




Cognitive Science 48 (2024) e70012

© 2024 The Author(s). *Cognitive Science* published by Wiley Periodicals LLC on behalf of Cognitive Science Society (CSS).

ISSN: 1551-6709 online

DOI: 10.1111/cogs.70012

Adults and Children Engage in Subtle and Fine-Grained Action Interpretation and Evaluation in Moral Dilemmas

Isa Blomberg,^{a,b}  Britta Schünemann,^a  Marina Proft,^{a,b} 
Hannes Rakoczy^{a,b} 

^aDepartment of Developmental Psychology, Institute of Psychology, University of Göttingen

^bLeibniz-ScienceCampus Primate Cognition, Göttingen

Received 11 April 2024; received in revised form 5 September 2024; accepted 16 October 2024

Abstract

Understanding the actions of others is fundamental for human social life. It builds on a grasp of the subjective intentionality behind behavior: one action comprises different things simultaneously (e.g., moving their arm, turning on the light) but which of these constitute intentional actions, in contrast to merely foreseen side-effects (e.g., increasing the electricity bill), depends on the description under which the agent represents the acts. She may be acting intentionally only under the description “turning on the light,” but did not turn on the light in order to increase the electricity bill. In preregistered studies ($N = 620$), we asked how adults and children engage in such complex subjective action interpretation and evaluation in moral dilemmas. To capture the deep structure of subjects’ representations of the intentional structures of actions, we derived “act trees” from their response patterns to questions about the acts. Results suggest that people systematically distinguish between intended main and merely foreseen side-effects in their moral and intentionality judgments, even when main and side-effects were closely related and the latter were harmful. Additional experimental conditions suggest that, when given ambiguous information, the majority of subjects assume that agents act with beneficial main intentions. This “good intention prior” was so strong that participants attributed good intentions even when the harmful action was no longer necessary to resolve the dilemma (Study 2). These methods

Correspondence should be sent to Isa Blomberg, Department of Developmental Psychology, Institute of Psychology, University of Göttingen, 37073 Göttingen, Germany. E-mail: isa.blomberg@uni-goettingen.de

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

provide promising new ways to investigate in more subtle and fine-grained ways how reasoners parse, interpret, and evaluate complex actions.

Keywords: Intentional action; Moral dilemma; Theory of mind; Act trees; Preregistered

1. Introduction

Understanding the subjective reasons and intentions with which people act is foundational for cooperation, communication, and social evaluation; it is an essential part of our meta-representational Theory of Mind (Perner, 1991). Theory of Mind develops gradually from early childhood on, and even infants have some basic grasp of intentional action, distinguishing, for example, voluntary actions from mere behavior (Behne et al., 2005; Gergely et al., 2002; Meltzoff, 1995; Woodward, 1998).

Often the actions that we care about practically and morally in our everyday lives are complex and interpreting them requires understanding their subtle internal intentional structures (Bratman, 1987; Searle, 1983). Most actions can be represented at many hierarchical levels. For example, the referee may blow the whistle, end the game, and end the tournament. She may do each of these things intentionally, and do one (end the tournament) by doing the others (by ending the game, by blowing the whistle). She may be doing many other things simultaneously, but not necessarily intentionally. Blowing the whistle may call a loose dog onto the field, chasing the players, and thus causing huge chaos. Since the referee did not know about the dog, she did not do these things intentionally. Such differentiation between intended effects and unknown and thus unintentional side-effects is an integral part of our everyday action understanding, and even 5-to-8-year-olds reliably make these distinctions (Kamawar & Olson, 2011; Proft & Rakoczy, 2019).

Another class of cases, however, is more challenging but crucial from everyday evaluative and moral points of view: many actions are ambiguous, with different effects all foreseen by the agent, but still the question arises which of these are intended. For example, a person can move her arm, turn the light switch, and thereby turn on the light. She is acting intentionally under these descriptions. At the same time, she is waking up the cat by turning on the light. She might be very well aware of that but it usually is less clear whether she turned on the light in order to wake up the cat—and thus whether she acted intentionally under that description. A particularly relevant case are moral dilemma scenarios in which good and bad effects conflict. To illustrate, consider the following thought experiment of cavers in an underground cave (Foot, 1967; Fuller, 1949; Levine, Leslie et al., 2018): The cave is filling with water. Caver X gets stuck in the only exit while trying to escape. Caver Y has dynamite with her. The dilemma is this: If Y does nothing, everyone dies. Alternatively, if she clears the entrance by killing the stuck caver X with her dynamite, everyone is saved (except X). Hence, the only available means of saving all the other cavers is blowing up and thus killing X.

Now, imagine that Y finally decides to use the dynamite. How are such actions interpreted? What can the agent be blamed for? This has been extensively discussed under the rubric of the “doctrine of double effect.” The doctrine of the double effect holds that it is sometimes

permissible to perform an act which causes a harmful side-effect, as long as it is not intended (McIntyre, 2019). Empirical studies have shown the link between intentional action and moral permission, so that actions may be permissible when the harm caused was not intentional but a foreseeable side-effect rather than a means to an end (Cushman, 2008; Greene et al., 2009; Levine, Mikhail et al., 2018; Mikhail, 2007; Proft et al., 2019; Proft & Rakoczy, 2019; Waldmann & Dieterich, 2007).

However, it is not trivial to decide which effects are intended and which are merely foreseeable. This becomes particularly clear in the discussion of the so-called “closeness problem”: Given two descriptions that both apply to a given action (to take an example from the cavers dilemma: “blowing up the thing” and “killing the man”), how close can action descriptions be conceptually such that observers can still meaningfully say the agent performed one action intentionally without performing the other intentionally? One influential response, the so-called “closeness argument,” is that in many morally charged dilemmas, the relations of means to good ends and the bad side-effects caused by these means are too close to warrant meaningful differentiation (Foot, 1967): In the cave example, “blowing up the man” and “killing him” are *conceptually* too close to seriously claim that the caver intended only the former (i.e., “blowing him up”) but not the latter (i.e., “killing him”).

Others have argued, in contrast, that observers can and do differentiate between intended main effects and merely foreseen side-effects even in such cases: “blowing up the man” is intended as a means to saving the others, but not all states of affairs caused by this act, in particular killing him (Masek, 2010). There are various descriptions under which actions can be represented. Differentiation of intentional action descriptions may be even more fine-grained (Levine, Leslie et al., 2018). For example, the caver might be seen to only act intentionally under the description “blowing up *the thing* blocking the exit,” which happens to be a living human, but he is not necessarily acting intentionally under the description “blowing up the man,” let alone “killing the man” (Levine, Leslie et al., 2018).

How can we empirically decide between these positions? One previous study has already investigated that question and found empirical support for a more fine-grained reading of the agent’s intentions (Levine, Leslie et al., 2018). In the present study, we aim to build on this evolving line of research, extend it to “closeness” cases, and make use of act tree formalism (Goldman, 1970; Levine, Leslie et al., 2018; Mikhail, 2007). Act trees are a means of representing how subjects represent actions as a whole and which parts are regarded as means or side-effects. In act trees, all act descriptions are represented as nodes (see Fig. 1). Basic acts (B), represented as lower nodes, generate means (M) located above them, from which side-effects (SE) branch off, and ends (E), which sit above the means. Lower nodes generate upper ones irrespective of the branch. In the example in Fig. 1, the basic act to detonate dynamite (B) generates the effect that everyone gets dirty (SE) and that the cave entrance is cleared (M). Nodes are connected in several ways: black dashed links indicate causal connections (“by detonating dynamite the agent made everyone dirty”) or constitutive ones (“by raising his hand, the agent voted”). However, from an intentional point of view, not all causally or constitutively linked nodes may be intended by the agent and thus lie on the same act tree branch. Red solid links indicate connections between nodes that are causally/constitutively,

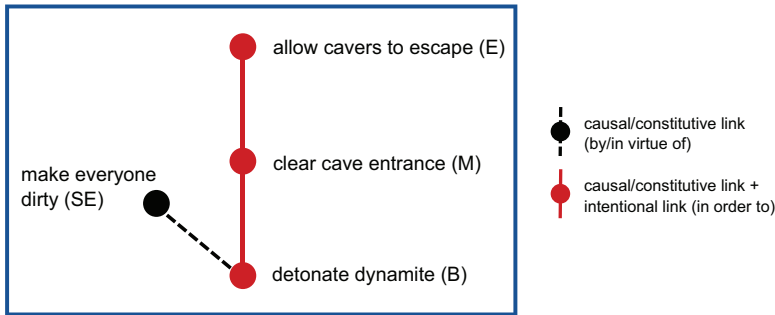


Fig. 1. Act tree representation of a simplified version of the cavers dilemma with a basic act (B), means (M), ends (E), and a side-effect (SE). All figures based on Levine, Leslie et al. (2018).

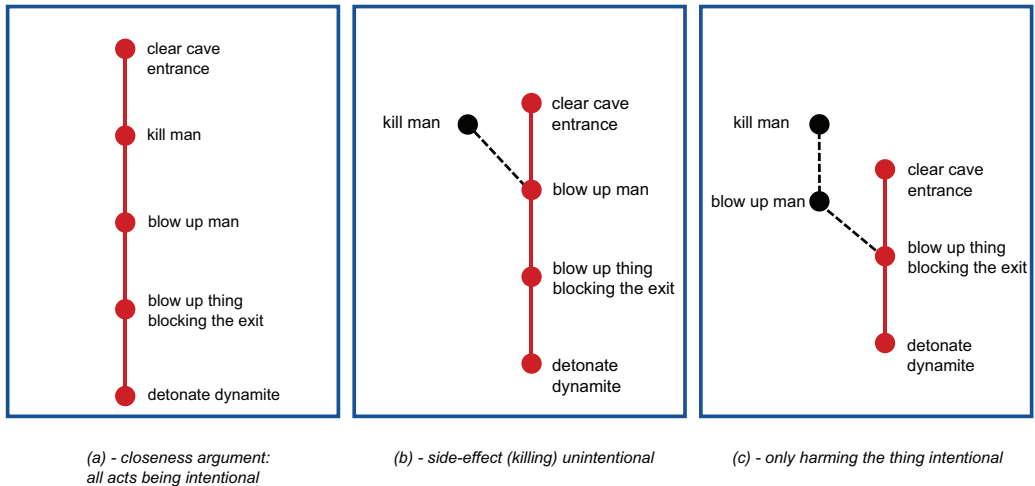


Fig. 2. Act tree representations of the cavers dilemma depicting various options that do (a) or do not (b, c) follow the “closeness argument.”

and additionally intentionally linked (“the agent detonated the dynamite in order to clear the cave entrance”).

To find out which act tree representations underlie subjects’ action interpretation, one can probe their linguistic use of ordinary expressions like “in order to” in describing a given scenario (Knobe, 2010; Levine, Leslie et al., 2018). Recent studies showed that subjects affirmed statements in which lower nodes of intentional acts are connected via “in order to” to upper nodes (Levine, Leslie et al., 2018). Subjects, however, rejected these statements when comprising unintended side-effects on a branch diverting from the main branch (e.g., “detonating dynamite” in order to “make everyone dirty”) or connecting upper to lower nodes (e.g., “clear cave entrance” in order to “detonate dynamite”). Possible underlying act trees can be distinguished by analyzing response patterns to clusters of such test questions. Regarding the closeness problem, the positions mentioned above translate into the act trees depicted in Fig. 2. Act

tree (a) follows the closeness argument: with respect to intentionality, no distinction can be made between the node “kill man” and “blow up man.” In contrast, act tree (b) and (c) represent the counter position to the closeness argument. In act tree (b), the node “kill man” is represented as distinct from “blow up man” and thus an unintended side-effect, not placed on the main branch. Act tree (c) represents the even stricter counter position to the closeness argument: the agent acts intentionally only under the description “blow up thing.” Both “kill man” and “blow up man” are placed on side branches.

1.1. Act tree construction: How do subjects represent the intentional structure of closeness cases?

Building on recent studies using the act tree method (Knobe, 2010; Levine, Leslie et al., 2018), and closely following up on Levine, Leslie et al. (2018), the first goal of the present studies was to test which model best describes how people represent such complex closeness cases. Do they categorically distinguish between main and side-effects in terms of their intentionality? That is, do subjects place main and side-effects on different act tree branches (options b or c)? Or do they map them onto one unitary act tree, according to the closeness argument (option a)? To address this question, we systematically analyzed response patterns to multiple tailor-made questions that probed the relation of various action descriptions via “in order to” mappings (Levine, Leslie et al., 2018).

1.2. Cross-branch linking in constructed act trees

Our first research question is thus which act trees are constructed. Put simply, act trees are constructed by asking questions starting from the basic action: “Did the agent do B [basic action] in order to do M/E [means/end]?”. If the answer is “yes,” an act is considered intentional, and it is put on the main branch of intentional actions. In contrast, if for a given act the answer to the question “Did the agent do B [basic action] in order to do M/E [means/end]?” is “no,” then this act is not considered intentional and is not placed on the main intentional branch, but on a side branch of nonintended acts and effects. Act trees are then constructed from there with diverging branches. Nodes on the main branch¹ are linked by both causal/conventional and intentional relations. Nodes on the side branch in contrast are only related in causal/conventional terms (the side-effect is brought about by doing the basic act, etc.) and not in intentional terms (the basic act is not done in order to do the side-effect).

