The effect of cognitive load on intent-based moral judgment

Justin W. Martin[1&], Marine Buon[2], Fiery Cushman[3]

[1] Department of Psychology, Boston College, Chestnut Hill, MA 02467

[2] Department of Psychology, Paul Valéry University, Montpellier, France 34090

[3] Department of Psychology, Harvard University, Cambridge, MA 02138

[&] To whom correspondence should be addressed.

Boston College Department of Psychology,
140 Commonwealth Avenue
Chestnut Hill, MA 02467
E-mail: justinwmartin@gmail.com

Word count: 8,366

Acknowledgments: We thank David Mely for translation of the audio files from French to English and Maya Manning for recording the audio files for use in the verbal shadowing task. We thank Katherine McAuliffe, Lindsey Powell, Michael Henry Tessler and Liane Young for feedback on prior versions of this manuscript. This research was supported by grant 61061 from the John Templeton Foundation to FC.

Abstract

When making a moral judgment, people largely care about two factors: Who did it (causal responsibility), and did they intend to (intention)? Since Piaget's seminal studies we have known that as children mature they gradually place greater emphasis on intention, and less on mere bad outcomes, when making moral judgments. Today we know that this developmental shift has several signature properties. Recently, it has been shown that when adults make moral judgments under cognitive load they exhibit a pattern similar to young children; that is, their judgments become notably more outcome-based. Here, we show that all of the same signature properties that accompany the outcome-to-intent shift in childhood characterize the "intent-to-outcome" shift obtained under cognitive load in adults. These findings hold important implications for current theories of moral judgment.

Keywords: Moral judgment, punishment, wrongness, deliberation

1. Introduction

Human moral judgment is largely organized around two factors: *causal responsibility* for harm, and *intent* to harm (Cushman et al., 2013; Malle et al., 2014; Piaget, 1932; Trémolière & Djeriouat, 2016; Young et al., 2006, 2007; Young, Camprodon, et al., 2010; Young, Nichols, et al., 2010; Young & Saxe, 2011). These factors structure moral judgment across diverse cultures (Barrett et al., 2016) and from an early age–most clearly by around 5 years of age but potentially within the first 14 months of life (Cushman et al., 2013; Hamlin, 2013; Hamlin, Mahajan, et al., 2013; Hamlin, Ullman, et al., 2013; Mulvey et al., 2020; Nobes et al., 2009, 2016; Sommerville et al., 2013). Thus, a major goal of contemporary research is to understand how causal and mental state representations are integrated during moral judgment (Buon et al., 2016; Cushman et al., 2013; Killen et al., 2011; Malle et al., 2014; Margoni & Surian, 2016).

Many key insights have come from studies of child development. In particular, children's moral judgments show increasing sensitivity to a harmdoer's intent as they mature. For instance, seminal studies by Piaget (1932) showed that 6-year-old children will blame a person for accidental harms (e.g., a large ink spot created accidentally) more than intentional harms (e.g., purposefully creating a small ink spot). Subsequent work has found this so-called "outcome-to-intent" shift consistently and extensively (Armsby, 1971; Baird & Astington, 2004; Costanzo et al., 1973; Cushman et al., 2013; Imamoglu, 1975; Killen et al., 2011; Margoni & Surian, 2020; Piaget, 1932; Shultz et al., 1986; Yuill & Perner, 1988; Zelazo et al., 1996; cf Nobes et al., 2016, 2009) while at the same time finding signatures of this shift at younger ages than Piaget, as early as 3 to 4

years of age (Margoni & Surian, 2020). Although originally characterized in relatively simple terms, we now understand that it is a quite complex pattern of developmental changes, as we review below.

Notably, recent work suggests that when adults make moral judgments under cognitive load they exhibit a more child-like pattern of response (Buon, Jacob, et al., 2013). Specifically, adults judge accidental harms much more harshly under cognitive load than at baseline. Importantly, this is not due to an inability to determine whether the agent in question behaved intentionally or caused harm. This pattern of results may reflect a decrement in the ability to incorporate information about intentions into adults' moral judgments induced by cognitive load. Alternatively, it might reflect the prioritizing of information about outcomes over information about intentions. Either way, these results clearly show that adults under load make moral judgments exhibiting a more child-like pattern of responding.

How should we interpret the similarities in moral judgment between young children and adults under cognitive load? One possibility is that the evaluative processes underlying moral judgment in early childhood remain intact in adulthood and express themselves unimpeded in the presence of cognitive load. This possibility is in conflict with Piaget's model of moral judgment according to which intent-based moral judgment fully replaces outcome-based moral judgment during development. Insofar as the early childhood mode of judgment is revealed under conditions of cognitive load, specifically, it is more in line with the general class of proposals which suggest a combination of automatic and controlled contributions to moral judgment (e.g. Greene, 2008), and to decision-making more generally (Evans, 2008; Sloman, 2014; Stanovich

& West, 2000). In this case, the influence of outcomes on moral judgment is apparently a relatively automatic process, and becomes dominated or modulated by a resource-dependent process which asserts the influence of intentions. Thus, adult moral judgment is dominated by the controlled, intent-based process under ordinary circumstances, but the enduring automatic, outcome-based process can be revealed under cognitive load.

