 inhabitants of Morhall to recover, but also not to cause any harm to the 10 inhabitants of Lorness. She decides to open the dam and redirect the mountain river. All of the 100 inhabitants of Morhall recover. However, all of the 10 inhabitants of Lorness die within a few days.

Here is an illustration of the two valleys and the state of their inhabitants after the river has been redirected (Fig. 4).

After completing all six scenarios, demographic variables were assessed, and participants were presented with the same attention check as in the previous study.

3.2. Results and discussion

Fig. 5 shows the mean moral permissibility ratings per condition, along with GSUM's predictions. The scenarios elicited judgments across the whole range of the rating scale. The action was judged as least permissible in the case of an unfavourable trade-off (saving one and killing 10) and when the affected entities were people ($M = 2.85$, $SD = 2.34$). It was judged as most permissible, nearly at ceiling, when the trade-off was favourable (saving 100 and killing 10) and the affected entities were plants ($M = 8.56$, $SD = 1.81$). In between, permissibility ratings increased as a function of the numerical ratio of saved compared to killed entities (more permissible with more entities saved compared to killed) and of the kind of affected entities (more permissible when plants were concerned than animals, and more permissible for animals than for people). This pattern indicates that people are more willing to trade off saving with harming when plants are involved than when the trade-offs concern animals. The strongest reluctance can be seen with humans.

A mixed 4 (number saved: hundred vs. twenty vs. five vs. one, between-subjects) × 6 (scenario: people/foodtruck, people/river, monkeys, fish, trees, roses; within subject) ANOVA confirmed the impression from the visual inspection. There was a large main effect of the number of saved entities, $F(3, 590) = 137.62$, $p < .001$, $\eta_p^2 = 0.41$ [0.36; 0.45],[5] as well as a somewhat smaller, but still large effect of scenario, $F(5, 2950) = 186.76$, $p_{GG} < 0.001$, $\eta_p^2 = 0.24$ [0.22; 0.26].

There was also an interaction effect, $F(15, 2950) = 6.78$, $p_{GG} < 0.001$, $\eta_p^2 = 0.03$ [0.02; 0.04], indicating that the number of saved entities did not have an equally strong effect on moral permissibility ratings in all scenarios (the ANOVA results do not change when adjusting p-values for multiple testing). We followed up on this interaction with contrasts checking for an overall linear trend for the number variable, and possible interactions of this trend with the scenario factor. As expected, moral permissibility ratings showed an overall linear trend, increasing with more entities saved compared to harmed ($D = 2.64$, $t = 13.46$, $p < .001$). The significant interactions revealed that this linear trend was stronger when the involved entities were fish rather than people ($D = 0.81$, $t = 2.92$, $p = .003$), trees rather than people ($D = 1.10$, $t = 3.97$, $p < .001$), and roses rather than people ($D = 0.80$, $t = 2.89$, $p = .003$). The strength of the trend did not differ between the two scenarios involving people ($D = -0.09$, $t = -0.32$, $p = .75$.), nor between people and monkeys ($D = 0.35$, $t = 1.25$, $p = .21$).[6] Thus, the number of saved compared to killed entities mattered less for moral permissibility ratings in scenarios involving trade-offs among human lives compared to those of nearly all other entities. Detailed descriptive statistics for all conditions can be found in the Supplementary Materials.

To test GSUM, we generated permissibility predictions for all experimental conditions (see Supplementary Materials for the code). The model predicts participants' judgments well. We compared the fit of linear, exponential, and sigmoid functions to describe the relationship between model predictions and participants' mean moral evaluations of the scenarios. An exponential function ($y = ax^b$, $a = 12.16$, $t_{22} = 10.28$, $p < .001$, $b = 1.19$, $t_{22} = 7.64$, $p < .001$, normalized[7] RMSE = 0.16, Cragg & Uhler $R^2 = 0.77$) described the relationship best. Instead of group means, the model can also be fit to the group medians, which results in a virtually identical fit (here, a linear model described the relationship best, $b = 16.13$, $t_{22} = 8.29$, $p < .001$, normalized RMSE = 0.16, $R^2 = 0.76$).

We also generated a separate set of predictions in which values for the dead states of all entities were replaced by zeroes. This model corresponds to the one proposed by Cohen and Ahn (2016) in which only alive/intact states were compared. This model fits the data of the present study on life-and-death dilemmas roughly equally well, regardless of whether means or medians were used as criterion (means: $y = ax^b$, $a = 15.74$, $t_{22} = 8.19$, $p < .001$, $b = 1.45$, $t_{22} = 8.06$, $p < .001$, normalized RMSE = 0.15, Cragg & Uhler $R^2 = 0.79$; medians: $y = ax^b$, $a = 21.20$, $t_{22} = 6.23$, $p < .001$, $b = 1.94$, $t_{22} = 7.80$, $p < .001$, normalized RMSE = 0.15, Cragg & Uhler $R^2 = 0.79$). Again, we compared the fit of linear, exponential, and sigmoid functions, and reported the best-fitting relation, which was the exponential function). The next experiment will provide a better test between the models.

The results of Experiment 1 show that our generalized subjective utilitarian model (GSUM) predicts people's moral permissibility judgments of the actions of *other* agents. The better the outcomes of acting compared to inaction in a scenario, the higher participants' ratings of

[5] We report 90% confidence intervals for all eta squared effect sizes, see Steiger (2004).

[6] There was also a significant negative cubic trend ($D = -0.47$, $t = 2.43$, $p = .015$) for the manipulation of the numbers (overall, no interactions with scenario). This trend is likely due to the fact that ratings increased more steeply between five and twenty than between the other numerical conditions. The trend analyses were not adjusted for multiple testing and should be regarded as exploratory. When Bonferroni-correcting for the number of trend tests involved in the polynomial contrasts of the numbers variable (18 tests), only the following trends remain significant: the overall linear trend ($p < .001$) and the interaction with the trees scenario ($p = .001$).

[7] RMSEs were normalized by the range of the criterion on all occasions where they are reported.

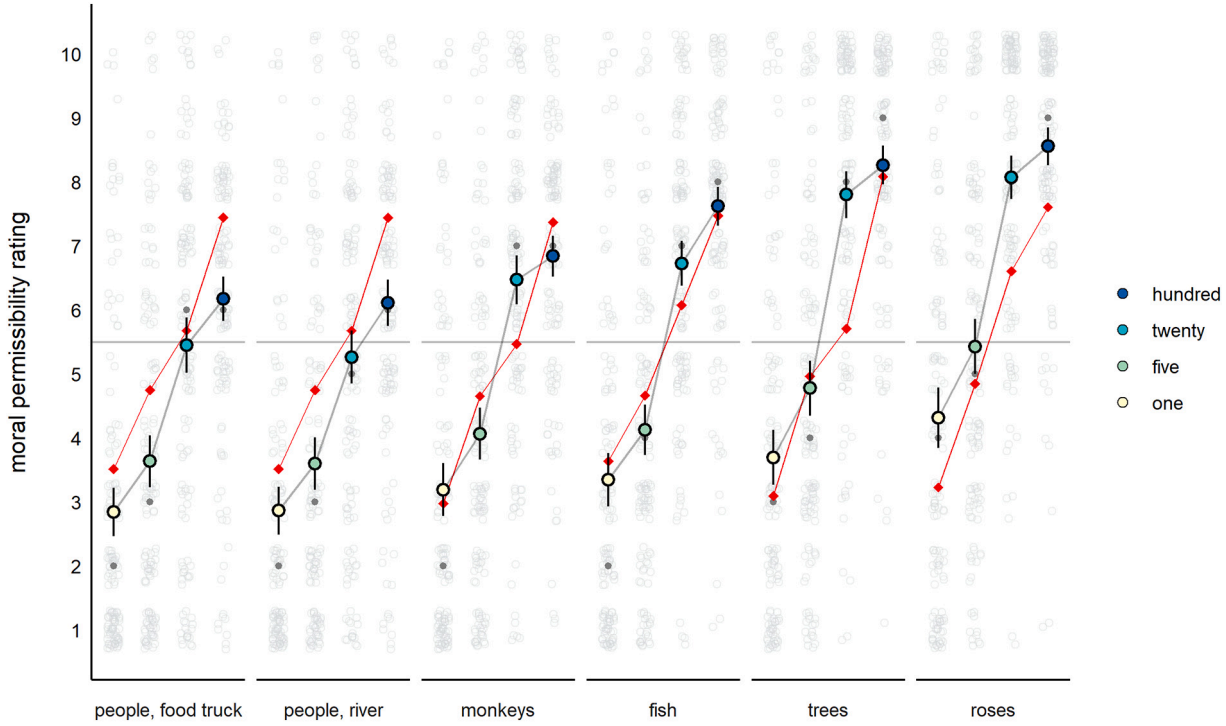


Fig. 5. Mean moral permissibility ratings (large points in blue colors) per condition in Experiment 1. Error bars are 95% confidence intervals. Medians are displayed in dark grey, individual data points (jittered) in light grey. GSUM predictions (fitted to means) are shown in red. The light grey line indicates the scale midpoint. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

moral permissibility. Thus, it seems that subjective utilities do not only predict judgments about what people think *they* would *do* in a dilemma (Cohen & Ahn, 2016), but also of how they morally evaluate *other* people's behaviour.

It is noteworthy that participants' moral judgments did not show a strict split (i.e., uniformly low whenever fewer lives are saved than lost, uniformly high otherwise). Instead, moral permissibility ratings linearly increased with higher numbers of saved lives, even though the strength of the trend differed between entities. This pattern suggests that people's intuitions about the cases may be driven by the subjective values of the outcomes (relative to the outcomes of inactions) rather than, say, by a categorical principle.

4. Experiment 2: Saving versus improving

The aim of the second experiment was to extend the scope of investigated situations to cases beyond simple life-and-death dilemmas. Many actions with multiple morally relevant outcomes are not just about trade-offs between life and death. Other cases can be understood in terms of state differences, too. For example, an action might improve the lives of 100 people, but cause the deaths of ten others. The gain that is obtained by making the lives of 100 people somewhat better has to be traded off against the loss of 10 lives. If the perceived gain is higher than the perceived loss, the action should be seen as morally permissible. While decisions like this are more common than life-and-death dilemmas, previous models like the one proposed by Cohen and Ahn (2016) do not address them. By explicitly modelling the state changes that all entities undergo due to an action, GSUM can fill this gap. If moral judgments about improving scenarios are also driven by the subjective value of outcomes, an action should be seen as more morally permissible, the stronger its outcomes outweigh the outcomes of inaction (in the case of improving scenarios, retaining the status quo). An alternative possibility is that such actions are categorically impermissible, independent of the relation between losses and gains. Such a constraint might be justified deontologically, for example by positing that causing death can never be allowed when the positive outcome is a mere improvement of other's lives. In this case, participants permissibility judgments about such cases should be uniformly low.

4.1. Methods

4.1.1. Participants and design

The design was identical with the one of Experiment 1, except for the addition of a new between-subjects condition (improving). Here, the scenario was not described as a life-and-death dilemma; rather, the agent in the scenario had to decide whether to perform an action that would improve the states of some entities (people, animals, or plants, whose numbers varied as in Experiment 1) while causing the deaths of ten others. Thus, the full design was 2 (saving vs. improving, between-subjects) × 4 (number saved: hundred vs. twenty vs. five vs. one, between-subjects) × 6 (scenario: people 1 (*foodtruck* case) vs. people 2 (*river* case) vs. monkeys vs. fish vs. trees vs. roses, within subject). We decided to aim for a sample size of 300 participants in both the saving and the improving condition ($N = 600$ in total). We invited 621 participants to take part in our survey via *prolific* (www.prolific.co), who had not participated in Experiment 1. Otherwise, the inclusion criteria were the same as in Experiment 1. Participants were paid £0.50 for an estimated five minutes of their time (6 GBP/h). 14 participants were excluded for failing the attention check, leaving data of 607 participants for all analyses (mean age = 37.4, $SD = 13.3$, ca. 55% female, ca. 45% male, < 1% no answer). With 303 participants (rounded down) in both the *saving* and the *improving* conditions, we achieved a power of approximately 80% to detect a between-subjects effect of numbers at a size of Cohen's $f = 0.20$ ($\eta_p^2 = 0.038$), and a power of approximately 90% to detect this effect at a size of Cohen's $f = 0.22$ ($\eta_p^2 = 0.046$) in each condition (determined with GPower 3.1.9.2, Faul et al., 2007, and Superpower, Lakens & Caldwell, 2021). Note that these effects are the

smallest effects of interest in our design (the power is even higher for the within-factor "kind of affected entities" and for the main effect of "saving vs. improving" on moral permissibility ratings in an overall ANOVA).

4.1.2. Materials and procedure

In the saving conditions, we used the same vignettes as in Experiment 1. In the improving conditions, a different positive primary effect was described. As in Experiment 1 and as in the saving conditions, the action in the improving scenarios was a re-allocation of resources. This feature allowed us to keep all scenario features comparable to the saving conditions, with the exception that the agent did not re-allocate the resources to save a threatened group from death, but to improve a non-threatened group's condition while causing another group's death due to a lack of a resource. In the case of inaction, both groups of entities would remain in their normal, non-threatened state. For the example presented earlier (in which 100 people were saved), the improving version of the vignette included the following changes (see Supplementary Materials for the full text of all scenarios):

(..) One day she learns that the health of the 100 inhabitants of one valley, Morhall, could be even better and their lifespan vastly extended if extra water was available to them. Olivia could order to open a dam that would redirect a mountain river toward Morhall. With a quick water supply, the 100 inhabitants of Morhall could improve farming and hygiene and thereby reach an even better level of health and longer life than before. (…)

Olivia is aware of all the facts. She wants the 100 inhabitants of Morhall to improve their health and extend their lifespan, but also not to cause any harm to the 10 inhabitants of Lorness. She decides to open the dam and redirect the mountain river. All of the 100 inhabitants of Morhall improve their health and extend their lifespan. However, all of the 10 inhabitants of Lorness die within a few days.

As in the saving conditions, the improving versions of the vignettes included illustrations of numbers and states. Moral permissibility ratings and demographics were assessed in the same manner as in Experiment 1.

4.2. Results and discussion

Fig. 6 provides an overview of results, along with model predictions (see Supplementary Materials for all descriptive statistics). The results in the *saving* condition showed roughly the same patterns as in Experiment 1. In the *improving* conditions, the permissibility ratings were generally low. In most conditions, participants found an *improving* action not permissible (i.e., ratings below scale midpoint). However, within the lower half of the rating scale, permissibility ratings in the improving conditions still tended to increase when more entities' conditions were improved, as would be expected by GSUM. Trading off human lives was again least permissible, followed by animals, and plants.

