Abstraction Modulates Judgments of Intentional Action

Angela Cao (ayc8004@nyu.edu), Mark Ho & Robert Rehder
New York University, Department of Psychology

Abstract

This paper investigates how the different levels of hierarchical plans modulate judgments of intentional action. While previous research has focused on how moral valence influences intentionality judgments of side-effects (outcomes that were not a part of the agent's original plan), we focus on sub-events, which are candidate decompositions of higher-level goals that *were* a part of the agent's plan. We address two central questions: (1) does the impact of moral valence on intentionality judgments differ between sub-events and side-effects? and (2) does the level of abstraction at which a sub-event is defined affect judgments of its intentionality? Our results show that the effect of moral valence is largely neutralized for sub-events, which instead exhibit a decrease in perceived intentionality as they are decomposed into lower-level actions. This indicates that future models of intentional action should take into account the asymmetry in judgments between higher- and lower-level actions.

Keywords: Intention; Abstraction; Event structure; Morality

Introduction

Judgments of intentional action are fundamental to our cognition. From legal courtrooms to everyday disputes, distinguishing between what an agent brought about intentionally versus unintentionally is central to how we assign blame, praise, and punishment (Halpern & Kleiman-Weiner, 2018; Kleiman-Weiner et al., 2015; Levine et al., 2018), and is related to how we perceive agents' plans (Bratman, 1987; Bratman, 2018). While there is a substantial body of literature examining how factors such as moral valence influence intentionality judgments of side-effects (Halpern & Kleiman-Weiner, 2018; Knobe, 2003, 2004; Levine et al., 2018; Nanay, 2010), less attention has been paid to judgments of the hierarchical components of plans themselves.

We argue that sub-events—the constitutive parts of a plan—are a distinct phenomenon from side-effects, and that ratings of their intentionality are governed by different factors, such as their level of depth. While side-effects are outcomes that are foreseen by the agent but are not instrumental to the achievement of the goal (Levine et al., 2018), and are thus causally related to the action, sub-events share a constitutive relationship with the super-ordinate event. For example, "Sally accruing interest" is a causal side-effect of "Sally saving money," whereas "Sally saving $30 on May 1st" is a constitutive part of *saving money*. A similarly constitutive part of "Sally saving $30 on May 1st" is "Sally opens the envelope that contains her paycheck," but we have the intuition that this latter action is perceived as less goal-related. We ask how different levels of hierarchical plans modulate intentionality judgments and explore how humans choose sets of sub-goals from the infinite number of possible event decompositions (Correa et al., 2020, 2023).

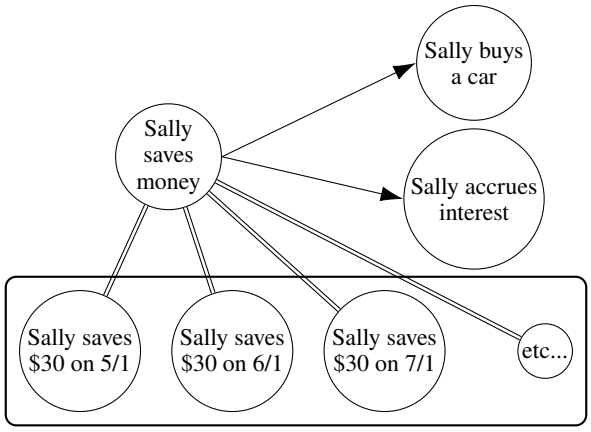

Figure 1: Representation of the scenario described in the Introduction. Directed arrows represent causal relations, while double lines (||) denote decomposition relations. Boxed components represent a potential plan. Importantly, nodes represent observable events and not mental states.

Constraints on how the achievement of goals are decomposed into sub-goals can be grounded in means-end coherence (Bratman, 1987). Means-end coherence requires that sub-plans are at least as extensive as what the agent believes is "required" to fulfill a goal, and that event components in sub-plans are defined to a level of abstraction that is "appropriate to [the agent's] habits and skills," but no further. Thus, according to this framework, sub-plans should satisfy two desiderata: (1) that the component sub-events are necessary for fulfilling some superordinate goal, and (2) that the level of abstraction is appropriate to the context. Notably, 'necessary' here refers to plan-internal necessity rather than logical necessity. While a higher-level goal can often be achieved through alternative paths, once an agent commits to a specific plan for action, the constituent sub-events become conditionally necessary for the execution of that specific plan. Furthermore, because hierarchical plans are developed recursively, they cannot be conceived of instantaneously. So, we might expect some asymmetry between higher-level components of

plans and lower-level components that are recursed to.