Following Goldman’s act tree formalism, the separation between main and side branches should be categorical and strict: First, side-effects in themselves are not brought about intentionally, and subjects should thus disagree with statements like:

I “Agent detonated dynamite [basic action] in order to kill man [side-effect].”

Second, side-effects are not intentionally related in any way with ends on the main branch. Subjects should thus disagree with statements like:

II “Agent made everyone dirty [side-effect] in order to clear the entrance [end].”

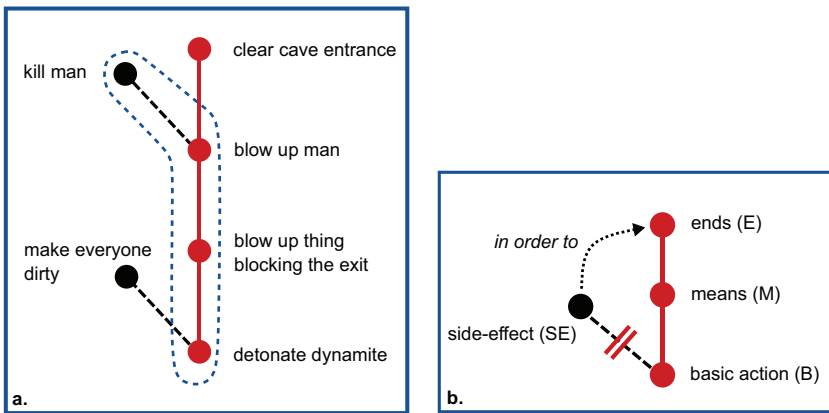


Fig. 3. (a) Act tree that indicates how “kill man” seems to be chunked as described in the section above. (b) shows the seemingly paradoxical pattern of cross-branch linking of the side-effect with the goal. On the one hand, subjects deny that the basic action is done in order to do the side-effect. On the other hand, they affirm that the side-effect is done in order to achieve the overlying goal.

Empirically, however, reality may be messier than that: subjects seem to link side branches with main branches for some purposes in various situations. For example, research on the side-effect effect suggests that for some side-effects, subjects would agree with statements of type (II), while rejecting statements like (I) (Knobe, 2003a, 2003b, 2010): The CEO of a company is implementing a policy [basic act] in order to make profit [end], thereby also (unintentionally) harming the environment [side-effect]. Subject might then deny statements of type (I) (“the CEO implemented a policy [basic act] in order to harm the environment [side-effect]”); while accepting statements of the type (II) (“the CEO harmed the environment [side-effect] in order to make profit [end]”). When asked in goal terms (i.e., whether the basic act was done in order to do the harmful act), people disagree with (I)-like statements. However, when asked in action terms (i.e., whether the harmful act was done in order to do the goal), subjects affirm these (II)-like statements.

What makes the combination of judgments about these two statements so intriguing is that they suggest opposite things about how participants conceptualize the structure of action (see Fig. 3b). Applied to the present “closeness case,” the most natural way to explain why people think statement (I) of the cavers dilemma is wrong is that they make a distinction between “removing the thing blocking the exit” and “killing the man.” They think the agent performed the action in order to “remove the thing blocking the exit,” but not in order to “kill the man.” However, the most natural way to explain why people think that statement (II) is right, is that they are *not* drawing a distinction between “removing the thing blocking the exit” and “killing the man.” They treat “removing the thing blocking the exit” and “killing the man” as basically the same thing (i.e., this one thing that was done in order to achieve the goal).²

How can we make sense of this paradoxical pattern? One recent idea is that subjects may flexibly chunk nodes and branches in context-relative ways (Levine, Leslie et al., 2018; see also Fig. 3). Like, one can represent elements individually or chunk them into bigger units in memory (Miller, 1956), one can focus on the elements (nodes) on a given branch, or on the

whole branch as a bigger unit. One can then mentally zoom-in on elements of the sequence when asked in goal terms and mentally zoom-out when asked in action terms. We will discuss these ideas in more detail in the General Discussion.

The phenomenon of cross-branch linking is interesting in general as it challenges the conception of unintended side-effects. It is interesting in closeness cases in particular, because it highlights how differently people could think about closely related and morally charged act descriptions in action- versus goal-frames. If such paradoxical patterns can be found in the cavers dilemma, it suggests that people make more fine-grained distinctions when representing goals, even if those goals come about by the same action (distinction between “use dynamite” in order to “remove thing” or “kill man”). On the other hand, when it comes to descriptions of different actions as means, people tend to engage in chunking and treat those action descriptions interchangeably (i.e., no distinction between “removing thing” and “killing man” in order to “save all”).

Empirically, existing studies have not yet investigated whether this paradoxical pattern of cross-branch linking emerges systematically in connection with and as a function of action interpretation and act tree construction (i.e., see our options (a)–(c) discussed above and Fig. 2). Once participants create act trees with main and side branches (i.e., act tree [b—side-effect unintentional] or [c—only harming *thing* intentional]), do they also engage in cross-branch linking? That is, do they agree with statements like (II)? Levine, Leslie et al. (2018) argue for and presented indirect evidence across separate studies suggesting that participants might engage in cross-branch linking. Here, we seek more direct evidence by applying the act tree method.

Therefore, we investigate if act trees constructed in closeness cases systematically include cross-branch linking. To do so, we examine participants’ responses to both traditional act tree construction questions, and to the additional questions probing cross-branch linking (i.e., “act SE [side-effect] in order to act E [end]?”), and we test if such patterns coemerge systematically.

1.3. Which prior assumptions go into disambiguating complex cases of act tree construction?

A second, related, question underlying pertaining to act tree construction was the following: In cases with several foreseen effects that remain ambiguous with respect to intentionality, how do observers determine the intentional structure of the action? Which prior assumptions underlie their act tree construction, action interpretation, and evaluation? A recent study on moral dilemmas claimed and presented evidence that adults and children adhere to a prior assumption that agents act with good intentions (Levine, Mikhail et al., 2018). In these moral dilemmas, the agent’s basic act had positive and negative effects. In the ambiguous baseline condition, no explicit motive of the agent was introduced; in two other conditions, the agent explicitly expressed either a motive for the positive or for the negative effect (see Fig. 4). Results revealed that when asked forced-choice questions about whether the agent acted in order to achieve A (positive effect) or B (negative effect), subjects rated the ambiguous baseline condition like the condition with an explicitly good motive (and rated both conditions differently from the condition with the explicitly bad motive).

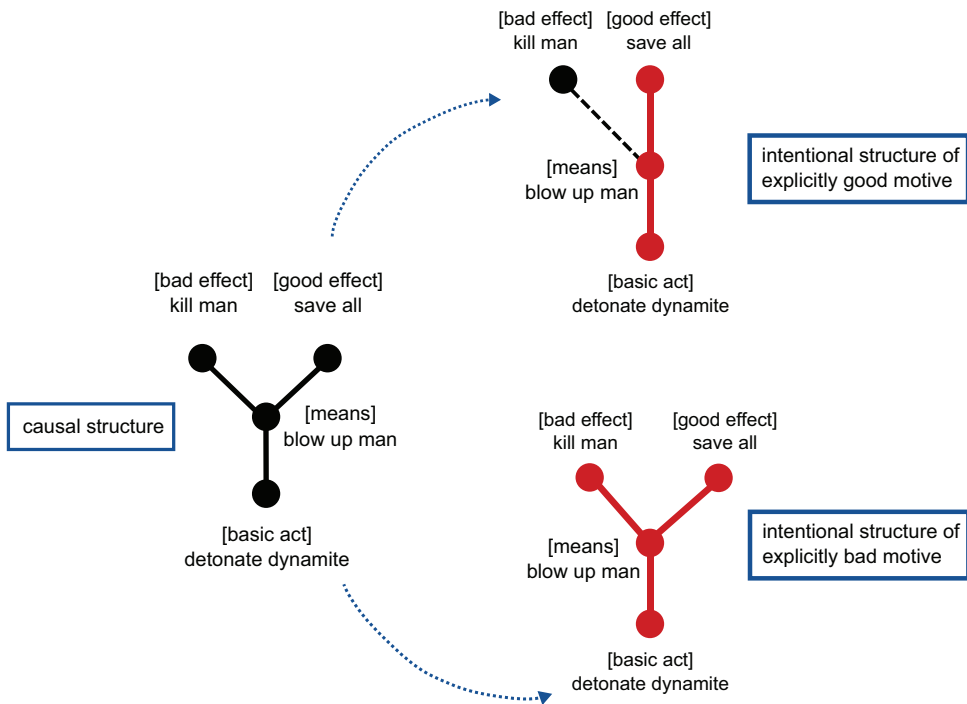


Fig. 4. Causal structure (left) and possible intentional structures (right) applied to the cavers dilemma. Please note that there are several ways in which the intentional structure could manifest in the explicitly bad motive condition. Importantly, the effect of killing the man should be considered an unintended, foreseen side-effect. See also footnote 3 for more details.

We applied a similar logic to our study design.³ Besides an ambiguous baseline condition in which an adaptation of the cavers dilemma was introduced, two additional conditions were implemented in which either a positive or a negative intention was stated explicitly. The rationale was to investigate whether the ambiguous baseline condition is interpreted similarly to the explicit bad or explicit good motive condition. The question was whether the negative effect (i.e., killing the man) was placed on a side-branch (as plausible in the explicit good motive condition) or not (as plausible in the explicit bad motive condition). This allowed us to investigate the relation of prior assumptions and act tree representations within one case. To test how robust participants' prior assumptions are, we further modified the moral dilemma so that the harmful action was no longer necessary to achieve the good effect and tested whether participants still attribute good intentions to the protagonist.

1.4. How do these forms of action interpretation develop?

The third question concerned cognitive development: One central developmental question is how the capacity for action interpretation and evaluation of such complex main effect/side-effect cases develops. We know that infants show some basic form of intentional action

interpretation (e.g., Gergely et al., 2002); and that later in development, children from around age 4–5 gradually develop more sophisticated forms of action interpretation such as distinguishing between main and side-effects (Leslie et al., 2006; Pellizzoni et al., 2009; Rakoczy et al., 2015) or evaluating actions as a function of underlying intentions rather than mere outcomes (Cushman et al., 2013; Helwig et al., 2001; Killen et al., 2011; Proft & Rakoczy, 2019). It is an open question how children come to represent the fine-grained structure of complex actions such as the closeness cases in the form of act trees. Since such act tree representation is a complex capacity with many cognitive presuppositions, we tested here children at elementary school age when the presupposed capacities can be expected to be firmly in place.

1.5. *The present study*

In sum, building on an evolving line of research (Knobe, 2010; Levine, Leslie et al., 2018; Levine, Mikhail et al., 2018) and combining the new method of mapping act trees as a window into the underlying action representations, moral evaluations, and prior assumptions in adults and children, we aim to address three questions:

1. By building on Levine, Leslie et al. (2018) we ask, how subjects represent the intentional structure of harmful, foreseen actions in closeness cases. In other words, we address which act trees are constructed: which act descriptions are considered intentional and represented on the main branch; which nodes are represented as unintended and put on a side-branch? If participants construct act trees with main and side branches, do they systematically engage in cross-branch linking by linking side-effects on a side branch to ends on the main branch?
2. In ambiguous cases with positive and negative effects, what default assumption do reasoners make about the intentionality when disambiguating complex actions? How robust are these assumptions? Essentially, we ask which priors underlie act tree constructions. How robust are these prior assumptions?
3. How do these capacities and reasoning patterns of action interpretation emerge in child development? Here, we address these different questions (pertaining to action interpretation of “close” action descriptions, the seemingly paradoxical phenomenon of cross-branch linking, and participants’ prior assumptions in act tree constructions) in one paradigm with adults and children.

For Study 1, we developed a child-friendly version of the covers dilemma to be able to compare children and adults. For Study 2, we used the original covers dilemma again and tested how robust participants’ good prior assumptions about the protagonist are. All materials, pre-registrations (only for Study 1 here: <https://osf.io/53zqh>), data, and analyses are uploaded to the Open Science Framework (OSF): <https://osf.io/29jqv/>.