The statistical analyses confirmed the descriptive patterns. In a

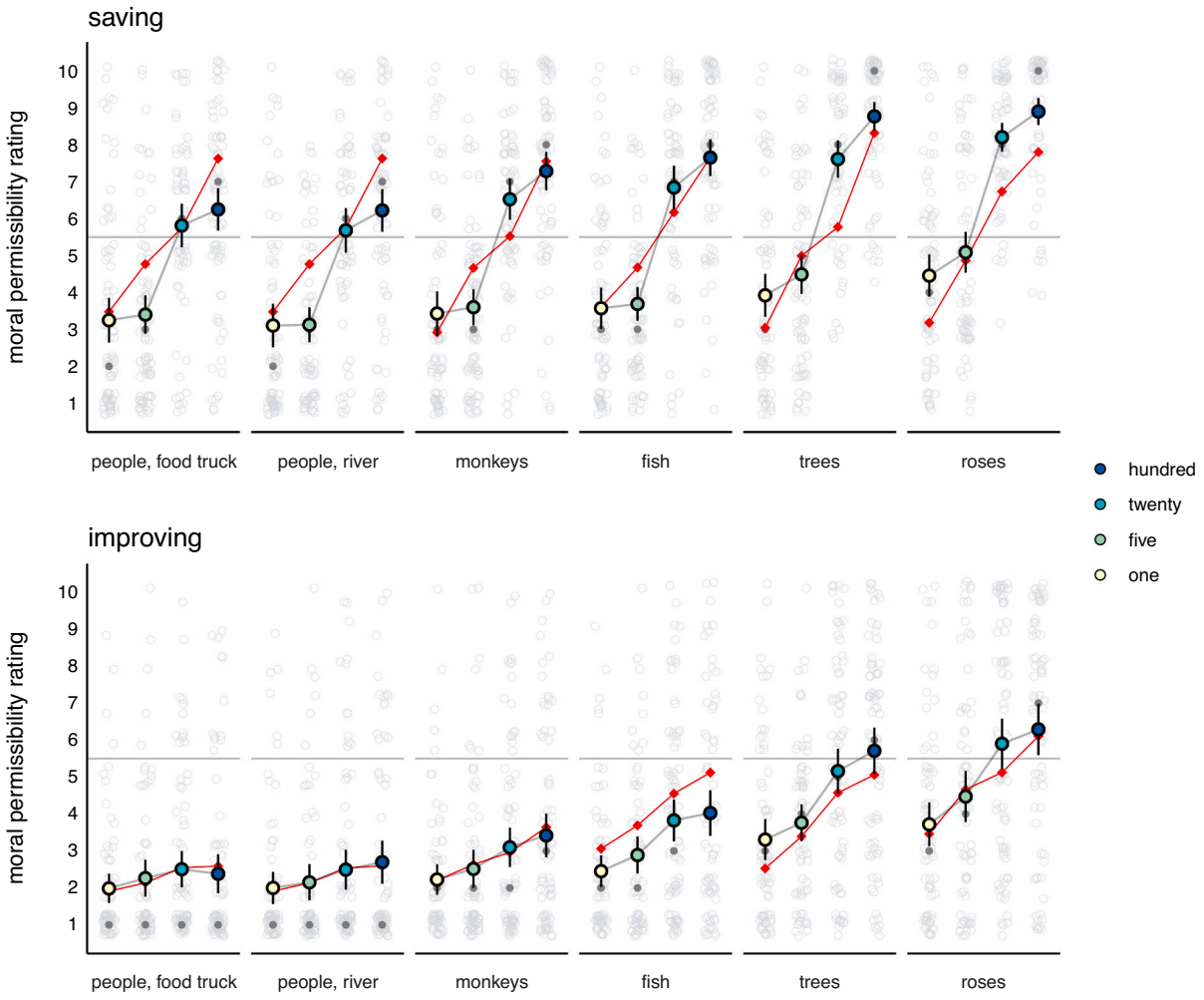


Fig. 6. Mean moral permissibility ratings (large points in blue colors) per condition in Experiment 2 (upper panel: saving conditions, lower panel: improving conditions). Error bars are 95% confidence intervals. Medians are displayed in dark grey, individual data points (jittered) in light grey. GSUM predictions (fitted to means) are shown in red. The light grey line indicates the scale midpoint. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Results of the overall ANOVA for Experiment 2. (*p*-values are Greenhouse-Geisser-corrected, degrees of freedom are unadjusted. Bonferroni-adjusting p-values for multiple testing did not change the results.)

Effect	*df*	*F*	*p*	η_p^2 [90% CI]
(Intercept)	1,599	3501.09	<0.001	
structure	1,599	192.34	<0.001	0.24 [0.20; 0.30]
number saved	3, 599	80.63	<0.001	0.29 [0.24; 0.33]
scenario	5,2995	224.49	<0.001	0.27 [0.25; 0.29]
structure:number saved	3, 599	19.33	<0.001	0.09 [0.05; 0.12]
structure:scenario	5,2995	6.84	<0.001	0.01 [0.01; 0.02]
number saved:scenario	15,2995	7.70	<0.001	0.04 [0.02; 0.04]
structure:number saved: scenario	15, 2995	0.41	0.937	<0.01

mixed 2 (structure: *saving* vs. *improving*, between-subjects) × 4 (number helped: hundred vs. twenty vs. five vs. one, between-subjects) × 6 (scenario: people/foodtruck vs. people/river vs. monkeys vs. fish vs. trees vs. roses, within subject) ANOVA, all main effects and two-way interactions were significant (all *p*'s < 0.001, see Table 2). To follow up on the differences between *saving* and *improving* cases (indicated by the main effect of structure and the two interactions involving structure), we conducted separate mixed ANOVAs for the two conditions. For the saving condition, we replicated the results from Experiment 1 with very similar effect sizes. People judged an action to be more permissible the more entities were saved compared to killed, $F(3, 299) = 86.79$, $p_{GG} < 0.001$, $\eta_p^2 = 0.47$ [0.40; 0.52], but again also differentiated between groups, with low permissibility ratings for the killing of people, higher permissibility ratings for animals, and the highest permissibility ratings for harming plants, $F(5,1495) = 90.28$, $p_{GG} < 0.001$, $\eta_p^2 = 0.23$ [0.20; 0.26]. Again, there was a small two-way interaction effect, $F(15,1495) = 3.61$, $p_{GG} < 0.001$, $\eta_p^2 = 0.03$ [0.01; 0.04], indicating that the number of saved compared to killed entities did not influence permissibility ratings equally for all groups (Bonferroni-adjusting *p*-values for multiple testing did not change the results). As in Experiment 1, contrasts revealed an overall positive linear trend in the moral permissibility ratings with increasing numbers of saved entities ($D = 2.55$, $t = 9.62$, $p < .001$), and this trend was stronger in the scenarios about fish ($D = 0.89$, $t = 2.37$, $p = .018$), trees ($D = 1.40$, $t = 3.72$, $p < .001$), and roses ($D = 1.12$, $t = 3.0$, $p = .003$) compared to people. The two people scenarios did not differ from each other ($D = 0.11$, $t = 0.30$, $p = .77$), and neither did the people and monkey scenarios ($D = 0.69$, $t = 1.84$, $p = .07$).[8]

The ANOVA for the *improving* condition confirmed that the number of affected entities also led to higher permissibility ratings in *improving* cases, although the effect was smaller than in the *saving* condition, $F(3,300) = 11.05$, $p_{GG} < 0.001$, partial $\eta^2 = 0.10$ [0.05; 0.15]. As in the saving condition, trade-offs among lives of people were seen as least permissible, followed by animals, and then plants, $F(5,1500) = 135.32$, $p_{GG} < 0.001$, $\eta_p^2 = 0.31$ [0.28; 0.34]. A small two-way interaction effect indicated that the influence of the number of improved entities did not affect moral judgments equally for all entities, $F(15,1500) = 4.39$, $p_{GG} < 0.001$, $\eta_p^2 = 0.04$ [0.02; 0.05] (these results did not change when Bonferroni-adjusting p-values for multiple testing). Follow-up contrasts showed that this time, there was no significant overall linear trend for the influence of numbers on permissibility ratings, but a linear trend emerged for the scenarios about fish ($D = 0.95$, $t = 2.41$, $p = .016$), trees ($D = 1.61$, $t = 4.08$, $p < .001$) and roses ($D = 1.72$, $t = 4.39$, $p < .001$), when compared to people.[9]

We generated predictions for the permissibility judgments in all experimental conditions using GSUM. Again, the model fit the data well. As in the previous study, we tested the fit of linear, exponential, and sigmoid functions. Exponential functions described the relationships best, and the fit was better for improving ($y = ax^b$, $a = 18.92$, $t_{22} = 7.35$, $p < .001$, $b = 1.12$, $t_{22} = 11.65$, $p < .001$, normalized RMSE = 0.11, Cragg & Uhler $R^2 = 0.90$) than for saving scenarios ($y = ax^b$, $a = 12.75$, $t_{22} = 9.62$, $p < .001$, $b = 1.25$, $t_{22} = 7.42$, $p < .001$, normalized RMSE = 0.17, Cragg & Uhler $R^2 = 0.76$). A similar fit is obtained overall when group medians were used as the criterion, with slightly better predictions of saving scenarios ($y = ax^b$, $a = 17.14$, $t_{22} = 8.62$, $p < .001$, $b = 1.72$, $t_{22} = 8.51$, $p < .001$, normalized RMSE = 0.13, Cragg & Uhler $R^2 = 0.82$), and slightly worse predictions of improving scenarios ($y = ax^b$, $a = 47.07$, $t_{22} = 3.82$, $p = .001$, $b = 1.90$, $t_{22} = 9.39$, $p < .001$, normalized RMSE = 0.12, Cragg & Uhler $R^2 = 0.85$).

The model predictions captured the patterns that we observed in the moral judgments. For improving scenarios, permissibility ratings and model predictions increased less steeply with higher numbers of entities benefitting from an action, compared to saving scenarios. Moreover, the model predictions reflected the differences between people, animals, and plants. The permissibility was generally lowest for people, and increased only very little with higher numbers of lives improved in this case. The predictions were higher for monkeys, trees, fish, roses (in this order), and also increased more steeply with numbers of lives improved for these groups.

To test GSUM against Cohen and Ahn's (2016) model, we again replaced all valuations of the dead states with zeroes, as their model solely took into account the valuations of the alive or intact states. The model was roughly equivalent to GSUM for the saving scenarios, regardless of whether means or medians were used as criterion (means: $y = ax^b$, $a = 16.63$, $t_{22} = 7.54$, $p < .001$, $b = 1.51$, $t_{22} = 7.70$, $p < .001$, normalized RMSE = 0.16, Cragg & Uhler $R^2 = 0.77$; medians: $y = ax^b$, $a = 24.62$, $t_{22} = 6.44$, $p < .001$, $b = 2.08$, $t_{22} = 8.56$, $p < .001$, normalized RMSE = 0.13, Cragg & Uhler $R^2 = 0.82$).[10] The model is not applicable for improving cases, as these cases require comparisons between more than just the two alive/intact states of entities. In the improving cases, three states are traded off against each other (dead, normal, improved), which is beyond the scope of the Cohen and Ahn model.

Interestingly, the predictions of GSUM were able to account for two patterns that might otherwise be attributed to deontological constraints. First, the model correctly predicted that in improving scenarios, acting was generally seen as impermissible. A possible account of this difference could have been that people regard causing death to merely improve other's lives as categorically impermissible, regardless of the extent of the benefit to one group. GSUM makes this prediction based on

[8] As in Experiment 1, there was also a significant negative cubic trend ($D = -0.95$, $t = 3.52$, $p < .001$) for the numbers factor (overall, no interactions with scenario). This trend is likely due to the fact that ratings increased more steeply between five and twenty than between the other numerical conditions. The trend analyses were not adjusted for multiple testing and should be regarded as exploratory. When Bonferroni-correcting for the number of trend tests (18 tests), only the following trends remain significant: the overall linear trend ($p < .001$), the overall cubic trend ($p = .008$), the interaction of the linear trend with the trees scenario ($p = .004$), and the interaction of the linear trend with the roses scenario ($p = .05$).

[9] No other trends for the numbers factor were significant. Trend analyses were not adjusted for multiple comparisons and should be regarded as exploratory. When Bonferroni-correcting for the number of trend tests (18 tests), only the following trends remain significant: the interaction with the trees scenario ($p < .001$), and the interaction with the roses scenario ($p < .001$).

[10] Only linear and exponential functions were compared for the relationship between the predictions by Cohen and Ahn's model and mean moral judgments, as sigmoid models did not converge here. Exponential functions described the relationship better for saving as well as improving scenarios and are therefore reported.

the fact that gains generally do not outweigh losses in improving cases, when the alternative state is normal. If a categorical constraint against acting in improving scenarios governed people's judgments, we should have observed equally low permissibility ratings in all numerical conditions and for all entities. We instead observed that permissibility ratings generally increased when larger numbers benefitted, suggesting that subjective utilities still influence permissibility judgments here. Second, this increase was weaker for higher-valued entities than for lower-valued entities, and not statistically detectable at all in scenarios about human lives. Again, this difference between species might be attributed to a deontological constraint shielding human lives from being traded off. Note however that our model predicts both the generally lower permissibility ratings for humans in improving scenarios, and the weak-to-absent increase of permissibility ratings with higher numbers in scenarios about human lives (see Fig. 6). GSUM makes these predictions based on the differences of the subjective utilities alone: Improving scenarios are generally fairly impermissible because here the losses (i.e., deaths) are not outweighed by the gains. However, they become gradually more permissible the larger the perceived gains are in relation to the perceived losses. Within the class of improving scenarios, acting is less permissible when people are concerned because losses are especially large for this group at all levels of the numerical manipulation, while at the same time the differences between normal and improved states (the gains) are more similar for all species groups (see Fig. 2).

5. Interindividual differences as a possible boundary condition

Based on the results we have described for the utility estimation data, the two experiments, and the fit between model predictions and data, we can derive additional hypotheses about subsets of participants for which better or worse correspondence between model predictions and data can be expected.[11] An inspection of the utility estimation data shows that participants differed in their use of the scales. This raises the question whether interindividual differences in the way the entities are valued may have generated noise that negatively affects the fit of our model. It is therefore interesting to test whether the predictions of GSUM change for different subsets of participants. We generated another set of predictions based on just the utilities of participants who assigned the minimal value of −1000 to any number of human deaths ($N = 66$). This corresponds to participants anchoring the scale at "dead people" = −1000, and determining the values of the other items from there. We also explored other anchors, such as "dead people = -1000 and improved people = 1000" (again for at least one of the numerical conditions). As for the relationship between the original GSUM predictions and the moral judgment data, we investigated the fit of several functions (linear, exponential, sigmoid), and we used both group means and medians as criterion.

The upshot of these analyses is that in four out of six cases, the best-fitting model based on the utilities of a homogeneous subset of participants fit the data better than the best-fitting model based on all participants' utilities (based on comparing normalized RMSE's, see Table 3). In two cases, the fit was identical, and there was only one case (improving, means as criterion) in which the predictions based on all participants' utilities fit the data slightly better. Thus, homogenizing the predictor variable improved model fit. Of course, these analyses should be regarded as exploratory, especially since the subgroups comprised just slightly more than between half and a third of participants in the utility estimation study.

A second focus of our analyses was on the minority of people who may have strict deontological constraints about intervening in a moral dilemma, even when more lives are saved than lost (cf. Thomson, 2008). GSUM's predictions will fail to describe the judgment of people who are insensitive to consequences in a moral dilemma. We used the data of Experiments 1 and 2 to estimate the upper bound of the proportion of such people. Typically, deontological constraints are applied to actions that harm or kill humans, not animals or plants. To use a lenient criterion, we thus determined the proportion of participants who thought that intervening was completely impermissible when human lives were at stake (rating = 1 on the scale ranging from 1 to 10), even though the ratio of lives saved compared to lost was favourable (conditions 100 vs. 10 and 20 vs. 10). 17% of participants in Experiment 1 and also 17% in Experiment 2 conformed to this criterion. Thus, 17% of the participants in our samples provided moral judgments that cannot be explained by GSUM. However, our experiments did not exhaust the space of possible outcome trade-offs. It may be the case that even though the threshold is higher for subjects classified as "deontologists"; they may ultimately waver in their judgment when outcome trade-offs in sacrificial dilemmas involve larger numbers of saved people than the "20 vs. 10" condition or even the "100 vs. 10" condition (i.e., disaster cases) (see Wiegmann & Waldmann, 2014, Experiment 5). Since we did not measure moral judgments in such disaster scenarios, we take the proportion of 17% to be an estimate of the upper bound of the true proportion of "deontologists" in our sample.

6. General discussion

It is generally undisputed that the foreseen outcomes of an action matter for its moral evaluation. Psychological theories of moral judgment acknowledge this, but how people reason about outcomes in morally charged situations has received little attention in the literature. Initially, one might be tempted to speculate that people do simple ordinal comparisons. When acting in a life-and-death dilemma saves more lives than not acting, the outcome trade-off may be registered as favourable, and it will factor into the action's global evaluation as a "pro" reason. However, such a simple notion of outcome comparisons quickly runs into problems, for example when different kinds of entities are compared, say, the life of one person against the lives of two fish, or against inanimate objects. A "common currency" is required. Subjective utility is a standard concept in decision theory, which has only recently been brought to bear on morally charged judgments and decisions (Cohen & Ahn, 2016).

In the present research we have shown that the contrast between subjective utilities of outcomes of an action, in comparison to inaction, predicts people's judgments of moral permissibility in different types of moral scenarios involving trade-offs between multiple outcomes. The contrasts also explain the different moral evaluation of dilemmas compared to cases in which one group's state is merely improved at another's expense. We observed a relatively high tendency to make trade-offs in life-and-death dilemmas, whereas the trade-off curves were flatter in improving situations. In these cases we discovered that subjects were more reluctant to trade-off a mere improvement against death when humans were involved compared to animals. For plants the willingness to make trade-offs was strongest.

While previous studies only assessed judgments about what participants would personally do, we demonstrated that our generalized subjective-utilitarian model (GSUM) can predict moral judgments about other people's actions in classic life-and-death dilemmas as well as for other multiple-outcome scenarios. In classical life-and-death dilemmas, GSUM's predictions converge with the predictions of earlier models (Cohen & Ahn, 2016). It apparently makes little difference whether the model considers negative valuations of dead states or assigns them a value of zero. However, as demonstrated in our improving scenarios, moral dilemmas do not only arise when life versus death is at stake, they may also require the considerations of different states of entities who remain alive. While the model of Cohen and Ahn (2016) is only applicable to situations that can be reduced to a comparison between the positive values of different entities (implicitly assuming that death or destruction can be represented by a constant, for example, zero), GSUM,

[11] We thank an anonymous reviewer for suggesting these additional analyses.