In this paper, we first present two questions regarding these judgments, then report two experiments designed to test related hypotheses, and finally discuss the implications of our results for models of planning and moral reasoning.

Sub-events versus side-effects

The first question we ask is whether sub-events and side-effects have the same effect on intentionality judgments. Side-effects are well-studied in the literature. Previously, Knobe (2003) demonstrated that people are more likely to deem harmful side-effects as intentional compared to helpful side-effects (the "Knobe effect"). Subsequent work specifies these valences as "moral considerations" (Knobe, 2004). Knobe argues that this effect occurs because folk judgments of intentional action are not purely about goal inference; rather, these judgments also take into account moral considerations.

Notably, Levine et al. (2018) make different predictions. They find experimental evidence for a "good intention prior," arguing that, modulo contextual information about the agent's mental states, people generally believe that good effects are more intended than bad ones. This prediction would be expected to hold for side-effects and sub-events. Levine et al. suggests that the Knobe effect occurs because the associated experiments included a declaration by the agent of not caring about the side-effect, which subsequent studies have suggested conveys added intention in the case of the negative effect and not the positive one (Guglielmo & Malle, 2010; Nanay, 2010; Sripada, 2010; Uttich & Lombrozo, 2010).

Bratman (1987)'s account makes yet another prediction: Only sub-events are intended, while side-effects are not. Because side-effects do not function as goals or means that guide an agent's future reasoning, Bratman argues that they fall outside the scope of what is intended.

We ask whether sub-events can be empirically distinguished from side-effects. We remain agnostic about the direction of any possible effect (i.e., either morally bad side-effects are rated as more intentional, à la Knobe; or, morally good events are rated as more intentional, à la Levine et al.). However, regardless of direction, we ask if the effect that moral valence has on intentionality judgments of side-effects differs from the effect that it has on intentionality judgments of sub-events.

We consider the possibility that sub-events are a distinct phenomenon from side-effects. Theoretically, this distinction relies on the difference between causal and constitutive relationships. While side-effects are merely consequences that flow causally from an action, sub-events are often components of the plan itself; steps without which the super-ordinate action would not exist. Under Bratman (1987)'s framework of means-end coherence, an agent cannot rationally intend a goal without also intending the means chosen to achieve it. Consequently, unlike side-effects, sub-events act as essential manifestations of the agent's will. Because sub-events are sometimes inseparable from the execution of the goal, they may be perceived as 'pre-authorized' by the initial intention.

In contrast, observers might not distinguish between sub-events and side-effects. On one hand, Figure 1 represents our default assumption regarding how an observer construes Sally's actions – that they understand that the sub-events (e.g, Sally saves $30 on 5/1) are constituents of Sally's plan to save money, whereas the side-effects (e.g., accruing interest) are not. Yet, there is no guarantee that an observer will make this distinction. Instead, the sub-events may be construed simply as causal consequences of Sally's plan in the same way that side-effects are. Furthermore, Rosas et al. (2020) relates these nodes under the concept of *downward causality*, and argues that the relation between certain nodes of different levels is in fact causal. Under this ascription, there is no reason to expect that an observer's intentionality judgments will differ for the two types of events.

Abstraction level of sub-events

Our second question is whether the level of abstraction at which an event is described affects its ratings of intentionality. On one hand, we reiterate that according to Bratman (2018), when an agent intends a goal, they are rationally committed to intending the means necessary to achieve it. Thus, the constitutive components of sub-plans are predicted to be (fully) intentional. This aligns with linguistic models that argue that the verb "intend" is non-gradable (Grano, 2017), suggesting that once an agent intends a goal, they must fully intend the necessary means, and that one intention can not be more intended than another. For example, Grano argues that the following sentences are infelicitous:

1. #What I intend the most is to leave.
2. #John intends to leave more than Bill does.