Currently, however, the evidence supporting this conclusion is rather circumstantial. In contrast to the complex pattern of child development characterized across numerous studies, we presently have only a narrow probe of one dimension of the adult pattern of judgment–the influence of intentions and outcomes as modulated by cognitive load–from a single study (Buon, Jacob, et al., 2013). Thus, our goal is to provide a systematic study of several key features of adult moral judgment under cognitive load in order to establish whether it tracks early childhood judgment in close detail.

1.1 Two features of the outcome-to-intent shift

We focus on two central features of the outcome-to-intent shift in early childhood (Fig. 1). First, this shift is strongest for cases of accidental harm, judgment of which changes substantially over development (Cushman et al., 2013; Killen et al., 2011; Nobes et al., 2009; Piaget, 1932; Shultz et al., 1986; Zelazo et al., 1996), though we note that some studies do not support this pattern (Margoni & Surian, 2017). As described above, at an early age (Cushman et al., 2013; Piaget, 1932; Shultz et al., 1986; Zelazo et al., 1996) moral judgment tends to show greater sensitivity to causal

responsibility and diminished sensitivity to mental states (e.g., intent). As a consequence, even accidental harms tend to be forcefully condemned, due to the agent's causal connection to a bad outcome. As children mature moral judgments are mostly dominated by sensitivity to culpable mental states, with a diminished role for mere causal responsibility. As a consequence, accidental harms are often exculpated by older children and adults. Thus, accidental harms transition from being strongly condemned to forgiven. (It is noteworthy that among older adults this trend reverses, and their judgments begin to show more child-like sensitivity to accidental outcomes, in contrast to the intent-based judgments of younger adults (Margoni et al., 2018)).

The outcome-to-intent shift also influences cases of attempted harm, in which the agent's intent is negative but no bad outcome obtains, though to a lesser extent (Baird & Astington, 2004; Costanzo et al., 1973; Cushman et al., 2013; Nobes et al., 2009; Zelazo et al., 1996). For instance, one study comparing both cases found that judgments of accidental harm shifted between 4 and 8 years of age around 40%, whereas judgments of attempted harms shifted around 20% (Cushman et al., 2013).

Second, recent evidence suggests that this shift does not occur equivalently for all types of moral judgments[1]. In particular, judgments of how "naughty" a harmdoer behaved are sensitive to mental state information earlier than are judgments of how much "punishment" the harmdoer should receive (Cushman et al., 2013). For instance, between 4 and 5 years of age, children shift from judging accidental harms as naughty on average (>50% of individuals), to judging them as not naughty on average. An

[1] Throughout the manuscript we use "moral judgment" to refer to the superordinate category under which judgments of punishment, wrongness, etc. fall.

equivalent shift is not seen in judgments of punishment until 7 years of age (Cushman et al., 2013). A similar pattern of the emergence of punishment relative to judgments of badness was obtained by Margoni & Surian (2017). The preliminary evidence from these two studies mirrors well-established patterns in adulthood, in which judgments of deserved punishment are strikingly more sensitive to accidental outcomes than are other kinds of moral judgment, like judgments of wrongfulness or bad character (Cushman, 2008; Martin & Cushman, 2016). Combined, these results suggest a tight and early-emerging relationship between intentionality and naughtiness or wrongness, with punishment influenced by mental states only later in life.

In sum, the development of moral judgment appears to follow an outcome-to-intent shift, in which outcomes caused carry greater weight early in life and agents' intentions carry greater weight with development. This shift occurs most strongly for cases of accidental harm and for judgments of wrongness. By adulthood, mental states continue to carry special weight in moral judgment (Barrett et al., 2016; Cushman, 2008; Gummerum & Chu, 2014; Koster-Hale, Saxe, Dungan, & Young, 2013; Martin & Cushman, 2015; Young, Camprodon, et al., 2010; Young et al., 2006, 2007; Young & Saxe, 2009a, 2009b, 2011), with judgments of wrongness for harm especially influenced by other's intentions and not outcomes caused (Chakroff et al., 2016; Cushman, 2008; Young & Saxe, 2011).

1.2 Adult judgment under cognitive load

As discussed earlier, an initial study suggests that adults' moral judgments under cognitive load resemble the ordinary moral judgments of young children (Buon, Jacob,

et al., 2013). Under load, as in early childhood, intentions exert less of an influence and outcomes exert an increased influence. However, this study did not investigate the influence of cognitive load on the two critical features of the development of moral judgment discussed above: that it is largest for (1) cases of accidental harm (compared with attempted harm), and (2) judgments of wrongness (compared with punishment).