Table 3
Overview of fit measures for GSUM predictions based on the utility estimates of all participants ("full set", N = 123), and based on subgroups of participants who used the valuation scale more similarly to each other ("subsets", d1 = utilities of N = 66 participants who valued dead people at − 1000 in any numerical condition, d2 = utilities of N = 42 participants who valued dead people at − 1000 and improved people = 1000 in any numerical condition).

			GSUM full set			GSUM subsets			
Exp.	Condition	Criterion	Model	(Pseudo-)R^2	NRMSE	data		(Pseudo-)R^2	NRMSE
Exp1	Saving	Means	Exponential	0.77	0.16	d2	Linear	0.79	0.14
		Medians	Linear	0.76	0.16	d1	Sigmoid	0.82	0.14
Exp2	Saving	Means	Exponential	0.76	0.17	d1	Exponential	0.79	0.16
		Medians	Exponential	0.82	0.13	d2	Linear	0.83	0.13
	Improving	Means	Exponential	0.90	0.11	d1	Linear	0.79	0.13
		Medians	Exponential	0.85	0.12	d1	Sigmoid	0.86	0.12

due to its sensitivity to all relevant actual, hypothetical or counterfactual states of entities, can also analyse other dilemmas. In our improving scenarios, for example, these states were dead, normal, and improved, but other cases can be construed. Such situations are beyond the scope of Cohen and Ahn's (2016) model.

6.1. Can deontological response patterns be explained by differences in subjective utilities?

We have seen that for improving scenarios, GSUM correctly predicts lower permissibility ratings for trade-offs involving the lives of higher-valued entities, such as people or monkeys compared to trade-offs involving lower-valued entities, such as trees, fish, or roses. This pattern makes intuitive sense, and indeed it was apparent in participants' moral judgments. These evaluations can also be predicted by psychological variants of deontological ethics, which ascribe special rights to humans and not to other forms of live – with some deontological positions even claiming that human lives may not be traded off at all (see Alexander & Moore, 2016, for an overview of variants of deontological ethics). However, we have seen that these evaluations do not necessarily require positing gradually weakening deontological constraints on harming. GSUM can explain them without resorting to deontological ethics. Specifically, a person being dead is considered to be much worse than animals or plants being dead, while the difference between normal and improved states is more similar for all groups. This explains why scenarios in which the improvement of one group is traded off against the death of others received constantly low permissibility predictions by GSUM when people were concerned, compared to other types of entities. Thus, our findings show that psychologically the valuation of outcomes alone can account for some intuitions that otherwise might be interpreted as supporting deontological ethics.

6.2. The moral status of animals and plants

Treating animals and plants as moral entities is a quite recent development in Western philosophy. Kant (1974) made a sharp distinction between humans who have rights and must not be treated as means and animals who are largely outside the realm of morality (Korsgaard, 2018). In the meantime, both consequentialist (Singer, 1975) and nonconsequentialist (Korsgaard, 2018) philosophers have acknowledged the moral worth of animals. This development seems to have been partly triggered by an increasing awareness that animals are sentient beings who have emotions and can feel pain. In psychology, there has been increased interest in the psychological foundations of speciesism in the past years (Caviola et al., 2021; Caviola, Everett, & Faber, 2019; Crimston, Bain, Hornsey, & Bastian, 2016; Goodwin & Benforado, 2015; Horta, 2010).

One way to explain people's greater readiness to approve of trade-offs between animals compared to human lives could be that in both cases observers realize that acting leads to a gain compared to inaction, for example, because a larger number of lives are saved or because some lives are vastly improved. In the case of humans however, additional deontological considerations are activated, for example the intuition that humans have special rights not to be sacrificed in such a way. These deontological constraints then reduce people's willingness to morally approve of the action. Our results, in contrast, suggest that effects of speciesism may manifest much earlier in the assessment of values. Specifically, our results show that especially in improving scenarios, losses are not considered as equally outweighed by gains in trade-offs among members of different species, even when the objective numbers are constant. An interesting avenue for future research will be to investigate why people value the lives of different species so differently. Recent research asking subjects to assess the cognitive and suffering capacity of humans versus animals found that even when these features were matched, people still granted special consideration to human lives that were not extended to other species (Caviola et al., 2021).

Less is known about where the moral value of plants comes from. Our utility study shows that they are valued less than animals but still show intuitively plausible value differences. Although some people believe in the sentience of trees (e.g., Wohlleben, 2017), we believe that a more plausible source of the valuation of plants is their relation to human interests. Roses, for example, are aesthetically pleasing and it pains us if we see a bulldozer running over them. Moreover, there is an increasing awareness that our well-being is connected to nature and the climate, and we realize that destroying the rain forest, for example, has widespread consequences for our lives.

While species differences may to some extent be explained by differences of the associated subjective utilities, there are also established patterns in moral judgment that GSUM cannot capture in its current form. For example, when keeping outcomes constant, it is generally seen as morally worse to cause harm intentionally, by action rather than omission, as a means rather than a side effect, and by so-called personal force rather than more indirectly (for overviews see May, 2018; Waldmann et al., 2012). It has been suggested that these factors should be subsumed under the concept of "agential involvement" (May, 2018). The more involved an agent is in bringing about a harm, by any of the ways listed above and possibly others, the more severe our moral judgment tends to be.

6.3. Conclusion

In its current version, we regard GSUM as the formalization of one important component in a larger network of factors that jointly produce moral judgment. It constitutes an outcome formalism that can be implemented within different psychological theories of moral judgment. Even though we have not focused on these larger issues in this article, we think that the causal model framework may be best suited for this task. Causal models connect outcomes of different valences to the actions that produce them, and these actions can in turn be connected to mental states and character dispositions (Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021; Sloman, Fernbach, & Ewing, 2009; Waldmann, 2017; Waldmann, Wiegmann, & Nagel, 2017). Given that a central component of causal models are outcomes generated by actions, a mechanism computing trade-offs between outcomes is central. GSUM

could serve as the mechanism that compares the alternative outcomes when different causal paths are instantiated. In sum, we have demonstrated that the subjective utilities of outcomes predict genuinely moral judgments about multiple-outcome structures, not just personal preferences between possible courses of action. A central future goal will be to embed the trade-off component in a more complex theory that is sensitive to other relevant factors of moral judgments, such as intentionality and causality.

CRediT authorship contribution statement

Neele Engelmann: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Michael R. Waldmann:** Conceptualization, Writing – review & editing, Supervision.

Acknowledgements

We have no known conflicts of interests to disclose. Findings related to this project were presented at the 2019 Annual Meeting of the Cognitive Science Society in Montreal, Canada (Engelmann & Waldmann, 2019).

We would like to thank Alex Wiegmann for helpful comments. We also thank Michał Białek and three anonymous reviewers.

References

Alexander, L., & Moore, M. (2016). Deontological ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., … Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64.

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.

Białek, M., Turpin, M. H., & Fugelsang, J. A. (2019). What is the right question for moral psychology to answer? Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science, 30*(9), 1383–1385.

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.

Caviola, L., Everett, J. A., & Faber, N. S. (2019). The moral standing of animals: Towards a psychology of speciesism. *Journal of Personality and Social Psychology, 116*(6), 1011.

Caviola, L., Kahane, G., Everett, J. A. C., Teperman, E., Savulescu, J., & Faber, N. S. (2021). Utilitarianism for animals, Kantianism for people? Harming animals and humans for the greater good. *Journal of Experimental Psychology: General, 150*, 1008–1039.

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General, 145*(10), 1359.

Crimston, D., Bain, P. G., Hornsey, M. J., & Bastian, B. (2016). Moral expansiveness: Examining variability in the extension of the moral world. *Journal of Personality and Social Psychology, 111*(4), 636.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences, 17*(8), 363–366.

Cromley, A. R., & Cohen, D. (2019). *Subjective values theory: The psychophysics of psychological value*. https://doi.org/10.31234/osf.io/wfd5s

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380.

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

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science, 35*(6), 1052–1075.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science, 17*(12), 1082–1089.

DeBruine, L. (2020). *faux: simulation for factorial designs*. Retrieved from. https://doi.org/10.5281/zenodo.2669586.

Engelmann, N., & Waldmann, M. R. (2019). Moral reasoning with multiple effects: Justification and moral responsibility for side effects. In *Proceedings of the 41st meeting of the cognitive science society* (pp. 1703–1709). Austin, TX: Cognitive Science Society.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Feltz, A., & May, J. (2017). The means/side-effect distinction in moral cognition: A meta-analysis. *Cognition, 166*, 314–327.

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

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Thousand Oaks: Sage.

Goodwin, G. P., & Benforado, A. (2015). Judging the goring ox: Retribution directed toward animals. *Cognitive Science, 39*(3), 619–646.

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. New York: Penguin.

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

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.

Grothendieck, G. (2013). nls2: Non-linear regression with brute force. Retrieved from https://CRAN.R-project.org/package=nls2.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language, 22*(1), 1–21.

Horta, O. (2010). What is speciesism? *Journal of Agricultural and Environmental Ethics, 23*(3), 243–266.

Kahane, G., & Shackel, N. (2010). Methodological issues in the neuroscience of moral judgement. *Mind & Language, 25*(5), 561–582.

Kant, I. (1974). *Anthropology from a pragmatic point of view. Translated by Mary Gregor*. The Hague: Martinus Nijhoff.

Kassambara, A. (2019). ggpubr: ggplot2 based publication ready plots. Retrieved from https://CRAN.R-project.org/package=ggpubr.

Kelley, K. (2019). MBESS: the MBESS R package. Retrieved from https://CRAN.R-project.org/package=MBESS.

Korsgaard, C. M. (2018). *Fellow creatures: Our obligations to the other animals*. Oxford University Press.

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science, 4*(1).

Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology, 129*, 101412.

Lawrence, M. A. (2016). ez: Easy analysis and visualization of factorial experiments. Retrieved from https://CRAN.R-project.org/package=ez.

Mangiafico, S. (2016). Summary and analysis of extension program evaluation in R (version 1.18.8). Retrieved from https://rcompanion.org/handbook/F_08.html.

Mangiafico, S. (2019). rcompanion: Functions to support extension education program evaluation. Retrieved from https://CRAN.R-project.org/package=rcompanion.

May, J. (2018). *Regard for reason in the moral mind*. Oxford University Press.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143–152.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press.

Nadelhoffer, T., & Feltz, A. (2008). The actor–observer bias and moral intuitions: Adding fuel to Sinnott-Armstrong's fire. *Neuroethics, 1*(2), 133–144.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2020). nlme: Linear and nonlinear mixed effect models. Retrieved from https://CRAN.R-project.org/package=nlme.

Plunkett, D., & Greene, J. D. (2019). Overlooked evidence and a misunderstanding of what trolley dilemmas do best: Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science, 30*(9), 1389–1391.

R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.R-project.org/.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*(5), 347–356.

Royzman, E., & Hagan, J. P. (2017). The shadow and the tree: Inference and transformation of cognitive content in psychology of moral judgment. In J.-F. Bonnefon, & B. Trémolière (Eds.), *Moral inferences* (pp. 64–82). London: Routledge.

RStudio Team. (2016). RStudio: Integrated development environment for R. Boston, MA. Retrieved from http://www.rstudio.com/.

Schein, C. (2020). The importance of context in moral judgments. *Perspectives on Psychological Science, 15*(2), 207–215.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron, 67*(4), 667–677.

Singer, P. (1975). *Animal liberation: A new ethic for our treatment of animals*. New York: HarperCollins.

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2009). Causal models: The representational infrastructure for moral judgment. *Psychology of Learning and Motivation, 50*, 1–26.

Slovic, P. (2007). "If I look at the mass I will never act": Psychic numbing and genocide. *Judgment and Decision making, 2*(2), 1–17.

Soter, L. K., Berg, M. K., Gelman, S. A., & Kross, E. (2021). What we would (but shouldn't) do for those we love: Universalism versus partiality in responding to others' moral transgressions. *Cognition, 217*, 104886.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*(1), 76–105.

Steele, K., & Stefánsson, H. O. (2020). Decision theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9*(2), 164.
Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology, 4*, 250.
Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*(6), 1395–1415.
Thomson, J. J. (2008). Turning the trolley. *Philosophy & Public Affairs, 36*(4), 359–374.
Torchiano, M. (2020). effsize: Efficient effect size computation. Retrieved from https://CRAN.R-project.org/package=effsize.
Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.
Waldmann, M. R. (Ed.). (2017). *The Oxford handbook of causal reasoning*. New York: Oxford University Press.
Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak, & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 364–389). New York: Oxford University Press.
Waldmann, M. R., Wiegmann, A., & Nagel, J. (2017). Causal models mediate moral inferences. In J.-F. Bonnefon, & B. Trémolière (Eds.), *Moral inferences* (pp. 37–55). London: Routledge.
Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*(12), 1–20. Retrieved from http://www.jstatsoft.org/v21/i12/.
Wickham, et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686.
Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition, 131*(1), 28–43.
Willemsen, P., & Reuter, K. (2016). Is there really an omission effect? *Philosophical Psychology, 29*(8), 1142–1159.
Wohlleben, P. (2017). *The hidden life of trees: What they feel, how they communicate*. Glasgow: William Collins.
Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120*(2), 202–214.
Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News, 2*(3), 7–10. Retrieved from https://CRAN.R-project.org/doc/Rnews/.

Appendix C Engelmann & Waldmann (2021)

A Causal Proximity Effect in Moral Judgment

Neele Engelmann (neele.engelmann@uni-goettingen.de)

Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology
University of Göttingen, Germany

Abstract

In three experiments (total $N = 1302$) we investigated whether causal proximity affects moral judgments. We manipulated causal proximity by varying the length of chains mediating between actions and outcomes, and by varying the strengths of causal links. We demonstrate that moral judgments are affected by causal proximity with longer chains or weaker links leading to more lenient moral evaluations. Moreover, we identify outcome foreseeability as the crucial factor linking causal proximity and moral judgments. While effects of causal proximity on moral judgments were small when controlling for factors that were confounded in previous studies, knowledge about the presence of causal links substantially alters judgments of permissibility and responsibility. The experiments demonstrate a tight coupling between causal representations, inferences about mental states, and moral reasoning.

Keywords: Causal Reasoning; Moral Judgment; Causal Proximity; Causal Strength; Causal Chains

Suppose someone is contemplating an action that, as an unintended side effect, could cause serious harm to another person. For example, imagine a doctor in an emergency situation who has to decide which one of two live-saving drugs to administer to an unconscious patient. Both drugs will have the same stabilising effect on the patient, but both also have a risk of causing blood clots as a side effect. The exact probabilities are unknown, but the doctor remembers that when drug A causes blood clots, it happens via several intermediate steps. Drug A first needs to cause a number of intermediate events in the body before blood clots can develop. By contrast, when drug B causes blood clots, it does so directly. Which drug should the doctor choose?

If you prefer drug A, your preference may be an instance of a so-called causal proximity effect. Causal proximity refers to the position of a target cause (such as an action) relative to a target effect (such as a harmful outcome). A cause is traditionally called more proximate when fewer intermediate events connect it to a target outcome. Arguably, a cause may also be perceived as more proximate when its link to the effect is stronger, as spatio-temporal co-occurrence and causal strength tend to be correlated. We will explore both facets of proximity here.

It has been suggested that the length of a causal chain matters for our moral evaluations of agents and their actions. For example, Sloman, Fernbach, and Ewing (2009) note: "Actions that are connected to bad outcomes through fewer intermediate causes are more blameworthy" (p.11). In our example, administering the drug that can cause blood clots directly would thereby be predicted to be morally worse than administering the drug which can cause the same outcome via several intermediate steps. Similar effects can be expected when the strength of causal links is increased.

Are causal proximity effects rational?

Proximity effects have sometimes been described as biases (e.g., Johnson & Drobny, 1985). But this does not have to be the case. Proximity effects can naturally arise from the way in which moral reasoning about agents, actions, and outcomes is mediated by causal models (Waldmann, Wiegmann, & Nagel, 2017; Sloman et al., 2009).

In a causal model framework (see Waldmann, 2017; Sloman, 2005, for overviews), representing a chain of causally connected events generally means representing a number of events that are connected by probabilistic links. If each event in the chain actually occurs, there is a certain probability that the next event in the chain will occur as well. Say that in our example, there are five intermediate links between administering drug A and the development of blood clots, each of them with a probability of 0.15 conditional on its direct cause. Then $p(\text{blood clots}|\text{drug A}) = 0.15^5 = 0.00008$. For drug B, there is just one probabilistic link with a strength of 0.15, thus $p(\text{blood clots}|\text{drug B}) = 0.15$. In such a case, administering drug B would thus be much more likely to cause harm than administering drug A.

