The Mental Representation of Human Action

Sydney Levine, a,b Alan M. Leslie, a,b John Mikhail c

a Department of Psychology, Rutgers University
b Center for Cognitive Science, Rutgers University
c Georgetown University Law Center

Received 6 March 2016; received in revised form 15 January 2018; accepted 18 January 2018

Abstract

Various theories of moral cognition posit that moral intuitions can be understood as the output of a computational process performed over structured mental representations of human action. We propose that action plan diagrams—“act trees”—can be a useful tool for theorists to succinctly and clearly present their hypotheses about the information contained in these representations. We then develop a methodology for using a series of linguistic probes to test the theories embodied in the act trees. In Study 1, we validate the method by testing a specific hypothesis (diagrammed by act trees) about how subjects are representing two classic moral dilemmas and finding that the data support the hypothesis. In Studies 2–4, we explore possible explanations for discrete and surprising findings that our hypothesis did not predict. In Study 5, we apply the method to a less well-studied case and show how new experiments generated by our method can be used to settle debates about how actions are mentally represented. In Study 6, we argue that our method captures the mental representation of human action better than an alternative approach. A brief conclusion suggests that act trees can be profitably used in various fields interested in complex representations of human action, including law, philosophy, psychology, linguistics, neuroscience, computer science, robotics, and artificial intelligence.

Keywords: Moral judgment; Act trees; Moral grammar; Human action; Intention

1. Introduction

Various theories of moral cognition have proposed that moral intuitions can be understood as the output of a computational process performed over structured mental representations of human action (Mikhail, 2000, 2011; see also Greene, 2014; Knobe, 2010;
Nichols & Mallon, 2006). These accounts each assume that subjects are representing specific information about human action (such as temporal sequence, causal structure, value and likelihood of outcomes, intention of the agent, and so forth), even though that information may not be explicitly available in the stimulus itself. The theories further predict that this information is then represented in a highly structured way so that a computation can be performed over the representation and a moral evaluation of the act in question can be formed (such as permissible, obligatory, forbidden, and so forth).

This notion—that moral intuitions can be understood as the output of a computational process conducted over structured mental representations of moral acts—has preliminary empirical support (see Greene, 2014; Knobe, 2010; Mikhail, 2007, 2011; Nichols & Mallon, 2006). Yet relatively little experimental work has been done to provide a direct test of the richness of these mental representations or their role in moral cognition. Moreover, many theorists have presupposed complex hypotheses about subjects’ mental representation of particular acts that are difficult to precisely communicate without ambiguity (e.g., Chemla, Egré, & Schlenker, 2015; Kamm, 2007) or have gone untested (e.g., Greene, 2014; Mikhail, 2007). Our goal in this paper is to fill this gap by developing a methodology that will support the systemic experimental investigation of the mental representations of human action—including but not limited to those that underlie moral judgment.

We do this by first introducing the idea of “act trees” (Goldman, 1970), a useful technique for cognitive scientists to diagram their hypotheses about the information contained in subjects’ mental representations of human action (Mikhail, 2011). We then describe a methodology for testing the hypotheses embodied in the act trees. Finally, we report the experimental results of six studies that put this methodology into practice and provide support for a series of specific hypotheses about how subjects are mentally representing human actions. While this paper mainly focuses on the mental representation of human actions underlying moral judgment (and uses morally-charged cases as experimental stimuli), the techniques we develop can be applied to action representations more generally. We demonstrate this extension explicitly in Study 5.

1.1. Act trees

Act trees are a particularly useful tool for diagramming a hypothesis about how subjects are mentally representing morally charged acts in the world. We will divide the superordinate category “act tree” into two subordinate categories: cause trees (e.g., Fig. 1a and b) and intention trees (e.g., Fig. 1c and d).

Cause trees are a way of diagramming subjects’ representations of causal sequences of events that happen in the world. Therefore, these trees represent hypotheses about the workings of cognitive systems deployed for understanding how objects behave (e.g., Bailargeon, Li, Gertner, & Wu, 2011; Leslie, 1994b; Spelke, 1990). Nodes (act tokens) located at the relatively lower end of the cause tree are represented by subjects as occurring causally prior to nodes farther up the tree. If trees contain branch points, the node at the branch point is represented by subjects as generating two different causal sequences.
Intention trees describe representations of the goal structure of an agent’s action plan. So, these trees can be understood to illustrate hypotheses about the workings of cognitive systems deployed for understanding and inferring these agents’ goals and mental states (e.g., Csibra & Gergely, 1998; Leslie, 1994a). In intention trees, the vertical line that connects the agent’s most basic act with his goal depicts the agent’s series of intended actions, or chain of means and ends—his intended action plan. Nodes connected to the main branch with diagonal lines represent side effects of the agent’s intended actions. Note that we put aside the (difficult and important) question of how to determine which act tokens are represented at all. In this paper we focus on understanding the relationships between act tokens that make up cause trees and intention trees.

A further point to note about cause trees and intention trees is that they can look very similar when they represent the same act. This is because an agent’s action plan is constructed on the basis of what that agent believes or understands will occur in the causal world, at least in the standard case. The critical difference between the two kinds of tree...
structures is that intention trees diagram goals, means, and side effects, whereas cause
trees do not.

Act trees can be drawn to contain an abundance of information about subjects’ mental
representations (Mikhail, 2011; also see Fig. 17), but in this paper we focus on develop-
ing a methodology that allows us to test hypotheses about how subjects are representing
two features of human actions in particular: (a) their causal structure and (b) their inten-
tion structure.

1.1.1. Test of causal structure: By-test

For any two acts, A and B, that can be attributed to the same agent, the “by-test” asks
whether it is appropriate to say of those acts that A was done by doing B. We use the by-test
to tap into subjects’ mental representations of causal sequence. Our predictions for how sub-
jects will respond to the by-test can be read right off of our cause trees. If we have drawn
the cause tree correctly, subjects should affirm that A was done by doing B if A sits above
B on the cause tree. In this sense, by-relations are true “reading down” the cause tree.

1.1.2. Test of intention structure: In-order-to-test

For any two acts, A and B, that can be attributed to the same agent, the "in-order-to"
(IOT-) test asks whether it is appropriate to say of those acts that A was done in order to
do B. The IOT-test captures the notion that the agent had in mind—as an element of her
action plan—doing A with the aim of bringing about the intended effect B. Our predic-
tions for how subjects will respond to the IOT-test can also be read right off of our inten-
tion trees. If we have drawn the intention tree correctly, subjects should affirm that A
was done in order to do B if A sits below B on the intention tree. In this sense, in-order-
to-relations are true “reading up” the intention tree. 1

1.1.3. Junction-crossing

Although both tests rely on the fact that key components of an action plan can be ordered
serially, there is a subtle but critical difference between the by-test and the IOT-test: the by-
test permits “junction crossing” but the IOT-test does not. A junction in an act tree is a node
where a branching occurs, that is, where two separate causal sequences diverge. As long as
the by-test “runs down" the act tree (i.e., takes the form “A by B” where A is above B on
the act tree), then the sentence will be true regardless of which branch of the tree node A is
on. This is not the case for the IOT-relation. The IOT-relation is true when it “runs up” the
act tree (i.e., takes the form “A in order to B” where A is below B on the act tree), but only
when it crosses the junction onto one branch (the branch depicting means/goals) and not the
other (the branch depicting side effects). It is this difference in junction-crossing that differ-
entiates subjects’ representations of causal trees and intention trees.

For example, consider the cause tree depicted in Fig. 1(a). The node “killing the man”
appears above the node “throwing the man.” Despite the fact that getting from “killing the
man” to “throwing the man” requires crossing the junction (the branching of the act tree),
we expect subjects to affirm the sentence “Ian killed the man by throwing the man.” Now
consider the intention tree depicted in Fig. 1(c). Again, the node “killing the man” appears
above the node “throwing the man.” Again, getting from the former node to the latter requires crossing the junction and (importantly) the junction-crossing occurs from the side effect branch to the means/goal branch. We therefore expect subjects to deny the sentence “Ian threw the man in order to kill the man.” If we detect this kind of junction-crossing difference in subject responses, it will suggest that the by- and IOT-tests can be used to probe subjects’ representations of the causal structure and intention structure of human actions.

2. Study 1

2.1. Methods

2.1.1. Stimuli

We used two classic versions of the “trolley problem” as stimuli (Mikhail, 2000; Thomson 1985). In the Sidetrack version, a train is out of control and about to kill five people who are stuck on the train tracks and cannot get off in time. A bystander can flip a switch and turn the train to a side-track, which will kill one person standing there, but save the five on the main track. In the Footbridge version, a train is again out of control and about to kill five people who are stuck on the tracks. A bystander and a heavy man are both standing on a footbridge that overlooks the tracks. The bystander can push the heavy man off the bridge into the path of the train, which would kill the man, stop the train, and save the five people on the tracks (see Appendix). In each case, five people are saved and one person is killed, yet subjects tend to judge the agent’s action in the Sidetrack case and being more permissible than in the Footbridge case (Cushman, Young, & Hauser, 2006; Mikhail, 2011).