Scenario	Jakob cuts the rope (B) to throw off the ballast/heavy thing (M_1) / the sheep (M_2) to prevent the balloon from hitting against the mountain (E), thereby killing/hurting the sheep (H)		
Condition	Baseline	Intention[+]	Intention[-]
			
Motive	–	"I want to save us! This is my chance. This way, I can drop ballast/the heavy thing!"	"I hate sheep! This is my chance. This way, I can kill/hurt the sheep!"
Intentional action questions	<p><i>Open-ended question:</i> Why did Jakob cut the rope? <i>Closed questions</i> (randomized order):</p> <ol style="list-style-type: none"> 1. B <i>in order to</i> M_1: Did Jakob cut the rope in order to throw off the ballast/heavy thing? 2. B <i>in order to</i> M_2: Did Jakob cut the rope in order to throw off the sheep? 3. B <i>in order to</i> H: Did Jakob cut the rope in order to kill/hurt the sheep? 4. B <i>in order to</i> E: Did Jakob cut the rope in order to prevent the hot air balloon from hitting against the mountain? 5. H <i>in order to</i> E: Did Jakob kill/hurt the sheep in order to prevent the hot air balloon from hitting against the mountain? 		
Moral evaluation	<p><i>Study 1a:</i> Adults evaluated the moral acceptability of the basic act cutting the rope on 7-point scale from 1 ("No, not acceptable at all") to 7 ("Yes, fully acceptable"). <i>Study 1b:</i> Children evaluated the basic act on a four-point smiley scale.</p>		

Fig. 5. Experimental logic.

2. Study 1a: Adults

2.1. Method

2.1.1. Participants

Two hundred twenty-two German speaking adults (age: $M = 32.49$, $SD = 12.95$, 146 women, 76 men, 0 diverse) took part in our preregistered online-study. Sample size was calculated a-priori based on a multiple regression analysis with medium effect size and preregistered. We conducted convenience sampling. All participants were recruited via social media and personal contacts. All participants gave informed consent to take part in the study.

2.1.2. Design

In a between-subjects design, subjects were randomly assigned to one of three conditions by the built-in randomization module of the study platform formr.org (Arslan et al., 2020). Randomization resulted in roughly equal group sizes in the three conditions (baseline: $n = 75$, intention[+] $n = 74$, intention[-] $n = 73$). Conditions differed only with respect to the expressed motive of the action presented in the test situation. Each subject saw one video of a moral dilemma. A detailed structure of the experiment is given in Fig. 5.

2.1.3. Materials

All videos were animated with Vyond (GoAnimate Inc., 2020). To increase reproducibility, all materials, including all test questions and code to recreate the survey, were uploaded to OSF.

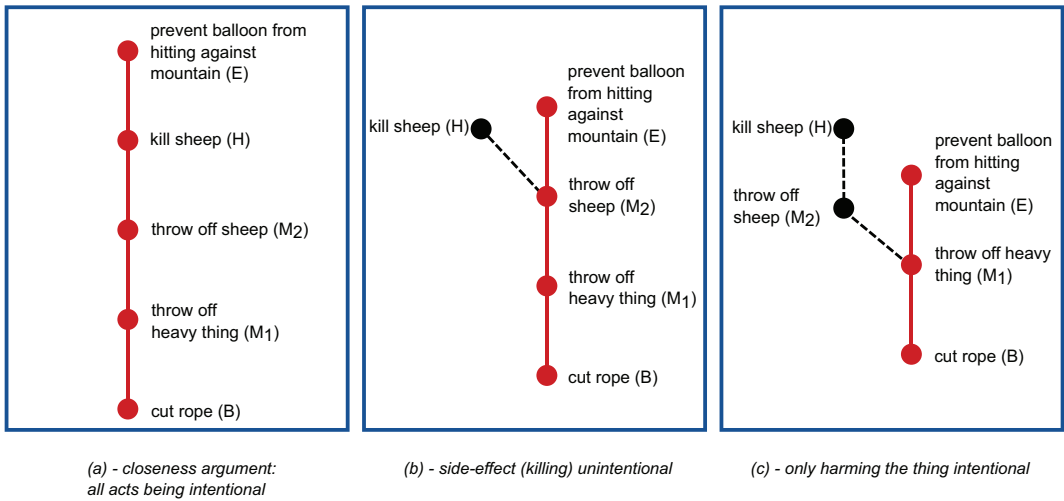


Fig. 6. Act tree representations of our sheep scenario depicting various options that do (a) or do not (b, c) follow the “closeness argument.”

2.1.3.1. Moral dilemma: To enhance the generalizability of previous findings, we devised a new scenario with the same logic as the cavers dilemma suitable for all age groups (adults but also children). Key features of the cavers dilemma were kept constant. The resulting scenario was the following: Jakob and Anna were in a hot-air balloon near a mountain which they could not cross. They could get rid of ballast (M₁) by cutting a rope (B) connecting baskets of sheep to the balloon and, therefore, throwing them off (M₂) and causing them to die (H)—the harmful effect. This prevented the balloon from hitting the mountain (E). If they did nothing, all (Anna, Jakob, and the sheep) would die. In every video, Jakob cut the rope. Moral dilemmas were presented via animated videos. Analogous to the cavers’ dilemma, we derived act trees depicting the positions on the closeness problem (see Fig. 6).

2.1.3.2. Conditions: Conditions (baseline, intention[+], and intention[-]) differed only with respect to the agent’s motive. In the baseline condition, subjects were presented with the scenario without any additional motivation for the basic action of cutting the rope. In the intention[+] condition, Jakob expressed a motive for the good effect while performing the critical basic action (“I want to save us! This is my chance. This way I can throw off ballast!”). In the intention[-] condition, Jakob expressed a motive for the bad effect (“I hate the sheep! This is my chance. This way I can kill the sheep!”). The causal and intentional structure of the conditions are depicted in Fig. 7.

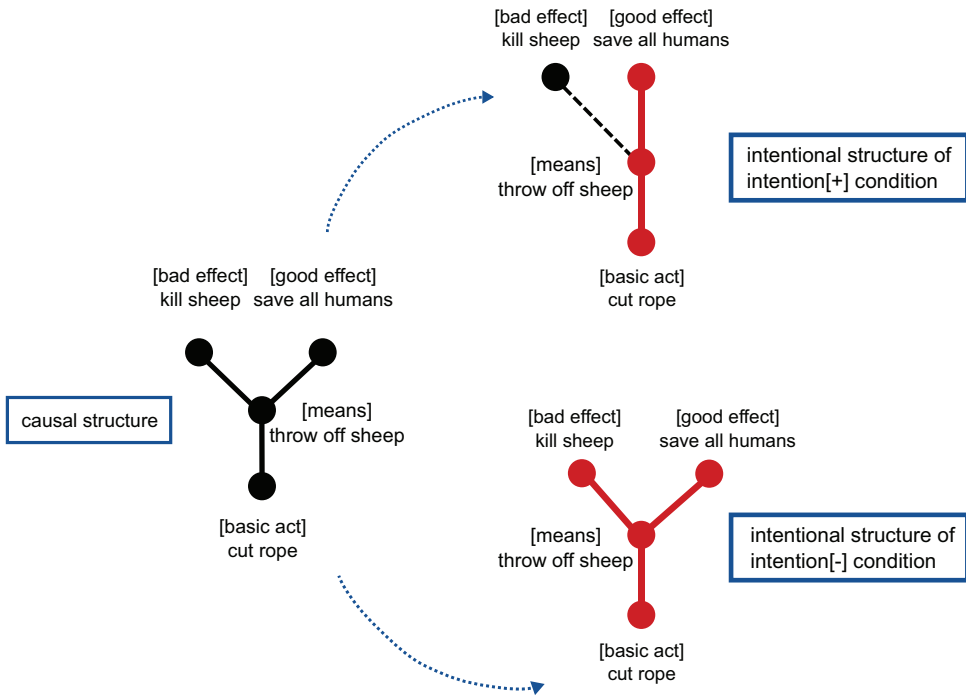


Fig. 7. Causal structure (left) and intentional structure (right) applied to the sheep dilemma. Please note that there are several ways in which the intentional structure could manifest in the intention[-] condition. Importantly, the effect of killing the man should be considered an unintended, foreseen side-effect. See also footnote 3 for more details.

2.1.4. Procedure

Adults watched the respective moral dilemma and answered a set of test questions (see Fig. 5) to elicit participants' mental representation of the intentional structure of the acts B, M_1 , M_2 , E, and H (see also Supplement B for interpretation of open-ended questions). We used a 7-point scale to measure whether the act was morally acceptable (7) or unacceptable (1). Subjects were debriefed after completion of the study. The procedure was in accordance with the recommendation of the ethics committee of the first author's university and the General Data Protection Regulation of the European Union.

2.1.5. Data analysis

Analyses followed the preregistered analysis protocol unless noted otherwise. All analyses were conducted using R (R Core Team, 2021). For each model, we used Likelihood Ratio Tests to compare the fit of the full model to that of a null model which was identical but lacked the predictors of interest. This way, we tested the overall effect of our fixed effects and avoided "cryptical multiple testing" (Forstmeier & Schielzeth, 2011). If not stated otherwise, the full-null model comparison was significant.

Table 1
Assigned act trees based in participants' response patterns

	Question	Act trees			other
		(a) - closeness argument: all acts intentional	(b) - side-effect (killing) unintentional	(c) - only harming the <i>thing</i> intentional	
decisive for differentiation between act trees					
1	B in order to M ₁	Yes	Yes	Yes	all other combinations of responses
2	B in order to M ₂	Yes	Yes	No	
3	B in order to H	Yes	No	No	
4	B in order to E	Yes	Yes	Yes	
additional: side-effect in order to goal					
5	H in order to E				

B = basic act (cut rope), M₁ = means (throw off ballast), M₂ = means (throw off sheep), H = harmful act (kill sheep), E = end (prevent the hot air balloon from hitting against the mountain).

Each act tree predicted a specific, preregistered pattern of responses to questions 1–4 following Goldman's theory on act trees. The analyses of all possible act trees (beyond those discussed under the closeness problem) are described in Supplement D. Each participant's response pattern was assigned to the respective act tree shown in Table 1. Take act tree (a—closeness argument) as an example. Act tree (a—closeness argument) represented that all descriptions of the named actions are too close to warrant distinction, and, therefore, are all intentional. Accordingly, a participant's response was assigned to this act tree when she affirmed the following statements: Jakob cut the rope in order to (1) throw off the ballast, (2) to throw off the sheep, (3) to kill the sheep, and (4) to prevent the balloon from hitting the mountain. In contrast, subjects who affirmed all statements but not (3) to kill the sheep were assigned to act tree (b—side-effect unintentional) in which the killing is represented as a side-effect. Participants who additionally rejected statement (2) to throw off the sheep were assigned to act tree (c—only harming *thing* intentional) because they represented the agent as acting only intentionally under the description “throwing off the ballast.” Response patterns that did not fit the predicted pattern were classified as “other.”

2.1.5.1. Act tree construction: How do adults represent the intentional structure of closeness cases? To answer our first research question, whether subjects distinguished between main and side-effects in their respective intentionality, we only analyzed the response patterns in the baseline condition, in which no additional motive was mentioned.⁴ If there is a default understanding of the situation, the corresponding act tree should be more frequent than expected by chance and more frequent than any other act tree. To this end, we compared the frequency of act trees in the baseline condition using an exact binomial test with a chance level of 0.25.⁵

2.1.5.2. Cross branch-linking in constructed act trees: After identifying the act tree structures underlying participants' representations, we investigated whether subjects system-

atically engage in cross-branch linking. We investigated whether the affirmation rates on statements like “agent killed the sheep [side-effect] in order to prevent the balloon from hitting against the mountain [end]” systematically occurred together with act trees with main and side branches (i.e., act tree [b—side-effect unintentional] or [c—only harming *thing* intentional]). We thereby looked at coherent patterns (i.e., act trees that are derived from theory and philosophical discussions) and the response to the additional questions. This is a more stringent test to tap the existence of cross-branch linking than comparing individual judgments of intentionality.

2.1.5.3. Which prior assumptions go into disambiguating complex cases of act tree construction? To test our second research question about which presumptions underlie action representations, we computed a multinomial logistic regression model.⁶ If subjects in the ambiguous baseline condition assumed that the agent acted out of good intentions, the distributions of the act trees should not differ between baseline and intention[+] since in this case expressing the positive motive is nothing new. Whereas in the intention[-] condition, the distribution of the act trees should be different from the other two conditions. The reverse would be true if subjects in the ambiguous baseline condition assumed bad intentions of the agent: same distributions of the act trees in baseline and intention[-] but different from intention[+] condition. To test for the good intention prior, we set up a-priori orthogonal contrasts: The first contrast compared baseline to intention[+] condition. The second contrast compared baseline and intention[+] to intention[-] condition.

2.1.5.4. Moral evaluation: Participants indicated from 1 (“No, not acceptable at all”) to 7 (“Yes, fully acceptable”) how morally acceptable act B (cutting the rope) was. Using this evaluation allowed us to test whether a good intention prior is reflected in people’s moral evaluations (Levine, Mikhail et al., 2018). Analogous to intentionality judgments, if subjects in the ambiguous baseline condition assumed that the agent acted out of good intentions, the moral evaluations should not differ between the baseline and intention[+] condition but from the intention[-] condition. The reverse pattern will be true if people have a presumption of malicious intentions: Moral evaluations should be similar in baseline and intention[-] condition, but different in intention[+] condition. To test whether moral evaluations varied as a function of the agent’s motive, we calculated a linear regression predicting moral evaluation by condition with the same contrasts mentioned above. All other preregistered exploratory analyses are described in detail in the Supplementary Material and are uploaded to OSF.