This calculation of course rests on the assumption that all single links, be it the direct relation or a component of the chain, are roughly equally strong[1] and that there are no alternative causes of the events in the chain. Nothing in our introductory example suggests that this needs to be the case. However, research has shown that a chain representation can indeed trigger the impression of a lower probabilistic dependency between a target cause and effect, and that this effect may be produced by people assigning roughly constant strength priors to verbally instructed probabilistic links (Stephan, Tentori, Pighin, & Waldmann, 2021; Bes, Sloman, Lucas, & Raufaste, 2012).

If people perceive a lower conditional probability of harm given action A than given action B, it naturally follows that

[1] Or, at least, that the links in the chain are sufficiently weak to lower $p(\text{outcome}|\text{action})$ relative to the direct relation.

action A is morally preferable. This should hold prospectively (before acting, as in our introductory example), but it may also be true for retrospective moral evaluations. For instance, an agent may be deemed less morally responsible or blameworthy for harm when their action produced the outcome via a chain rather than directly (Sloman et al., 2009). We posit that such proximity effects on *moral* judgments are mediated by the agents' foreseeability of the harmful consequences (see Lagnado & Channon, 2008, Kirfel & Lagnado, 2020, for effects of foreseeability in other contexts). If an action causes harm via a longer chain, the harm is seen as less likely and thus less foreseeable than in a direct relation (assuming roughly equal strength of causal links), justifying a more lenient moral evaluation of action and agent. Thus, whenever the assumption of a lower probability of harm in a chain is justified (e.g., roughly constant link strengths), proximity effects in causal and moral judgments are not a bias. Direct causal relations can also vary in strength, which we view as a different way of manipulating proximity. Again, we predict a harsher moral evaluation of action and agent the stronger the causal link between their action and a harmful outcome is.

Proximity effects in causal and moral reasoning

Surprisingly, cases like our example have rarely been investigated in the context of moral judgment. Research on causal and moral judgments about chains involving human actions has largely focused on comparisons *within* chains. For example, a debate has revolved around the question whether the first or the last element in a causal chain is selected as "the" (main or most important) cause of a final outcome, and how such judgments are affected by features of causes (such as being intentional actions vs. physical events), or by how much they raise the probability of the outcome (Lagnado & Channon, 2008, Spellmann, 1997, Hilton, McClure, & Sutton, 2010).

Our focus, in contrast, are comparisons *between* two causal chains with the same start (an action) and end (a harmful outcome), but a different number of intermediate events (none vs. several). We found just one study that directly investigated such cases. Johnson and Drobny (1985) presented participants with a case in which a truck driver forgets to replace a safety pin in the steering column of his truck. In the "simple chain" condition, the steering fails and results in an accident. Subsequently, gasoline spills and ignites causing a house to burn down. In the "complex chain" condition, the gasoline first pours into a river, floats across it, ignites grass on the other side, then a field, and finally also burns down the house. In the condition with the longer chain, participants considered the truck driver to be less liable for the damage to the house, and they also indicated that he could foresee the outcome to a lesser extent. However, Johnson and Drobny were interested in legal rather than moral judgments, described negligent omissions instead of actions, and provided their participants with extensive jury instructions. Moreover, the two conditions vary in several confounded aspects. The complex condition presents a chain whose elements are both spatially and temporally more extended than the simple condition. Plus, background knowledge or assumptions about the described events may have had an influence. If participants for example assigned a very low probability to burning gasoline floating across a river, the obtained effects may have been produced by the perception of one very weak but necessary link, instead of being generated by the chain representation as such.

In the moral domain, so-called deviant causal chains have been shown to attenuate judgments of blame (Pizarro, Uhlmann, & Bloom, 2003). In these chains, an actor brings about an intended harmful outcome, but in an unexpected and unusual way. Coincidentally, the described deviant chains are often also longer than their "regular" counterparts. Moreover, in deviant chains foreseeability is often altered because the agent achieves the goal in an unforeseeable fashion. An example would be the case of an unpractised gunman who intends to shoot someone. His shot misses the victim, but startles a herd of pigs that trample the victim to death (cf. Davidson, 2001, p.72). Thus, while these studies investigate moral judgments, it is not clear whether their results are due to length, foreseeability, or deviancy of the chains.

Experiment 1: Varying Chain Length

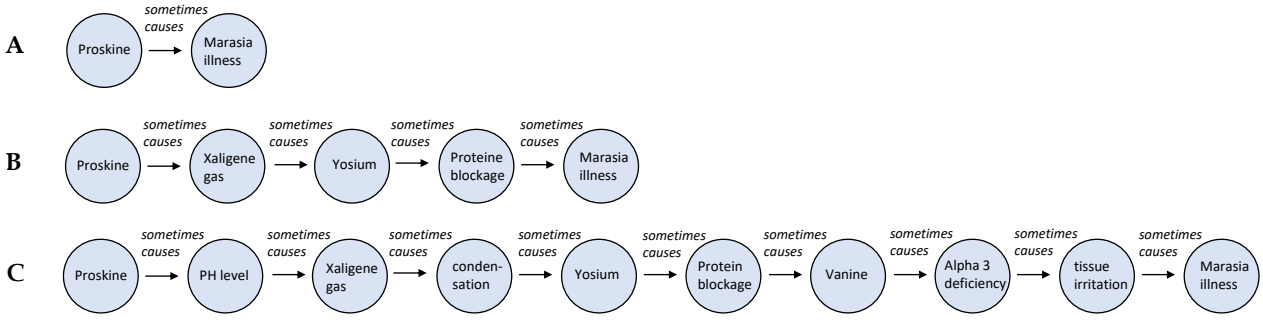
The aim of this experiment was to test for causal proximity effects in chains in a controlled setting, with little to no background knowledge about the events. If people assigned strength priors to links that are roughly constant, we should observe a lower estimated $p(\text{outcome}|\text{action})$ in chains compared to direct relations, along with a more lenient moral evaluation of action and agent in chains. We asked participants to morally evaluate action and agent both prospectively ("is it okay to act?") and retrospectively ("to what extent is the agent morally responsible for the harmful outcome?"). Unlike Johnson and Drobny, we used artificial materials that did not draw on prior knowledge about causal strength, or spatial and temporal relations. The strengths of the links were instructed using verbal labels suggesting equal link strengths (see Figure 1 for an illustration).

Methods

Design, Material and Procedure We created three cover stories about agents causing undesired harm to another person.[2] In each cover story, there was a chain version (three intermediate events and four probabilistic links between action and outcome) and a direct version (no intermediate events and just one probabilistic link). Each participant saw the direct version of one cover story, and the chain version of a different story (in random order). In total, there were six possible Latin square combinations of cover story and structure. This

[2] see https://osf.io/85s23/ for material, data, and code of analyses and figures for all studies reported in this paper. All analyses were conducted and all figures created using R (R Core Team, 2020) and RStudio (RStudio Team, 2020), as well as the following packages: *effsize* (Torchiano, 2020), *ez* (Lawrence, 2016), *ggpubr* (Kassambara, 2020), *MBESS* (Kelley, 2020), *reshape2* (Wickham, 2007), and the *tidyverse* (Wickham et al., 2019).

Figure 1: Example illustrations for direct relations (A), chains in Experiment 1 (B), and chains in Experiment 3 (C).



and all following experiments were implemented online using Unipark Questback.

In the beginning of each story, generic information about the relationship between action and outcome was presented. Here's an example: "A group of scientists is investigating the effects of exposure of to a certain chemical called Proskine. In their lab studies, they found that when Proskine is produced and stored, the following mechanism can unfold: Exposure to Proskine sometimes causes Marasia illness, a new and severe respiratory condition." In the chain condition, the second part of this story read: "Proskine sometimes causes Xaligene gas to develop in its environment. When Xaligene gas develops, it sometimes reacts and causes another chemical, Yosium, to form as well. When Yosium is present, it sometimes causes certain proteins in the body to be blocked upon exposure. When these proteins are blocked, this sometimes causes Marasia illness, a new and severe respiratory condition." On the same page, an illustration of the causal structure was provided, depicting the cited events as nodes, with arrows between them representing the causal links. The arrows were labelled with "sometimes causes" (see Figure 1).

After the generic information, we presented participants with the case of an agent who plans to carry out the action in question (in the example: creating and storing Proskine). The agent was always described as aware of the information from the previous page, but not desiring the negative outcome (in the example, the agent is a chemist who needs to create and store Proskine for research). The stories also mentioned the presence of potential victims of the harmful action. In the chemical scenario, we stated: "The lab is shared with several colleagues". Before giving any information about the occurrence of the harmful outcome, we asked participants to answer the following *prospective* moral question: "From a moral point of view, is it okay for [agent] to [perform the target action]?" Ratings were given on a scale ranging from 1 ("not at all") to 10 ("fully"). On a subsequent page, the actual occurrence of the negative outcome was described (in the example: a colleague in the same lab develops Marasia illness), and participants were asked to indicate the extent to which the agent was morally responsible for this outcome (i.e., a retrospective moral evaluation) on an identical scale. After moral judgments for both cases were recorded, the cases were presented anew and participants were asked to estimate the probability of the harmful outcome given the action. Answers were given on a slider ranging from 0 to 100%.

We predicted the following pattern of results: 1) the action should be seen as more allowed ("okay") in the chain condition compared to the direct condition, 2) the agent should be held less morally responsible for the outcome in the chain compared to the direct condition, and 3) $p(\text{outcome}|\text{action})$ should be estimated as lower in the chain compared to the direct condition.

Participants To achieve 90% power for observing all three effects at a minimum effect size of $d = .20$ each, we planned for a power of 97% for each of three one-sided t-tests ($0.97^3 \approx 0.91$ power to detect all three effects). This resulted in a required sample size of 300 participants. We recruited 304 participants on the platform *prolific.co*. Inclusion criteria (identical for all further experiments) were being a native English speaker, not having participated in previous studies using similar material, an acceptance rate of at least 90% of previous tasks on the platform, and not completing the survey via smartphone. Participants received a compensation of £0.45 for an estimated four minutes of their time. Five participants were excluded due to failing a simple attention check[3], leaving data of 299 participants for the analyses ($M_{age} = 34$, $SD_{age} = 12.1$, 63% women, 37% men, <1% no answer).

Results and Discussion

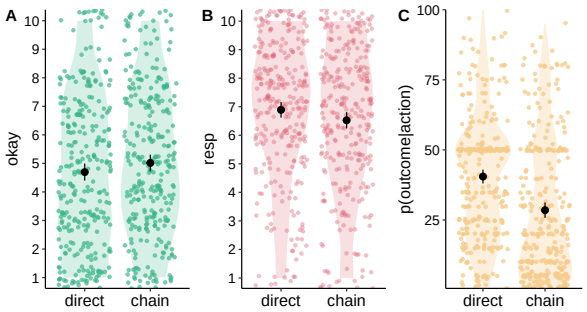
Figure 2 shows the results for all three measures. For the two moral questions, we observed significant, albeit small effects in the predicted directions (*okay*: $M_{direct} = 4.70$, $SD_{direct} = 2.62$, $M_{chain} = 5.02$, $SD_{chain} = 2.55$, $t_{298} = 1.96$, $p = .025$, $d = 0.12$ [0; 0.25]; *moral responsibility*: $M_{direct} = 6.89$, $SD_{direct} =$

[3] "If Peter is taller than Alex, and Alex is taller than Max, who is the shortest among them?" This attention check was used in all studies reported here.

2.39, M_{chain} = 6.53, SD_{chain} = 2.51, t_{298} = 2.41, p = .008, d = 0.15 [0.03; 0.27]). For the causal measure on the other hand, we observed a medium-sized effect in the predicted direction, M_{direct} = 40.51, SD_{direct} = 21.7, M_{chain} = 28.51, SD_{chain} = 24.19, t_{298} = 10.61, $p < .001$, d = 0.52 [0.42; 0.62]. No corrections of p-values were applied (although, given our conjunctive hypothesis, we could have increased the alpha-level per test).

Thus, while both kinds of moral judgment were indeed more lenient when a longer chain was described between action and outcome, the effects were relatively small. However, participants clearly perceived a difference in the strengths of the causal relations connecting action and outcome between the direct and the chain conditions. In chains, outcomes were estimated as less likely to occur than in direct relations. A possible explanation is that while p(outcome|action) matters for moral judgments, a relatively large difference is required. We test this hypothesis in the next experiment by directly manipulating causal strength.

Figure 2: Mean ratings for whether it is okay to act (A), agents' moral responsibility (B), and p(outcome|action) (C) per structure condition in Experiment 1. Error bars are 95% CIs.



Experiment 2: Varying Strength

In Experiment 1, we compared a direct probabilistic causal relation with a chain that contained several probabilistic links of equal strength, which entails lower overall strength in the chain than in the direct condition given equal link strengths. Thus, both the causal strength of the relation between action and outcome and chain length was varied. In Experiment 2, we focused only on direct causal relations while manipulating their strength. Moreover, strength is conveyed more saliently here by presenting numeric values.

Methods

Design, Material and Procedure We varied p(outcome|action) in three levels: .30, .60, and .90. The manipulation was delivered within subject. The cover stories were otherwise identical to the ones in Experiment 1. Each participant saw each link strength in the context of a different cover story (in random order), in one of three possible Latin square combinations.

Instead of learning about direct relations versus chains, participants in this experiment were presented with generic information about p(outcome|action) for each case. To convey strength information, we presented relative frequencies. For the "chemical" example, the instruction read: "A group of scientists is investigating Marasia illness, a new and severe respiratory condition. They suspected that it may be related to exposure with Proskine, a newly developed chemical that is sometimes used in pharmaceutical labs. The scientists therefore reviewed the health records of 1000 employees of pharmaceutical companies who have been in contact with Proskine. For comparison, they also reviewed the records of 1.000 employees who do the same job, but have not been in contact with this specific chemical. These are their results: Of the 1000 people who have been in contact with Proskine, [300/600/900] contracted Marasia illness. Of the 1000 people who have not been in contact with Proskine, no one contracted Marasia illness." On the subsequent pages, the task proceeded exactly as in Experiment 1, with identical measures.

We predicted the following pattern of results: 1) The agent's action should be assessed as more allowed ("okay") the less likely its negative effect (.30 > .60 > .90), and 2) the agent should be held less morally responsible for the negative outcome the less likely it was to result from their action (.30 < .60 < .90). We also expected participants to accurately infer p(outcome|action) from the presented numbers, which can be seen as a manipulation check in this case.

Participants We aimed for a sample size of 292 valid participants in this experiment. In three one-way, repeated-measures ANOVAs, this will yield a power of 91% to detect a small effect (η_p^2 = .01) on all measures. We invited 300 participants to take part in the experiment. Participants received a compensation of £0.60 for an estimated six minutes of their time. Nine participants were excluded due to failing a simple attention check, leaving data of 291 participants for the analyses (M_{age} = 31.59, SD_{age} = 11.47, 59% women, 39% men, 2% another identity or no answer).