According to this view, we would not expect to see any intentional action to be rated as more intended than another.

On the other hand, it is possible that intentionality ascriptions "decay" as a sub-event is decomposed into more granular, lower-level steps. For instance, we might expect that as we zoom in on low-level, mechanical details required to execute a plan (e.g., moving your feet in order to get to the grocery store), the intuition that those specific details are each "intended" weakens. This would suggest that folk psychology treats intention not as a boolean property that transfers perfectly down the tree of means, but as one that dilutes as it travels deeper. Recall that Bratman argues that plans are hierarchical and partial, meaning that agents reason about achieving sub-components of plans *only* to some "appropriate level of abstraction". Bratman notes that this level is determined by whether a component is *necessary* for achieving a means and whether it is appropriate to the agent's specific *skills*. While Bratman is primarily concerned with the agent's perspective, we can reasonably infer that observers are aware that agents' plans are only partial, and that this would affect their intentionality judgments. For instance, when we watch a skilled pianist plan to and in fact play a sonata, it is certain that we believe that the pianist *intends* to play the sonata. In contrast, it is more questionable whether observers would

judge that the pianist intends the specific flexion of their finger for the thirty-second note of the second measure. The lower-level action is necessary, but it falls below the threshold of the plan's 'appropriate level' because it is handled by a well-learned skill (in this example, a motor movement), rather than conscious deliberation.

To investigate these two questions, we ran one experiment each. Both experiments used similar protocol and only varied by the sentences included in the vignette. Experiments were hosted on Qualtrics. All materials associated with this paper are available here.

Experiment 1

Stimuli

We created thematic vignettes that vary on two dimensions, resulting in four conditions:

C1: Moral valence of side-effect:

(a) morally good side-effect

(b) morally bad side-effect

C2: Moral valence of sub-effect:

(a) morally good sub-effect

(b) morally bad sub-effect

Moral valence was operationalized using Kant's Categorical Imperative (CI) to distinguish morally permissible (good) from morally impermissible (bad) actions (Kant, 1997; O'Neill, 2017). Thus, each sentence used in our stimuli that is associated with a moral valence of bad fails the CI, while each sentence associated with a moral valence of good passes.

We instantiate these four conditions in eight different themes, resulting in 32 different vignettes. Each vignette has the same structure as the following example (using the theme of 'Bennett', a morally good side-effect, and a morally bad sub-event):

> "The mayor of a city, Bennett, intends to run advertisements with the goal of getting re-elected next term. He knows that running advertisements will result in the side effect of helping local advertisement businesses. In order to successfully run advertisements, Bennett plans to coerce two of his employees, Anna and David, into designing posters for free."

Then, the associated questions that are asked to participants about the side-effect and sub-event, respectively, are:

1. How much does Bennett intend to help local advertisement businesses?
2. How much does Bennett intend to coerce Anna into designing posters for free?

Procedure

After consenting, participants were shown a summary of the slider task with an example vignette, associated questions, and potential responses, in which the left endpoint of the scale (0) is described as "the person did not intend the outcome at all", while the right endpoint of the scale (100) is described as "the person fully intended the outcome".

Next, participants are shown a trial question in which the answer is clearly above 50. Specifically, the trial vignette describes a situation in which "An AGENT intends to ACTION to reach GOAL. Part of the process for achieving ACTION is SUB-EVENT." As part of the trial, participants are asked, "How much does AGENT intend GOAL?" and "How much does AGENT intend SUB-EVENT?" Participants are only allowed to continue after rating the former question as > 50. Notably, however, we did not constrain participants' responses to the latter question as to not bias their later responses about sub-events since we do not want to assume *a priori* the range of intentionality ascribed to sub-events.

Participants then began the main section in which they are presented with eight vignettes, each paired with two questions. One question asks about the side-effect and the other asks about the sub-event, and the order of the questions is randomized. Additionally, an attention check designed to appear as a normal question was also presented.