2.2. Results

2.2.1. Act tree construction: How do adults represent the intentional structure of closeness cases?

Based on responses to the intention questions, participants were assigned to patterns matching underlying act trees. Table 2 shows the frequencies of act trees and other patterns. Act tree (b—side-effect unintentional) was most frequent across conditions (50%), followed by 22%

Table 2
Frequency of act trees in Study 1a

	Act trees			other
	(a) - closeness argument: all acts intentional	(b) - side-effect (killing) unintentional	(c) - only harming the <i>thing</i> intentional	
baseline ($n = 75$)	4	47	20	4
intention[+] ($n = 74$)	5	46	19	4
intention[-] ($n = 73$)	24	18	9	22
Total ($N = 222$)	33	111	48	30

act tree (c—only harming *thing* intentional), and 15% act tree (a—closeness argument). In addition, 14% of the participants answered in nonsystematic patterns.

To test the default understanding of the dilemma, the frequencies of act tree patterns in the baseline condition were compared to chance (0.25) using a binomial test. Act tree (b—side-effect unintentional) was more frequent than expected by chance ($N = 75$, $k^7 = 47$, probability = 0.63, $p < .001$). Act tree (a—closeness argument) patterns, however, were less likely than expected by chance ($N = 75$, $k = 4$, probability = 0.05, $p < .001$) and act tree (c—only harming *thing* intentional) patterns were not significantly different from chance ($N = 75$, $k = 20$, probability = 0.27, $p = .790$). In addition to the preregistered analyses, we found that in the baseline condition, act tree (b—side-effect unintentional) was more likely than act tree (a—closeness argument) ($\chi^2(1, n = 51) = 36.25, p < .001$) and act tree (c—only harming *thing* intentional) ($\chi^2(1, n = 67) = 10.88, p < .001$). This indicated that subjects' default understanding was mostly in line with option b on the closeness problem: the harmful effect was considered as an unintended side-effect.

2.2.1.1. Cross-branch linking in constructed act trees: In exploratory ways, we analyzed whether subjects engage in cross-branch linking. Most participants (89%) constructed act trees representing the harmful effect as not intentionally brought about (i.e., act tree [b—side-effect unintentional] and act tree [c—only harming *thing* intentional]). These subjects engaged in cross-branch linking if they now affirmed that the harmful side-effect was done in order to achieve the goal. Indeed, most subjects (83%) agreed (i.e., agreed with “kill sheep [side-effect] in order to prevent balloon from hitting against the mountain [end]”). This indicated that the vast majority answered in seemingly paradoxical ways of cross-branch linking. Overall, 90% of the subjects who answered in act tree (b—side-effect unintentional) patterns and 83% who answered in act tree (c—only harming *thing* intentional) patterns affirmed this question. The harmful effect was considered as an unintended side-effect when asked in goal terms (i.e., whether basic act was done in order to do the harmful act). When asked in action terms (i.e., whether the harmful act was done in order to do the goal act), subjects affirmed it with high rates and thus, did not treat it as a clear side-effect. We will discuss these findings later.

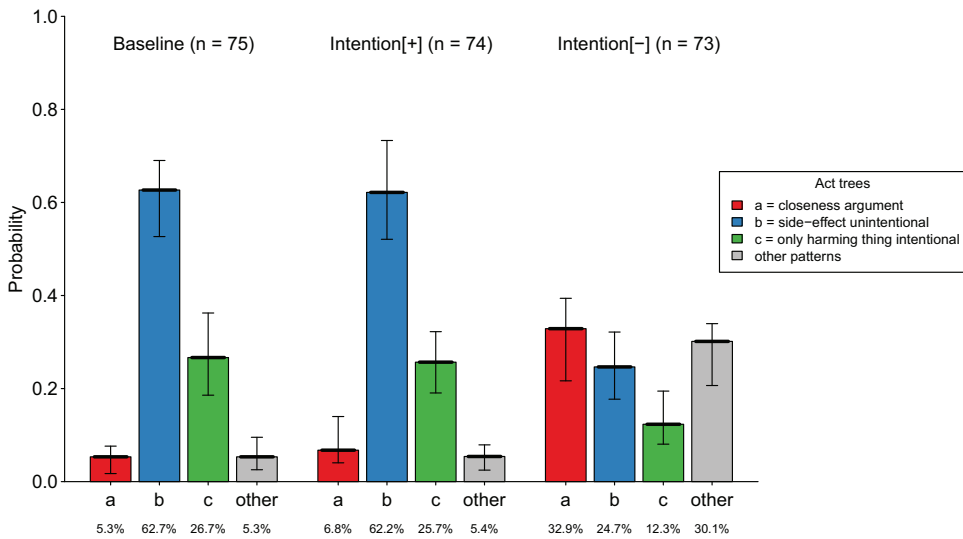


Fig. 8. Distribution of act trees in Study 1a across conditions. Bars and percentages below represent actual proportions of act trees observed. Horizontal lines indicate predicted probability of an act tree by the multinomial model and vertical lines their 95% confidence intervals. Predicted values and their confidence intervals have been obtained via bootstrapping with 1000 boots.

2.2.2. Which prior assumptions go into disambiguating complex cases of act tree construction?

Next, we compared conditions. Multinomial logistic regression analysis revealed that the distribution of act trees differed between conditions as was predicted by a good intention prior (see Fig. 8). No significant differences were found between baseline and intention[+] condition (act tree [b—side-effect unintentional] vs. act tree [a—closeness argument]: $OR = 0.88$, 95% CI [-0.81, 0.57]; vs. act tree [c—only harming thing intentional]: $OR = 1.01$, 95% CI [-0.36, 0.39]). That means that participants interpreted the baseline condition as the intention[+] condition. Thus, they placed the negative effect of killing the sheep on the side-branch. However, participants in intention[-] condition (compared to baseline and intention[+]) were more likely to represent the killing as intentionally brought about than representing the harmful effect (killing the sheep) as an unintended side-effect. That is reflected in the decreased likelihood of answers in the pattern that capture act tree (b—side-effect unintentional) compared to act tree (a—closeness argument) ($OR = 0.42$, 95% CI [-1.18, -0.57], $p < .001$). In other words: the likelihood to regard the harmful effect (killing the sheep) as intended, reflected in act tree (a—closeness argument), was substantially higher when the person was in intention[-] condition than in baseline or intention[+] condition. It should be noted that based on this analysis, it is not yet clear whether the harmful effect was considered as intended means or as a simultaneous goal.

2.2.3. Moral evaluation

The majority of participants morally evaluated the critical act (i.e., cut the rope) as relatively acceptable in all conditions ($M = 5.11$, 95% CI [4.89, 5.33], range = 1–7; see Fig. 9).

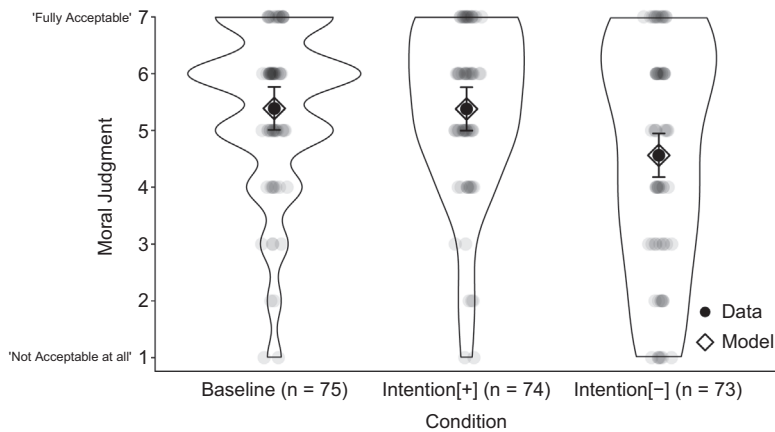


Fig. 9. Adults' mean moral evaluation across conditions. Subjects evaluated how morally acceptable the action to cut the rope was. Each gray point represents the moral evaluation of a participant. Filled black dots indicate the mean moral evaluations per condition and diamond shapes the predicted values by the linear regression model predicting moral evaluations by condition. Vertical lines indicate 95% confidence intervals of the model.

Regression analysis confirmed that moral evaluations in the baseline condition and the intention[+] condition were not different and were evaluated as acceptable ($Mdn = 6$ both, $M_{baseline} = 5.39$, 95% CI [5.06, 5.72]; $M_{intention[+]} = 5.38$, 95% CI [5.03, 5.71]; $b = 0.00$, 95% CI [-0.26, 0.27], $t(219) = 0.03$, $p = .976$) and significantly higher than in the intention[-] condition ($M_{intention[-]} = 4.56$, 95% CI [4.1, 5.02]; $b = 0.27$, 95% CI [0.12, 0.43], $t(219) = 3.46$, $p < .001$; overall small effect of Cohen's $f^2 = 0.035$). The results indicated that moral evaluations of the critical act were influenced by a good intention prior, too. Subjects rated the act as equally morally acceptable when no or a beneficial motive was expressed but as less morally acceptable when the agent stated a malicious intention.

In exploratory ways, we looked at how moral evaluations might be influenced by act trees and condition. The following interaction effects were found. First, participants in the baseline condition who answered in act tree (c—only harming *thing* intentional) patterns gave significantly lower moral acceptability evaluations than participants in the intention[+] condition who answered in act tree (c—only harming *thing* intentional) patterns ($b = -0.68$, 95% CI [-1.28, -0.08], $t(210) = -2.22$, $p = .027$). Second, participants in baseline and intention[+] condition who were classified as “other” patterns gave significantly higher moral acceptability evaluations than participants in the intention[-] condition ($b = 0.73$, 95% CI [0.22, 1.24], $t(210) = 2.81$, $p = .005$). No main effects were significant.

3. Study 1b: Children

In Study 1a, we found that the default understanding of the moral dilemma was mainly in line with option b, and to a lesser degree with option c. When no additional motive was expressed, adults considered the harmful effect of killing the sheep as not directly intended.

Results also suggested that adults are guided by good intention priors when interpreting the scenario. Only when the agent explicitly stated his bad intentions of harming the sheep, adults did change their ascription of intentionality (i.e., considering the harmful effect as intended) and responsibility (i.e., lower moral acceptability evaluations of the critical act). To investigate these findings further taking a developmental perspective, we conducted the preregistered Study 1b.

3.1. Method

3.1.1. Participants

Sample size calculations were based on a power simulation using the data of Study 1a (see OSF for more details). One hundred sixteen 8- to 10-year-old German speaking (96–131 months, $M = 112.87$, $SD = 9.94$, 64 girls, 51 boys, 0 diverse) children took part in the noninteractive online study. Additional five children were tested but excluded because of technical problems ($n = 1$), a parent-reported developmental disorder ($n = 1$), wrong answers in familiarization trials ($n = 2$), and wrong answers in a control question ($n = 1$). For a detailed description of the exclusion criteria, see the preregistration. Children were recruited through our database, to which parents had previously given consent, our website, and social media. Parents gave their informed consent for their children to participate in the study. We administered the same one-factorial between-subjects design as in Study 1a.

3.1.2. Materials and procedure

Study 1b was conducted using the online platform LabVanced (Finger et al., 2017). Children were randomly assigned to one of three conditions using a built-in module which selected for every new participant the group with the fewest participants at that time (baseline: $n = 39$, intention[+]: $n = 38$, intention[-]: $n = 39$). We used the same materials as in Study 1a (see preregistration). The procedure was adapted to make sure that parents and children can navigate through the online study. All used materials can be found on OSF.

3.1.3. Act tree construction and moral evaluation

Test questions indicative of the act trees were the same as in Study 1a. Only the format was adjusted to be age-appropriate. Thus, all questions were read out or asked by an avatar and could be repeated as often as desired. Children's responses to the open-ended question of why the agent acted as he did were recorded via webcam (see Supplement B). The moral acceptability evaluation was made after the explanation of the scale using a 4-point smiley scale. The scale ranged from a smiley with a red background and the corners of the mouth pulled down all the way (1) to a smiley with a green background and a friendly smile (4). The assessment was given by clicking on one of the smileys.

3.1.4. Data analysis

The same analyses as in Study 1a were computed. They were based on the preregistration of the current study and unless noted otherwise.

Table 3
Frequency of act trees in Study 1b

	Act trees			other
	(a) - closeness argument: all acts intentional	(b) - side-effect (hurting) unintentional	(c) - only hurting the <i>thing</i> intentional	
baseline ($n = 39$)	0	23	14	2
intention[+] ($n = 38$)	1	21	16	0
intention[-] ($n = 39$)	29	5	1	4
Total ($N = 116$)	30	49	31	6

3.2. Results

Overall, we could replicate our findings regarding sophisticated intention ascriptions with 8- to 10-year-old children. In fact, response patterns were even more clear-cut.