Results and Discussion

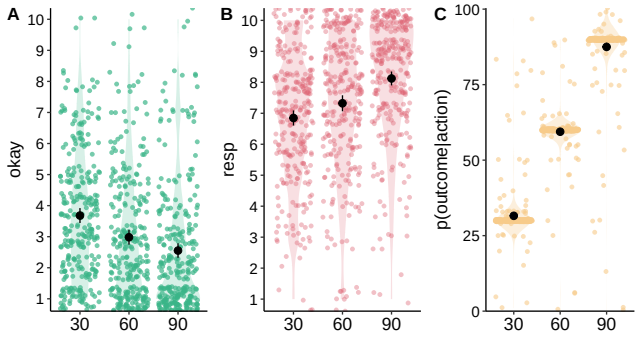
See Figure 3 for results. Prospectively, actions were regarded as more permissible ("okay") the lower the probability of harm was ($M_{30\%}$ = 3.68, $SD_{30\%}$ = 2.09, $M_{60\%}$ = 2.98, $SD_{60\%}$ = 2.07, $M_{90\%}$ = 2.55, $SD_{90\%}$ = 2.02, $F_{2,580}$ = 29.39, $p < .001$, η_p^2 = .09 [0.06; 0.13][4], which is confirmed by a negative linear trend in group means (t_{870} = -6.60, $p < .001$), and no detectable quadratic trend (t_{870} = 0.91, p = .365). Retrospectively, agents were held less morally responsible for harm the weaker the probabilistic relation between their action and the outcome was ($M_{30\%}$ = 6.85, $SD_{30\%}$ = 2.11, $M_{60\%}$ = 7.32, $SD_{60\%}$ = 2.24, $M_{90\%}$ = 8.12, $SD_{90\%}$ = 1.95, $F_{2,580}$ = 43.34, p

[4]We report 90% confidence intervals for all η_p^2, see Steiger (2004))

$< .001$, $\eta_p^2 = 0.13$ [0.09; 0.17]). There was a positive linear trend in group means $t_{870} = 7.32$, $p < .001$), and no detectable quadratic trend ($t_{870} = 1.08$, $p = .28$). Finally, responses to the query about p(outcome|action) confirmed that the strengths of probabilistic relations between action and outcome were accurately inferred ($M_{30\%} = 31.56$, $SD_{30\%} = 9.31$, $M_{60\%} = 59.41$, $SD_{60\%} = 7.68$, $M_{90\%} = 87.49$, $SD_{90\%} = 11.96$,, $F_{2,580} = 2919.31$, $p < .001$, $\eta_p^2 = 0.91$ [0.90; 0.92]), with a positive linear trend in group means ($t_{870} = 68.76$, $p < .001$), and no detectable quadratic trend ($t_{867} = 0.17$, $p = 0.87$). No corrections of p-values were applied (although, given our conjunctive hypothesis, we could have increased the alpha-level per test for the moral questions).

In sum, we demonstrated that the probabilistic strength of the relationship between action and outcome clearly influenced moral evaluations. However, a very large effect on the causal measure, p(outcome|action) only led to medium-sized effects on the moral measures. Thus, the chain manipulation in Experiment 1 may not have decreased the perceived p(outcome|action) enough to produce a large effect on moral judgments. In Experiment 3 we went back to comparing direct causal relations with chains but increased the length of the chain hoping for a stronger effect. Moreover, we tested the hypothesis that foreseeability mediates the effect.

Figure 3: Mean ratings for whether it is okay to act (A), agents' moral responsibility (B), and p(outcome|action) (C) per strength condition in Experiment 2. Error bars are 95% CIs.



Experiment 3: The Role of Foreseeability

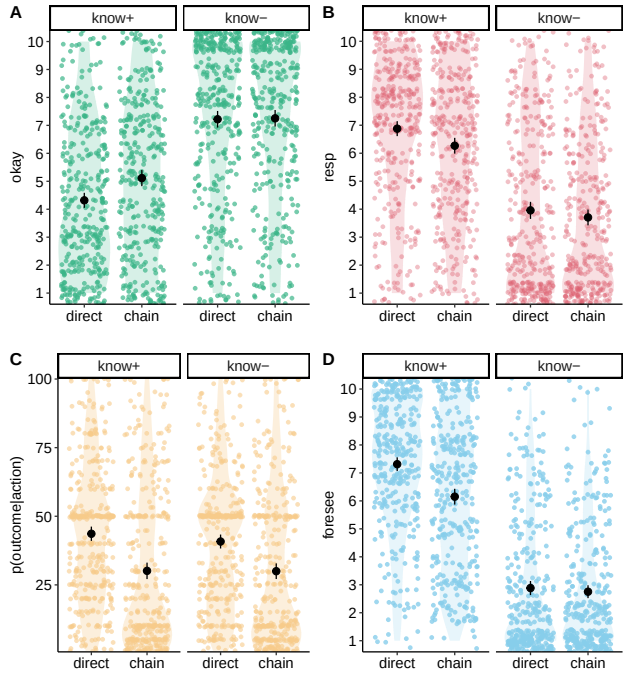
In this experiment, we used the same task as presented in Experiment 1, but with a stronger chain manipulation (nine probabilistic links instead of four, see Figure 1). In addition, we added new conditions in which agents were unaware of the possible harm that may result from their action. A longer causal chain does not only entail lower causal strength, but should normally also decrease the foreseeability of the negative outcome compared to a direct relation. However, this difference in outcome foreseeability between chains and direct relations depends on agents' awareness of the relation. When someone is unaware of any relation between their action and a harmful outcome, it seems hardly morally relevant whether action and outcome are related directly or by a longer chain. If proximity effects on moral judgments are mediated by outcome foreseeability, they should be eliminated without proximity knowledge.

Methods

Design, Materials, and Procedure We implemented a 2 (structure: direct vs. chain, within-subjects) x 2 (proximity knowledge: yes vs. no, between-subjects) mixed design. In the chain conditions, nine probabilistic links were instructed instead of four (see OSF for the full material). The direct conditions were identical to the ones in Experiment 1. In the"proximity knowledge" conditions (*know+*), agents were aware of the relation between their action and the possible harmful outcome (as in Experiments 1 and 2). In the "no proximity knowledge" conditions (*know-*), they were not. In the "chemical" cover story, for instance, we stated in *know-*: "Since the scientists studying Proskine have not published their results so far, Mary cannot be aware of them. To the best of her knowledge, there are no special risks associated with producing and storing Proskine." The cover stories were combined with the levels of the structure manipulation in a Latin square as described in Experiment 1. Procedure and measures also were the same as in Experiment 1, with the addition of a question about foreseeability at the end of the experiment, presented along with the question about p(outcome|action). The new question read: "To what extent could [agent] foresee that someone would be harmed by [her/his] action?", and responses were provided on a scale ranging from 1 ("not at all") to 10 ("fully"). This question was intended primarily as a manipulation check.

Participants We aimed for a sample size of 700 valid participants in this experiment. The sample size was determined by a simulation (see OSF for code), focusing on the two moral questions (*okay* and *resp*). Based on pilot studies, we planned for proximity effects of $d = 0.22$ for *okay* and of $d = 0.28$ for *resp*, in one-sided, paired t-tests in *know+*. We predicted null effects in two-sided paired t-tests for both measures in *know-*. If these patterns obtained, we should thus also observe a significant structure x proximity knowledge interaction in mixed ANOVAs for both *okay* and *resp*. With 700 participants, we achieve a power of >90% to detect the full set of the specified effects. We invited 720 participants to take part in the experiment. Participants received a compensation of £0.65 for an estimated six minutes of their time. Sixteen participants were excluded due to failing a simple attention check or due to completing the survey from a smartphone against instructions, leaving data of 704 participants for the analyses ($M_{age} = 34.81$, $SD_{age} = 13.14$, 50% women, 49% men, 1% non-binary or no answer).

Figure 4: Mean ratings for whether it is okay to act (A), agents' moral responsibility (B), p(outcome|action) (C), and agents' outcome foreseeability (D) per structure and proximity knowledge condition in Experiment 3. Error bars are 95% CIs.



Results and Discussion

Figure 4 shows the results. In the *know+* conditions, we found the predicted proximity effects on the moral questions, with larger effect sizes than in Experiment 1 (*okay*: M_{direct} = 4.32, SD = 2.56, M_{chain} = 5.11, SD = 2.75, t_{351} = 5.08, $p < .001$, d = 0.30 [0.18; 0.42], *resp*: M_{direct} = 6.88, SD = 2.55, M_{chain} = 6.27, SD = 2.69, t_{351} = 3.8, $p < .001$, d = 0.23 [0.11; 0.35]). As predicted, the effects largely disappeared in the *know-* conditions, although a very small significant effect remained for attributions of moral responsibility (*okay*: M_{direct} = 7.22, SD = 2.83, M_{chain} = 7.25, SD = 2.83, t_{351} = 0.26, p = 0.793, *resp*: M_{direct} = 3.96, SD = 2.91, M_{chain} = 3.70, SD = 2.71, t_{351} = 2.03, p = 0.043, d = 0.09 [0; 0.18]). The predicted interaction between structure and foreseeability was thus found for the *okay* question ($F_{1,702}$ = 14.01, $p < .001$, η_p^2 = .02 [0.01; 0.04]), but not for moral responsibility ($F_{1,702}$ = 3.11, p = 0.078). Independent of proximity knowledge, participants thought that the final outcomes were less likely to occur in the long chains than in the direct causal relation (t_{703} = 14.22, $p < .001$, d = 0.46 [0.39; 0.53], no interaction). However, the effect size was similar to the one we found in Experiment 1 (d = 0.52). As expected, participants only ascribed less outcome foreseeability with increased chain length to agents who were aware of the relation between action and outcome but not to agents without knowledge about the causal relation (interaction: $F_{1,702}$ = 36.89, $p < .001$, η_p^2 = .05 [0.03; 0.08], see OSF for the full analysis and all descriptive statistics). No corrections of p-values were applied (although, given our conjunctive hypothesis, we could have increased the alpha-level per test for the moral judgment questions).

General Discussion

We set out to test the hypothesis that (1) instructing a chain of probabilistically linked events between an action and a harmful outcome would lead to a more lenient moral evaluation of the agent and the action, compared to instructing a direct relation. We expected this pattern because (2) participants should perceive the harmful outcome as less likely to actually occur in a chain than in a direct relation. Moreover, we predicted that (3) the effect will be mediated by participants' attributions of outcome foreseeability to agents.

We found evidence for (1) in two experiments, but with surprisingly small effect sizes. The actions were only seen as slightly more permissible, and agents only judged as slightly less responsible in the chain compared to the direct conditions. Our data in all experiments are *consistent* with (2). The probability of the final outcome given action was indeed perceived as lower in chains than in direct relations. In all experiments, a lower p(outcome|action) was also associated with a more lenient moral evaluation. However, it is unclear whether these effects were caused by the difference in the structure of the causal models (direct vs. chain), or by the lowered strenghts of the relation between action and final outcome. It is possible to experimentally dissociate these two factors. For a more rigorous test, we would need to keep chain length constant and vary p(outcome|action) independently (see Stephan et al., 2021). We have conducted such a study in the meantime and found that the effect of chain length on moral judgments is at least substantially mediated by p(outcome|action) (Engelmann & Waldmann, *manuscript in preparation*). Further experiments are ongoing. Finally, Experiment 3 provides support for (3), the mediating role of outcome foreseeability. Chain length largely ceased to affect moral judgments when agents were unaware of the presence of the chain or of the direct relation. A very small effect persisted for moral responsibility, reminiscent of the moral luck literature (Young, Nichols, & Saxe, 2010). We will explore this puzzling effect further in future research.

An unexpected and interesting observation in all experiments was that medium (Experiment 1, Experiment 3) or large (Experiment 2) differences in p(outcome|action) only translated into small (Experiment 1, Experiment 3) or medium (Experiment 2) effects on the moral judgment measures. The only manipulation that pushed moral judgments across the scale midpoints (from permissible to impermissible and from responsible to not responsible) was the knowledge manipulation in Experiment 3. In the causal reasoning literature, it is sometimes claimed that people care more

about causal structure than about causal strength (e.g., Bes et al., 2012). Possibly, a similar effect obtains in moral reasoning, where causal reasoning is combined with inferences about others' mental states: Once agents know about the mere existence of a causal link between an action and a harmful outcome (as is the case in all our scenarios except the *know*-conditions of Experiment 3), we may be reluctant to judge their actions as permissible or blameless, even when harm becomes increasingly unlikely. While causal strength is clearly not irrelevant, a negative impression based on the causal link between an action and harm may prevail. A current example is the reluctance of some people to get vaccinated against Covid-19 because of extremely unlikely side-effects of some of the available vaccines, despite those risks being dramatically outweighed by the benefits.

Given that our chain manipulation here did not dissociate causal strength from causal structure (as explained above), it follows that the knowledge manipulation in Experiment 3 also did not dissociate *knowledge about causal strength* from *knowledge about causal structure*. It is clearly possible for agents to be aware of the presence of a causal link without knowing its strength. Likewise, agents might know about a statistical association between events without knowing if and how they are causally connected. We are presently conducting further experiments that aim to illuminate how these components combine to inform moral judgments.

References

Bes, B., Sloman, S., Lucas, C. G., & Raufaste, (2012). Non-bayesian inference: Causal structure trumps correlation. *Cognitive Science*, *36*(7), 1178–1203.

Davidson, D. (2001). *Essays on actions and events: Philosophical essays* (Vol. 1). Oxford University Press.

Engelmann, N., & Waldmann, M. R. (*manuscript in preparation*). Causal structure, causal strength, and moral judgment.

Hilton, D. J., McClure, J., & Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology*, *40*(3), 383–400.

Johnson, J. T., & Drobny, J. (1985). Proximity biases in the attribution of civil liability. *Journal of Personality and Social Psychology*, *48*(2), 283–296.

Kassambara, A. (2020). ggpubr: 'ggplot2' based publication ready plots [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=ggpubr (R package version 0.4.0)

Kelley, K. (2020). Mbess: The mbess r package [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=MBESS (R package version 4.8.0)

Kirfel, L., & Lagnado, D. A. (2021). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, *212*, 104721.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770.

Lawrence, M. A. (2016). ez: Easy analysis and visualization of factorial experiments [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=ez (R package version 4.4-0)

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology*, *39*(6), 653–660.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

RStudio Team. (2020). Rstudio: Integrated development environment for r [Computer software manual]. Boston, MA.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press.

Sloman, S., Fernbach, P. M., & Ewing, S. (2009). Causal models: The representational infrastructure for moral judgment. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 50, pp. 1–26). Academic Press.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, *126*(4), 323.

Steiger, J. H. (2004). Beyond the f test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological methods*, *9*(2), 164.

Stephan, S., Tentori, K., Pighin, S., & Waldmann, M. R. (2021). Interpolating causal mechanisms: The paradox of knowing more. *Journal of Experimental Psychology: General*.

Torchiano, M. (2020). effsize: Efficient effect size computation [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=effsize (R package version 0.8.1)

Waldmann, M. R. (2017). *The Oxford handbook of causal reasoning*. Oxford University Press.

Waldmann, M. R., Wiegmann, A., & Nagel, J. (2017). Causal models mediate moral inferences. In J. Bonnefon & B. Tremolière (Eds.), *Moral inferences* (pp. 37–55). London: Routledge/Taylor Francis Group.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*, *1*(3), 333–349.

Appendix D Engelmann & Waldmann (2019)

Moral Reasoning with Multiple Effects: Justification and Moral Responsibility for Side Effects

Neele Engelmann (neele.engelmann@uni-goettingen.de)
Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen, Germany

Abstract

Many actions have both an intended primary effect and unintended, but foreseen side effects. In two experiments we investigated how people morally evaluate such situations. While a negative side effect was held constant across conditions in Experiment 1, we varied features of the positive primary effect. We found that judgments of moral justification of actions were sensitive to the numerical ratios of helped versus harmed entities as well as to the kind of state change that was induced by an agent's action (saving entities from harm versus improving their status quo). Judgments of moral responsibility for side effects were only sensitive to the latter manipulation. In Experiment 2, we found initial support for a subjective utilitarian explanation of the moral justification judgments.

Keywords: Moral Reasoning, Causal Reasoning

Introduction

Research on moral judgments often probes people's intuitions about moral dilemmas. One of the most famous and well-studied dilemmas is the so-called trolley problem (Foot, 1967). In the side effect variant of trolley dilemmas, agents have a choice between letting a runaway trolley kill several people or an action that redirects the trolley to a different track where it would kill fewer people. The primary question in these studies is typically whether it is morally permissible to act. Many factors have been identified that influence people's intuition about this question (for an overview see Waldmann, Nagel, & Wiegmann, 2012).

The two dominant normative ethical approaches, utilitarianism and nonconsequentialism, largely agree in this situation. According to utilitarian recommendations, the action should be performed whenever its positive consequences outweigh the negative effects. Nonconsequentialist theories, such as the *Doctrine of Double Effect* (DDE, see Mikhail, 2011), arrive at similar conclusions for this case. The focus of the DDE and nonconsequentialism in general lies on the causal structure mediating acts and outcomes. In the side effect variant of the trolley dilemma, acting is considered permissible because the negative effect is not an intended means, but merely a foreseen side effect, and is not out of proportion to the positive effect. Psychological research on the side effect dilemma has shown that subjects indeed take the alternative outcomes into account when assessing the action's permissibility (e.g., Mikhail, 2011; Cohen & Ahn, 2016).

Evaluating Actions and their Side Effects

The focus of research on trolley dilemmas is on how people evaluate the permissibility of an action that causes two outcomes. All theories assume that in the side effect dilemma, both outcomes are compared and affect the moral evaluation, but little is known about the functional form of this comparison. A typical claim is that harming is permissible if the good outweighs the bad, but it is unclear whether this decision is just based on a simple categorical decision about which value is larger, or whether gradual differences between outcome values affect the decision. Few studies have systematically manipulated the numbers of victims that are saved or harmed in moral dilemmas (but see Cohen & Ahn, 2016; Waldmann & Wiegmann, 2012).