For each of the scenarios, Sidetrack and Footbridge, we built a cause tree and an intention tree reflecting our hypothesis about how subjects will mentally represent the two cases (Fig. 1). The by- and IOT-tests were created in the following way: All nodes of interest (NOIs) from the act trees in Fig. 1 were connected in pairs with “by” and “in order to.” The NOIs from Sidetrack included the following: “throwing the switch,” “turning the train,” “preventing the train from killing the men,” “causing the train to hit the man,” and “killing the man.” The NOIs from Footbridge included the following: “throwing the man,” “causing the train to hit the man,” “preventing the train from killing the men,” and “killing the man.” This process of connecting the NOIs in pairs resulted in 64 sentences total; 32 by-sentences and 32 in-order-to-sentences.

Of the 64 stimuli sentences, 52 of them can be categorized into four main sentences types: (a) correct causal structure, (b) incorrect causal structure, (c) correct intention structure, (d) incorrect intention structure (see Table S1 in Supplementary Materials for complete sentence categorization scheme). Sentences in these categories reflect clear directional hypotheses: Sentences that fall into categories (a) and (c) are those for which the act trees predict that subjects will judge the sentences “true.” Sentences that fall into categories (b) and (d) are those for which the act trees predict that subjects judge the sentences “false.”
For example, sentence [1] falls into the “correct causal structure” category because its description of causal structure is consistent with the cause tree.

[1] Hank turned the train by throwing the switch.

The node “throw the switch” occurs lower on the tree than “turn the train,” indicating that the former act-token should be represented by subjects as causally generating the latter (see Goldman, 1970, for a more detailed explanation of “causal generation”). Subjects should therefore judge this sentence *true*.

Sentence [2] falls into the “incorrect causal structure” category:

[2] * Hank threw the switch by turning the train.

Subjects should therefore judge this sentence “false.”

Sentence [3] falls into the “correct intention structure” category:

[3] Ian threw the man in order to prevent the train from killing the five men.

The node “prevent the train from killing the five men” appears above “threw the man” on the vertical chain of means and ends. This indicates that the former node is causally generated by the latter and that the former is directly connected to the latter by the vertical chain of means and ends in the agent’s action plan.

Sentences [4] and [5] fall into the “incorrect intention structure” category:

[4] * Ian prevented the train from killing the five men in order to throw the man.

[5] * Ian threw the man in order to kill the man.

These sentences describe a reversed causal order of events (Sentence [4]) or connect a node on the main branch with a node on the side-branch (Sentence [5]), thereby entailing an incorrect intention structure.

Of the 64 sentences in our stimuli, 12 of them do not have directional hypotheses given by the act trees. These sentences are those that connect the “branches” of the act trees. They can be divided into four categories: (a) side effect by goal, (b) goal by side effect, (c) side effect IOT goal, (d) goal IOT side effect. We discuss these sentences further in the analysis and discussion sections.

2.1.2. Subjects

Subjects were recruited through Amazon’s MTURK website and paid for participating. Subjects were recruited until 40 subjects who passed the attention check had completed each of the two conditions. One hundred and one total subjects completed the experiment; of these, 21 were excluded for failing the attention check. All subjects read both Footbridge and Sidetrack stories. (See Appendix for text of the stimuli.) Subjects were randomly assigned to the “by-test” condition or the “in-order-to-test” condition. Based on their condition, subjects responded “true” or “false” to each of 32 by-test sentences or 32 IOT-test sentences.
2.2. Results

2.2.1. Main sentence types

The 52 sentences with directional predictions were collapsed into two categories: correct structure (including correct causal structure and correct intention structure sentences) and incorrect structure (including incorrect causal structure and incorrect intention structure sentences). Average responses for each sentence type were computed for each subject in the following way: If a subject responded “true” to a sentence, that item was scored with a 1. If a subject responded “false,” that item was scored with a 0. The mean score for the sentences in each sentence type was then computed for each of the 80 subjects. This mean (ranging from 0–1) is the subject’s probability of responding “true” to a sentence in a given type. That is, if a subject receives a probability score of 1 for a sentence category, this indicates that the subject responded “true” to every sentence in that category. The distribution of the subject means for each sentence category was compared. Subjects were significantly more likely than chance to respond “true” to sentences in the correct structure category ($M = 0.88, SD = 0.17, t(79) = 19.9, p < .0001, \text{two-tailed}$) and less likely than chance to respond “true” to sentences in the incorrect structure category ($M = 0.21, SD = 0.21, t(79) = 12.4, p < .0001, \text{two-tailed}$). Subject responses to the two sentence types were significantly different from each other ($t(79) = 21.7, p < .0001, \text{two-tailed}$; see Fig. 2).

The 52 sentences were then divided into four categories: (a) correct causal structure, (b) incorrect causal structure, (c) correct intention structure, and (d) incorrect intention structure (see Table S1 in Supplementary Material for full list of sentences in each category). Average responses for each sentence type were computed for each subject as described above. The distribution of subjects’ probabilities of responding “true” to a sentence in a given type was compared. Subjects were significantly more likely than chance to respond “true” to sentences in the correct causal structure category ($M = 0.91, SD = 0.15, t(39) = 17.3, p < .0001, \text{two-tailed}$) and the correct intention structure category ($M = 0.85, SD = 0.19, t(39) = 11.8, p < .0001, \text{two-tailed}$). Subjects were significantly less likely than chance to respond “true” to sentences in the incorrect causal structure category ($M = 0.25, SD = 0.24, t(39) = 6.5, p < .0001, \text{two-tailed}$) and the incorrect intention structure category ($M = 0.22, SD = 0.22, t(39) = 7.2, p < .0001, \text{two-tailed}$).

Fig. 2. Subject responses to "correct" and "incorrect" structure sentences. Average subject probabilities of judging a sentence “true” (accuracy scores) are graphed for each sentence type. An asterisk denotes a significant difference between sentence types ($p < .0001$).
structure category \((M = 0.18, SD = 0.15, t(39) = 13.3, p < .0001, \text{ two-tailed})\). Subject responses to the sentences in the correct causal structure category were significantly different than responses to sentences in the incorrect causal structure category \((t(39) = 13.5, p < .0001, \text{ two-tailed})\). Subject responses to the sentences in the correct intention structure category were significantly different than responses to sentences in the incorrect intention structure category \((t(39) = 17.8, p < .0001, \text{ two-tailed})\). See Fig. 3.

2.2.2. Junction-crossing

Of the 52 sentences with directional hypotheses, 20 of them were included in the junction-crossing analysis. These were sentences that (1) passed through a junction in the act tree (where two causal sequences diverged) and (2) were of the form \(A \textit{by } B\) where A is above B on the action tree or \(A \textit{in order to } B\) where A is below B on the action tree. In other words, if junction crossing does not matter, then all the sentences should be correct. If junction crossing does matter, then all the \textit{by}-sentences should be correct, regardless of which junction-crossing category they fall in. However, only the IOT sentences that fall in the non-junction-crossing category should be correct. These sentences were divided into categories based on whether they (1) described the causal structure or intention structure of the action and (2) whether they crossed between the side branch and the main branch (crossed the junction) or stayed on the main branch (did not cross the junction). This division created the following four categories: (a) causal structure, crosses junction, (b) causal structure, does not cross junction, (c) intention structure, crosses junction, (d)

<table>
<thead>
<tr>
<th>Causal Structure, Crosses Junction</th>
<th>Intention Structure, Crosses Junction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hank killed the man by turning the train</td>
<td>Hank turned the train in order to kill the man</td>
</tr>
<tr>
<td>Hank killed the man by throwing the switch</td>
<td>Hank threw the switch in order to kill the man</td>
</tr>
<tr>
<td>Hank caused the train to hit the man by turning the train</td>
<td>Hank turned the train in order to cause the train to hit the man</td>
</tr>
<tr>
<td>Hank caused the train to hit the man by throwing the switch</td>
<td>Hank threw the switch in order to cause the train to hit the man</td>
</tr>
<tr>
<td>Ian killed the man by throwing the man</td>
<td>Ian threw the man in order to kill the man</td>
</tr>
<tr>
<td>Ian killed the man by causing the train to hit the man</td>
<td>Ian caused the train to hit the man in order to kill the man</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Causal structure, does not cross junction</th>
<th>Intention structure, does not cross junction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hank prevented the train from killing the five men by turning the train</td>
<td>Hank turned the train in order to prevent the train from killing the five men</td>
</tr>
<tr>
<td>Hank prevented the train from killing the five men by throwing the switch</td>
<td>Hank threw the switch in order to prevent the train from killing the five men</td>
</tr>
<tr>
<td>Ian prevented the train from killing the five men by throwing the man</td>
<td>Ian threw the man in order to prevent the train from killing the five men</td>
</tr>
<tr>
<td>Ian prevented the train from killing the five men by causing the train to hit the man</td>
<td>Ian caused the train to hit the man in order to prevent the train from killing the five men</td>
</tr>
</tbody>
</table>
intention structure, does not cross junction. (See Table 1 for a full list of sentences in each category.)