3.2.1. Act tree construction: How do children represent the intentional structure of closeness cases?

Children's responses to the intention questions were assigned to patterns matching underlying act trees (see Table 3 for the frequencies). As in Study 1, act tree (b—side-effect unintentional) was most prevalent (42%), followed by 27% act tree (c—only harming *thing* intentional), and 26% act tree (a—closeness argument) patterns. In total, only 5% of the children answered in nonsystematic matters.

Again, the frequencies of act tree patterns in the baseline condition were compared to chance level of 0.25.⁸ Replicating Study 1a, act tree (b—side-effect unintentional) was more frequent than chance ($N = 39$, $k = 23$, probability = 0.59, $p < .001$), act tree (a—closeness argument) patterns were less likely than chance ($N = 39$, $k = 0$, probability = 0, $p < .001$), and act tree (c—only harming *thing* intentional) patterns were not significantly different from chance ($N = 39$, $k = 14$, probability = 0.36, $p = .137$). Beyond the preregistration, we found that act tree (b—side-effect unintentional) was more likely than act tree (a—closeness argument) ($\chi^2(1, n = 23) = 23.00, p < .001$) but as likely as act tree (c—only harming *thing* intentional) ($\chi^2(1, n = 37) = 2.19, p = .139$).

3.2.1.1. Cross-branch linking in constructed act trees: As in Study 1a, most children (95%) affirmed the question whether the agent performed the harmful act H (hurt sheep) in order to E (prevent the balloon from hitting against the mountain) indicating that children answered in the described seemingly paradoxical cross-branch linking, too. Overall, 94% of the subjects who answered in act tree (b—side-effect unintentional) patterns and 94% who answered in act tree (c—only harming *thing* intentional) patterns affirmed this test question.

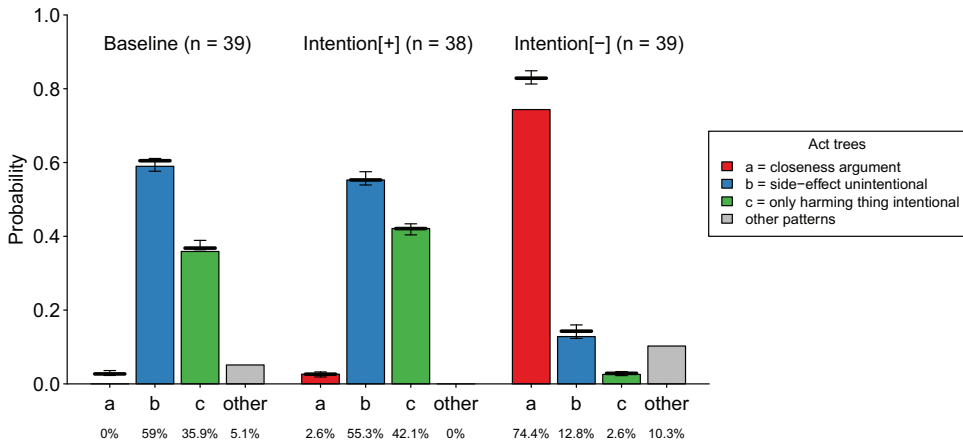


Fig. 10. Distribution of act trees in Study 1b across conditions. The bars and percentages below represent the actual distribution of act trees observed. Horizontal lines indicate predicted probability of an act tree by aggregated multinomial models and vertical lines their 95% confidence intervals. No predicted values and confidence intervals were obtained for the category “other.” Predicted values and their confidence intervals have been obtained via bootstrapping with 1000 boots.

3.2.2. Which prior assumptions go into disambiguating complex cases of act tree construction?

As in Study 1a, act tree distributions differed between conditions (see Fig. 10). Indeed, analyses yielded an even clearer and more extreme pattern. Due to the even more extreme distribution of the response patterns, it was necessary to deviate from the preregistered analysis in the following ways: To avoid complete separation in the baseline condition with act tree (a—closeness argument), data were simulated based on case-wise replacements. The category “other” was excluded from the analysis due to its small proportion (5%; see also analyses on OSF for more details).

Aggregated multinomial logistic regression analysis of these data confirmed that the distribution visible in the original data (see Fig. 10) was significantly different between conditions, as was the case for the adult sample. As predicted by the good intention prior, no significant differences emerged between baseline and intention[+] condition (act tree [b—side-effect unintentional] vs. act tree [a—closeness argument]: $OR = 0.97$, 95% CI [-1.45, 1.385]; vs. act tree [c—only harming thing intentional]: $OR = 0.89$, 95% CI [-0.581, 0.356]). Participants in the intention[-] condition (compared to baseline and intention[+]), however, were less likely to answer in patterns that capture act tree (b—side-effect unintentional) than act tree (a—closeness argument) ($OR = 0.20$, 95% CI [-2.18, -1.04], $p < .001$). This means that children were more likely to regard the negative effect of the act (harming the sheep) as an unintended side-effect in the baseline or intention[+] condition, reflected in the increased likelihood to answer in act tree (b—side-effect unintentional) patterns. At the same time, children in the intention[-] condition were more likely to regard the harmful effect as intended compared to the other two conditions, reflected in substantially

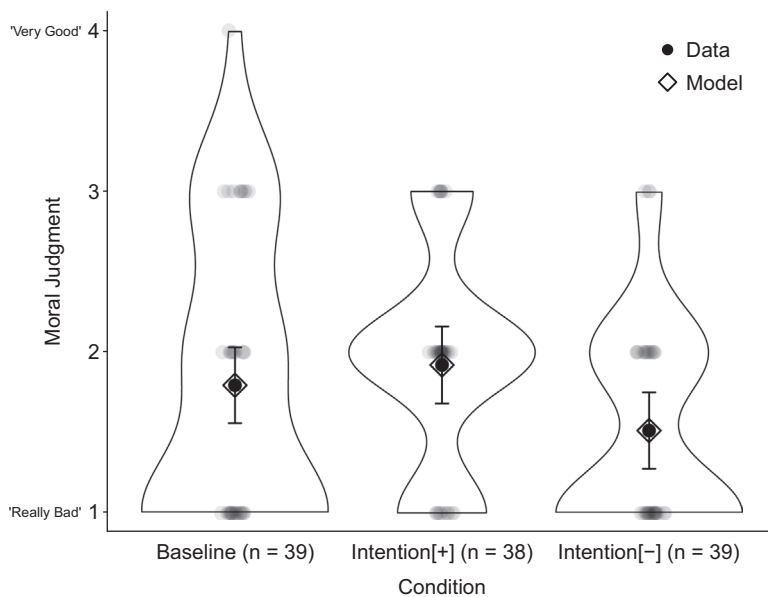


Fig. 11. Children's mean moral acceptability evaluations across conditions. Subjects answered whether the action of Jakob was "very good," "a little good," "a little bad," or "really bad." Each gray point represents the moral evaluation of a participant. Filled black dots indicate the mean moral evaluation per condition and diamond shapes the predicted values by the linear regression model predicting moral evaluation by condition. Vertical lines indicate 95% confidence intervals of the model.

more act tree (a—closeness argument) patterns. Thus, overall, these results clearly replicated Study 1a.

3.2.3. Moral evaluation

Children morally evaluated the critical action (i.e., cutting the rope) as a bad action on average ($M = 1.74$, 95% CI [1.6, 1.88], range = 1–4; see Fig. 11). Means differed slightly between conditions ($M_{\text{baseline}} = 1.79$, 95% CI [1.5, 2.08]; $M_{\text{intention[+]}} = 1.92$, 95% CI [1.7, 2.14]; $M_{\text{intention[-]}} = 1.51$, 95% CI [1.3, 1.72]). The full-null model comparison of the regression analysis predicting moral acceptability evaluations by condition was not significant ($F(2, 113) = 3.03$, $p = .052$). The effect was small (Cohen's $f^2 = 0.035$). The differences in moral judgments between baseline and intention[+] versus intention[-] ($b = 0.12$, 95% CI [0.02, 0.21], $t(113) = 2.35$, $p = .020$) and between baseline versus intention[+] condition ($b = -0.06$, 95% CI [-0.23, 0.11], $t(113) = -0.74$, $p = .460$) should not be interpreted due to the nonsignificance of the full-null model comparison.

Exploratively, we investigated how moral evaluations were influenced by act trees and condition.⁹ There was no significant effect of condition or act trees on children's moral evaluation.

3.3. Discussion

The main findings of Study 1 were the following: First, subjects made fine-grained conceptual distinctions in their action representations between foreseen intended main effects on one branch, and foreseen but unintended side-effects on another branch. The same adults and children, however, engaged in cross-branch linking in constructed act trees indicating a contextual sensitivity of the intentionality of the side-effect. Second, subjects operated with prior assumptions of good intentions: Ambiguous cases with no information regarding different good and bad effects were interpreted and morally evaluated like disambiguated cases in which explicit information about good intentions were given; and the two were treated differently from cases with explicit information about bad intentions. Third, these patterns were similarly found in 8- to 10-year-old children. Our results clearly showed that adults and children did not answer in ways predicted by the closeness argument (Foot, 1967): they differentiated in their intentionality judgments between act descriptions. While our results were similar in that regard to Levine, Leslie et al. (2018), they differed essentially in strictness in intentionality judgments of act descriptions. In the baseline condition, we found that act tree (b—side-effect unintentional) was most prevalent compared to act tree (c—only harming *thing* intentional) (which was the default act tree in Levine, Leslie et al., 2018). The difference between these act trees lies in affirmation versus rejection that the means (throwing off sheep) were intentional (act tree b) or not (act tree c). Of course, it is possible that the differences also result from the different scenarios. To address this possibility, we used the original cavers dilemma again in Study 2.

4. Study 2

Following up on the issues left unresolved by the first Study, Study 2 addressed two open questions. First, were diverging results in Study 1 and previous research (act tree [b—side-effect unintentional] in Study 1 vs. act tree [c—only harming *thing* intentional] in Levine, Leslie et al. (2018) as the “default” act tree) due to differences in the moral dilemmas used? Second, how robust are participants’ prior good assumptions?

The sheep and cavers dilemmas that were used in Study 1 face potential limitations when it comes to assessing the good intention prior. First, harming someone to save lives when the harmed individual would die anyway if no action was taken might be perceived as a rational and morally justifiable choice. Second, and relatedly, refraining from harming the stuck man would mean accepting one’s own death. Third, based on the question used in Study 1, it was not possible to differentiate whether people judged the harmful effect as an intended means or as a simultaneous goal. To overcome these limitations, we devised new manipulations that rendered the acts leading to harmful effects unnecessary. In the cavers dilemma, an alternative exit was discovered, eliminating the need to blow up the man stuck in the exit as a means to save lives. However, despite this, the protagonist chose to proceed with the action. Do participants still think the agent acted out of good intentions (such that, e.g., she had not understood that there was another exit)? In this case, we would expect them to ascribe no

intentionality to the harmful effects (such as killing him) in ambiguous (baseline) conditions, but to attribute intentionality to the harmful effects in disambiguated conditions where the agent explicitly states their malicious motive. Additional questions were created to investigate what kind of overarching goal is ascribed to the agent in all conditions—still a beneficial or a malicious end? However, if individuals attribute intentionality to harmful effects even in the ambiguous baseline conditions, this points to a lack of a robust good intention prior.

4.1. Method

Sample size calculations were based on a power simulation using the data of Study 1. Two hundred eighty-two German-speaking adults (age: 35.41, $SD = 15.74$, 177 female, 102 male, 3 diverse) took part in the preregistered online study. We conducted convenience sampling. All participants were recruited via social media and personal contacts. All participants gave informed consent to take part in the study. Study 2 was conducted using the online platform formr.org (Arslan et al., 2020). All used materials can be found on OSF.

We administered six conditions between subjects. Three were structurally analog to Study 1 (baseline, intention[+], intention[−]) and three included the new alternative exit manipulation. Adults were randomly assigned to one of six conditions using a built-in randomization module (baseline: $n = 50$, intention[+]: $n = 44$, intention[−]: $n = 51$, baseline alternative exit: $n = 50$, intention[+] alternative exit: $n = 41$, intention[−] alternative exit: $n = 46$).

4.1.1. Moral dilemma and conditions

We used the original cavers dilemma (see Introduction and Fig. 2). In addition to conditions analogous to Study 1, we developed conditions for Study 2 in which there was an alternative exit of the cave. Just as in Study 1, the protagonist said “I want to save us! This is my chance. This way I can clear the cave exit!” while performing the critical action (i.e., detonating the dynamite) in the regular intention[+] condition. In the regular intention[−] condition, the protagonist said “I hate this man! This is my chance. This way I can kill the man!” while performing the critical action as in Study 1. The alternative exit conditions differed from the regular conditions only insofar as another caver left, found another exit of the cave, and said before the protagonist used the dynamite: “Hey look! I found another exit back there!”. In this scenario, it was not explicitly noted that the agent heard and understood this statement. Thus, one could apply a good intention prior here too by assuming that the agent must have overheard the report about the other exit.