Cohen and Ahn (2016) postulate a subjective utilitarian analysis. For each item or set of items (e.g., 5 people) subjects provided an estimate of their personal value. The personal values were affected by the type of item and their number, although the number turned out to have a relatively small effect. These estimates of the personal values were then used to predict subjects' judgments about choice situations in which one set of items is about to be destroyed (or killed) when no action is taken but saved when the agent acts, which in turn would destroy (kill) a second set of items. According to the categorical utilitarian decision strategy, the action is chosen that saves items with the higher personal value. The model also predicts reaction times: Given that the comparison is typically influenced by uncertainty, a faster reaction time is predicted when the difference between values becomes larger.

One key goal of our project is to provide further tests of the subjective utilitarian model. A salient problem of the current version of the model is that it lacks generality. Its predictions are based on the personal values of the items involved in the outcomes but this model neglects that actions cause transitions between states. An evaluation of an action thus needs to take into account the values of the states of the items in the presence versus the absence of the action. Cohen and Ahn (2016) did not consider how subjects assess the personal values of the items in their destroyed or dead states, probably because this was the standard state in the absence of an action across all item sets. However, actions can also improve the state of items that otherwise would be in a normal state, or they could be saved from a disease that would harm, but not kill them. To provide a full utilitarian account of how outcomes of actions should be evaluated we suggest that people compute contrasts between the personal values of the outcomes in the presence versus the absence of the target action. We will also argue that sometimes more than two states need to be considered. We will present an experiment that presents a wider range of actions, which allows us to test our subjective utilitarian model against theories that are not sensitive to different types of states in the presence and absence of the

target action.

A further focus of our study is to investigate how the relation between the number of people that are positively or negatively affected by the action influences the degree to which people find the action morally justifiable and the agent morally responsible for the outcomes, especially the negative side effect. We systematically manipulated the numbers involving the positive primary effect while holding the negative side effect constant (see also Waldmann & Wiegmann, 2012, for a similar design but different tasks). For example, in one of our experimental conditions, ten members of a tribe are harmed by an action that would save a varying number of members of a different tribe. According to Cohen and Ahn's (2016) model, an act involving a negative side effect should lead to faster reaction times the more entities are helped compared to harmed. If reaction times indicate certainty about an act's permissibility, one can also derive from this theory the prediction that justification ratings should be affected in a similar manner.

One limitation of trolley studies is that so far they have focused on a particular type of situation in which the primary goal is to save victims that otherwise would be killed. It may well be that acts that lead to negative side effects are only considered justified when the primary effect targets entities that, prior to the intervention, are threatened to be harmed. The primary effect may be less effective as a justification when the act is supererogatory and just improves the states of entities that prior to the act are in a normal state. For example, instead of saving varying numbers of victims from grave harm, the people may be fine prior to the act, with the act just improving their health and living conditions. The theory proposed by Cohen and Ahn (2016) does not make predictions here because it only takes into account the personal values of the entities in their intact state. We will in Experiment 2 test a modified account that postulates that subjects take into account personal values of states in both the presence and the absence of an action. This account makes predictions for the difference between saving entities or improving their states.

Another limitation of the typical trolley dilemma studies is that they have focused on situations in which saving and harming are causally achieved by redirecting a harmful entity (the runaway trolley). In order to widen the range of studied dilemmas and to be able to manipulate the prior state of the entities involved in the primary goal, we tested a different causal structure in which a helpful act rather than a threat was redirected (see also Ritov & Baron, 1999; Bartels & Medin, 2007). For example, in the condition involving two tribes, a dam may be opened that redirects water from one tribe to the other. Redirecting might save tribe members from a negative state or improve their normal situation.

Finally, a limitation of previous research is that the test question typically focuses only on the act leading to two outcomes. We are also interested in how people evaluate the two outcomes individually. We therefore added as test questions requests to judge moral responsibility for the negative side effect. Our goal was to test whether these judgments are also influenced by the value of the primary effect (e.g., number of victims). If subjects just focus on the side effect, the primary effect should not have an influence. However, if the status quo or the number of affected entities are used as exonerating factors, their impact should also be seen in moral responsibility ratings for the side effect.

Together, these manipulations and the studied judgments widen the focus of previous work on people's moral intuitions about cases with multiple effects. The aim of the first experiment was to test whether the relation between primary and side effect of an action influences moral justification assessments. Moreover we were interested in whether the primary effect influences moral responsibility assessments for a bad side effect. We tested whether these two types of moral queries are affected by the kind and number of entities that are potentially harmed or saved, and by their state change due to a possible intervention. Experiment 2 inquires to what extent the results of Experiment 1 can be explained by a subjective-utilitarian framework.

Experiment 1

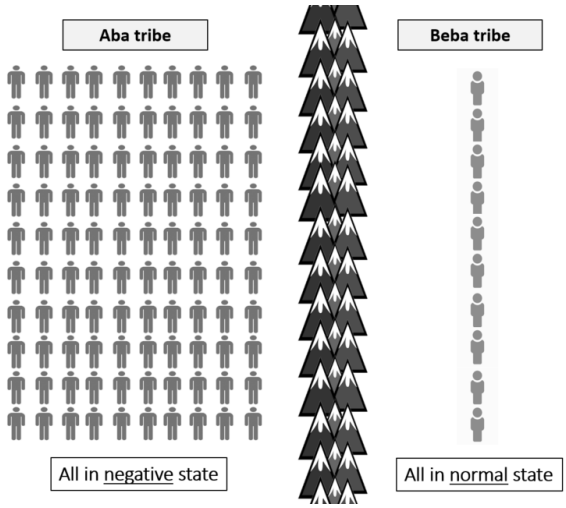
We constructed three scenarios in which an agent decides to perform an action with a positive, intended primary effect and a negative, unintended (but foreseen) side effect. The negative side effect was held constant across conditions and always consisted in killing 10 entities (people, animals, or plants). We varied whether 1, 5, 20 or 100 entities benefitted from the action. Furthermore, we manipulated whether these entities were in a negative or a neutral state prior to the action. In the situations in which the entities were in a negative state, they would have died without the agent's action; in the contrasted normal state condition, the action would merely cause additional benefit (e.g., people improving their living conditions or plants growing better).

Design, Material and Procedure[1] 450 participants were recruited via the UK based platform Prolific Academics for a compensation of £0.25 (£6 per hour). Inclusion criteria were a minimum age of 18 years, English as a first language, a study approval rate on the platform of at least 90%, and not having participated in previous studies with similar material. Participants were randomly allocated to one of 24 conditions (primary effect: saving vs. improving; number of helped entities: 1 vs. 5 vs. 20 vs. 100; affected entities: people vs. animals vs. plants). Here is an example vignette from the *saving* conditions. The example describes a condition in which 100 people are saved by the action, who otherwise would die:

Suzy is the prime minister of Tolosia, a mountainous country with many distant and small villages. The villages are populated by different indigenous tribes. She is authorised to make all decisions about the country's welfare that she deems appropriate. One day, she learns that a mountain village has

[1]The full material and data for both experiments are available under `https://osf.io/jcux6/`

suffered from an ongoing drought that left its inhabitants, the Aba tribe, in poor health due to lack of water. Exactly 100 people belong to the Aba tribe, all of whom are in critical condition and will die if nothing is done. Suzy could order to open a dam that would redirect a mountain river towards the Aba tribe. With a quick water supply, the 100 members of the Aba tribe could recover. However, the redirection of the river could also cause a lack of water in another mountain village, home to the Beba tribe, causing its 10 members to die of thirst within a few days. All of the 10 members of the Beba tribe are fine at the moment. Since both mountain villages are inaccessible to any means of transport, redirecting the river is the only currently available measure to influence the well-being of the two tribes. Here is a schematic representation of the two tribes and the current state of their members:



Suzy is aware of all the facts. She wants the 100 members of the Aba tribe to recover, but also not to cause any harm to the 10 members of the Beba tribe. She decides to open the dam and redirect the mountain river. All of the 100 members of the Aba tribe recover. However, all of the 10 members of the Beba tribe die within a few days.

The figure was followed by the instruction: "Here is a schematic representation of the tribes and their state after the river has been redirected" along with the same figure as above in which the lower labels now read "all in normal state" for the Aba tribe and "all dead" for the Beba tribe. In the corresponding *improving* condition, the vignette stated that the Aba tribe could vastly improve their health and lifespan with an extra water supply (no threat by a drought was mentioned). In the subsequent test phase participants were asked to rate the extent to which they saw the agent's action as morally justified ("To what extent was Suzy's action morally justified ?"). The moral responsibility question focused on the side effect ("To what extent is Suzy morally responsible for the members of the Beba tribe dying?"). As a control, we also asked about the primary goal ("To what extent is Suzy morally responsible for the members of the Aba tribe improving their health?"). Ratings were given on a 10-point Likert scale with the endpoints labelled "not at all" (1) and "fully" (10). Justification and responsibility questions were presented on two separate pages, with page order counterbalanced between participants; order of the two responsibility questions within the respective page was randomized. Subsequently, two manipulation check questions assessed whether people had correctly understood how many entities were harmed and helped in the scenario.

Results and Discussion 18 participants were excluded for failing at least one of the manipulation check questions, leaving data of 432 participants for the analysis (mean age = 34.4, SD = 11.93). We conducted a 2 (primary effect) x 3 (entity) x 4 (numbers) x 2 (test question order) ANOVA for each of the three dependent variables. Since our study is partly exploratory, we used a conservative significance threshold that takes into account the number of tests in the models (here: $p<.003$). Results for the 432 valid subjects can be seen in Figure 1.

Moral justification ratings were higher the more entities were helped compared to harmed, $F_{(3,\,384)} = 8.81$, $p<.001$, $\eta^2 = .06$. Additionally, a large effect was obtained between the conditions saving and improving, $F_{(1,384)} = 130.74$, $p<.001$, $\eta^2 = .25$. The interaction was not significant (p=.37). Participants gave the highest justification ratings when the primary effect was an instance of saving and more entities were saved than killed.

Post hoc tests (Newman-Keuls) for the saving condition revealed that the case in which only one entity was saved as a primary effect was judged significantly less morally justified than the cases in which twenty or a hundred entities were saved. The other cases did not differ significantly from each other. In the *improving* condition, post hoc tests showed no significant differences.

There was also a main effect of vignette. Subjects considered the action as most morally justified when the affected entities were plants (M = 5.23, SD = 2.6), followed by animals (M = 4.41, SD = 2.52), and people (M = 3.84, SD = 2.77), $F_{(2,384)} = 14.39$, $p<.001$, $\eta^2 = .07$. A possible reason for this ordering might be that harming people may be seen as a harsher moral violation than harming plants and therefore less justifiable by good effects. Animals seem to be in the middle.

Additionally, a small unexpected order effect was found. Ratings were slightly higher when the moral justification question was presented after the moral responsibility questions (M = 4.88, SD = 2.71) compared to before (M = 4.12, SD = 2.62), $F_{(1,384)} = 12.51$, $p<.001$, $\eta^2 = .03$.

Moral responsibility ratings for the negative side effect were generally high, but not detectably influenced by the number of helped entities, $F_{(3,384)} = 0.35$, $p = .79$ (see Fig. 1). However, the ratings were lower when the action's primary effect was an instance of saving (M = 8.09, SD = 2.23) rather than improving (M = 9.12, SD = 1.59), $F_{(1,384)} = 33.51$, $p<.001$, $\eta^2 = .08$. The interaction was not significant (p=.61). *Moral responsibility ratings* for the positive primary effect were

high (M = 8.23, SD = 2.31) and not influenced by any manipulation.

In sum, the moral justification ratings of the action were sensitive to the relation between the primary and the side effect. The more entities were helped as a primary effect, the more justified the action was judged. This pattern shows that moral justification is a continuous quantity that is sensitive to the relative size of the outcomes. A novel result concerns the comparison between different status quos, which generated the largest effect. If entities are saved from a threat, the action was seen as substantially more justified than when the primary goal is just to improve states starting from a neutral state.

The fact that subjects took into account both the primary and the side effect in their justification judgments is predicted by both nonconsequentialist and utilitarian accounts. However, the specific theory proposed by Cohen and Ahn (2016) does not predict the largest effect in our experiment: Subjects clearly differentiated between saving entities versus improving their state. Simply using assessments of personal values of the entities does not predict these effects without taking into account the personal values of the states of the entities in the absence of the action. We will test a modified model that is sensitive to state changes in Experiment 2.

An interesting unexpected finding was that moral responsibility ratings proved insensitive to the number of helped entities, but were reduced when the action's primary effect was an instance of saving rather than improving. This latter effect makes it unlikely that the lack of an effect of number is due to a ceiling effect. A possible interpretation of this pattern may be that subjects tried to focus on the side effect alone but were influenced by features of the primary effect that have a large impact on justification, such as the status quo, rather than only a small effect, such as the numbers.[2]

Experiment 2

The aim of the second experiment is to investigate to what extent the effects observed in Experiment 1 could be explained by a a variant of a subjective utilitarian theory that in crucial aspects differs from the one proposed by Cohen and Ahn (2016). Cohen and Ahn (2016) modeled choices as decisions based on the personal values of the entities involved in the alternative outcomes. For example, the task in their second study was to choose which of two sets of items should be saved and which destroyed in a dilemma. The model claims that the differences between the personal values of the two sets of items predict judgments. The focus on the personal values of the items seems appropriate here because all actions represented a choice between leaving the items intact or destroying (or killing) them. This restriction of the task allowed Cohen and Ahn (2016) to focus on the personal values of the affected items. However, the model is a too restrictive as a general model of moral reasoning. We suggest that the focus should be on actions, which can cause transitions between various states, not only between the states dead and alive or intact and destroyed. For example, in our Experiment 1 we presented cases in which actions improved states of entities that prior to the intervention were in a normal state.

To overcome the limitations of the model proposed by Cohen and Ahn (2016), we here propose a variant of a subjective utilitarian theory that focuses on actions and models them as state changes. When people evaluate an action, they should be sensitive to both the outcomes in the presence of the action but also to what happens in the absence of the action. For example, an action that improves the state of an entity can be represented as the difference between the personal values of the improved state and the normal state prior to the action. More complex state transitions are conceivable, and in fact in Experiment 1 we presented scenarios in which the entities shifted between four possible states (normal, threatened, improved, dead). In the present study we collected assessments of personal values of all the entities for these four states and used these assessments to predict the justification judgments obtained in Experiment 1.

Figure 2 shows how we adapted our model to the cover stories in Experiment 1. In the example in Figure 2, 100 people are under the threat of dying prior to any action. In the absence of an action (i.e., omission) they would die, which is modeled here as the contrast of the personal values between death and a critical state (second component of Figure 2a). In the presence of the action, the people in critical state would be shifted into a normal, healthy state, here represented as the difference between the personal values of a critical versus a normal state (first component of Figure 2a). The overall utility of saving the people is modeled as the sum of these contrasts because the action both prevents the people from being killed and puts them from a critical into a healthy state. Thus, the representation of the saving action considers both the effects of the potential action and of its omission. In the case of improving (not depicted), the model simplifies to a contrast between the values of the improved versus the normal states. The second component in the equation in Figure 2a would amount to 0 in this case because there is no threat to the normal state. Finally, Figure 2b shows how we model the total utility of the action in a scenario with multiple effects: It is the sum of the median utilities of the primary effect (saving) and the harmful side effect (killing 10 people).

Design, Material and Procedure The design of our basic value estimation task largely follows the methodology described in Cohen and Ahn (2016) but assesses a wider range of possible states of entities. Like Cohen and Ahn (2016), we tested the influence of the numbers of entities (1 vs. 5 vs. 10 vs. 20 vs. 100) on personal value assessments in

[2] In this experiment, moral justification was assessed globally (i.e., for a whole action), while responsibility was assessed separately for the single effects. One might worry that this does not allow us to tell whether the differences between the two judgments are driven by the type of judgment or by the focus of the question on global or separate outcomes. We therefore conducted a follow-up study in which we fully crossed these two factors. We found that the type of judgment seems to be the driving factor. The study is available online along with materials and data.