Average responses for each sentence type were computed for each subject as described above. The distribution of subjects’ probabilities of responding “true” to a sentence in a given type were compared. Subjects were significantly more likely than chance to respond “true” to sentences in the following categories: causal structure, crosses junction (\(M = 0.90, SD = 0.19, t(39) = 13.1, p < .0001, \) two-tailed), causal structure, does not cross junction (\(M = 0.92, SD = 0.16, t(39) = 16.6, p < .0001, \) two-tailed), and intention structure, does not cross junction (\(M = 0.91, SD = 0.19, t(39) = 13.2, p < .0001, \) two-tailed). Subjects were significantly less likely than chance to respond “true” to sentences in the intention structure, crosses junction category (\(M = 0.26, SD = 0.30, t(39) = 4.9, p < .0001, \) two-tailed). Subject responses to the two causal structure categories were not significantly different (\(t(39) = 1.0, \) n.s). Subject responses to the two intention structure categories were significantly different (\(t(39) = 11.4, p < .0001, \) two-tailed). See Fig. 4.

![Fig. 3](image1.png)

**Fig. 3.** Subject responses to four main sentence categories. Average subject probabilities of judging a sentence “true” (accuracy scores) are graphed for sentences in four sentence types. An asterisk denotes a significant difference between sentence types (\(p < .0001\)).

![Fig. 4](image2.png)

**Fig. 4.** Subject responses to “junction crossing” sentences. Average subject probabilities of judging a sentence “true” (accuracy scores) are graphed for sentences in four sentence types. An asterisk denotes a significant difference between sentence types (\(p < .0001\)).
2.2.3. Branch hopping

Of the 64 stimuli sentences, 12 did not fit into categories with directional predictions. Those sentences were divided into four types: side effect by goal, goal by side effect, side effect IOT goal, goal IOT side effect (see Table S2 in Supplementary Material for full list of sentences in each category). Average responses for each sentence type were computed for each subject as described above. The distribution of subjects’ probabilities of responding “true” to a sentence in a given type was compared. Subjects were significantly more likely than chance to respond “true” to sentences in the goal by side effect ($M = 0.93, SD = 0.15, t(39) = 17.7, p < .0001, \text{two-tailed}$), side effect IOT goal ($M = 0.92, SD = 0.12, t(39) = 12.5, p < .0001, \text{two-tailed}$), and side effect by goal ($M = 0.68, SD = 0.37, t(39) = 3.0, p < .01, \text{two-tailed}$) categories and significantly less likely than chance to respond “true” to sentences in the goal IOT side effect category ($M = 0.20, SD = 0.32, t(39) = 6.0, p < .0001, \text{two-tailed}$). See Fig. 5. “Goal by side effect” is accepted at a significantly higher rate (83.3%) than “side effect by goal” (67.5%), Upton’s $\chi^2 (1, n = 240) = 8.08, \phi = 0.18, p = .0045, \text{two-tailed}$. “Side effect IOT goal” is accepted at a significantly higher rate (91.7%) than “Goal IOT side effect” (20.0%), Upton’s $\chi^2 (1, n = 240) = 124.4, \phi = 0.72, p < .0001, \text{two-tailed}$.

2.2.4. Individual items

An individual item analysis was conducted with sentence as the random variable. Of the 64 stimuli sentences, 52 of them were categorized into one of the four sentence types, which had predictions about the direction of subjects’ responses. Thirty-eight of those 52 (73.1%) were evaluated in the predicted direction significantly more likely than chance (corrected for multiple comparisons, Binomial tests, $p < .0001$, two-tailed, see Table S1 in Supplementary Materials).

2.3. Discussion

Our first question of interest was whether subjects’ performance on the by- and IOT-tests aligned with our predictions. We found that subjects overwhelmingly judged
sentences that are in the “correct causal structure” and “correct intention structure” categories as “true” and sentences that are in the “incorrect causal structure” and “incorrect intention structure” categories as “false.” Thus, we may conclude that (a) subjects spontaneously represent the causal and intention structure of human actions, (b) the by- and IOT-tests are effective measures for revealing that competence, and (c) our hypothesized act trees capture something meaningful about the way subjects represent the causal and intentional structure of the typical trolley task.

2.3.1. Junction-crossing
The main finding was that the by-relation permits junction-crossing, whereas the IOT-relation prohibits junction-crossing. The by-relation captures causal representations of the world and is insensitive to the goal/side effect distinction, whereas the IOT-relation captures representations of intention structures, where the goal/side effect distinction is paramount. This is the first direct and systematic experimental evidence that the act trees we presented above (Fig. 1; c.f. Mikhail, 2007) may be accurate diagrams of subjects’ representations of the complex moral actions described in the trolley problem stimuli. Moreover, this finding suggests that subjects can infer properties of human actions that are not directly given in the stimulus.

2.3.2. Branch hopping
In contrast to junction-crossing, we had no specific predictions concerning sentences that were generated via “branch hopping.” It was not obviously apparent how the two terminal nodes of the act trees stand in relation to each other when investigated using the by- and IOT-tests. For example, it was unclear to us at the outset of this study whether subjects would represent that a side effect is done “by” achieving a goal (or vice versa) or that a side effect is done “in order to” achieve a goal (or vice versa). It seems possible to read Goldman (1970) as suggesting that by-relations work exclusively for relations that go directly down branches. If so, then this would suggest that subjects should judge “false” relations that hop across branches. Earlier we postponed addressing this question, but now we take it up directly.

In our study, subjects accepted sentences of the form “goal by side effect” (e.g. “Hank prevented the train from killing the five men by killing the man.”) and “side effect by goal” (e.g. “Hank killed the man by preventing the train from killing the five men.”). However, “goal by side effect” was accepted at a significantly higher rate than “side effect by goal.” Subjects accepted sentences of the form “side effect in order to goal” (“Hank killed the man in order to prevent the train from killing the five men.”) at a higher rate than sentences of the form “goal in order to side effect” (“Hank prevented the train from killing the five men in order to kill the man.”). Put another way, the side effect in our stories seems to be behaving as if it is below the goal node on the act tree (see Fig. 6). That is, sentences of the form X by Y and Y in order to X are accepted just in case Y is below X on the act tree, and that is how we see the side effect and goal nodes behaving. What could explain this effect?
To answer this question, we draw on an insight of Eric d’Arcy (1963), whose philosophical account of action representation seems to be supported by recent research on model-based hierarchical reinforcement learning (e.g., Botvinick & Weinstein, 2014). His insight is that some sequences of actions and action descriptions can be mentally “chunked” together for purposes of processing. d’Arcy (1963) describes a case (attributed to J.J. Smart, although ultimately derived from John Austin and others; see Hall, 1960) in which a sheriff frames, convicts, and executes a person for a crime he did not commit to avoid a mob’s vigilante justice that would kill five people. D’Arcy points out that the act committed by the sheriff could be described thusly:

1. He tensed his forefinger.
2. He pressed a piece of metal.
3. He released a spring.
4. He pulled the trigger of a gun.
5. He fired a gun.
6. He fired a bullet.
7. He shot a bullet at a man.
8. He shot a bullet towards a man.
9. He shot a man.
10. He killed a man.
11. He committed judicial murder.
12. He saved four lives. (p. 3).

D’Arcy’s point in generating this list is to show that “there is not necessarily one, and only one, correct description of a given act” (p. 10) and that the act can be described differently on different occasions given the interest of the inquirer. Moreover, an action may

![Fig. 6. The act tree on the left is a simplified version of the act tree for the Hank story (side-track case). The “killing the man” node is represented on a side-branch, but based on subject judgments of the by and IOT sentences, the node seems to have some of the properties we would expect if it were located below the goal node (depicted in the act tree on the right). The dotted arrow shows the apparent movement of the “killing the man” node from its original hypothesized location to its “collapsed” location. [Color figure can be viewed at wileyonlinelibrary.com]]
sometimes simply be referred to by its consequences. So the action of firing a gun may be felicitously described as killing a man. However, D’Arcy argues that there are some restrictions on which consequences may be used to re-describe an act. In particular, when an act results in a consequence of a certain level of “significance” then that consequence may not be ignored in describing the act. These special consequences include “killing, maiming, slander, torturing, deceiving” and so on (p. 18). In the example above, then, the sheriff’s action cannot be referred to as an act of “saving four lives” given that the action had the significant consequence of killing a man.