4.1.2. Act tree construction and moral evaluation

The test questions in Study 2 were similar to those in Study 1, with the exception of two new intentional action questions. Seven intention-based questions were presented in a random order (see Table 4). The initial four questions and question 7 were similar to those in Study 1. The fifth question aimed to clarify whether the agent had an overarching negative goal. The sixth question asked participants to assess the necessity of the agent’s action to detonate dynamite from the agent’s perspective.

Table 4
Assigned act trees based in participants' response patterns

Question	Act trees						
	(a)	(b)	(c)	(c2)	(2b1s)	(h)	
decisive for differentiation between act trees							
1	Did Jakob detonate the dynamite in order to blow up the thing that was blocking the exit?	Yes	Yes	Yes	No	Yes	Yes/No
2	Did Jakob detonate the dynamite in order to blow up the other man?	Yes	Yes	No	No	Yes	Yes
3	Did Jakob detonate the dynamite in order to kill the other man?	Yes	No	No	No	Yes	Yes
4	Did Jakob detonate the dynamite in order to prevent the people in the cave from drowning?	Yes	Yes	Yes	Yes	Yes	No
decisive for other plausible patterns							
5	Did Jakob kill the other man to get rid of the one he hates?	No	No	No	No	Yes	Yes
6	Did Jakob believe that it was necessary to detonate the dynamite in order to save the people?	Yes	Yes	Yes	Yes	Yes/No	No
additional: side-effect in order to goal							
7	Did Jakob kill the other man in order to prevent the people in the cave from drowning?						

Through the introduction of these new test questions, we were able to analyze more meaningful response patterns. We could thus investigate whether subjects considered the harmful effect as an indented means or a simultaneous goal in the intention[-] conditions. First, we examined whether participants perceived the agent's action as achieving multiple goals simultaneously, denoted as the "2 birds with 1 stone" pattern (2b1s). This pattern was assigned when participants ascribed both the goal of preventing people from drowning and the goal of eliminating someone he despised. Second, we identified a pattern that indicated the agent intended solely the harmful consequences. This pattern emerged when participants rejected the positive goal of preventing people from drowning but ascribed the negative goal of getting rid of someone they hated, referred to as pattern (h—intending only harm). Additionally, we observed another pattern, labeled as (c2—not even harming *thing* intentional). A significant number of participants exhibited patterns similar to (c—only harming *thing* intentional), but they also denied that the agent detonated the dynamite in order to remove the *thing* blocking the exit. This pattern (c2—not even harming *thing* intentional) can be seen as a more stringent version of option c, where the agent's sole purpose was to accomplish the positive goal of rescuing the other cavers. Finally, participants provided moral evaluations of the acceptability of the act of detonating the dynamite using the same 7-point scale employed in Study 1.

4.1.3. Data analysis

Analyses were based on the preregistration of the current study unless noted otherwise. As in Study 1, the frequency of act trees (a), (b), and (c) in the baseline condition were

Table 5
Frequency of act trees in Study 2

	Act trees				
	(a) - closeness argument: all acts intentional	(b) - side-effect (killing) unintentional	(c) - only harming the thing intentional	(c2) - not even harming the thing intentional	other
baseline ($n = 50$)	2	9	25	12	2
intention[+] ($n = 44$)	3	6	24	9	2
intention[-] ($n = 51$)	23	4	6	2	16
Total ($N = 145$)	28	19	55	23	20

compared. Act tree patterns and moral evaluations were predicted by the regular conditions in two models. The following additional exploratory analyses were conducted. Based on the two new test questions, more act tree patterns are distinguishable. The new assigned act trees were predicted in one model by the regular conditions and in another model by the new alternative exit conditions using a multinomial logistic regression. Finally, moral evaluations were predicted by act trees.

4.2. Results

4.2.1. Act tree construction: How do adults represent the intentional structure of closeness cases?

Based on responses to the intention questions, participants were assigned to patterns matching underlying act trees. Table 5 shows the frequencies of act trees and other patterns. Act tree (c—only harming *thing* intentional) was most frequent across conditions (38%), followed by 19% act tree (a—closeness argument), 16% act tree (c2—not even harming *thing* intentional), and 13% act tree (b—side-effect unintentional). In total, 14% of the participants answered in nonsystematic matters following the act tree assignment from Study 1.

To test the default understanding of the dilemma, the frequencies of act tree patterns in the baseline condition were compared to chance level (0.25). Only act tree (c—only harming *thing* intentional) was more frequent than expected by chance ($N = 50$, $k = 25$, probability = 0.5, $p < .001$). Act tree (a—closeness argument) patterns were less likely than expected by chance ($N = 50$, $k = 2$, probability = 0.04, $p < .001$) and act tree (b—side-effect unintentional) patterns were not significantly different from chance level ($N = 50$, $k = 9$, probability = 0.18, $p = .327$). We found that in the baseline condition, act tree (c—only harming *thing* intentional) was more likely than act tree (b—side-effect unintentional) ($\chi^2(1, n = 34) = 7.53$, $p = .006$). This indicated that subjects' default understanding was mostly in line with option c on the closeness problem—in contrast to the pattern found in Study 1.

4.2.1.1. Cross-branch linking in constructed act trees: In our exploratory analysis, we investigated whether subjects engage in cross-branch linking in their constructed act trees. Despite denying that the protagonist intended to cause harmful side-effects (i.e., act tree pat-

Table 6
Frequency of new act trees in Study 2

	Act trees					other
	(a)	(b)	(c)/(c2)	(2b1s)	(h)	
baseline ($n = 50$)	2	8	35	0	0	5
intention[+] ($n = 44$)	3	6	32	0	0	3
intention[-] ($n = 51$)	1	1	7	22	5	15
baseline alt. exit ($n = 50$)	1	5	34	0	1	9
intention[+] alt. exit ($n = 41$)	4	1	19	2	1	14
intention[-] alt. exit ($n = 46$)	1	1	2	16	21	5
Total ($N = 282$)	12	22	129	40	28	51

terns [b—side-effect unintentional] and [c—only harming *thing* intentional]), would subjects still link the harmful side-effect to the achievement of the goal. The majority of subjects (86%) agreed with the statement “kill man [side-effect] in order to prevent people from drowning in the cave [end].” This pattern of responses was observed in 89% of subjects with act tree (b—side-effect unintentional) patterns and 98% with act tree (c—only harming *thing* intentional) patterns. When framed in terms of goals, the harmful effect was considered an unintended side-effect. However, when framed in terms of actions, subjects strongly affirmed the connection between the harmful act and the achievement of the goal, indicating that they did not perceive it as a clear side-effect.

4.2.1.2. Exploratory analysis: By introducing new test questions, we were able to examine more meaningful response patterns. First, we assessed whether participants perceived the agent’s action as fulfilling multiple goals simultaneously, referred to as the “2 birds with 1 stone” pattern (2b1s). Indeed, a significant percentage of response patterns in the regular intention[-] condition (43%) could be better explained by this new response pattern than by the act tree (a—closeness argument) pattern (see also Table 6).

Second, we identified a pattern indicating that the agent intended only the harmful effect, referred as pattern (h—intending only harm). This pattern accounted for 46% of the responses in the intention[-] alternative exit condition.

Consistently, 75% of the subjects who answered in (2b1s) patterns affirmed the “side-effect in order to [positive] goal” question. In contrast, 93% of subjects who responded with (h—intending only harm) patterns denied the same question, indicating that they did not exhibit this seemingly paradoxical response pattern of cross-branch linking. These individuals consistently viewed the harmful actions as a means to fulfill the agent’s malicious goal and not in order to achieve a beneficial goal.

4.2.2. Which prior assumptions go into disambiguating complex cases of act tree construction?

Similar to Study 1, there were variations in act tree distributions across conditions (see Fig. 12a). Since act tree (c2—not even harming *thing* intentional) accounted for approxi-

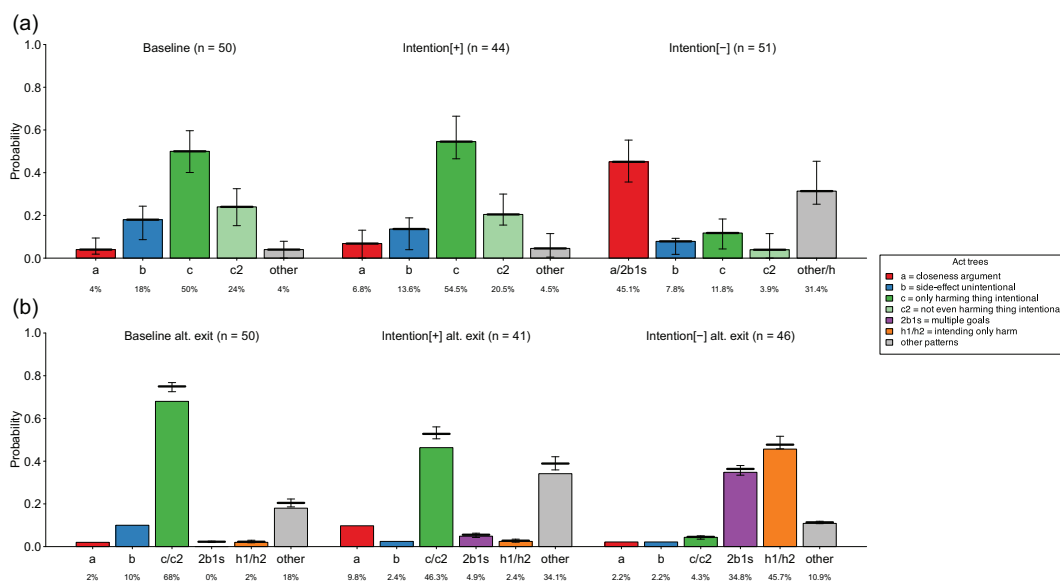


Fig. 12. (a) Distribution of act trees in Study 2 across conditions baseline, intention[+], intention[-]. (b) Distribution of act trees in Study 2 in alternative exit conditions. Bars and percentages below represent proportions of act trees observed. Horizontal lines indicate predicted probability of an act tree by the multinomial model and vertical lines their 95% confidence intervals. Predicted values and their confidence intervals have been obtained via bootstrapping with 1000 boots. Due to low frequency of act trees (a), and (b) in alternative exit conditions, they were excluded from the analysis, thus no predicted values and confidence intervals were obtained. To avoid complete separation in the baseline condition with (2b1s) patterns, data were simulated based on case-wise replacements.

mately 50% of the “other” patterns, we decided to incorporate this noteworthy response pattern into our analysis, deviating from our initial preregistration. Given that act tree (c—only harming *thing* intentional) was the most prevalent pattern observed, it was chosen as the reference category for the analysis.

The multinomial logistic regression analysis of the data confirmed that the distribution of act trees observed in the data differed significantly between conditions, similar to the findings in Study 1. Consistent with the predicted influence of the good intention prior, no significant differences were found between the baseline and intention[+] conditions (act tree [c—only harming *thing* intentional] vs. act tree [a—closeness argument]: $OR = 0.80$, 95% CI [-1.16, 0.71]; vs. act tree [b—side-effect unintentional]: $OR = 1.20$, 95% CI [-0.40, 0.77]; vs. act tree [c2—not even harming *thing* intentional]: $OR = 1.13$, 95% CI [-0.39, 0.64]). In the intention[-] condition, participants were less likely to respond with act tree (c—only harming *thing* intentional) patterns compared to act tree (a—closeness argument) patterns ($OR = 0.30$, 95% CI [-1.65, -0.78], $p < .001$), when compared to the baseline and intention[+] conditions. The likelihood of perceiving the harmful effects (killing the man or blowing him up) as intentional, as reflected in act tree (a—closeness argument), was considerably higher in the regular intention[-] condition compared to the regular baseline or intention[+] conditions.

Therefore, overall, these results clearly replicated the findings of Study 1 in relation to the influence of the good intention prior.¹⁰

4.2.2.1. Exploratory analysis: Aggregated multinomial regression models confirmed again that the distributions were different between the additional conditions (see Fig. 12b). In the intention[−] alternative exit condition, participants were less likely to respond with act tree (c—only harming *thing* intentional)/(c2—not even harming *thing* intentional) patterns compared to (21bs) patterns ($OR = 0.19$, 95% CI [−2.29, −1.01], $p < .001$) and (h—intending only harm) patterns ($OR = 0.16$, 95% CI [−2.53, −1.18], $p < .001$), when compared to the baseline and intention[+] alternative exit conditions. The likelihood of perceiving the harmful effects as intentional was considerably higher in the intention[−] alternative exit condition compared to the baseline or intention[+] alternative exit conditions.

Furthermore, an intriguing finding was that the influence of the good intention prior persisted even in the alternative exit conditions. In both the baseline and intention[+] conditions with an alternative exit, the majority of participants answered with (c—only harming *thing* intentional) or (c2—not even harming *thing* intentional) patterns, indicating their rejection of the notion that the agent intentionally killed the obstructing individual or *thing*, despite it no longer being necessary.