We begin by describing some details of the original study. The authors presented adults with videos that varied whether a protagonist intended harm or not, and whether they succeeded in causing harm or not, focusing on three cases: intentional harm, accidental harm and coincidental harm (a case in which a victim suffers but this suffering occurs "coincidentally", and is not related in any way, either causally or intentionally, to an actor's behavior). This allowed the authors to investigate the influence of intentions and outcomes independently, by comparing judgments of intentional versus accidental harm (cases differing in intent but not whether harm was caused) and by comparing judgments of accidental versus coincidental harm (cases differing in causal connection to harm but not intent). The fact that the coincidental case included a victim who was harmed but not by the agent in question was particularly important in that both cases involve a victim and no negative intent on the part of the agent, meaning any difference in judgment must be a result of the agent's causal connection to harm in the accidental case. After viewing these videos, participants made a series of socio-moral judgments ("Who is the good guy", "Who is the bad guy", "Who do you want to play with?", "Who do you want to give a gift to?"), which were aggregated into one index. Critically, participants viewed these videos and made these judgments either normally or under cognitive load (the authors used a form

of verbal shadowing, which impairs aspects of working memory and executive function). For those not under cognitive load, judgments mirrored past work: intentional harm was viewed as much worse than accidental harm and accidental harm was viewed as moderately worse than cases in which the protagonists had a benign intent and did not cause harm (though it occurred through other means). Thus, participants cared both about the protagonist's intent and whether they caused harm or were merely present when it occurred. For those under cognitive load, a very different pattern emerged. Now, accidental harm was viewed as much worse than cases involving coincidental harm. Most surprisingly, participants now no longer viewed cases of intentional and accidental harm as different–they viewed them as equally bad. Importantly, the authors also showed that this pattern of results was specific to moral judgment and was not due to an inability to decode information about intentions under load.

Our goal is to extend this paradigm to cover the two key features of the outcome-to-intent shift described above. First, as noted above, the authors included cases of intentional harm, accidental harm, and coincidental harm, but not cases of attempted harm. Cases of attempted harm are a critical comparison case, as they show less of a developmental shift than cases of accidental harm. Thus, their results provide evidence that cognitive load alters judgments of accidental harm, but not that it alters them relatively more than judgments of attempted harm.

Moreover, including cases of attempted harm allows us to further clarify the impact of cognitive load on the processing of intentions versus outcomes. As discussed above, while prior results regarding the impact of cognitive load have been

interpreted as demonstrating that cognitive load reduces the impact of intentions on moral judgment (Buon, Jacob, et al., 2013), an alternative interpretation is that cognitive load increases the impact of outcomes on moral judgment. If judgments of attempted harms are unaffected by load (while judgments of accidental harm remain affected by it), this would indicate that load alters sensitivity to the presence of harmful outcomes, while leaving sensitivity to intent intact (this is because sensitivity to harmful outcomes is required to condemn accidental harms, while harmful intent is absent). Alternatively, if load affects judgments of accidents and attempts equally, this would indicate that load alters sensitivity to intent.

Second, the socio-moral judgments asked of participants in Buon, Jacob, et al. were simple and designed to be able to be used with both adults and children (i.e. “Who is the good guy”, “Who is the bad guy?”, “Who do you want to play with?”, “Who do you want to give a gift to?”). They did not include precise assessments of wrongness, punishment, etc. Thus, whether cognitive load would have a particularly strong influence of judgments of wrongness is unclear. By making both of these changes–including cases of accidental and attempted harm and asking participants to make judgments of wrongness or punishment–we can specifically ask whether cognitive load causes adult moral judgment to revert to a child-like state, “undoing” the process of moral development.

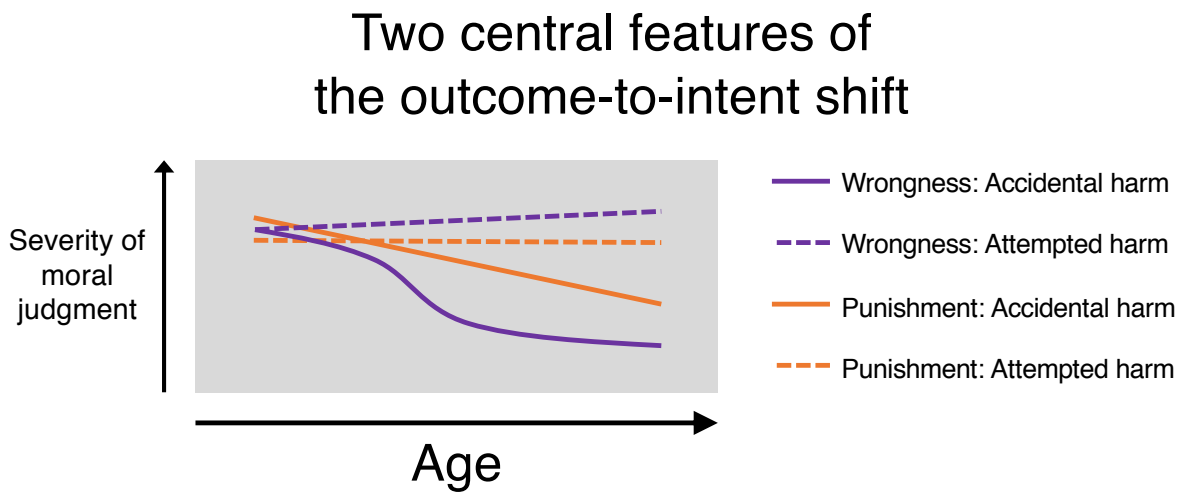

Fig. 1. An illustration of two central features of the outcome-to-intent shift in the development of moral judgment. One feature is that the shift in moral judgment is strongest for cases of accidental harm (solid lines) relative to cases of attempted harm (dashed lines). The other feature is that the shift is strongest for judgments of wrongness (purple lines) relative to judgments of punishment (orange lines).