One way of reading D’Arcy’s claim is that certain salient effects of an action cause mental “chunking” such that the sub-sequence of acts can be re-described in terms of the one more salient consequence. This is a view explored in the field of model-based hierarchical reinforcement learning (MB-HRL; Botvinick & Weinstein, 2014). The thought is that the mind represents “options” for purposes of decision-making, which are planning sub-sequences (a group of interrelated actions) that can be reused and recombined across planning episodes. Planning then occurs in a salutatory fashion, jumping from action to effect without needing to necessarily represent intermediate, low-level outcomes. Put another way, a series of actions is mentally “recast” as their more abstract (sometimes temporally extended) effect (Botvinick, Niv, & Barto, 2009; Sutton, Precup, & Singh, 1999).

A major area of research in the field of MB-HRL is attempting to solve the “option discovery problem,” that is, to figure out which action sub-sequences get chunked. D’Arcy’s suggestion is that chunking occurs when an action results in a “morally charged” effect. The main point for our purposes is that when chunking takes place, an entire sub-sequence of actions can be referred to by a more general description of the action. In our case, chunking occurs for the event of killing the man on the side-track, so the descriptions of the act that occur on the action tree prior to (below) that node can be re-described by the event of killing the man. As a result, the node “killing the man” appears to be lower on the act tree than it actually is because “killing the man” can stand in for any of its subparts. This is why subjects affirm the sentences such as “Hank prevented the train from killing the five men by killing the man” and “Hank killed the man in order to prevent the train from killing the five men” (see Fig. 7).

One key feature of the chunking approach is that it is still possible to zoom in on the chunked action sub-sequence when prompted to. That is, subjects can recognize that Hank flipped the switch, turned the train, caused the train to hit the man, and killed the man, even though they can also re-describe that entire action sequence as an action of killing the man. In other words, chunking does not erase the ability to represent the internal structure of the action sub-sequence. One implication of this view is that if the effects in our story were replaced with those that did not encourage chunking, then we would expect the branch-hopping phenomenon to disappear and subjects to reject by-relations of the form “side effect by goal” and “goal by side effect.”

It is important to note that morally charged effects are probably only a subset of the effects that elicit chunking. If so, then this method of investigating the mental
representation of acts can be fruitfully applied to morally indifferent acts as well as morally salient ones. We will take up this point in more depth in Section 5 below, where we also compare our “chunking” proposal to an alternative account of “branch collapsing” proposed by Knobe (2010).

3. Unexpected findings

In our first study, we emphasized that our theory does an excellent job predicting the data. Overall, the data are explained by act trees we drew for the moral scenarios used in Study 1 and the rules we posited for how subjects would respond to by- and IOT-tests. We now turn our attention to individual sentences that did not elicit the predicted response from our subjects. Given that our hypothesis predicted subject responses to most of the sentences so accurately, why does it make an incorrect prediction for some of them?

Fig. 8 shows the distribution of sentences by the percent of subjects that responded to the sentence with the predicted response (“true” for correct causal and intention structures and “false” for incorrect causal and intention structures). The distribution is heavily skewed towards the right, indicating that subjects respond to most sentences as the theory predicts. However, there are two sentences on the left side of the distribution which subjects respond to at around chance levels, which seems to indicate that subjects are confused or uncertain about them:

![Fig. 7. The act tree for the Hank story (side-track). The dotted line indicates the “chunked” action. This is an action sub-sequence that can be referred to by the more general action description “killing the man.” The general term can stand in for any of the component parts, including the ones that are lower on the action tree. This may be why it makes sense to say that Hank prevented the train from killing the men by killing the man. Here, “killing the man” may be standing in for a node that is closer to the basic action (e.g., “throwing the switch”). [Color figure can be viewed at wileyonlinelibrary.com]
1. Ian caused the train to hit the man by preventing the train from killing the five men.
2. Ian threw the man in order to cause the train to hit the man.

In what follows, we will attempt to determine why these sentences do not elicit the behavior predicted by the theory using two different strategies. The first (taken up by Study 2) is to investigate whether the language that is being used in each sentence fails to capture what the act tree is supposed to depict. The second (taken up by Studies 3 and 4) is to investigate whether the act tree we have drawn (and tested) does not accurately reflect subjects’ mental representations of the moral action in the story. Put another way, the first approach considers whether our linguistic probe is misleading, while the second considers whether our hypothesis about subjects’ mental representation (the act tree) is mistaken.

3.1. Study 2: Revising the linguistic probe

What do the two target sentences have in common? The main similarities are that both sentences describe the Ian story (footbridge case) and include the node “cause the train to hit the man.” We can call these sentences “Ian+Cause” sentences.

It seems possible that subjects equivocate about the truth value of these sentences because it sounds strange to say that Ian caused the train to do something when Ian was not acting on the train at all. Rather, Ian throws the man in order to enable the train to hit the man or to cause the man to be hit by the train. (Hank, on the other hand, the protagonist in the side-track version, does act on the train more directly by flipping the switch.) On this view, the words we chose to describe the action do not form an accurate match with how, in fact, we suspect subjects are representing the action. So, we should modify the language in the Ian sentences accordingly.

3.1.1. Methods
3.1.1.1. Stimuli: The stories and questions were the same as in Study 1, with the following exception: In the Ian questions, all sentence parts that read “cause the train to hit the man” were replaced with “cause the man to be hit by the train.”
3.1.1.2. Subjects: Subjects were recruited through Amazon’s MTURK website and paid for participating. One hundred subjects completed the experiment (50 subjects in the “by” condition and 50 subjects in the “in order to” condition). Of these, 9 subjects in the by-condition and 15 subjects in the IOT condition were excluded for failing an attention check. All subjects read both Footbridge and Sidetrack stories and responded “true” or “false” to each of 32 by-test sentences or 32 IOT-test sentences, depending on their condition.

3.1.1.3. Results: Overall, we replicated the main results of Study 1. (See Fig. S1 and Table 2 in the Supplementary Material. Our main question was whether modifying the language of the Ian+Cause sentences would move the distribution of the sentences, and in particular whether subjects’ performance on the two target sentences would improve. The answer we uncovered was negative. As shown in Fig. 9, the distribution of sentences looks similar to the distribution before the revision.

We analyzed this data with a linear mixed effects model that allows for participants and items (i.e., questions) to be modeled as random factors. This sort of analysis is designed to be able to have greater confidence that results will generalize beyond the specific prompts used and subjects recruited (Baayen, Davidson, & Bates, 2008; Jaeger, 2008). The model was implemented using the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2016). To determine whether the effect of study was significant, we compare a model that includes the “study” term and item type as fixed factors with a model that includes only item type as a fixed factor. In both models, subject and sentence were treated as random variables (Barr, Levy, Scheepers, & Tily, 2013). This analysis confirms that there is no significant effect of study on subjects’ judgments ($\chi^2 (1, n = 156) = 0.096, \phi = 0.025, p = .76$, two-tailed), and this finding holds when the analysis is restricted to the Ian+Cause sentences ($\chi^2 (1, n = 156) = 0.59, \phi = 0.061, p = .44$).

Fig. 9. Histogram of the percentage of subjects responding as predicted to each sentence for Study 1 and Study 2. The distributions of subjects’ responses for the two studies are not significantly different.
Moreover, we conducted planned comparisons which show that there was no significant difference between Study 1 and this study on subjects’ willingness to affirm either of the target sentences (see Fig. 10).

1. “Ian caused the train to hit the man by preventing the train from killing the five men”: Upton’s $\chi^2 (1, n = 75) = 0.024$, $\varphi = 0.018$, $p = .88$, two-tailed.

2. “Ian threw the man in order to cause the train to hit the man”: Upton’s $\chi^2 (1, n = 81) = 0.59$, $\varphi = 0.085$, $p = .44$, two-tailed.

3.2. Study 3: Revising the act tree

Changing the wording of the sentences to more accurately reflect our initial hypothesis about how subjects are representing the cases did not improve subject performance on the outlier sentences. It therefore seems possible that our hypothesis about how subjects are mentally representing the action described by each of these sentences needs to be modified. Could each of the Ian+Cause sentences require its own more particular explanation? Here, we attempt to explain only one of them to illustrate how such an investigation might proceed.

About half of subjects say that the sentence “Ian threw the man in order to cause the train to hit the man” (or “cause the man to be hit by the train”, as in Study 2) is true and half say it is false. This sentence describes a critical piece of Ian’s action plan; without an intention to cause/allow the train to hit the man, Ian’s action of throwing the man is nonsensical. It is the critical means by which Ian plans to prevent the train from killing the five. Subjects seem to understand all of this (as evinced by their competence at responding to the large majority of other sentences describing the story), yet they do not seem to want to affirm the proposition that Ian caused the man to be hit by the train. What’s going on?

![Fig. 10. Percentage of subjects responding in the predicted direction for each of the target sentences for Study 1 and Study 2. The black line indicates chance performance. The dotted line indicates performance significantly better than chance.](image-url)
One possibility worth considering is that “in order to” picks out only ends and subjects are not willing to use “in order to” to pick out means. This explanation cannot be quite right, however, because subjects are willing to endorse the sentence “Hank flipped the switch in order to turn the train” (97.5% of subjects respond that this sentence is true in Study 1), which uses “in order to” to pick out a means.