To explain the randomization process of our eight target questions, consider that we want each participant to see each theme once and each combination of the levels of **C1** and **C2** twice. To achieve this, we used two layers of an 8×8 Williams Latin square design (Williams, 1949). This design organizes the themes to account for possible first-order carryover effects, ensuring that specific neighboring theme-pairings (e.g., Theme A following Theme B) and neighboring combinations of the levels of **C1** and **C2** are balanced across the experiment.

Finally, participants are asked to write a free-response to the prompt "Thank you for taking our survey! Please tell us how you felt about this study (using a minimum of 25 words)! Your feedback is greatly appreciated." This question was included to ensure that the experiment flow paralleled Experiment 2, although AI usage was not a concern in this experiment since the university students took the studies in person.

Participants

With the goal of achieving > 5 ratings for each of the 64 vignette-question pairs (32 vignettes paired with 2 questions each), 26 university undergraduates participated in return for course credit. Participants were native English speakers residing in the US. One participant failed the attention check and was not included in the analysis.

Results

Recall that our first question asks how morality affects side-effects, compared to sub-events. Looking at Figure 2A, we first observe that the absolute difference between mean ratings of bad and good side-effects seems much larger than that of bad and good sub-events. The pattern in side-effects seems to align with Levine et al., 2018's good intention prior. In contrast, bad sub-events and good sub-events do not seem to also show this pattern.

To assess the effects, we fit a Bayesian zero-one inflated beta (zoib) regression to the participant judgments. This model decomposes the response variable into a continuous beta com-

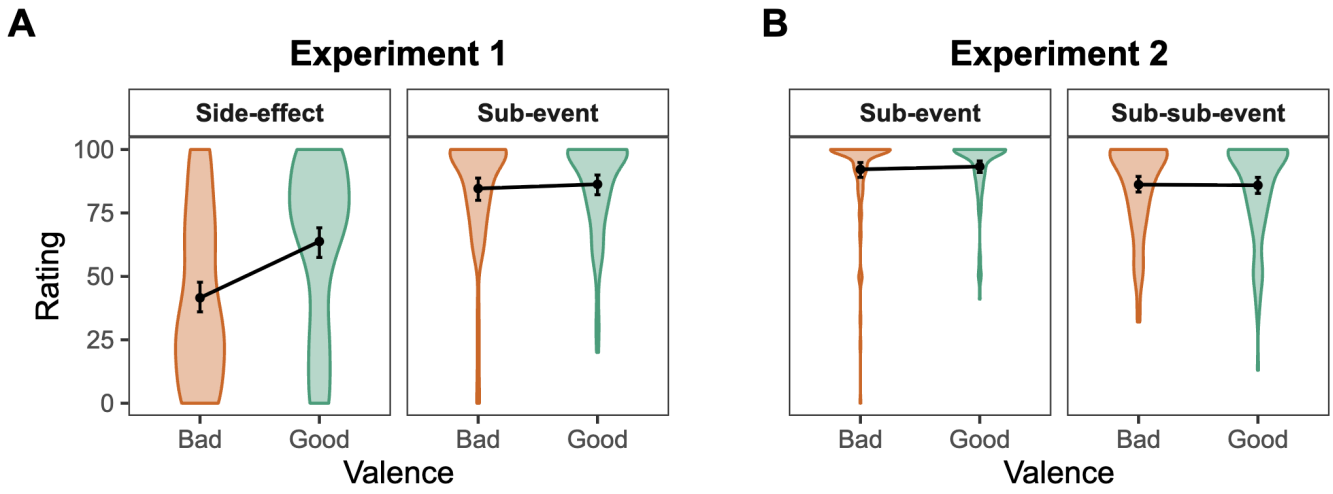

Figure 2: Participants' judgments of intentional action in both experiments. Violins represent the distribution of data, black points indicate empirical mean ratings, and error bars represent 95% bootstrapped confidence intervals.