4.2.3. Moral evaluation

Participants evaluated the critical act (i.e., detonate the dynamite) descriptively as morally less acceptable than adults in Study 1 ($M = 3.9$, 95% CI [3.62, 4.19], range = 1–7; see Fig. 9). Regression analysis revealed that moral acceptability evaluations did not differ between conditions ($M_{\text{baseline}} = 3.98$, 95% CI [3.53, 4.43]; $M_{\text{intention[+]}} = 4.07$, 95% CI [3.53, 4.52]; $M_{\text{intention[-]}} = 3.69$, 95% CI [3.17, 4.21]). The full-null model comparison was not significant ($F(2, 142) = 0.64$, $p = .528$).

4.2.3.1. Exploratory analyses: Moral evaluations were found to be influenced by the different act tree patterns while controlling for the effect of condition. Specifically, adults who responded with act tree (c2—not even harming *thing* intentional) patterns provided significantly higher evaluations of moral acceptability compared to adults who responded with act tree (c—only harming *thing* intentional) patterns ($b = 0.87$, 95% CI [0.07, 1.68], $t(138) = 2.15$, $p = .033$). On the other hand, individuals who answered in other patterns gave significantly lower evaluations of moral acceptability than those who answered with act tree (c—only harming *thing* intentional) patterns ($b = -1.87$, 95% CI [−2.87, −0.87], $t(138) = -3.71$, $p < .001$).

Across all six conditions (baseline, intention[+], intention [−] without and with alternative exit), moral evaluations differ as a function of type of condition and whether an alternative exit was present or not (see Fig. 13). Evaluations of moral acceptability were significantly lower in intention[−] conditions ($t(278) = -3.25$, $p = .001$) and when an alternative exit was present ($t(278) = -4.02$, $p < .001$). The effect was small (Cohen's $f^2 = 0.085$). Another model, predicting moral evaluations by the alternative exit factor, condition, and assigned act trees revealed that act trees also influenced subject's moral evaluations. These results should,

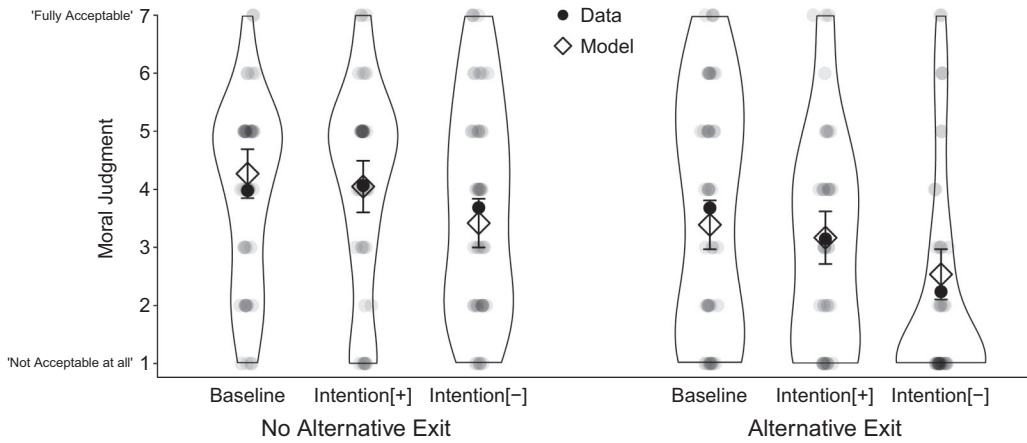


Fig. 13. Moral evaluation across conditions in Study 2 as a function of the agent's motives and whether an alternative exit was present. Subjects evaluated how morally acceptable the action to detonate the dynamite was. Each transparent point represents the moral evaluation of a participant. Filled black dots indicate the mean moral evaluations per condition and diamond shapes the predicted values by the linear regression model predicting moral evaluations by condition. Vertical lines indicate 95% confidence intervals of the model.

however, be interpreted with great caution since the assigned act trees have very different sample sizes and were not experimentally manipulated. Consistently, lower moral acceptability evaluations were given when participants answered in act tree (a—closeness argument) ($t(273) = -2.25, p = .025$), (h—intending only harm) ($t(273) = -5.53, p < .001$), (2b1s) ($t(273) = -2.02, p = .044$), and other patterns ($t(273) = -3.66, p < .001$) than act tree (c—only harming *thing* intentional)/(c2—not even harming *thing* intentional) patterns. Act tree (b—side-effect unintentional) patterns were not significantly different from act tree (c—only harming *thing* intentional)/(c2—not even harming *thing* intentional) patterns ($t(273) = 1.14, p = .254$). Generally, moral judgments were more likely to be lower when another exit was present ($t(273) = -2.05, p = .042$) than when there was no other exit. There was no main effect of condition. The overall effect of the model was medium (Cohen's $f^2 = 0.229$).

4.3. Discussion

The key findings of Study 2 can be summarized as follows: First, by utilizing the original cavers dilemma, we successfully replicated the results reported by Levine, Leslie et al. (2018). Participants were able to distinguish in their intentionality judgments between the main effects and foreseen side-effects of actions, contradicting the predictions of the closeness argument (Foot, 1967). Interestingly, participants demonstrated an even stricter notion of intentionality compared to Study 1, potentially influenced by the severity of the dilemma itself. Second, we observed the robustness of the good intention prior established by Levine, Mikhail et al. (2018) across new manipulations. Even when the actions leading to harmful effects became unnecessary, participants consistently attributed good intentions to the protagonist. Lastly,

through the analysis of response patterns to additional test questions, we detected indications of participants perceiving the agent as acting with both good and bad goals simultaneously, as well as perceiving a solely malicious overarching goal.

5. General discussion

The present studies investigated the cognitive foundations and development of complex action interpretation and evaluation. Children and adults were confronted with and probed about action scenarios with closely related main and side-effects as they have been discussed under the rubric of the closeness problem. In the current studies, we asked three questions. First, how do subjects represent the intentional structure of closeness cases? Second, which prior assumptions go into disambiguating complex cases and how robust are they? Third, how do these forms of action interpretation develop? In the following, we first summarize the main findings and then discuss them in light of our main research questions.

First, adults and children demonstrated the ability to make nuanced conceptual distinctions in their representations of actions which clearly contradict the closeness argument (Foot, 1967). They distinguished between intended main effects on one branch and unintended side-effects on another branch. Second, participants operated under the assumption of good intentions. In cases where no information was provided about the potential beneficial or malicious motive of the agent, ambiguous scenarios were interpreted and morally evaluated similarly to scenarios where explicit information about the agent's good intentions were provided. These cases were distinct from situation with explicit information about the agent's malicious intentions. The good intention prior phenomenon was robust across an even more dramatic manipulation which made the agent's actions not necessary anymore. Third, these observed patterns were also evident among children aged 8-10 years old.

In addition, in more exploratory ways, we found interesting relationships between the ways in which subjects parsed and interpreted action—indicated in their act tree patterns—and their moral evaluations of the acts. In Studies 1a and 2, act trees which incorporated no intentionality ascription to the harmful effects were associated with higher moral acceptability evaluations. Likewise, act trees capturing that harmful effects and goals were intended were associated with lower moral acceptability evaluations. These preliminary findings, however, should be treated with caution as we have not enough statistical power (low and high frequencies of act trees are compared). Future research should investigate this pattern more systematically by experimentally inducing both moral evaluation and act trees.

5.1. Representation of intentional structure in closeness cases

At the same time, the present findings leave open and raise many questions. We and others (Levine, Leslie et al., 2018) have argued that people do not follow the logic of the closeness argument when interpreting such scenarios. But how far do they deviate? In our Study 1, the majority answered in line with option b—that the killing is not causally necessary, so the death is not intended (Masek, 2010) and a substantial proportion answered in line with option

c—an even finer distinction. In Study 2, we could replicate the original finding by Levine, Leslie et al. (2018) such that the majority answered in line with option c or an even stricter notion (pattern c2). Is there a principle behind such patterns? In the philosophical literature, several suggestions have been made (e.g., that one can claim distinctions for causal, but not for constitutive relations; Fitzpatrick, 2006). When action descriptions are causally connected such as “blowing up the man” and “killing the man,” there are still possible (even if unlikely) scenarios in which the blowing up does not lead to the death. This uncertainty gap is not present in constitutive connections between different act descriptions such as “blowing up the thing blocking the exit” and “blowing up the man”: since in this situation, the man is the thing blocking the exit, you cannot perform one without the other action. From an empirical perspective, it remains to be clarified which principles play a role under which conditions in closeness cases.

Another explanation for the discrepancies observed between the sheep and cavers dilemma could be the intuitive differences in the strength of the connection between certain act descriptions. Despite our efforts to maintain structural similarity in the act descriptions, we acknowledge that the intuitive connection between blowing someone up and killing them (Study 2) may be closer than the connection between throwing someone off a hot-air balloon and killing them (Study 1). To explore this possibility further, it would be valuable to have participants rate the likelihood of causally and constitutively connected act descriptions, as well as their moral valence. In general, future studies should investigate the determinants that influence the strictness with which individuals attribute intentionality in various moral dilemmas.

Similarly, what factors influence the degree to which individuals interpret an agent’s intentions in a moral dilemma? Could the severity of the dilemma itself play a role? One possibility is that in more severe dilemmas, the descriptions of the actions taken to “resolve” the dilemma are inherently more morally charged, leading individuals to be less inclined to attribute intentionality to the agent. It is thus an open question whether the good intention prior also influences this tendency.

5.1.1. *Cross-branch linking in constructed act trees*

Overall, subjects in the present studies clearly kept apart, in their constructed act trees, main branches of intended effects, and side branches of merely foreseen effects. Interestingly, however, most subjects did engage also in seemingly paradoxical cross-branch linking: while denying that agents did basic actions to bring about merely foreseen side-effects, they affirmed that the agent brought about the foreseen side-effects in order to bring about the intended main effect. For example, subjects claimed that the agent cut the rope [basic action] in order to prevent the balloon from hitting against the mountain [end], but did not cut the rope [basic action] in order to kill the sheep [side-effect]; yet, also claimed that the agent killed the sheep [side-effect] in order to prevent the balloon from hitting against the mountain [end].

How can we make sense of this seemingly paradoxical pattern? One recent suggestion is the following: Depending on the context, subjects flexibly “chunk” nodes on a given branch of an act tree (Levine, Leslie et al., 2018). “Chunking,” as we know from working memory research, is a form of grouping information that reduces the need to remember individual elements. For instance, instead of recalling four digits (2-0-2-3), one only needs to remember one date—“2023” (Miller, 1956). Chunks can be represented as one unit (“2023”), but can

still be examined and their components can be reported if necessary (e.g., one can still judge that “2” appears twice).

Chunking in the context of act descriptions means that the action in question can be variously described with each node description (e.g., “detonate dynamite,” “blow up man,” “clear cave entrance”) in such ways that one can still mentally “zoom” in and out as required (also called “the accordion effect”; Feinberg, 1970 as cited in Bratman, 2006). And so, sometimes, the action performed will be referred to under the description of the side-effect; thereby, chunking the nodes (bits of information) together that led to the main goal (detonating dynamite, blowing up thing, blowing up man, and killing him; see also Fig. 3a).

Applied to the present case, the chunking construal explains this apparent paradox: when asked in goal terms (i.e., was the basic act performed in order to bring about the harmful effect?), they “zoom-in” and deny that the harmful effect was brought about intentionally. When asked in action terms (i.e., was the harmful effect brought about in order to bring about the good goal?), they “zoom-out” and affirm the statement.

The present findings add to a growing body of evidence that chunking can play a role in flexible action parsing and interpretation (Knobe, 2010; Levine, Leslie et al., 2018). But *when* and *why* does chunking come into play? Currently, not much is known about this. One suggestion is that chunking occurs in cases where a given act tree branch describes what a person “most essentially does” (Knobe, 2010; Levine, Leslie et al., 2018). What exactly this means remains disputed, however. Knobe (2010) argues that mostly morally bad actions describe more what a person is most essentially doing. Scholars (Knobe, 2010; Levine, Leslie et al., 2018) argue and present evidence that chunking might not be limited to morally charged cases. In their examples, a chef moves his arms in order to make an omelet in order to make breakfast while thereby getting some exercise (side-effect). Here, participants chunked the main act tree branch (move arms → make omelet → make breakfast) but not the side branch. Future research will thus need to investigate more systematically when chunking occurs, and whether there is a notion of “actions done most essentially” that may carry some explanatory weight in this context.

5.2. *Good intention prior in disambiguating complex cases of act tree construction*

The present studies extend previous results on the “good intention prior” in act tree constructions (Levine, Mikhail et al., 2018): Through act tree analyses across various moral dilemmas and age groups, this phenomenon has demonstrated remarkable robustness. In cases with ambiguous information about the agent’s motive, participants consistently interpreted and morally evaluated them in a manner similar to cases where explicit information about good intentions was provided. This effect held even when the dilemma was manipulated in such a way that the necessity of the action leading to harmful effects was eliminated. Notably, cases involving explicit information about bad intentions were treated differently such that subjects ascribed more malicious intentionality and morally evaluated these acts more harshly.