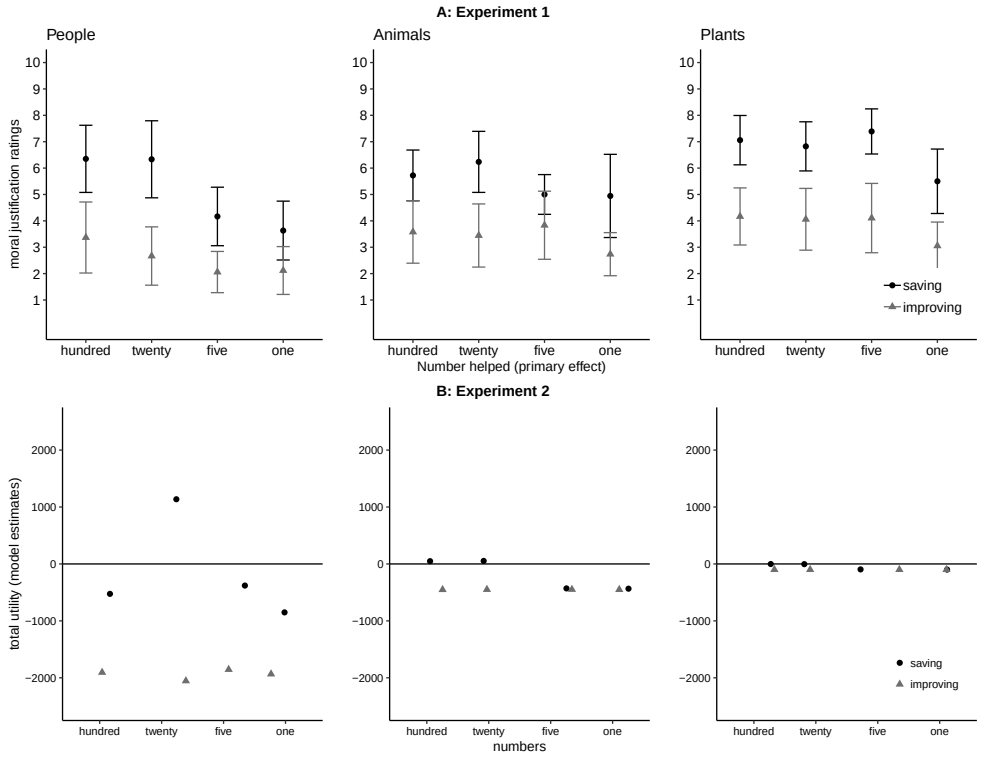


Figure 1: A: Means and 95% confidence intervals for moral justification ratings in Experiment 1, B: Total utility estimates generated by our model in Experiment 2.

separate experimental groups to avoid demand characteristics (i.e., participants feeling pressured to assign exactly five times the value of one entity to a group of five of the same entities). Within each group, we presented instances of people, fish and roses, each of them in all of the states that were described in Experiment 1 (normal vs. threatened vs. improved vs. dead). Thus, each participant judged 12 stimuli, in randomised order.[3] Like Cohen and Ahn, we presented people with a measuring standard to calibrate their value estimates. They were told that "one healthy chimpanzee" should be taken to have a value of 1000. If they valued any item half (or twice or any other ratio) as much as one healthy chimpanzee, they should assign the corresponding value to the item (e.g., a value of 500 if they value an item half as much as the chimpanzee). Participants were further instructed that "personal value" does not necessarily correspond to monetary value and that they should judge the entities' value in their *current* state. 250 participants (mean age = 36.6, SD = 13.5, 67% female, 32% male, 1% other) were recruited on Prolific Academics and completed the survey for a compensation of £0.40 (£6 per hour). Inclusion criteria were identical to Experiment 1, and not having participated in Experiment 1.

Results and Discussion To test our model, we used the value estimates of the four states of the entities to generate predictions for the justification assessments. Following the rationale outlined in Figure 2 we generated predictions for all 24 experimental conditions.The results are shown in Figure 1B. The total utilities overall capture the patterns found in Experiment 1, even though the maximal range of values was much wider for people cases compared to animals and plants (see Fig. 1A). Most importantly, the total utility estimates reflected the differences between improving versus saving, at least for people (Kruskal-Wallis $\chi^2 = 6.14$, $p = .01$) and animals (Kruskal-Wallis $\chi^2 = 6.14$, $p = .01$)[4]. In both cases the total utility for saving was larger than for improving, which mirrors the effects in Experiment 1. The corresponding effect for plants was not significant when correcting for multiple testing. Moreover, we did not find significant effects for the manipulation of the number of the affected entities for either people, animals or plants. But note that this effect was fairly small in Experiment 1 (and also in Cohen & Ahn, 2016). Also, this factor was the only one manipulated between subjects, which may have led to reduced sensitivity to this factor.

As an overall test of the fit of our model to the data of Experiment 1, we conducted a linear regression analysis with to-

[3] With the exception of the "10 entities" condition, which referred to the constant side effect. Here, we only needed estimations of each set of entities in their normal and dead states since the side effect entities never were in other states.

[4] We used again a conservative significance threshold that takes into account that we tested each factor separately for each entity category (here: $p < .017$).

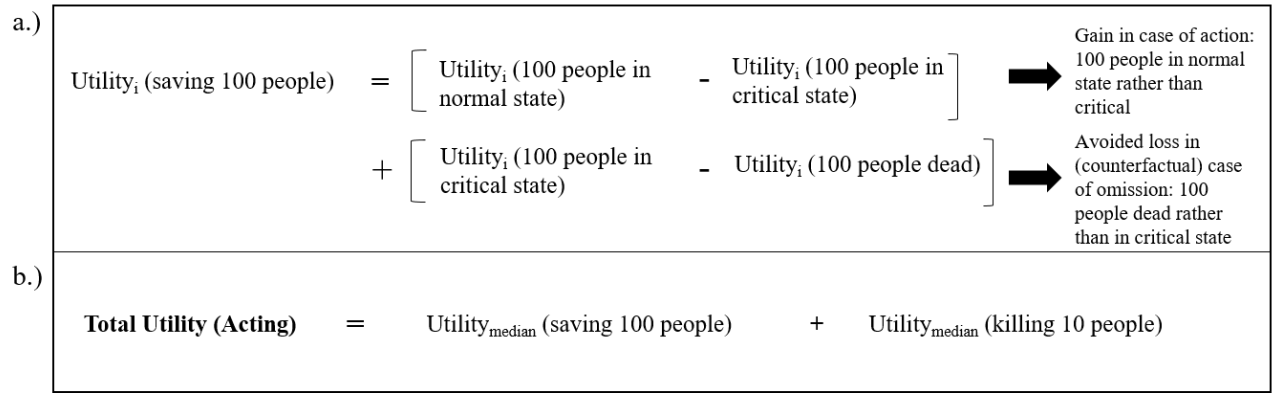


Figure 2: Rationale of our calculation of an action's total utility, spelled out for the example of the *saving 100 people* scenario. See text for explanation.

tal utilities estimated by our model as the predictor and mean moral justification ratings obtained in Experiment 1 as the criterion. The model fit the data well and explained a substantial amount of variance in the criterion, $F_{(1,22)} = 16.31$, $p<.001$, $R^2 = .43$, $RMSE = 1.14$.

General Discussion

The main goal of our study was to provide more fine-grained evidence on how moral judgments are influenced by characteristics of multiple effects of an action in dilemma situations. Experiment 1 showed that judgments of moral justification for the agent's action increased with more favourable ratios of helped compared to harmed entities, but were even more influenced by the change of state that was induced by the agent's action (saving vs. improving). Moral responsibility judgments for the negative side effect were only affected by the latter manipulation but not by the number of affected entities.

In Experiment 2 we tested a novel subjective utilitarian model that goes beyond previous proposals. Whereas Cohen and Ahn (2016) claimed that moral decisions are based on the personal values of the affected entities in their healthy or intact states, we argued that this assumption restricts their model to a small set of situations in which actions destroy or kill entities. Our goal was to propose a model that is more general. A basic assumption of our model is that actions can be modelled as state changes and that moral judgments are sensitive to both the states that entities are in prior and following a target action. This model allowed us to not only model cases of killing and saving but also, for example, cases of improvement.

Although our results in Experiment 2 showed that the new model explains a substantial amount of variance, it does not capture all effects. One reason for this may have been the necessary differences in the designs of Experiments 1 and 2. But there may be other reasons: For example, to demonstrate the increase of expressiveness of our model, we suggested a model for the cover stories of Experiment 1 that captures transitions between the four possible states mentioned there. Given that utility measurements are unreliable and influenced by additional factors, making the model more complex will certainly reduce its fit to the data.

Future research will also have to investigate whether there are alternative models that may also capture the results. As in the case of improving, we could, for example, generally use a more basic utilitarian model that only compares the two states in the presence versus absence of the action (e.g., dead vs. alive in the case of saving). Future research will need to test in greater detail the assumptions entering the different variants of the model.

We labeled our model "subjective utilitarian" because it was inspired by the theory of Cohen and Ahn (2016). However, we mentioned in the introduction that both utilitarian and nonconsequentalist theories predict that in side effect dilemmas the outcomes should be compared. Thus, our model may also be viewed as a component of a nonconsequentialist account. One possible way to test the two alternative theoretical possibilities is to take a closer look at the assumption that actions can be modeled as state changes. This assumption embodies the utilitarian claim that it is only the outcomes that matter, not the type of action leading to the outcomes. We suspect, however, that the type of action and the type of causal relations leading to the changes may also matter (see Kamm, 2007; Waldmann, Wiegmann, & Nagel, 2017). Future research will have to further explore these issues.

Acknowledgements

We thank Alex Wiegmann for helpful discussions about our utility model.

References

Bartels, D. M., & Medin, D. L. (2007). Are morally motivated decision makers insensitive to the consequences of their choices? *Psychological Science*, *18*(1), 24–28.

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General*, *145*(10), 1359–1381.

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

Kamm, F. M. (2007). *Intricate ethics: Rights, responsibilities, and permissible harm*. New York: Oxford University Press.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York: Cambridge University Press.

Ritov, I., & Baron, J. (1999). Protected values and omission bias. *Organizational behavior and human decision processes*, *79*(2), 79–94.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. *The Oxford handbook of thinking and reasoning*, 364–389.

Waldmann, M. R., & Wiegmann, A. (2012). The role of the primary effect in the assessment of intentionality and morality. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34, pp. 1102–1107). Austin, TX: Cognitive Science Society.

Waldmann, M. R., Wiegmann, A., & Nagel, J. (2017). Causal models mediate moral inferences. In J.-F. Bonnefon & B. Tremolière (Eds.), *Moral inferences* (pp. 37–55). London: Routledge/Taylor Francis Group.

Appendix E Abstract of Stephan, Engelmann, & Waldmann (2021)

Psychological dependency theories of causal cognition, such as Causal Bayes Net accounts, postulate that the strength of individual causal links is independent of the causal structure in which they are embedded. Strength is inferred from dependency information such as statistical regularities. We propose a hybrid representation account that postulates that people's concept of causality is richer, and predicts a systematic influence of causal structure knowledge on causal strength intuitions. Our view incorporates the notion held by dispositional theories that causes produce effects in virtue of an underlying continuous causal capacity or power. Going beyond existing dispositional theories, we argue that people's concept of causality involves the idea that causal powers behave similarly to phenomena studied in fluid dynamics: People assume that continuous causes, but not genuinely binary causes, spread their power across and along causal pathways, akin to fluids running through pipe systems, leading to the prediction of a structure-dependent perceived dilution of causal strength. A series of experiments (N = 3, 733) and a meta-analytic summary corroborate the theory. For common causes, people think that link strength decreases with the number of links served by that cause. In causal chains, people perceive a negative strength gradient between initial and terminal links. This dilution effect is robust across various contexts, but disappears if the causal variables are represented as genuinely binary. We discuss the theoretical and empirical implications of our findings.

Stephan, S., Engelmann, N., & Waldmann, M. R. (2021). The perceived dilution of causal strength [Manuscript submitted for publication].

Appendix F Abstract of Wiegmann & Engelmann (2022)

Consider the following case:

Dennis is going to Paul's party tonight. He has a long day of work ahead of him before that, but he is very excited and can't wait to get there. Dennis's annoying friend Rebecca comes up to him and starts talking about the party. Dennis is fairly sure that Rebecca won't go unless she thinks he's going, too. Rebecca: Are you going to Paul's party?

(1) Dennis: No, I'm not going to Paul's party.

(2) Dennis: I have to work.

Rebecca comes to believe that Dennis is not going to Paul's party. In (1), Dennis tricks Rebecca into a false belief by explicitly expressing a falsehood. By contrast, in (2) Dennis achieves his aim in a less direct way, namely by means of a conversational implicature. Cases of the first kind are usually described as cases of lying, while cases of the second kind are characterized as merely misleading. Philosophers have discussed such pairs of cases with regard to the question of whether lying is morally different from misleading. In this paper, we report the results of approaching this question empirically, by presenting 761 participants with ten matched cases of lying versus misleading in separate as well as joint evaluation designs. By and large, we found that cases of lying and misleading were judged to be morally on a par, to have roughly the same consequences for future trust, and to elicit roughly the same inferences about the speaker's moral character. When asked what kind of deception participants would choose if they had to deceive another person, the clear majority preferred misleading over lying. We discuss the relevance of our findings for the philosophical debate about lying and misleading, and outline avenues

for further empirical research.

Wiegmann, A., & Engelmann, N. (2022). Is lying morally different from misleading? An empirical investigation. In L. Horn (Ed.), *From lying to perjury: Linguistic and legal perspectives on lies and other falsehoods*. De Gruyter. https://doi.org/10.1515/9783110733730-005

Appendix G Abstract of Viebahn, Wiegmann, Engelmann, & Willemsen (2021)

In several recent papers and a monograph, Andreas Stokke argues that questions can be misleading, but that they cannot be lies. The aim of this paper is to show that ordinary speakers disagree. We show that ordinary speakers judge certain kinds of insincere questions to be lies, namely questions carrying a believed-false presupposition the speaker intends to convey. These judgements are robust and remain so when the participants are given the possibility of classifying the utterances as misleading or as deceiving. The judgements contrast with judgements participants give about cases of misleading or deceptive behaviour, and they pattern with judgements participants make about declarative lies. Finally, the possibility of lying with non-declaratives is not confined to questions: ordinary speakers also judge utterances of imperative, exclamative and optative sentences carrying believed-false presuppositions to be lies.

Viebahn, E., Wiegmann, A., Engelmann, N., & Willemsen, P. (2021). Can a question be a lie? An empirical investigation. *Ergo*, *8*(7). https://doi.org/10.3998/ergo.1144

Appendix H Abstract of Wiegmann & Engelmann (2020)

Die experimentelle Untersuchung des moralischen Denkens, Urteilens und Verhaltens hat sich in den vergangenen zwanzig Jahren zu einem dynamischen Forschungsfeld mit vielversprechenden Zukunftsaussichten entwickelt. Ein besonderer Reiz des Feldes liegt in seiner Interdisziplinarität: Philosophie, Psychologie, Biologie, Neurowissenschaften, Linguistik und Anthropologie sind einige der Disziplinen, deren Beiträge zu einem besseren Verständnis unseres moralischen Kompasses beigetragen haben. Besonders zwischen Psychologie und Philosophie ergeben sich dabei immer wieder interessante Wechselbeziehungen. Entsprechen die alltäglichen moralischen Urteile von Menschen den Ansprüchen normativer ethischer Theorien, und wenn ja, welchen? Und inwiefern – falls überhaupt - sind empirische Befunde für moralphilosophische Fragen relevant? Das vorliegende Kapitel kann diese Fragen zwar nicht abschließend beantworten, wohl aber einen Überblick über die wichtigsten empirischen Arbeiten und theoretischen Entwicklungen der Moralpsychologie im 21. Jahrhundert geben. Dies ist ein hoffentlich hilfreicher Wegweiser für alle, die spezifischere Fragestellungen im fruchtbaren Spannungsfeld zwischen Psychologie und Philosophie verfolgen möchten. Im ersten Teil widmen wir uns globalen Theorien in der Moralpsychologie. Globale Theorien haben den Anspruch, die moralische Urteilsbildung auf allgemeine Weise zu charakterisieren. Unsere Beschreibung dieser Theorien orientiert sich dabei grob am historischen Verlauf. Wir zeichnen zunächst die Debatte um den respektiven Anteil von Kognition und Emotion im moralischen Denken nach, die mit den rationalistischen Ansätzen von Jean Piaget und Lawrence Kohlberg in der Entwicklungspsychologie beginnt. Diese Ansätze charakterisieren moralisches Urteilen primär als das Produkt von bewussten Denkprozessen, in dem emotionale unbewusste Prozesse keine große Rolle spielten. Um die Jahrtausendwende wurde diese

Sichtweise durch das „sozial-intuitionistische“ Modell von Jonathan Haidt auf den Kopf gestellt. Anschließend diskutieren wir Joshua Greenes sogenannte „Zwei-Prozesse-Theorie" [dual process theory], die als eine Art Mittelweg zwischen Kohlberg und Piagets rationalistischem und Haidts emotionsdominiertem Ansatz verstanden werden kann. Darauf folgt Cushmans und Crocketts Weiterentwicklung von Greenes Zwei-Prozesse-Theorie, bevor der erste Teil durch Mikhails universale Moralgrammatik abgeschlossen wird, die den Fokus auf die kausale und intentionale Struktur von moralischen Szenarien legt. Im zweiten Teil (Schlaglichter) widmen wir uns ausgewählten enger gefassten Themenbereichen.3 Wir beginnen mit Entwicklungen der Moralpsychologie in die Richtung mathematisch formalisierter Theorien, die eine präzisere Annäherung an relevante psychologischen Prozesse in Aussicht stellen (Komputationale Ansätze). Im Anschluss diskutieren wir die Frage, ob sich aus moralpsychologischen Befunden vorhersagen lässt, wie sich Menschen in realen moralischen Situationen verhalten und ob sich moralische Expertise in moralisch besserem Entscheiden und Handeln niederschlägt (Externe Validität und moralische Praxis). Der darauffolgende Abschnitt steht im Einklang mit der generellen Stoßrichtung des vorliegenden Bandes: Wir beschreiben Befunde über den Einfluss moralisch irrelevanter Faktoren auf Moralurteile und diskutieren deren Bedeutung für die philosophische Diskussion (Moralisch irrelevante Faktoren). Beschlossen wird dieses Kapitel mit ein paar Bemerkungen zur Replikationskrise in der Psychologie und welche Schlüsse aus ihr gezogen wurden.