Effect of cognitive load

We therefore investigated the influence of cognitive load on adult judgments of wrongness and punishment in cases of accidental and attempted harm. Our methods borrow heavily from past work (Buon, Jacob, et al., 2013). See Fig. 2 for a schematic of our procedure. To manipulate the intentional and causal status of agents, we use short cartoon videos depicting accidental harm and intentional harm (employed in Buon, Jacob et al. 2013) as well as new videos depicting attempted harm. All participants saw the video depicting intentional harm and were randomly assigned to additionally view either the video depicting accidental or attempted harm. Thus, participants were randomly assigned to one of two "contrasts". The intentional contrast (comparing accidental and intentional harm) compared two cases that differed only in the agent's intentions. The causal contrast (comparing attempted and intentional harm) compared two cases that differed only in the outcome produced. We manipulated two other factors between-subjects. First, participants were randomly assigned to either judge punishment or to judge wrongness. Second, participants were randomly assigned to either the load condition or the no-load condition. In the no-load condition, participants watched the videos as described below. In the load condition, we implemented a form a cognitive load called verbal shadowing (Buon, Jacob, et al., 2013; Forgeot d'Arc & Ramus, 2011; Hermer-Vazquez et al., 1999; Newton & Villiers, 2007; Orwig, 1979; Peschke et al., 2009; Ratliff & Newcombe, 2008). Participants were presented with recorded sentences and asked to repeat them back concurrently (i.e. to follow along). Once proficient at this task, they then watched the videos and performed the verbal shadowing task simultaneously.

2. Methods

Effect of cognitive load

2.1 Participants

Participants (N = 315) were university students recruited through an online study pool and participated for partial course credit or a small monetary payment. Sample size was determined in advance to include 30 completed participants per cell of our design, or 240 total participants who were able to complete the task without failing the load task (see below). We excluded participants from analysis who reported being not being native speakers of English (*n* = 3), and those who could not reach sufficient performance on the load task or failed to do the task (*n* = 15). Thus, our initial dataset includes 297 participants (87 Male, 210 Female; mean age 19.9 years [SD = 1.32 years]). We recruited 120 participants in the no-load condition, and a total of 177 participants in the load condition, recruiting until we reached 120 who performed the load task without failure (see below), for a final total of 240 participants. All procedures were approved by the Harvard University Committee on the Use of Human Subjects and all participants provided informed consent.

2.2 Stimuli & materials

Stimuli were a combination of short (10 second) Flash videos used in previous research (Buon, Jacob, et al., 2013)as well as additional videos unused previously. All videos depict an interaction between a protagonist and a victim. Protagonists were either a blue character (“Mr. Blue”) or a green character (“Mr. Green”) and the victim was always a red character (“Mr. Red”). Videos were constructed to vary the protagonist’s intentions and the outcome of the situation. Three types of videos were included. In the *intentional* condition, the protagonist sees the victim and proceeds to

harm them. In the *accidental* condition, the protagonist is facing away from the path the victim takes and thus has no knowledge the victim is present. The protagonist then proceeds to harm the victim, while continuing to look away, resulting in an accidental harm. In the *attempt* condition, the protagonist sees the victim and then attempts but fails to harm the victim. In addition, the type of context was manipulated: The protagonist could either swing into the victim, or could hit the victim with a rock. Thus, there were 6 possible combinations of condition and context. Two versions of each combination were constructed, one for the blue protagonist and one for the green protagonist. In total, this yields 12 videos.

2.3 Procedure

All participants were placed in a quiet room and tested individually. Each participant was randomly assigned to one of two contrasts (described below), with the experimenter blinded to this assignment. Within each pair of videos, one had the blue character as the protagonist and the other had the green character as the protagonist. Order of characters and videos was counterbalanced across participants. Context was counterbalanced across the full design, such that each of the 3 harm type conditions and each of the protagonists (blue vs. green) was roughly equally likely to appear in the swinging or throwing context, and was equally likely to be the first vs. second set of videos seen by the participant. Participants were randomly assigned to either the no-load or load condition and to the punishment or wrongness condition. Thus, our experiment employed a 2 (contrast) x 2 (no-load, load) X 2 (punishment, wrongness) fully between-subjects design.

Effect of cognitive load

All participants were seated in front of a computer, with the experimenter seated on the opposite side of the table. Thus, the experimenter had no knowledge of which video the participant saw or which contrast they were assigned to at any point during the experimental session. Participants were given both written and verbal instructions about their task. As the main task proceeded quite similarly in both the no-load and load conditions, we begin by describing the no-load version of the task, and then proceed to describe the additional features of the load version. See Fig. 2 for schematic of our experimental procedure.