In this context, the act tree analyses proved to be a valuable tool for investigating the good intention prior both in representation (i.e., how people think about intentional structures of acts) and inference (i.e., under which conditions they update their representations).¹¹ In

Study 1, only one sentence was changed from the baseline version of the dilemma: whether the agent explicitly intended some effects or not and gave a reason (e.g., “I hate sheep”). That drastically changed which parts of the action were perceived as brought about intentionally and which act trees were constructed. In Study 2, we additionally manipulated the necessity of the harmful act. Here, only in combination with the changed motive, participants’ “default” representation was updated. Taken together, act tree analysis allows for a coherent analysis of plausible response patterns and their possible updating through an experimental manipulation. We believe that, depending on the research question, such an act tree analysis is more meaningful than comparing individual intentionality judgments.

5.3. *Developmental perspective*

Generally, children and adults showed qualitatively analogous patterns of action interpretation and moral evaluation. However, descriptively, adults’ and children’s response patterns differed slightly in Study 1. Children made clearer and more extreme judgments than adults. They evaluated the critical action as less morally permissible than adults and ascribed less intentionality to (more) harmful act descriptions. This may be due to the fact that in Study 1, animals were hurt instead of humans and children tend to prioritize humans over animals less than adults do (Wilks et al., 2021)—an exciting possibility that deserves to be investigated more systematically in future research.

More generally, future research should investigate in more systematic ways the developmental origins and trajectories of complex forms of action individuation, interpretation, and evaluation before the ages tested here. To this end, the method of act tree questions should be radically simplified to make it suitable for younger children. One interesting question would be whether chunking in act tree interpretation develops in tandem with chunking capacities in other areas such as working memory (e.g., Stahl & Feigenson, 2014) or action planning (Blankenship & Kibbe, 2023). Similarly, Levine and Leslie (2022) have investigated children’s understanding of means in relation to goals within the framework of the “means principle,” which is highly compatible with the doctrine of the double effect. In this study, children were able to correctly assess when harm was used as a means and situate these means with superordinate goals (Levine & Leslie, 2022). Children were also able to take these judgments into account in their moral judgments (Levine & Leslie, 2022).

Another exciting question is whether and how intention ascription and act tree construction are related to counterfactual reasoning.¹² In general, we know from many studies that Theory of Mind and counterfactual reasoning are closely related in development (Rafetseder et al., 2021; Rafetseder & Perner, 2018; Rasga et al., 2016; Riggs et al., 1998; see also Jacob, 2020, for a critical view). Regarding action interpretation, there seem to be close conceptual connections between ascription of complex intentions and counterfactual considerations. For example, an agent typically would not have performed a basic action B if (she believed) it had not led to the desired end E; but she would have performed B irrespective of whether B had led to an unintended neutral side-effect. From an ontogenetic perspective, how counterfactual reasoning and complex action interpretation are developmentally related will thus be an important topic for future research.

5.4. Conclusion


Overall, the adapted act tree analysis used in the present studies proved to be a profitable method to investigate the cognitive foundations and the development of complex action individuation, interpretation, and evaluation. Taken together, the present findings suggest that everyday action interpretation and evaluation, elucidated by act tree analyses, involves a sophisticated distinction between intended and merely foreseen effects even if the two are very closely related and the latter involve considerable harm. This distinction, fundamental to moral, legal, and other forms of evaluative thought and discourse, is even in place relatively early in development.

Acknowledgments

We thank Joshua Knobe, Simon Stephan, Alex Wiegmann for helpful discussions, Andreas Cordes and Roger Mundry for statistical advice, and Imke Koch and Monja Krücke for help with data collection and reliability coding. This work was supported by Evangelisches Studienwerk Villigst, Studienstiftung des Deutschen Volkes, and the Deutsche Forschungsgemeinschaft (DFG); SFB 1528: Cognition of Interaction, and (254142454/GRK 2070).

Open access funding enabled and organized by Projekt DEAL.

Open Research Badges

 This article has earned Open Data and Open Materials badges. Data and materials are available at <https://osf.io/vgu27/> and <https://osf.io/p47sk/>.

Notes

- 1 For actions with more than one intended effect, the structure becomes more complex, with more than one main intentional branch. We ignore these complications here for simplicity's sake.
- 2 We thank an anonymous reviewer for spelling out this point more clearly.
- 3 Due to our moral dilemmas, the structure of the explicit negative condition is different than in the original study by Levine, Mikhail et al. (2018). In our version, it is plausible that participants ascribe the protagonist two goals simultaneously: to kill the man and to save everyone else (including themselves). At the same time, it is unlikely that the protagonist saved themselves merely as a side-effect of their doing. Another possibility is that saving oneself is seen as a means to kill the caver stuck in the exit (in the intention[-] condition). We examined which of these possibilities applies in Study 2. We thank an anonymous reviewer for drawing our attention to this important point.
- 4 For a simple comparison with the study by Levine, Leslie et al. (2018), you will also find the cross-tabulations of the answers to the individual questions in the Supplement.
- 5 We decided that four possible answer options (three action tree patterns and the category of other plausible patterns) should be considered for the calculation of the chance level

- of 1/4 (= 0.25). This criterion is stricter than the pure maximum combination of all possible answers (yes/no) to all four questions ($4^2 = 16$).
- 6 We used the package `nnet` (Venables & Ripley, 2002) to calculate the multinomial logistic regression. Note that we preregistered to use `mlogit` package. However, with `nnet`, it was possible to perform the same analyses without the need to transform the data and easily obtain the predicted values and their 95% bootstrapped confidence intervals.
 - 7 k denotes the number of times that the event occurred. In our case, the number of one act tree in the baseline condition.
 - 8 We thereby deviate from our preregistered chance level of 1/6 and chose the more conservative chance level as in Study 1a.
 - 9 An anonymous reviewer correctly noted that when analyzing moral evaluations as a function of act trees, the condition should be included as a predictor. Otherwise, only the effect of the experimental manipulation might become visible. Therefore, we decided not to report the model that best fits the data, but the model just described. Other exploratory analyses can be found on OSF.
 - 10 As Tables 5 and 6 show, most of the patterns described in this analysis with act tree a are best explained by response pattern “2b1s.” That is, most individuals considered the negative effect (kill man) as a simultaneous goal alongside saving everyone else in the intention[-] condition.
 - 11 We are grateful to an anonymous reviewer who inspired us to make this point clearer.
 - 12 We are grateful for an anonymous reviewer who inspired us to spell this idea out.

References

- Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). Formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behavior Research Methods*, 52(1), 376–387.
- Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling versus unable: Infants’ understanding of intentional action. *Developmental Psychology*, 41(2), 328–337.
- Blankenship, T. L., & Kibbe, M. M. (2023). “Plan chunking” expands 3-year-olds’ ability to complete multiple-step plans. *Child Development*, 94(5), 1330–1339.
- Bratman, M. E. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.
- Bratman, M. E. (2006). What is the accordion effect? *Journal of Ethics*, 10(1–2), 5–19.
- Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2), 353–380.
- Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, 127(1), 6–21.
- Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). *LabVanced: A unified JavaScript framework for online studies*. International Conference on Computational Social Science.
- Fitzpatrick, W. J. (2006). The intend/foresee distinction and the problem of “closeness”. *Philosophical Studies*, 128(3), 585–617.
- Foot, P. (1967). *The problem of abortion and the doctrine of double effect*. Oxford University Press.
- Forstmeier, W., & Schielzeth, H. (2011). Cryptic multiple hypotheses testing in linear models: Overestimated effect sizes and the winner’s curse. *Behavioral Ecology and Sociobiology*, 65(1), 47–55.
- Fuller, L. L. (1949). The case of the speluncan explorers. *Harvard Law Review*, 62(4), 616–645.

- Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, *415*(6873), 755–755.
- GoAnimate Inc. (2020). *Vyond*. <https://www.vyond.com/>
- Goldman, A. I. (1970). *Theory of human action*. Princeton University Press.
- 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.
- Helwig, C. C., Zelazo, P. D., & Wilson, M. (2001). Children's judgments of psychological harm in normal and noncanonical situations. *Child Development*, *72*(1), 66–81.
- Jacob, P. (2020). *How relevant to the psychology of mindreading is knowledge-first epistemology?* <http://cognitionandculture.net/blogs/pierre-jacob/how-relevant-to-the-psychology-of-mindreading-is-knowledge-first-epistemology/>
- Kamawar, D., & Olson, D. R. (2011). Thinking about representations: The case of opaque contexts. *Journal of Experimental Child Psychology*, *108*(4), 734–746.
- Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition*, *119*(2), 197–215.
- Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.
- Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*(2), 309–324.
- Knobe, J. (2010). Action trees and moral judgment. *Topics in Cognitive Science*, *2*(3), 555–578.
- Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science*, *17*(5), 421–427.
- Levine, S., & Leslie, A. M. (2022). Preschoolers use the means principle in their moral judgments. *Journal of Experimental Psychology: General*, *151*(11), 2893–2909.
- Levine, S., Leslie, A. M., & Mikhail, J. (2018). The mental representation of human action. *Cognitive Science*, *42*(4), 1229–1264.
- 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.
- Masek, L. (2010). Intentions, motives and the doctrine of double effect. *Philosophical Quarterly*, *60*(240), 567–585.
- McIntyre, A. (2019). Doctrine of double effect. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019). Metaphysics Research Lab, Stanford University, <https://plato.stanford.edu/archives/spr2019/entries/double-effect/>
- Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, *31*(5), 838–850.
- Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.
- Pellizzoni, S., Siegal, M., & Surian, L. (2009). Foreknowledge, caring, and the side-effect effect in young children. *Developmental Psychology*, *45*(1), 289–295.
- Perner, J. (1991). *Understanding the representational mind*. MIT Press.
- Proft, M., Dieball, A., & Rakoczy, H. (2019). What is the cognitive basis of the side-effect effect? An experimental test of competing theories. *Mind & Language*, *34*(3), 357–375.
- Proft, M., & Rakoczy, H. (2019). The ontology of intent-based normative judgments. *Developmental Science*, *22*(2), e12728.
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

- Rafetseder, E., O'Brien, C., Leahy, B., & Perner, J. (2021). Extended difficulties with counterfactuals persist in reasoning with false beliefs: Evidence for teleology-in-perspective. *Journal of Experimental Child Psychology*, 204, 105058.
- Rafetseder, E., & Perner, J. (2018). Belief and counterfactuality: A teleological theory of belief attribution. *Zeitschrift Für Psychologie*, 226(2), 110–121.
- Rakoczy, H., Behne, T., Clüver, A., Dallmann, S., Weidner, S., & Waldmann, M. R. (2015). The side-effect effect in children is robust and not specific to the moral status of action effects. *PLOS ONE*, 10(7), e0132933.
- Rasga, C., Quelhas, A. C., & Byrne, R. M. J. (2016). Children's reasoning about other's intentions: False-belief and counterfactual conditional inferences. *Cognitive Development*, 40, 46–59.
- Riggs, K. J., Peterson, D. M., Robinson, E. J., & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development*, 13(1), 73–90.
- Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.
- Stahl, A. E., & Feigenson, L. (2014). Social knowledge facilitates chunking in infancy. *Child Development*, 85(4), 1477–1490.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. Springer.
- 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.
- Wilks, M., Caviola, L., Kahane, G., & Bloom, P. (2021). Children prioritize humans over animals less than adults do. *Psychological Science*, 32(1), 27–38.
- Woodward, A. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1), 1–34.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Coding System

Table S2. Named categories in adults' free-text responses corresponding to act trees and conditions

Table S3. Named categories in children's open responses corresponding to act trees and conditions

Table S4. Cross-tabulated counts of subjects answering yes/no to the questions whether the protagonist cut the rope in order to kill the sheep or to throw off the sheep in Study 1a

Table S5. Cross-tabulated counts of subjects answering yes/no to the questions whether the protagonist cut the rope in order to kill the sheep or in to throw off ballast in Study 1a.

Table S6. Cross-tabulated counts of subjects answering yes/no to the questions whether the protagonist cut the rope in order to throw off the sheep or in to throw off ballast in Study 1a.

Table S7. Cross-tabulated counts of subjects answering yes/no to the questions whether the protagonist cut the rope in order to kill the sheep or in to throw off the sheep in Study 1b.

Table S8. Cross-tabulated counts of subjects answering yes/no to the questions whether the protagonist cut the rope in order to kill the sheep or in to throw off ballast in Study 1b.

Table S9. Cross-tabulated counts of subjects answering yes/no to the questions whether the protagonist cut the rope in order to throw off the sheep or in to throw off ballast in Study 1b.

Table S10. Cross-tabulated counts of subjects answering yes/no to the questions whether the protagonist used the dynamite in order to kill the man or in to blow up the man in Study 2.

Table S11. Cross-tabulated counts of subjects answering yes/no to the questions whether the protagonist used the dynamite in order to kill the man or in to blow up the thing blocking the exit in Study 2.

Table S12. Cross-tabulated counts of subjects answering yes/no to the questions whether the protagonist used the dynamite in order to blow up the man or in to blow up the thing blocking the exit in Study 2.