Wiegmann, A., & Engelmann, N. (2020). Entwicklungen und Probleme der Moralpsychologie zu Beginn des 21. Jahrhunderts. In N. Paulo & J. C. Bublitz (Eds.), *Empirische Ethik - Grundlagentexte aus Psychologie und Philosophie* (pp. 139–175). Suhrkamp Verlag.

Appendix I Abstract of Hagmayer & Engelmann (2020)

In this paper, we give a brief overview of current, cognitive-psychological theories, which provide an account for how people explain facts: causal model theories (the predominant type of dependence theory) and mechanistic theories. These theories differ in (i) what they assume people to explain and (ii) how they assume people to provide an explanation. In consequence, they require different types of knowledge in order to explain. We work out predictions from the theoretical accounts for the questions people may ask to fill in gaps in knowledge. Two empirical studies are presented looking at the questions people ask in order to get or give an explanation. The first observational study explored the causal questions people ask on the internet, including questions asking for an explanation. We also analyzed the facts that people want to have explained and found that people inquire about tokens and types of events as well as tokens and types of causal relations. The second experimental study directly investigated which information people ask for in order to provide an explanation. Several scenarios describing tokens and types of events were presented to participants. As a second factor, we manipulated whether the facts were familiar to participants or not. Questions were analyzed and coded with respect to the information inquired about. We found that both factors affected the types of questions participants asked. Surprisingly, participants asked only few questions about actual causation or about information, which would have allowed them to infer actual causation, when a token event had to be explained. Overall the findings neither fully supported causal model nor mechanistic theories. Hence, they are in contrast to many other studies, in which participants were provided with relevant information upfront and just asked for an explanation or judgment. We conclude that more empirical and theoretical work is needed to reconcile the findings from these two

lines of research into causal explanations.

Hagmayer, Y., & Engelmann, N. (2020). Asking questions to provide a causal explanation – Do people search for the information required by cognitive psychological theories? In E. A. Bar-Asher Siegal & N. Boneh (Eds.), *Perspectives on Causation: Selected Papers from the Jerusalem 2017 Workshop* (pp. 121–147). Springer International Publishing. https://doi.org/10.1007/978-3-030-34308-8_4

Appendix J Curriculum Vitae

Neele Engelmann, M.Sc.

neeleengelmann.com
neele.engelmann@ruhr-uni-bochum.de
ORCiD: 0000-0002-0000-9940

Current Position

Postdoctoral researcher — Bochum, Germany
Ruhr-University Bochum — From 04/2022
Center for Law, Behaviour, and Cognition

- DFG project: "Experimental legal philosophy: The concept of law revisited". Principal investigator: Stefan Magen

Education

Ph.D. candidate — Göttingen, Germany
Georg-August-University Göttingen — 2017 - 2022
Georg-Elias-Müller Institute for Psychology
Department of Cognitive and Decision Sciences

- Thesis: "The role of causal representations in moral judgment". Advisor: Michael R. Waldmann
- Expected graduation: August 29, 2022

M.Sc. Psychology — Göttingen, Germany
Georg-August-University Göttingen. Grade: 1.3 ("very good") — 2014–2017

- Thesis: "An empirical investigation of the effects of prescriptive and statistical normality on judgments of actual causation and accountability". Grade: 1.0 ("very good"). Advisor: Michael R. Waldmann
- Additional courses in philosophy and German philology
- Research stay at University College London, UK (Summer 2016), Department of Experimental Psychology, Causal Cognition Lab (David Lagnado)

B.Sc. Psychology — Göttingen, Germany
Georg-August-University Göttingen. Grade: 1.3 ("very good", degree awarded with distinction) — 2010–2014

- Thesis: "Foraging with threatened options". Grade: 1.0 ("very good"). Advisor: Hansjörg Neth
- Additional courses in philosophy and French/Gallo-Roman studies
- Study stay at Université Paris Est Créteil Val de Marne, France (Summer 2014), Studies of French philology, literature, philosophy and history

Publications

Engelmann, **N.**, & Waldmann, M. R. (2022a). How causal structure, causal strength, and foreseeability affect moral judgments. *Cognition*, *226*, Article 105167.

Engelmann, **N.**, & Waldmann, M. R. (2022b). How to weigh lives. A computational model of moral judgment in multiple-outcome structures. *Cognition*, *218*, Article 104910.

Wiegmann*, A., & **Engelmann*, N.** (2022). Is lying morally different from misleading? An empirical investigation. In L. Horn (Ed.), *From lying to perjury: Linguistic and legal perspectives on lies and other falsehoods* (pp. 89–111).

Engelmann, **N.**, & Waldmann, M. R. (2021). A causal proximity effect in moral judgment. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd meeting of the cognitive science society* (pp. 2330–2336).

Viebahn*, E., Wiegmann*, A., **Engelmann**, **N.**, & Willemsen, P. (2021). Can a question be a lie? An empirical investigation. *Ergo*, *8*, Article 7.

Hagmayer, Y., & **Engelmann**, **N.** (2020). Asking questions to provide a causal explanation – Do people search for the information required by cognitive psychological theories? In E. A. Bar-Asher Siegal & N. Boneh (Eds.), *Perspectives on causation: Selected papers from the Jerusalem 2017 workshop* (pp. 121–147). Springer International Publishing.

Wiegmann*, A., & **Engelmann*, N.** (2020). Entwicklungen und probleme der moralpsychologie zu beginn des 21. jahrhunderts [Developments and problems in moral psychology in the 21st century]. In N. Paulo & J. C. Bublitz (Eds.), *Empirische ethik - Grundlagentexte aus psychologie und philosophie* (pp. 139–175). Suhrkamp Verlag.

Engelmann, **N.**, & Waldmann, M. R. (2019). Moral reasoning with multiple effects: Justification and moral responsibility for side effects. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 41th meeting of the cognitive science society* (pp. 1703–1709).

Hagmayer, Y., & **Engelmann**, **N.** (2014). Causal beliefs about depression in different cultural groups — What do cognitive psychological theories of causal learning and reasoning predict? *Frontiers in Psychology*, *5*, Article 1303.

Neth, H., **Engelmann**, **N.**, & Mayrhofer, R. (2014). Foraging for alternatives: Ecological rationality in keeping options viable. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36rd meeting of the cognitive science society* (pp. 1078–1083).

* = shared first authorship

Manuscripts under Review

Stephan, S., **Engelmann**, **N.**, & Waldmann, M. R. (2021). The perceived dilution of causal strength [Manuscript submitted for publication].

Scholarships and Awards

- Erasmus+ Scholarship for research stay at University College London, UK, ca. 1500€ 2016
- Erasmus Scholarship for stay at Université Paris Est Créteil Val de Marne, Paris, France, ca. 1500€ 2014
- "Niedersachsenstipendium" (Scholarship awarded by the German Federal Ministry of Education and Research for outstanding grades and political engagement, 500€) 2014
- "Deutschlandstipendium" (Scholarship awarded by the German Federal Ministry of Education and Research for outstanding grades and political engagement, 12 x 300€) 2012–2013

Invited Talks and Workshops

- "A causal proximity effect in moral judgment", *Chicago/Michigan Psychology and Law Studies Lab , September 29, 2021.*
- "A causal proximity effect in moral judgment", *Georgetown Law & Language Lab, September 13, 2021.*
- Invited discussant, *Workshop "The Moral Psychology of War." Oxford Institute for Ethics, Law, and Armed Conflict (ELAC) & Nuffield College Oxford, May 13 - 14, 2021.*
- Invited author, *book workshop "Empirische Ethik". Center for Interdisciplinary Research at Bielefeld University, Germany, August 29 - 30, 2019.*
- "Norms and causation - Current empirical findings" (with Lara Kirfel), *Workshop "The Experimental Philosophy of Morality and Causation - Perspectives from Philosophy, Psychology, and Law." Ruhr University Bochum, Germany, June 13, 2017.*

Conference Talks

- Wiegmann, A. & Engelmann, N., "Is lying morally different from misleading? An empirical investigation", *Workshop "Commitments in Grammar and Discourse" at the Annual Meeting of the German Society for Linguistics, February 23, 2022.*
- Engelmann, N. & Waldmann, M.R., "Moral reasoning with multiple effects: Exploring the scope of a subjective-utilitarian approach", *Annual Meeting of the European Society for Philosophy and Psychology (ESPP), Titania Hotel, Athens, Greece, September 6, 2019.*
- Engelmann, N. & Waldmann, M.R., "Moral reasoning with multiple effects: Exploring the scope of a subjective-utilitarian approach", *European Conference for Cognitive Science, Ruhr University Bochum, Germany, September 3, 2019.*
- Engelmann, N. & Waldmann, M.R., "Moral reasoning with multiple effects", *Tagung experimentell arbeitender Psychologen (Teap), London Metropolitan University, UK, April 17, 2019.*

Poster Presentations

- Engelmann, N. & Waldmann, M.R., "How to weigh lives. A computational model of moral judgment in multiple-outcome structures.", *Poster presentation at the workshop "Engineering and Reverse-Engineering Morality" at the Annual Meeting of the Cognitive Science Society, July 26, 2021 (online)*
- Engelmann, N. & Waldmann, M.R., "A causal proximity effect in moral judgment", *Poster presentation at the Annual Meeting of the Cognitive Science Society, July 26 - 29, 2021 (Online)*
- Engelmann, N. & Waldmann, M.R., "A causal proximity effect in moral judgment", *Poster presentation at the Annual Meeting of the Society for Philosophy and Psychology (SPP), June 28 - July 02, 2021 (Online)*
- Engelmann, N. & Waldmann, M.R., "Moral reasoning with multiple effects", *Poster presentation at the Annual Meeting of the Cognitive Science Society, Palais des Congrès, Montréal, Canada, July 27, 2019.*
- Engelmann, N. & Waldmann, M.R., "Moral reasoning with multiple effects", *Poster presentation at the RTG 2070 conference "Understanding Social Relationships", German Primate Center, Göttingen, Germany, October 17, 2018.*
- Engelmann, N. & Waldmann, M.R., "The interaction between causality, foreseeability, and outcome valence in attributions of moral responsibility", *Poster presentation at the UK Xphi Conference, University College London, UK, June 14, 2018.*
- Viebahn, E., Wiegmann, A., Engelmann, N., & Willemsen, P., "Can a question be a lie? An empirical investigation", *Poster presentation at the UK Xphi Conference, University College London, UK, June 14, 2018.*
- Stephan, S., Engelmann, N., & Kirfel, L., "The influence of learned statistical abnormality on singular causation judgments", *Poster presentation at the UK Xphi Conference, University College London, UK, June 14, 2018.*
- Engelmann, N., Stephan, S., & Lagnado, D., "Functional norms and causal judgment", *Poster presentation at the Annual Meeting of the European Society for Philosophy and Psychology (ESPP), University of Hertfordshire, UK, August 16, 2017.*

Teaching

Courses

Quantitative Methods I (Bachelor Psychology): basics of research design and hypothesis testing, data visualisation, descriptive and inferential data analysis, data visualisation, power analyses. Practical exercises in MS Excel.

- Winter 2017/18, Winter 2018/19, Winter 2019/20, Winter 2020/21, Winter 2021/22

Quantitative Methods II (Bachelor Psychology): General Linear Model and applications (multiple linear regression and its assumptions, ANOVA, contrast analyses, logistic regression, multilevel models). Practical exercises in R and RStudio.

- Summer 2018, Summer 2019, Summer 2020, Summer 2021

Thesis Supervision

- **Moral intuitions about death.** M.Sc. thesis, A. Rinn, Summer 2022 (co-supervision with Michael R. Waldmann).
- **Moral intuitions about death.** M.Sc. thesis, F. Koniusch, Summer 2022 (co-supervision with Michael R. Waldmann).
- **Killing as a side effect: Investigating the violability of soldiers and civilians in war.** B.Sc. thesis, K. McElyea, Summer 2021 (co-supervision with Juan Carlos Marulanda).
- **AI and moral decisions: Can experience learning and process transparency increase acceptability?** M.Sc. thesis, A. Zobott, Summer 2020.
- **Lying, deceiving, misleading: Moral evaluation and the influence of perspective.** M.Sc. thesis, E. Köhler, Summer 2020 (co-supervision with Dr. Alex Wiegmann).
- **The influence of causal structure and beliefs about causal structure on judgments of causation and moral responsibility.** B.Sc. thesis, S. Ernst, Summer 2020.
- **Moral reasoning with multiple effects: Intentionality and moral responsibility for side effects.** B.Sc. thesis, M. Rocholl, Summer 2019.
- **The role of outcome valence for causal attributions in disjunctive causal structures.** B.Sc. thesis, L. Ahlwes, Summer 2019.
- **Causal questions and explanation.** B.Sc. thesis, L. Bertz, Summer 2018 (co-supervision with Prof. York Hagmayer).

Other Academic Experience

- **Research assistant to York Hagmayer** 2012–2017
  Georg-Elias-Müller-Institute for Psychology, Department of Cognitive and Decision Sciences.
- **Teaching assistant to Hansjörg Neth** 2013
  Georg-Elias-Müller-Institute for Psychology, Department of Cognitive and Decision Sciences, Seminar "Basic Academic Skills".
- **Research assistant to Miriam Ellert** 2012–2013
  Seminar of German Philology, Lab for Psycholinguistics.
- **Teaching assistant to Willi Hager** 2011–2012
  Georg-Elias-Müller-Institute for Psychology, Department of Cognitive and Decision Sciences, Seminar "Quantitative Methods I & II".

Service and other Activities

- **Member of search committee for a professorship in work and organizational psychology** 2021/22
  Institute for Psychology, Georg-August-University Göttingen.
- **Participant in Dorothea Schlözer mentoring programme** 2021 - 2022
  Career mentoring for female PhD candidates and postdocs
- **Co-organizer of R-Ladies Göttingen** since 2021
  Organization to promote gender diversity in the R community
- **Member of the ethics board** since 2020
  Institute for Psychology, Georg-August-University Göttingen.

Ad Hoc Reviews

- *Acta Psychologica*
- *Cognition*
- *Cognitive Science*
- *Journal of Experimental Social Psychology*
- *PLoS one*
- *Proceedings of the Annual Meeting of the Cognitive Science Society*

Languages

- **German:** Native speaker
- **English:** Fluent
- **French:** Conversational

Software and Programming Skills

- R (advanced)
- LaTex (intermediate)
- MatLab (basics)
- Python (basics)

References

- Prof. Dr. Michael R. Waldmann, Department of Psychology, University of Göttingen, Germany.
  michael.waldmann@bio.uni-goettingen.de
  +49551 3933784
  https://www.psych.uni-goettingen.de/en/cognition/team/waldmann
- Prof. Dr. York Hagmayer, Department of Psychology, University of Göttingen, Germany.
  york.hagmayer@bio.uni-goettingen.de
  +49551 398293
  https://www.psych.uni-goettingen.de/de/cognition/team/hagmayer
- Dr. Dr. Alex Wiegmann, Institute for Philosophy II, Ruhr University Bochum, Germany.
  Alexander.Wiegmann@ruhr-uni-bochum.de
  +49 234 32 28207