Following the procedure used by Buon et al. (2013) participants in the no-load version saw each of the two videos three times total. First, they saw the first video two times in a row, with a 6 second buffer after each video (mirroring the timing of the load condition). After this buffer a beep sounded, which prompted the experimenter to ask the individual judgment question (as the experimenter could not see the computer monitor). This question was asked aloud by the experimenter and depended on judgment condition: Those in the punishment condition were asked “Should he be punished?”, while those in the wrongness condition were asked “Was his behavior wrong?” In both conditions, a screen appeared simultaneous with the beep indicating that the question to be asked by the experimenter was about the protagonist of the video and presenting an image of the protagonist. Participants responded verbally, without having been given instruction on acceptable responses, and responses were recorded by the experimenter. Then, the second video was presented two times in a row, occurring in the same manner as the first video. This ended the first phase. Following their response to the after the 2nd video, the final phase began. Participants

were presented with each video once more, in the same order as during the first phase. Following the presentation of the 2nd video, participants were shown a screen indicating that the question they were about to be asked was about both protagonists, and an image of each protagonist was displayed, side-by-side. Simultaneously, a beep was sounded, which prompted the experimenter to ask the comparative judgment question. As before, this question varied by condition: Those in the punishment condition were asked "Who should be punished?", while those in the wrongness condition were asked "Whose behavior was wrong?" Participants were again not instructed on acceptable answers, and the experimenter recorded their response. Participants were asked to respond using simple, open-ended questions so as to minimize the complexity of making a response (given that responses were made out loud and because half of participants had just been under cognitive load) and since this is the procedure used with prior cognitive load paradigms (Buon, Jacob, et al., 2013).

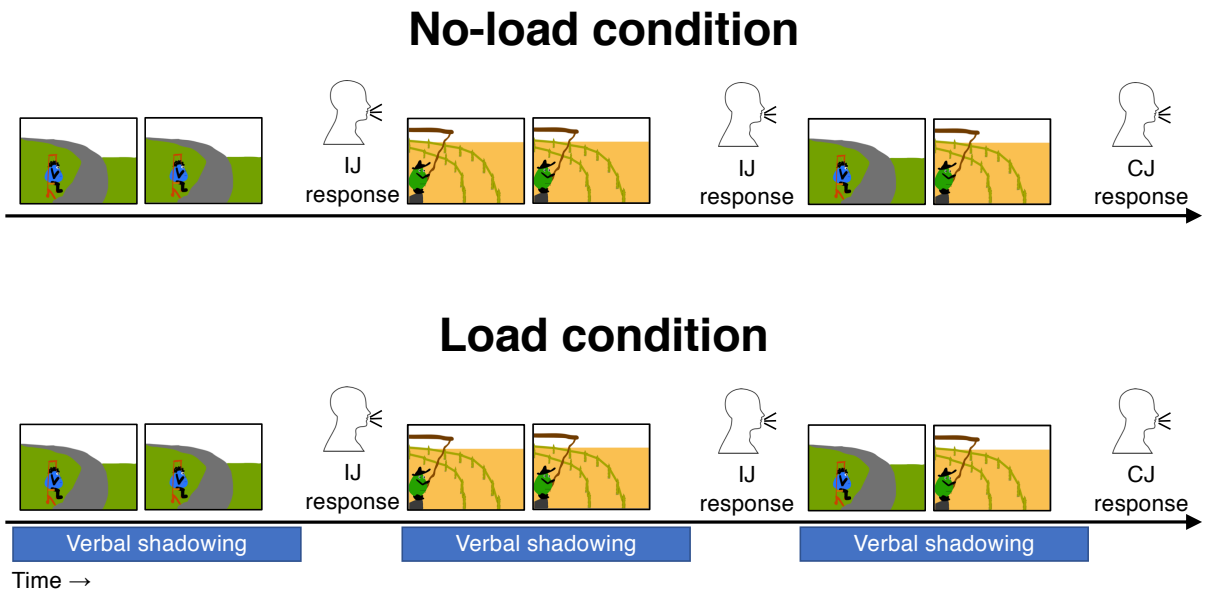

Fig. 2. A schematic of the task. Participants were assigned to either the load or no-load condition. In either condition, participants saw 2 videos a total of 3 times each. They saw the first video twice in a row and then made a moral judgment aloud about that agent (an individual judgment ["IJ"]). They then saw the second video twice in a row and then made a moral judgment aloud about that agent (an individual judgment ["IJ"]). Finally, they saw each video once more and made a final moral judgment comparing the two agents (the comparative judgment ["CJ"]). In the load condition, a verbal shadowing task was also performed while the videos were displayed. Verbal shadowing began 500 ms before the videos began and continued until 7 seconds after the videos ended.

Effect of cognitive load

For participants in the load condition, the first phase was preceded by a training session. During this session, participants were acquainted with the verbal shadowing task, which was designed to tax the cognitive resources available while viewing the stimuli. Participants were instructed to repeat aloud as accurately as possible a series of sentences that were presented over headphones. They were instructed to minimize any pauses or breaks between or during sentences. Sentences were adapted from those used by Buon and colleagues (2013), translated into English. Participants practiced verbal shadowing in sets of five sentences, with breaks between, until two sets of five sentences were completed perfectly. Once the participant achieved desired performance, the practice phase concluded and the familiarization phase began. The first and second phases proceeded as in the no-load condition except that the participant completed the verbal shadowing task while viewing the videos. Specifically, the audio clips began 500 ms before the start of the first video and continued 6 seconds after the conclusion of each video. Immediately upon cessation of the audio clips, the experimenter asked the relevant question (either the familiarization question or the test question) in the same manner as administered in the no-load condition.