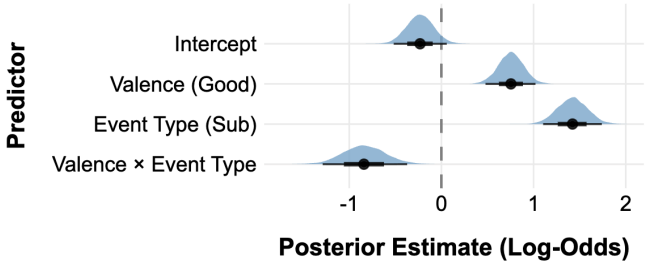

Figure 3: Posterior distributions for mean components of the zoib model in Experiment 1. The plot displays population-level coefficients (logit scale) for the effects of Valence and Event Type on ratings. Points represent posterior medians; shaded areas represent 66% and 95% highest-density credible intervals. The intercept represents the baseline for morally bad side-effects.

ponent for values between the endpoints and Bernoulli components that model the probability of responding exactly 0 or 100, which accounts for both the bounded nature of the 0–100 scale and the high frequency of responses at the endpoints, as seen in Figure 2A. We include a two-way interaction between moral valence and question type, as well as random intercepts for participant and theme. We used the default priors provided by the `brms` package. These consist of flat priors for the population-level coefficients and weakly informative Student-t or Logistic priors for the intercepts and group-level parameters.[1] Model posteriors are visualized in Figure 3.

We consider an effect supported/credible if 0 is not included in the credible interval (CrI). The model estimates that there was a supported main effect of question type when moral valence is held constant at 'bad', such that participants ascribed greater intentionality to bad sub-events than to bad side-effects ($\beta = 1.42, CrI = [1.10, 1.74]$). There was also a supported main effect of moral valence when question type is held constant at the 'side-effect' level, with participants ascribing greater intentionality to good side-effects compared to bad side-effects ($\beta = 0.75, CrI = [0.48, 1.02]$). Taken together, these results mean that sub-events are generally perceived as more intentional than side-effects, and that side-effect judgments are sensitive to moral valence, with good side-effects receiving higher intentionality ratings than bad ones.

Critically, we found a supported negative interaction between moral valence and question type ($\beta = -0.84, CrI = [-1.29, -0.37]$). This means that the effect of moral valence – the tendency to rate good actions as more intentional – was smaller for sub-events than for side-effects. If we look at just side-effects, a valence of 'good' credibly increases ratings compared to 'bad' ($\beta = -0.75$; $CrI = [-1.01, -0.48]$). However, for sub-events, this effect was neutralized, with no evidence of a difference of effects by moral valences ($\beta = 0.08$; $CrI = [-0.29, 0.44]$). This means that the asymmetric 'good intention prior' – which observers seem to use to interpret side-effects – is not applied to actions that are already represented as constitutive parts of a hierarchical plan.

Discussion

The results of Experiment 1 provide an answer to our first question, suggesting that sub-events are indeed treated as a distinct category of action, one that is less susceptible to the moral asymmetry found with side-effects.

[1] Given the shape of the data seen in Figure 2, we fit an alternative version using more informative priors consisting of Normal(0, 1) distributions for the Intercept and population-level effects (b), as well as for the inflation components (zoi and coi). Additionally, we applied an Exponential(1) prior for the precision parameter (ϕ). However, comparing the models using leave-one-out cross-validation (Vehtari et al., 2017) yields an ΔELPD of 0.0 ($SE_{diff} = 0.0$), indicating that there is no detectable difference in predictive performance between the models. Thus, we use the model with default priors for simplicity.

Note that one potential concern regarding this conclusion is that the ratings for sub-events were close to ceiling. On the one hand, the average ratings (84.6 and 86.3 for the bad and good sub-events, respectively) were well below the ceiling of 100. Yet, one might argue that the effective ceiling was lower for some subjects on the grounds that people sometimes avoid using the extreme ends of scales. Experiment 2 will present additional evidence regarding this possibility.

Experiment 2

Experiment 2 is concerned with whether the depth of a sub-event affects judgments of how intentional that sub-event is.