Another possibility is that subjects are not representing that Ian intends for the train to hit the man, but they are representing that Ian intends for the train to hit something. Of course, the “thing” that Ian can use to stop the train is the man, but on this view the “man-ness” or personhood of this object is not part of Ian’s action plan; instead, he merely conceives of it as a physical object capable of stopping the train. This hypothesis is depicted in the modified act tree in Fig. 11. We tested the hypothesis that subjects are representing Ian’s action in this way in Study 3.

3.2.1. Methods

Stimuli: Subjects read the Footbridge story from Study 1 and responded to the “in order to” sentences which fell into the “correct” or “incorrect” category (no “branch hopping” sentences were used). All those sentences (10 in total) were the same as those in Study 1 with the following exception: All sentence parts that read “cause the train to hit the man” were replaced with “cause the train to hit something.” Subjects responded “true” or “false” to each sentence.

Subjects: Subjects were recruited through Amazon’s MTURK website and paid for participating. Subjects were recruited until 40 subjects completed the experiment and passed the attention check (62 subjects completed the experiment; of these 22 were excluded for failing the attention check).

3.2.2. Results

As a preliminary analysis, sentences were divided into two categories, “correct” and “incorrect” based on the hypothesis depicted in the act tree (Fig. 11, see Table S1 of Supplementary Material for sentence categorization). For sentences in the “correct”
category, a response of “true” was the predicted response (and “false” for “incorrect” sentences). A value of 1 was assigned to each response that was predicted and a value of 0 to each response that was not predicted. An overall percent of predicted responses was computed for each sentence. Overall, subjects responded in the predicted direction 81.7% of the time. For these sentences, in Study 1, subjects responded in the predicted direction 83.8% of the time. A mixed effects model (treating study—Study 1 or 3—and item type—correct/incorrect—as fixed effects and subject and sentence as random effects) confirms that there is no significant effect of study on subjects’ judgments of the 10 sentences included in this study ($\chi^2 (1, n = 80) = 0.25, \phi = 0.056, p = .62$, two-tailed). (See Table S4 in Supplemental Material for sentence-by-sentence breakdown.)

Our main question was whether subjects would affirm the modified sentence, “Ian threw the man in order to cause the train to hit something.” Out of 40 subjects, 25 (62.5%) responded “true” to the target sentence (Binomial test, one-tailed, $p = .08$). This was not a significant difference from the percentage of subjects that responded “true” to the target sentence in Study 1 (52.5%; Upton’s $\chi^2 (1, n = 80) = 0.808, \phi = 0.10, p = .37$, two-tailed).

3.2.3. Discussion

Our hypothesis was that subjects are representing Ian’s intention in terms of one description of an action (“causing the train to hit something”) and not another (“causing the train to hit the man”). Our modification of the “cause the train to hit the man node” seems to have somewhat improved subject performance from chance to marginally better than chance. While this improvement does not meet the criteria of statistical significance and therefore should be interpreted with caution, we return to explore this idea about whether it is sensible to claim that a person can direct his intention so “strictly” in Section 4.

3.3. Study 4: Making things explicit

The target sentence contains the phrase “cause the train to hit the man/something.” While many of the nodes that are drawn on the act trees in Fig. 11 are explicitly given in the stimulus (e.g., throw the man, prevent the five men from being killed, killing the one man), the fact of the train’s hitting the man after he is thrown is not. Instead, it is a node that subjects must infer. Of course, in order for the agent’s plan to make sense, the subject must recognize that Ian intends for the train to hit the man. However, given the implied nature of that fact, we wondered if subjects were somehow managing to suppress representing the actual contact of the train and the man. To test this hypothesis, we created a version of the Footbridge story where the train’s hitting the man is explicitly mentioned.

3.3.1. Methods

3.3.1.1. Stimuli: Subjects read the Footbridge story, which was the same as the one used in Study 1, with the following exception: the story explicitly mentioned that the agent
would cause the train to hit the man wherever in the story that fact was implied in Study 1. (Exact stimuli can be found in the Appendix.) Subjects responded to same sentences as in Study 3. (As in Study 3, all sentence parts that read “cause the train to hit the man” were replaced with “cause the train to hit something.”) Subjects responded “true” or “false” to each sentence.

3.3.1.2 Subjects: Subjects were recruited through Amazon’s MTURK website and paid for participating. Sixty subjects were recruited and of these 16 were excluded for failing the attention check.

3.3.2. Results

As in Study 3, sentences were divided into two categories, “correct” and “incorrect.” For sentences in the “correct” category, a response of “true” was the predicted response (and “false” for “incorrect” sentences). A value of 1 was assigned to each response that was predicted and a value of 0 to each response that was not predicted. An overall percent of predicted responses was computed for each sentence. Overall, sentences were responded to in the predicted direction 79.5% of the time. A mixed effects model (treating study—Study 1 or 4—and item type—correct/incorrect—as fixed effects and subject and sentence as random effects) shows that the effect of study on subjects’ judgments of the 10 sentences included in this study failed to reach significance ($\chi^2 (1, n = 84) = 3.17, \phi = 0.19, p = 0.08$, two-tailed). (See Table S5 in Supplemental Material for sentence-by-sentence breakdown.)

Our main question was whether subjects would affirm the modified sentence, “Ian threw the man in order to cause the train to hit something.” Thirty-four out of 44 subjects (77.3%) responded “true” to the target sentence (Binomial test, $p < .0001$, two-tailed). Significantly more subjects responded in the predicted direction to the target sentence in this study than did so in Study 1 (Upton’s $\chi^2 (1, n = 84) = 5.62, \phi = 0.26, p < .05$, two-tailed).

![Fig. 12. Percentage of subjects responding in the correct direction to the sentence “Ian threw the man in order to cause the train to hit the man [something]” in studies 1, 3, and 4. The black line indicates chance performance. The dotted line indicates performance significantly better than chance.](image-url)
tailed; see Fig. 12). So the “explicit cause manipulation” of this study seems to have a targeted effect on our sentence of interest, rather than generally making it easier for subjects to respond to the sentences overall.

3.3.3. Discussion

Following the manipulations in Studies 3 and 4, subjects are willing to affirm the sentence “Ian threw the man in order to cause the train to hit something.” This involved both abstracting away from the person-hood of the man being thrown (referring to his object-hood instead in the “something” manipulation) as well as explicitly pointing out in the stimulus that this would occur. Of course, it remains unclear whether the effect is due to the explicitness manipulation alone or the interaction of the two manipulations. Regardless, the findings in this study suggest that subjects do in fact see the harm caused to the man as part of Ian’s action plan (a means to his end). Moreover, it seems that subjects are somewhat resistant to affirming (or representing) one of the batteries that will result from Ian’s action. We take up this point more fully in Section 4.

4. Extending the act tree method to other moral cases

This paper is concerned with developing an important tool—act trees—for diagramming complex hypotheses about the causal and intention structures of human actions and a method of testing those hypotheses: the by- and IOT- tests. Thus far, we have validated these tools on the well-studied cases of the side-track and footbridge trolley dilemmas, and we have demonstrated how the tool could be used to further investigate segments of the data that were not predicted by the hypothesis. We now turn to showing how this method can contribute to ongoing debates about the mental representation of human actions.

The Principle of Double Effect (PDE) is a non-consequentialist moral principle that holds that it is sometimes permissible to bring about a foreseen harm if that harm is a side effect and not specifically intended (see Mikhail, 2011, for a more complete definition). It is therefore critical for proponents of the PDE to explain whether a given effect counts as intended (as a means or an end) or merely foreseen (as a side effect). A famous thought experiment derived from Lon Fuller (Fuller, 1949; Shapiro & Fuller, 1999) is often used to highlight the difficult nature of this question. In the thought experiment, a group of cave explorers is in a cave as water in the cave starts to rise. A large member of the caving party attempts to leave the cave through the small exit and gets stuck. If the cavers do nothing, they will all drown. The cavers have dynamite with them, however, and they can detonate it and blow up the man who is stuck in the exit, killing the man, and allowing the rest of the caving party to escape (see also Foot, 1978). Philosophical debate persists about whether the cavers who detonate the dynamite intend the death of the man or merely intend to blow him to bits. Those in the latter category argue for a “strict” definition of intention. For example, Masek (2010) argues that because the death of the man per se is not causally necessary for bringing about the goal of the cavers, the death of the man can be regarded as collateral (a side effect). After all, if the large man somehow survived being blown up, that would not
interfere with the cavers’ goal of escaping the cave and saving themselves. On the other side, Foot (1978) argues that the events of blowing the man to bits and the man’s being killed are conceptually “too close” to claim that a person can aim at one without also aiming at the other; so, if it is clear that the cavers intend to blow the man to bits, they also intend to kill him (see Bennett, 1998, for an elaboration and critique of this view).