In order to investigate the influence of individual variability in ability to perform our cognitive load task on responses in both the no-load and load conditions (discussed further below), following completion of the main task participants in the no-load condition completed the same verbal shadowing training session as participants in the load condition. Participants in the no-load condition then completed a second round of the moral judgment task, identical to the first round of this task with the following exceptions: (1) participants now completed the verbal shadowing version of

the task, identical to participants in the load condition, (2) two novel videos were used and (3) participants made a novel moral judgment, with these latter changes made to avoid familiarity effects. We compare responses from those who fail this subsequent load task with those who pass it. While we find a difference in judgment of cases of attempted harm, this difference is driven by those assessing punishment (see Supplemental Results). Because our main findings regarding the influence of cognitive load are driven by judgments of wrongness in cases of accidental harm, as we discuss below, this difference in punishment of attempted harms, and any associated selection bias, is unlikely to explain our main findings. Given this, and in order to maximize statistical power, we include all participants from the no-load condition in our main analyses.

Upon completion of the test phase, participants filled out a demographic form and were debriefed.

Stimuli, experimental scripts and data can be found here: https://osf.io/5wxck/

3. Results

We focus on two dependent variables: Participants' comparative judgments (assessing which of two agents behaved wrongly or should be punished) and participants' individual judgments of each agent. For the comparative judgments (hereafter "CJ" results), participants' verbal responses were coded on a -1 to 1 scale, with -1 indicating harsher moral judgment of the intentional agent and 1 indicating harsher moral judgment of the agent who committed accidental or attempted harm. Specifically, responses indicating that the intentional agent was either more deserving of punishment or behaved more wrongly were coded as -1. Responses indicating that

either attempted or accidental harm were more deserving of punishment or more wrong were coded as 1. Responses indicating either both agents, neither agent, or indifference (e.g. "I don't know") were coded as 0. Thus, lower scores indicate harsher moral judgment toward the intentional agent (vs. accidental or attempted harmer), higher scores indicate harsher moral judgment toward the accidental or attempted harm and null scores indicate equal judgment toward the two agents.

For the individual judgments (hereafter "IJ" results), responses were coded on a -1 to 1 scale, with 1 indicating that the participant did not think the agent should be punished or behaved wrongly, and -1 indicating that the participant did think the agent should be punished or behaved wrongly. Responses indicating indifference (e.g. "I don't know") were coded as 0. Thus, lower scores indicate harsher moral judgment, and higher scores indicate less harsh moral judgment.

Turning to our main analyses, we begin by focusing on participants' comparative moral judgments in the no-load condition. As a reminder, for these CJ results, lower responses indicate harsher moral judgment (whether punishment or wrongness) of intentional harm. Consistent with past research (Cushman, 2008; Hebble, 1971; Imamoglu, 1975; Wellman et al., 1986; Young et al., 2007), when assessing wrongness, judgments depend almost exclusively on intentions (see Supplemental Fig. 1), with agents causing intentional harm viewed as behaving more wrongly than accidental harm (CJ: -0.97 [0.03]; one sample t-test against 0 $t(29) = -29$, $p < .001$, Cohen's $d = 5.29$, 95% CI = 3.89–6.69)[2] but no more wrongly than agents attempting to harm (CJ: -

[2] Our main inferential statistics come from the regression models for CJ and IJ responses. We provide these comparisons to more fully understand the pattern of results indicated by those models. Because

0.23 [0.11]; one sample t-test against 0 $t(29) = -2.04$, $p = .050$, Cohen's $d = 0.37$, 95 CI = -0.0007–0.74; we note that the significance of this comparison drops to $p = 0.09$ when correcting for multiple comparisons). For punishment, judgments reflect a combination of influence from intentions and outcomes, with intentional harm punished more than both accidental harm (CJ: -0.73 [0.11]; one sample t-test against 0 $t(29) = 6.89$, $p < .001$, Cohen's $d = 1.26$, 95% CI = 0.77–1.73) and attempted harm (CJ: -0.40 [0.11]; one sample t-test against 0 $t(29) = -3.53$, $p = .001$, Cohen's $d = 0.64$, 95% CI = 0.25–1.03).