Stimuli

The stimuli used in Experiment 2 included the condition **C2**, and a novel condition **C3**. **C3** is defined as the abstraction level of a sub-event, and included two levels: sub-event and sub-sub-event. However, since we are re-using vignette sentences associated with **C2**, each vignette used in this experiment also encodes a moral valence, although that is not our focus in this experiment. Notably, good sub-events were only paired with good sub-sub-events, and bad sub-events were only paired with bad sub-sub-events, unlike the variations in Experiment 1. Because of this, there are only 2 different combinations of conditions, which results in 16 total vignettes after accounting for our 8 different themes. Each vignette has the same structure as the following example (using the theme of 'Bennett', a morally bad sub-event, and a sub-sub-event):

> "The mayor of a city, Bennett, intends to run advertisements with the goal of getting re-elected next term. In order to successfully run advertisements, Bennett plans to coerce two of his employees, Anna and David, into designing posters for free. In order to coerce Anna, Bennett will need to go to Anna's desk. In order to go to Anna's desk, Bennett will need to take steps to get to Anna's desk."

Then, the associated questions that are asked to participants about the sub-event and sub-sub-event, respectively, are:

1. How much does Bennett intend to coerce Anna into designing posters for free?
2. How much does Bennett intend to take steps to Anna's desk?

Procedure

The main procedure used in Experiment 2 was identical to that of Experiment 1. However, the free response question was changed such that we asked for a minimum length of 50 words. While we are interested in the participants' experiences, the primary purpose of this question is to check for AI usage using Pangram (Emi & Spero, 2024), which requires a minimum length of 50 words. We found this to be a valid concern during pilot studies using Prolific.

Participants

Because we expect the effect to be smaller than in the first experiment, we aim for >10 ratings for each of the 32 vignette-question pairs (16 vignettes paired with 2 questions each). So, 31 participants (each compensated $1.50) were collected on Prolific. Participants were native English speakers residing in the US, with a pass rate of at least 99%, and with a prior completion rate of 200+. Two participants were flagged for AI usage and not included in the analysis.

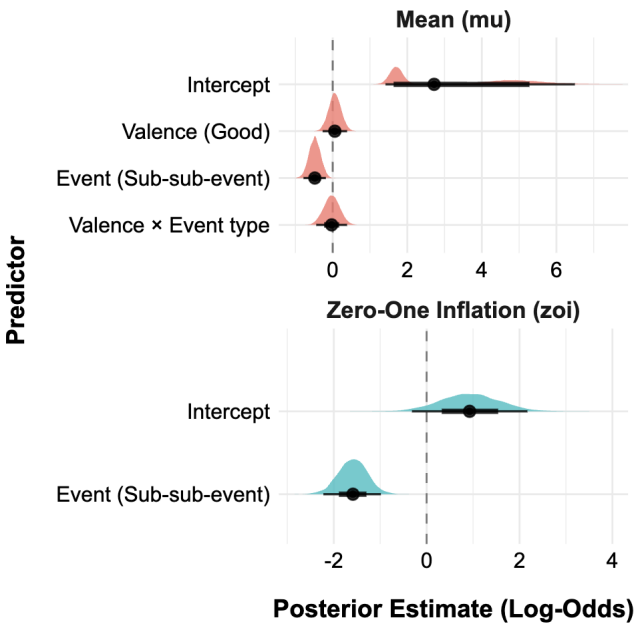

Figure 4: Posterior distributions of the zoib model used in the analysis of Experiment 2. The top panel shows effects on the mean intentionality rating (μ), and the bottom panel shows effects on the probability of boundary responses (0 or 100). Predictors are shown in log-odds. Points represent posterior means; shaded areas represent 66% and 95% highest density intervals. The intercept represents the baseline for morally bad sub-events.

Results

Recall that our second question is interested in whether the level of abstraction at which an event is described affects how intentional it is rated as. As shown in Figure 2B, the means seem to follow the trend that we expect, albeit the potential effect is not large. Notably, five participants were "straightliners" and rated everything as 100. After excluding these straightliners, the mean rating for questions about sub-events is 90.36 and the mean rating for questions about sub-sub-events is 81.88. We include the straightliners in the following analysis.

We once again fit a Bayesian zoib regression to the participant judgments. We include a two-way interaction between moral valence and question type (sub-event vs. sub-sub-event), as well as random intercepts for participant and theme. We also modeled the zero-one inflation (zoi) component as a function of question type.[2] This allows the model to

[2]We found that letting zoi vary by question-type provided a substantially better fit than a model assuming constant inflation across

estimate whether the level of abstraction affects the probability of participants providing boundary responses separately from its effect on the mean of the distribution. We used default priors. Model posteriors are visualized in Figure 4.