In the philosophical literature, this argument proceeds with each side pointing out that the position of the other side seems absurd or at least mistaken. Unsurprisingly, no consensus about this matter has been achieved. When viewed as an empirical question, however, this dispute can be informed by the experimental techniques that we have been developing in this paper. So conceived, the issue turns on whether subjects mentally represent the death of the trapped caver as a side effect (as diagrammed in Fig. 13a) or as a goal of the caving party (as diagrammed in Fig. 13b). We also test a third alternative: that the caver strictly intends to blow up the thing blocking the exit, rather than the man per se (as diagrammed in Fig. 13c). The act tree corresponding to this alternative assumes that subjects may represent the caver’s intention in an even stricter sense than Masek suggests.

4.1. Study 5: Another moral case

4.1.1. Methods

All subjects read the following vignette: “Suppose a group of explorers are in an underground cave that is filling with water. A large member of the caving party attempts to leave the cave through the small exit and gets stuck in the exit. If the cavers do nothing, they will all drown. The cavers have dynamite with them, however, and one of them detonates it and blows up the man who is stuck in the exit, killing the man, and allowing

![Fig. 13. Three candidate act trees that describe different hypotheses for how subjects are mentally representing the action of the caver in Study 5. Foot’s view is best described by model (a), namely, that the caver intends the death of the man when he blows him to bits. Masek’s view—a “strict” sense of intention, where killing the man is not intended—is best described by model (b). We include an even stricter notion of intention in model (c) in which both blowing up and killing the man are side effects of the intention of blowing up the thing blocking the exit. [Color figure can be viewed at wileyonlinelibrary.com]
the rest of the caving party to escape.” Because we were interested in comparing three
candidate act trees (Fig. 13), we restricted the sentences we used to those that would dif-
f erentiate most clearly between the trees. We queried the “in order to” relation between
the node “the caver detonated the dynamite” and the other nodes. All sentences were of
the form “the caver detonated the dynamite in order to [second node].”

Subjects were asked to indicate whether the following descriptions of the caver’s
action are true or false. Sentences were presented in a randomized order.

1. The caver detonated the dynamite in order to kill the man.
2. The caver detonated the dynamite in order to blow up the man.
3. The caver detonated the dynamite in order to blow up the thing that was blocking
   the exit.
4. The caver detonated the dynamite in order to clear the entrance to the cave.
5. The caver detonated the dynamite in order to allow the other cavers to escape.

We were interested in which of the three model intention trees in Fig. 13 best predicted
subjects’ true/false judgments. Sentences (4) and (5) do not differentiate between models
and served only to confirm that subjects generally agreed in their understanding of the story.

4.1.1.1 Subjects: Forty subjects were recruited through Amazon’s MTURK website and
paid for participating.

4.1.2. Results

Eighty percent of subjects judged question (4) above true (Binomial $p < .001$, two-
tailed) and 95% considered question (5) above true (Binomial $p < .001$, two-tailed) con-
firming a high degree of agreement on overall story understanding.

The main aim of this study was to determine which of the act tree models (Fig. 13) best
captures subjects’ mental representation of the intention of the caver. Comparing subjects’
responses across sentences (1) through (3) shows significant differences between sentences
(Cochran’s $Q = 22.9, p < .001$). Model (a) predicts subjects will agree with sentence (1),
yet only 17.1% did so (Binomial $N = 40, x = 7, p < .001$, two-tailed). Likewise, model
(b) predicts that subjects will agree with sentence (2), yet only 30% did so (Binomial
$N = 40, x = 12, p = .017$, two-tailed). Only model (c) is so far consistent with these data.
All three models predict agreement with sentence (3) and indeed a majority of subjects
(60%) did agree with (3), yet this number fails to reach significance (Binomial $N = 40,
x = 24, p = .27$, two-tailed). However, these tests are inadequate follow-ups for Cochran’s
$Q$ because they treat subjects’ answers as if three independent groups were tested. In fact,
each subject actually responded to all sentences in a repeated-measures design. The appro-
priate test is McNemar’s Test of Changes. This examines how often subjects changed their
answers across pairs of sentences (e.g., answered “true” to Sentence 1 and “false” to Sen-
tence 2), comparing changes from “true” to “false” with changes in the opposite direction.
Cross tabulating responses to sentences (1) and (2) shows that 25 subjects responded
“false” to both and 4 responded “true” to both (no change). The remaining 11 subjects
who changed answers were evenly split between 3 who thought only “kill the man” (S1)
was true and 8 who thought only “blow up the man” (S2) was true (McNemar binomial, \( p = .227 \), two-tailed). Comparing S1 with S3 in the same way shows that of the 17 subjects who changed their answers, all 17 judged “kill man” false but “clear exit” (S3) true, while no subject showed the opposite pattern (McNemar binomial, \( p < .001 \), two-tailed). Crucially, cross tabulation of responses to S2 and S3, shows that, of the 12 subjects who changed their answers, all 12 judged “blow up the man” to be false and “clear the exit” true; no subject showed the opposite pattern (McNemar binomial, \( p < .001 \), two-tailed), providing strong support for model (c) over the others (see Tables 2a–2d).

4.1.3. Discussion

A long-standing debate around the PDE concerns whether it is reasonable (or even possible) to aim at one effect as a means and to consider a very closely related description of that same effect (often a harmful description) as a foreseen side effect. This debate has often taken as its focal point the scenario about the trapped cavers that we tested here. Some have argued that the claim that the caver can aim to blow up the man while not aiming to kill him is absurd and therefore “too strict” a delineation of what

Table 2A-D

Experiment 5. Results testing predictions of three model act trees.

<table>
<thead>
<tr>
<th>Act Tree Models</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement 1</td>
<td>The caver detonated the dynamite in order to <strong>kill the man.</strong></td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td>Statement 2</td>
<td>The caver detonated the dynamite in order to <strong>blow up the man.</strong></td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>Statement 3</td>
<td>The caver detonated the dynamite in order to <strong>blow up the thing</strong> that was blocking the exit.</td>
<td>True</td>
<td>True</td>
</tr>
</tbody>
</table>

Prediction (true or false) made by each act tree for each statement. The key prediction for model (c) is that subjects will agree with statement 3 while rejecting statements 1 and 2. Percentages in parentheses show observed “True” judgments for each statement. Looking down columns, model (a) falls at the first hurdle, statement (1), model (b) falls at the second hurdle, statement (2), leaving only model (c) as consistent with observation (Cochrane’s Q = 22.9, \( p < .001 \)).

Table 2B

Statements 1 vs 2

<table>
<thead>
<tr>
<th>Blow up man?</th>
<th>False</th>
<th>True</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kill man?</td>
<td>False</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>True</td>
<td>3</td>
</tr>
</tbody>
</table>

Cross-tabulated counts of subjects answering true/false by statements 1 (kill the man) and 2 (blow up the man). Over 60% of subjects rejected both statements. The critical differential comparison is between subjects who answer differently for the two statements, 8 versus 3 (McNemar Binomial, \( p = .08 \), n.s.).
action descriptions can be intended (Bennett, 1998; Foot, 1978). Our results point to an even stricter understanding of intention that is at play in the way subjects represent these cases. Not only are subjects willing to say that the caver did not intend the death of the man, they are also unwilling to say that the caver intended to blow the man to bits (62.5% deny both). Instead, they prefer to say that he acted only to clear the exit to the cave. This finding may call for a re-analysis of the sort of event descriptions that are widely assumed to be intended in cases of harm as a means.

The act tree that presupposed that the cavers intended to act on an impersonal entity (“the thing blocking the exit”) rather than “a man” best predicted subjects’ judgments of the caver’s intention. The “thing” that is to be harmed is a person and we do not doubt that subjects are aware of this. Their unwillingness to affirm sentences that describe the harm as happening to a person seems to be an expression of the tension between the tendency to impute only good intentions to agents when possible (Levine, Mikhail & Leslie, in press) and causal feasibility. On the one hand, subjects do not want to admit or contemplate that the agent intends to harm the victim. On the other hand, it is impossible to deny that harming the victim is causally necessary to achieving the agent’s end. So, while subjects may grasp that the harm was unavoidable and in this sense intended, they express some unwillingness to admit this fact, and seem to be figuratively averting their (mind’s) eye from the gruesome nature of what must be done.

### Table 2C
Statements 1 vs 3

<table>
<thead>
<tr>
<th></th>
<th>Blow up thing blocking exit?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>False</td>
</tr>
<tr>
<td>Kill man?</td>
<td></td>
</tr>
<tr>
<td>False</td>
<td>16</td>
</tr>
<tr>
<td>True</td>
<td>0</td>
</tr>
</tbody>
</table>

Cross-tabulated counts of subjects answering true/false by statements 1 (kill the man) and 3 (blow up the thing blocking the exit). Forty percent of subjects rejected both statements. The critical differential comparison is between subjects who answer differently for the two statements: 17 subjects agree with the latter statement and reject the former, while no one showed the opposite pattern (McNemar Binomial, $p<.001$, two-sided).