Turning to participants' individual judgments, we find a similar pattern (see Fig. 3). When assessing wrongness, judgments depend almost exclusively on intentions, with agents causing intentional harm (IJ: -0.63 [0.10]) and agents attempting to cause harm (IJ: -0.53 [0.16]) behaving more wrongly than agents accidentally causing harm (IJ: 0.70 [0.13]; Intentional vs accidental Welch $t(63.9) = -8.18$, $p < .001$, Cohen's $d = -1.06$, 95% CI = -1.33 – -0.78; Attempted vs accidental Welch $t(55.76) = -6.08$, $p < .001$, Cohen's $d = -0.79$, 95% CI = -1.05 - -0.52) but no more wrongly than each other (Welch $t(53.33) = -0.54$, $p = .59$, Cohen's $d = -0.07$, 95% CI = -0.32 – 0.19). For punishment, judgments reflect a combination of influence from intentions and outcomes, with intentional harm (IJ: -0.43 [0.12]) punished more than accidental harm (IJ: 0.07 [0.19]; Welch $t(52.75) = -2.28$, $p = .03$, Cohen's $d = -0.29$, 95% CI = -0.55 – -0.04) and more than attempted harm (IJ: 0.13 [0.18]; Welch $t(53.06) = -2.60$, $p = .01$, Cohen's $d = -0.34$,

there are a large number of these comparisons, we investigated whether the pattern of results for these comparisons changes when correcting for multiple comparisons. Setting the False Discovery Rate to 0.05 and using the Benjamini-Hochberg procedure, the statistical conclusion for all tests remained the same, with one exception, which is noted.

95% CI = -0.59 – 0.08), with no difference between attempted and accidental harm (Welch $t(58.0) = 0.26$, $p = .80$, Cohen's $d = 0.03$, 95% CI = -0.22 – 0.29). Thus, our results from participants not under load match past work on the influence of intentions and outcomes for judgments of punishment and wrongness: For wrongness, judgments are almost exclusively influenced by information about an agent's mental states, whereas punishment decisions are influenced by an agent's mental state as well as the harm caused.

Our main interest, of course, is the potential for differences between normal judgments and those made under load: Do we see a different pattern of results under load? Looking both at comparative and individual judgments, we do. First, we conducted a linear regression predicting CJ responses from contrast (Intentional VS Accidental, Intentional VS Attempted), judgment (wrongness, punishment), load condition (no load, load) and context (swinging, throwing) as well as all interactions between contrast, judgment, load condition and context. A sensitivity analysis conducted using G*power (Faul et al., 2009) with sample size = 240, alpha = 0.05 and 16 predictors indicated that this model had 80% power to detect a minimum f^2 effect size of 0.03, with a critical t of 1.97 and non-centrality parameter of 2.81. We find a significant interaction between contrast, judgment and load (ß = -1.48, SE = 0.71, $t = -2.09$, $p = .04$). We find no other interactions or main effects (all t's > -1.58 or < 1.71, all p's > .08). When assessing punishment, judgments are quite similar with and without cognitive load (see Supplemental Fig. 1). In both cases, they depend on a mix of the agent's mental state and the harm caused, with intentional harm punished more than either attempted harm (Load condition: CJ = -0.47 [0.13]; one sample t-test against 0

t(29) = -3.5, *p* = .002, Cohen's *d* = 0.64, 95% CI = 0.24–1.03; No load condition: CJ = -0.40 [0.11]; one sample t-test against 0 *t*(29) = -3.53, *p* = .001, Cohen's *d* = 0.64, 95% CI = 0.25–1.03) or accidental harm (Load condition: CJ = -0.50 [0.13]; one sample t-test against 0 *t*(29) = -3.75, *p* < .001, Cohen's *d* = 0.68, 95% CI = 0.28–1.08; No load condition: CJ = -0.73 [0.11]; one sample t-test against 0 *t*(29) = -6.89, *p* < .001, Cohen's *d* = 1.26, 95% CI = 0.77–1.73). In contrast, judgments of wrongness were affected by cognitive load. When not under load, the results presented above show that those judgments are principally sensitive to mental states, with little effect of accidental outcomes. When under load, however, agents intentionally causing harm are viewed as behaving more wrongly than either agent attempting to cause harm (CJ = -0.43 [0.13]; one sample t-test against 0 *t*(29) = -3.26, *p* = .003, Cohen's *d* = 0.60, 95% CI = 0.20–0.98) or agents accidentally causing harm (CJ = -0.47 [0.13]; one sample t-test against 0 *t*(29) = -3.5, *p* = .002, Cohen's *d* = 0.64, 95% CI = 0.24–1.03). In other words, cognitive load made adult wrongness judgments look relatively more like those observed in young children, and relatively more like those observed when judging punishment (both among adults and children).

We find similar results when look at the individual judgments. Using glmer in the lme4 package (Bates et al., 2014) in R we run a mixed-effects logistic regression[3] with judgment (punishment, wrongness), load condition (load, no load), harm type (intentional, attempted, accidental) and context (throwing, swinging) as predictors, as well as all possible interactions between judgment, load condition, harm type and

[3] Out of 588 IJ responses, only 4 (0.6%) indicated indifference (i.e. "I don't know"). To make use of the advantages of logistic regression, we drop those responses for this analysis.