To begin, the model estimates that there was substantial variation between raters, indicating that participants used the scale very differently ($\sigma = 0.54, CrI = [0.35, 0.80]$). However, the effect of different themes was both small and unsupported ($\sigma = 0.09, CrI = [0.00, 0.27]$), indicating that the different vignettes demonstrated similar effects using the other predictors. Furthermore, the effect of moral valence was unsupported, indicating the lack of an effect of moral valence on participants' ratings ($\beta = 0.06, CrI = [-0.27, 0.39]$). This aligns with our results from Experiment 1. There was also no evidence for an interaction between question type and moral valence ($\beta = -0.03, CrI = [-0.44, 0.39]$), indicating that the effect of moral valence on participant judgments did not depend on whether we were asking about a sub-event or a sub-sub-event. Critically, the model estimates that there was a supported main effect of question type, meaning that participants ascribed more intentionality to sub-events than to sub-sub-events (in log-odds space unless otherwise indicated: $\beta = -0.48, CrI = [-0.78, -0.18]$). Our choice to model the zoi component as a function of question type also credibly influenced the probability of boundary responses; thus, participants were less likely to provide '0' or '100' ratings for sub-sub-events than for sub-events ($\beta = -1.59, CrI = [-2.22, -0.99]$). [3]

Discussion

The supported main effect of question type addresses our second question, demonstrating that events described at a lower level of abstraction are rated as less intentional than logically related events at higher levels. We also note that the unsupported effect of moral valence aligns with what we observed in Experiment 1; mainly, that there is no evidence that moral valence modulates judgments of sub-events the same way that it modulates judgments of side-effects.

With regards to our previous discussion about a possible ceiling effect, we note that in Experiment 2, the mean rating for sub-events are closer to ceiling (93.1). However, the mean of ratings for sub-sub-events are credibly lower at 87.1, and still do not show any effect of valence.

General Discussion

The results of our experiments provide empirical support for the hypothesis that sub-events are a distinct cognitive phenomenon from side-effects, and are governed by different cognitive factors. In Experiment 1, we observed that while moral valence credibly influenced judgments of side-effects – with "good" side-effects being rated as more intentional than "bad" ones – this effect was absent when participants evaluated sub-events. This asymmetry in side-effect judgments aligns with the good intention prior found in previous literature (Levine et al., 2018; Margoni & Surian, 2017; Mikhail, 2011), suggesting that raters generally presume agents to intend positive outcomes unless evidence suggests otherwise. However, for sub-events, the structural relationship of being a constitutive part of a higher-level plan appears to override these moral biases. This empirical difference suggests that sub-events are cognitively distinct from side-effects, and future work should look into factors that specifically affect intentionality judgments of sub-events.

To look into what factors might affect intentionality judgments of sub-events, Experiment 2 demonstrates that the level of abstraction used to describe an action modulates judgments of intentionality. We found that sub-events described at a lower level of abstraction (sub-sub-events) were rated as significantly less intentional than those described at higher levels. This supports the view that observers of agents infer that agents reason about plans only down to an "appropriate level of abstraction," as determined by the agent's habits and skills. When an action is decomposed too far – such as describing the individual steps to reach a desk – the perceived intentionality decreases, possibly because observers become less certain about the representation of such granular steps in the agent's mental representation.

Relatedly, while Experiment 2 only varied the level of depth of the sub-event, future work will look specifically into skill level. Recall our example about the skilled pianist, in which an observer is uncertain about whether the pianist intends "the specific flexion of their finger for the thirty-second note of the second measure" when playing a sonata. In contrast, imagine that we instead ask about a child who is new to the piano. In this case, it is more reasonable to state that the child intends the specific flexion of their finger for a specific note – because they are less skilled at the task, actions are less 'chunked', and the child must exert more effort in order to accomplish the same goal of performing the sonata. If expertise allows an agent to cache complex sequences into a single mental primitive, observers may perceive the individual components of those sequences as unintentional 'background' processes rather than active objects of the agent's deliberation. Thus, we speculate that the agent's skill level of the action's domain will affect ratings of how intentional that action is.