### Table 2D
Statements 2 vs 3

<table>
<thead>
<tr>
<th></th>
<th>Blow up thing blocking exit?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>False</td>
</tr>
<tr>
<td>Blow up the man?</td>
<td></td>
</tr>
<tr>
<td>False</td>
<td>16</td>
</tr>
<tr>
<td>True</td>
<td>0</td>
</tr>
</tbody>
</table>

Cross-tabulated counts of subjects answering true/false by statements 2 (blow up the man) and 3 (blow up the thing blocking the exit). Again, 40% percent of subjects rejected both statements. The critical differential comparison is again between subjects who answer differently for the two statements. Strikingly, given that the man is the thing blocking the exit, 12 subjects agreed with the latter statement while rejecting the former, while no one showed the opposite pattern (McNemar Binomial, $p<.001$, two-sided).
In this section, we return to our analysis of the “branch hopping” sentences that we discussed in Section 2. “Branch hopping” sentences are those that connect nodes on the side-effect and goal branches of an act tree with “by” and “in order to.” Knobe (2010) offers a different explanation for the “branch hopping” phenomenon than the “chunking” explanation we suggest. Like us, Knobe points out that there are some cases in which branch-hopping sentences are endorsed by subjects as adequate descriptions of the causal or intention structure of an agent’s action and some cases in which those sentences are denied. In Section 2, we explained this data by suggesting that action sub-sequences can be “chunked” and points in the action sub-sequence can be referred to by a more general description of the sequence.

Knobe explains the data by appeal to a theory of “branch collapsing.” On this view, there are some action descriptions that pick out different actions and some action descriptions that pick out “the very same action.” For example, sometimes an action that we think of as side effect is better understood as the very same action as the more basic action that brought it about. For example, Knobe describes a case (Ulatowski, 2008) in which Smith pumps poisoned water into a cistern which acts as a house’s water supply. Knobe points out that people perceive the action of operating the pump to be the very same action as the action of poisoning the people in the house. In cases like these, instead of drawing an act tree as being composed of a basic action, a goal, and a side effect, Knobe suggests that it may be more accurate to collapse the side effect node into the basic action and depict them in the same location on the action tree (see Fig. 14).

Knobe goes on to argue that branch collapsing occurs when an effect can be seen as “what a person is most essentially doing.” Morally bad effects are more likely to be seen as “what a person is most essentially doing” than morally good effects, but there are

Fig. 14. The act tree on the left depicts a basic action and a goal (get paid) with a morally bad side effect of poisoning the people. Knobe’s (2010) suggestion is that the bad side effect is seen as “what a person is most essentially doing” and therefore collapses into the more basic node, as depicted by the act tree on the right. [Color figure can be viewed at wileyonlinelibrary.com]
other kinds of action descriptions that can count as “what a person is most essentially doing” that have nothing to do with morality (this will be taken up below).

Knobe’s “branch collapsing” theory explains a good deal of the data, but it falls short insofar as it fails to account for the ability to perceive relationships among intermediate nodes between a basic action and a side effect. Consider the following case that Knobe uses to test his “branch collapsing” thesis: “Imagine a chef in a kitchen who is making some breakfast in the ordinary way. His arms are moving in just the way one would expect for a chef in his position. He is thereby making an omelette. He is also getting some exercise.” Knobe suggests that the basic action of the chef’s “moving his arms” results in two effects, which can be depicted as nodes on an act tree: the chef’s “making an omelette” and “getting some exercise.” However, “making an omelette” is what the chef is “most essentially doing,” so this node collapses into the “moving his arms” node (see Fig. 15). This collapsing explains why subjects are willing to say that the chef gets some exercise by making an omelette and why they are unwilling to say that the chef makes an omelette by getting some exercise.

However, consider the following modified version of the case: “Imagine a chef in a kitchen who is making some breakfast in the ordinary way. His arms are moving in just the way one would expect for a chef in his position. He is thereby making an omelette, which will be Jane’s breakfast. He is also getting some exercise.” If branch collapsing occurs in this case, then we would expect subjects to endorse the sentence:

(1) “He got some exercise by making breakfast for Jane”

And we would expect subjects to reject the sentence:

(2) * “He made breakfast for Jane by getting some exercise.”

This indicates that the conditions for collapsing seem to obtain (see Fig. 16, left panel). However, the collapsing view fails to account for the fact that we continue to

![Fig. 15. The cause tree on the left indicates that the chef moves his arms and produces two effects, getting some exercise and making an omelette. The cause tree on the right shows how, on Knobe’s (2010) view, the “making omelette” node collapses into the more basic node. [Color figure can be viewed at wileyonlinelibrary.com]](fig15.png)
represent the relationship between the chef’s moving his arms, making an omelette, and making breakfast. The chef makes breakfast by making an omelette, makes breakfast by moving his arms, makes an omelette by moving his arms but does not move his arms by making breakfast and forth (see Fig. 16, right panel). The branch collapsing view fails to capture this feature of the mental representation of the action. In Study 6, we tested this feature of the mental representation of acts.

5.1. Study 6

5.1.1. Methods

The method of this study closely follows the method of Experiment 3 in Knobe (2010). All subjects read the following vignette: “Imagine a chef in a kitchen who is making some breakfast in the ordinary way. His arms are moving in just the way one would expect for a chef in his position. He is thereby making an omelette, which will be Jane’s breakfast. He is also getting some exercise.” The addition of the phrase “which will be Jane’s breakfast” was the only change from the story given to subjects in Knobe’s Experiment 3.

The following nodes were connected in a pair-wise manner to create all combinations of by-test sentences: “got some exercise,” “made an omelette,” “moved his arms,” “made breakfast for Jane.” Subjects were asked to rate their agreement of the resulting sentences on a scale from 1 (disagree) to 7 (agree). (To maintain consistency with Knobe’s procedure, subjects rated their agreement rather than judging sentences true or false as in our previous studies.) Sentences were presented in randomized order.

1. He got some exercise by making an omelette.
2. He made an omelette by getting some exercise.
3. He made an omelette by moving his arms.
4. He moved his arms by making an omelette.
5. He got some exercise by moving his arms.
6. He moved his arms by getting some exercise.
7. He moved his arms by making breakfast for Jane.
8. He made breakfast for Jane by moving his arms.
9. He got some exercise by making breakfast for Jane.
10. He made breakfast for Jane by getting some exercise.
11. He made an omelette by making breakfast for Jane.
12. He made breakfast for Jane by making an omelette.

5.1.1.1 Subjects: Forty subjects were recruited through Amazon’s MTURK platform and were paid a small amount for participating. Following Knobe (2010), no attention check was used in this study.

5.1.2. Results
First, we replicated Knobe’s effect that he attributes to “branch collapsing,” namely that subjects agree more with the sentence “He got some exercise by making an omelette” (M = 5.4; SE=.32) than “He made an omelette by getting some exercise” (M = 2.88; SE = 0.33), paired-sample t-test, two-tailed, t(39) = 6.14, p < .0001. Likewise, subjects agree more with the sentence “He got some exercise by making breakfast for Jane” (M = 5.38; SE = 0.33) than “He made breakfast for Jane by making an omelette” (M = 6.6; SE = 0.14), paired-sample t-test, two-tailed, t(39) = 6.92, p < .0001. This indicates that our modified omelette story exhibits the effect attributed to “branch collapsing.”

The main result of this study is that subjects represent the nodes on the supposedly “collapsed branch” as having an ordered structure. That is, “made breakfast for Jane” is represented as above the node “made an omelette” and “made an omelette,” as above “moved his arms.” We can verify this because the sentence “He made breakfast for Jane by making an omelette” is rated as having significantly higher agreement (M = 6.62; SE = 0.14) than the sentence “He made an omelette by making breakfast for Jane” (M = 5.03; SE = 0.37), paired-sample t-test, two-tailed, t(39) = 4.15, p < .001. Likewise, “He made an omelette by moving his arms” was rated as having significantly higher agreement (M = 6.33; SE = 0.17) than “He moved his arms by making an omelette” (M = 5.05; SE = 0.38), paired-sample t-test, two-tailed, t(39) = 3.14, p < .005. Finally, “He made breakfast for Jane by moving his arms” (M = 6.05; SE = 0.25) was rated with higher agreement than “He moved his arms by making breakfast for Jane” (M = 5.38; SE = 0.34), paired-sample t-test, two-tailed, t(39) = 1.9, p = .06. A full list of the results from this study is given in Table S5 of the Supplementary Materials.

5.1.3. Discussion
On the chunking view, the felicity of (1) and non-felicity of (2) is explained by the fact that the chef’s moving his arms, making an omelette, and thereby making breakfast can be mentally chunked and described by the general description of “making breakfast” (as depicted in the act tree in Fig. 16, right panel). However, when the intermediate
nodes are queried, we can mentally “zoom in” to that branch and respond to sentences that describe the relationship between the nodes.

There is another feature of our mental representation of acts that the chunking view captures that the collapsing view does not, which is the temporal sequence of the nodes of the action. Returning to the case of Smith poisoning the people (Ulatowski, 2008), it seems right to say that Smith poisons the people after he operates the pump but not that Smith operates the pump after he poisons the people. Likewise, the chef gets some exercise after he moves his arms but does not move his arms after he gets some exercise. This indicates that there is a structured asymmetry between the two nodes that is revealed by a test of temporal sequence. Collapsing the nodes together (i.e., treating operating the pump and poisoning the people as the very same action) fails to capture that structure. The chunking view, on the other hand, would support the test of temporal sequence because it maintains a structural, ordered distinction between the nodes.

The broader point that this example brings to light is that a test of temporal sequence (e.g., a before/after test) can reveal the structure of the mental representations of acts in addition to the tests (by and IOT) that we have focused on in this paper. (See Fig. 17 for examples of how temporal markings could be added to act trees to reflect hypotheses about the temporal sequence of acts tokens.)

A shortcoming of both the chunking and collapsing views is that no worked-out account of the conditions for chunking or collapsing have yet been given. d’Arcy (1963) and Knobe (2010) both agree that morally bad acts seem to generate this phenomenon, but Knobe’s addition of the omelette case seems to suggest that collapsing/chunking

![Figure 17](image.png)

Fig. 17. Act trees representing additional information about the mental representations of two trolley problems, Footbridge (Ian) and Sidetrack (Hank). In addition to causal structure and intention structure of the actions of the agent, these act trees encode two additional features, temporal sequence (indicated by the superscript letters next to each node) and deontic units (indicated by the insertion of nodes at the location where a “battery” occurs).
occurs more broadly. Future work should aim to delineate when chunking occurs and when it does not.

6. Conclusion

Confirming earlier predictions (Mikhail, 2000: 178–179, Mikhail, 2011; 175–178), we developed a novel method of displaying and testing complex hypotheses about the mental representation of human action and demonstrated how these hypotheses can be tested using the by-test and IOT-test. Our findings thus pose a challenge to those researchers who either ignore the problem of how moral intuitions arise from eliciting situations (e.g., Haidt, 2001) or who uncritically assume that the mental representations of human action underlying moral judgment are exceedingly simple and can be adequately described in terms of heuristics and biases (e.g., Sunstein, 2005).

The act trees discussed in this paper highlight causal and intention information. With respect to the latter, it seems likely that theory of mind capacities (Csibra & Gergely, 1998; Leslie, 1994b) are at work when subjects make IOT judgments. On the other hand, our results leave open whether theory of mind capacities are also at play when subjects make by-judgments. Do subjects’ causal representations reflect assumptions about the agent’s own interpretation of the world or do they reflect the subject’s own interpretation of the world in which the agent is merely situated? Future research should investigate these and similar questions.

The act trees in Fig. 17, adapted from Mikhail (2011), include additional features beyond the causal and intentional properties of human action, including information related to temporal sequence, value of outcomes, and deontic structure. However, it is important to emphasize the limited and preliminary nature of all these proposals. We encourage other researchers in psychology, philosophy, linguistics, computer science, law, and other fields whose work touches on the representation of human action to adapt the act tree method for their own purposes. To name just a few examples, we think that this method can be used to inform existing debates and to promote further research on each of the following topics: the doctrine of triple effect and the role of so-called motivating side effects (Kamm, 2007; Masek, 2010; Shaw, 2005); which action descriptions are “conceptually close” to one another and whether that makes a moral difference (Bennett, 1998; Fitzpatrick, 2006; Foot, 1978; Hart, 1968; Quinn, 1989); whether “cause to die” is part of the conceptual structure of “to kill” (Fodor, 1970); the distinction between direct and indirect causes and effects (Fischer, Ravizza, & Copp, 1993; Quinn, 1989; Royzman & Baron, 2002); the role of the probability of an intention, desire, norm, or effect (Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015); and the “same offense” requirement in the law of double jeopardy (Goldman, 1994). These are just a small subset of possible applications of our act tree method, all of which sit at the intersection of human action and moral judgment.

Outside of the domain of moral psychology, moreover, we believe that there are many cases in which act trees can also be profitably used to investigate complex hypotheses.
about human action representation. For this reason, the theories and methods developed in this paper have much broader implications for cognitive science than the specific illustrations on which we have focused here. For example, neuroscientists interested in the representation of human action (e.g., Allison, Puce, & McCarthy, 2000; Blakemore & Decety, 2001) might rely on the act trees used in cognitive and behavioral studies to develop their own detailed hypotheses about the neural implementation of cognition. Likewise, developmentalists interested in how causal perception and intention inference develop (see Woodward, 2013, for a review) and become integrated into moral judgment (see Hamlin, 2015, for a review) might profitably use act trees to specify their hypotheses and highlight puzzles more clearly (see, e.g., Levine, 2016; Levine, Mikhail, & Leslie, In Press).

Finally, we note that a computational approach to action representation is quickly becoming indispensable in the fields of computer science, robotics and artificial intelligence, as the race to develop models of judgment and decision-making for use by artificial agents accelerates in domains as diverse as transportation, medicine, commerce, health, and military operations. In order to successfully navigate the social world, these agents need to be equipped with a human-like ability to compute fine-grained representations of human actions. The act tree method is an invaluable means for achieving this end.

Author contributions

All authors contributed to the study concept. S.L. and J.M. designed the studies and S.L. carried out data collection. S.L. conducted data analysis under the guidance of A.L. S.L. drafted the manuscript, with critical revisions by A.L. and J.M. All authors approved the final version of the manuscript for submission.

Acknowledgments

Parts of this research was supported by NSF grant BCS-0922184 to A.L., the DoD NDSEG Fellowship to S.L., and Georgetown Summer Writing Grants to J.M.

Notes

1. In explaining the by- and IOT-tests, the test is formulated in terms of a backward look at the action, that is, in the past tense. “A was done by doing B” and “A was done in order to do B.” This backward-looking perspective is not central to either test; the same tests and analysis can be applied in a forward-looking manner, that is, to acts that are being contemplated (or evaluated) but that have not yet been
performed. Action plans are temporally neutral in this respect and can be given both ex post and ex ante.

2. The attention check used for experiments in this paper was as follows: The following text appeared at the beginning of the experiment: “Thank you for taking the time to complete this survey. For our research, we are interested in knowing certain facts about you. Specifically, we are interested in whether you read directions carefully. So, in order to show us that you have read the instructions, please ignore this first question about attending university (don’t answer it). Also, please copy and paste the words ‘I have read the instructions’ without the quotation marks in the box labeled ‘Any comments or questions?’ Thank you very much.” The question “Have you attended university?” then appeared with four answer choices followed by a box labeled “Any comments or questions?” Subjects were excluded if they answered the question about university or did not write “I am paying attention” in the box. Subjects were not excluded if they wrote in the box that they were paying attention but had accidentally clicked a radio button on the university question and were unable to clear the response.

References


**Supporting Information**

Additional Supporting Information may be found online in the supporting information tab for this article:

**Table S1.** Subject responses to sentences formed from the 52 pairs of interest with specific predictions about the by- and IOT-tests.
**Table S2.** Individual item analysis of each of the branch-hopping sentences.
**Table S3.** Subject responses to sentences in Study 2.
**Table S4.** Subject responses to each sentence in Study 3.
**Table S5.** Subject responses to each sentence in Study 4.
**Table S6.** Mean agreement ratings and standard error of the mean for each sentence in Study 6.
**Fig. S1.** Percentage of subjects in Study 2 responding “true” to sentences in each of the four major categories.

**Appendix**

**Stimuli for Study 1**

The names of the scenarios are provided for reference only. Subjects were not shown the names of scenarios. Stimuli were taken from Mikhail (2007).

**Sidetrack case**

Hank is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Hank sees what has happened: The driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Hank is standing next to a switch, which he can throw, that will turn the train onto a side-track, thereby preventing it from
killing the men. There is a man standing on the side-track with his back turned. Hank can throw the switch, killing him; or he can refrain from doing this, letting the five die.

**Footbridge case**

Ian is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Ian sees what has happened: The driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Ian is standing next to a heavy object, which he can throw onto the track in the path of the train, thereby preventing it from killing the men. The heavy object is a man, standing next to Ian with his back turned. Ian can throw the man, killing him; or he can refrain from doing this, letting the five die.

**Stimuli for study 4**

The differences between the stimuli used in Study 1 and those used in Study 4 are bolded for reference here. No words were bolded for subjects in the study. The name of the scenario is provided for reference only. Subjects were not shown the name of the scenario.

**Footbridge case**

Ian is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Ian sees what has happened: The driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Ian is standing next to a heavy object, which he can throw onto the track in the path of the train, thereby causing the train to hit the object, which in turn will prevent it from killing the men. The heavy object is a man, standing next to Ian with his back turned. Ian can throw the man, causing the train to hit the man and killing him; or he can refrain from doing this, letting the five die.