Regarding limitations, a reviewer points out that several sub-events in our stimuli were not logically required to reach the goal, raising the possibility that it is exactly the logical non-necessity that causes participants to rate the event as more intentional. While this is plausible, we argue that the inclusion of an action in a specific plan is what signals a deliberate intentional commitment, regardless of alternative paths. This should be investigated in future work, in which stimuli are varied based on whether or not the actions chosen to be included in plans are logically necessary.

conditions (ΔELPD= -156.2, $SE_{diff} = 13.3$), based on leave-one-out cross-validation.

[3] Excluding the straightliners does not affect our analysis of key effects. Specifically, none of the model intercepts changed, and none of the credible intervals' boundaries changed by more than ± 0.01 or resulted in contacting 0.

References

Bratman, M. (1987). *Intention, plans, and practical reason*. MA: Harvard University Press.

Bratman, M. E. (2018, August). Intention, belief, and instrumental rationality. In *Planning, time, and self-governance: Essays in practical rationality*. Oxford University Press. https://doi.org/10.1093/oso/9780190867850.003.0003

Correa, C. G., Ho, M. K., Callaway, F., & Griffiths, T. L. (2020). Resource-rational task decomposition to minimize planning costs. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *42*. https://escholarship.org/uc/item/72s0v38j

Correa, C. G., Ho, M. K., Callaway, F., Daw, N. D., & Griffiths, T. L. (2023). Humans decompose tasks by trading off utility and computational cost (T. U. Hauser, Ed.). *PLOS Computational Biology*, *19*(6), e1011087. https://doi.org/10.1371/journal.pcbi.1011087

Emi, B., & Spero, M. (2024). Technical report on the pangram ai-generated text classifier. https://arxiv.org/abs/2402.14873

Grano, T. (2017). The logic of intention reports. *Journal of Semantics*, *34*(4), 587–632. https://doi.org/10.1093/jos/ffx010

Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? resolving a controversy over intentionality and morality [PMID: 21051767]. *Personality and Social Psychology Bulletin*, *36*(12), 1635–1647. https://doi.org/10.1177/0146167210386733

Halpern, J. Y., & Kleiman-Weiner, M. (2018). Towards formal definitions of blameworthiness, intention, and moral responsibility. https://arxiv.org/abs/1810.05903

Kant, I. (1997). *Groundwork of the metaphysics of morals* (M. Gregor, Ed. & Trans.) [Original work published 1785]. Cambridge University Press.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. (2015). Inference of intention and permissibility in moral decision making.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(279), 190–194. https://doi.org/https://doi.org/10.1111/1467-8284.00419

Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis*, *64*(2), 181–187. Retrieved December 1, 2025, from http://www.jstor.org/stable/3329125

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

Margoni, F., & Surian, L. (2017). Children's intention-based moral judgments of helping agents. *Cognitive Development*, *41*, 46–64. https://doi.org/10.1016/j.cogdev.2016.12.001

Mikhail, J. (2011). Moral grammar and intuitive jurisprudence: A formal model. In *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment* (pp. 123–180). Cambridge University Press.

Nanay, B. (2010). Morality or modality? what does the attribution of intentionality depend on? *Canadian Journal of Philosophy*, *40*(1), 25–39. Retrieved December 12, 2025, from http://www.jstor.org/stable/27822074

O'Neill, O. (2017). A simplified account of kant's ethics. In *Applied ethics*. Routledge. https://doi.org/10.4324/9781315097176-3

Rosas, F. E., Mediano, P. A. M., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., & Bor, D. (2020). Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLOS Computational Biology*, *16*(12), 1–22. https://doi.org/10.1371/journal.pcbi.1008289

Sripada, C. S. (2010). The deep self model and asymmetries in folk judgments about intentional action. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *151*(2), 159–176. Retrieved December 12, 2025, from http://www.jstor.org/stable/40926817

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116*(1), 87–100. https://doi.org/https://doi.org/10.1016/j.cognition.2010.04.003

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Williams, E. J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Research, Series A*, *2*, 149–168.