outcome type and a random intercept per subject. The pattern of our statistical results is unchanged if the random intercept for subject is removed. A sensitivity analysis conducted in R using the simr package (Green & Macleod, 2016) with 1000 bootstrap samples indicated that this model had 80% power to detect a coefficient for a 3-way interaction (our main result) of at least 3.30. As before, we use a model comparison approach, running a 5000-sample parametric bootstrap Likelihood Ratio Test (LRT), using the pbkrtest package (Halekoh & Højsgaard, 2014)in R. Whereas we do not find that the 4-way interaction improves model fit (LRT $X^2 = 2.76$, $p = .33$), including 3-way interactions does (LRT $X^2 = 16.5$, $p = .04$). Looking further at the model including 3-way interactions, we find a significant 3-way interaction between judgment, load condition and assessment of the accidental agent, relative to the intentional agent (ß = 2.48, SE = 1.17, $z = 2.13$, $p = .03$). Re-leveling the predictor for agent with the accidental agent as the reference level, we also find a significant 3-way interaction between judgment, load condition and assessment of the attempted agent relative to the accidental agent (ß = -2.57, SE = 1.28, $z = -2.01$, $p = .04$). We find no other interactions at this level with load condition. We do, however, find an interaction between judgment, context and assessment of the attempted agent relative to the intentional agent (ß = 3.15, SE = 1.27, $z = 2.48$, $p = .01$). See Table 1 for full model results).

Effect of cognitive load

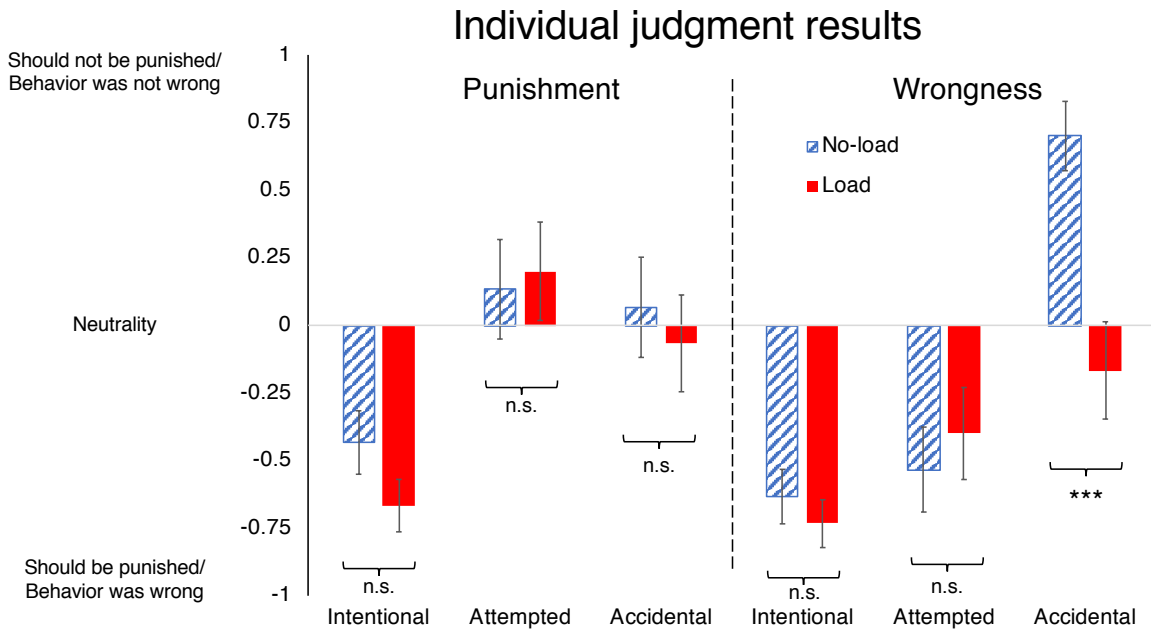

Fig. 3. Individual judgment results. Plotted is average agreement that the agent should be punished or that the agent's behavior was wrong, for both the load and no-load conditions and for each of the three harm types. Lower values indicate consistent agreement that the agent should be punished or behaved wrongly (worse moral evaluation), higher values indicate consistent disagreement that the agent should be punished or behaved wrongly (better moral evaluation) and values close to 0 indicate a lack of agreement. Error bars are SEM. *** indicates $p < 0.001$.

Effect of cognitive load

Table 1. Model results for regression predicting individual judgment (IJ) responses.

Mixed-effects logistic regression results

	Agent ref. level: Intentional	**Agent ref. level: Accidental**
Judgment	0.47 (0.36)	-0.77 (0.46)
Load condition	0.46 (0.36)	1.32 (0.46)**
Context	0.82 (0.36)*	0.65 (0.45)
Agent:Acc.	-1.94 (0.32)***	
Agent:Att.	-1.14 (0.29)***	
Agent:Int. (ref. = Acc.)		1.94 (0.32)***
Agent:Att. (ref. = Acc.)		0.80 (0.32)*
Load cond. X Agent:Acc.	0.86 (0.58)	
Load cond. X Agent:Att.	-0.83 (0.56)	
Load cond. X Agent:Int. (ref. = Acc.)		-0.86 (0.58)
Load cond. X Agent:Att. (ref. = Acc.)		-1.70 (0.63)**
Load cond X Context	-0.57 (0.72)	-1.48 (0.90)
Judgment X Load cond.	-0.45 (0.71)	2.03 (0.92)*
Judgment X Agent:Acc.	-1.25 (0.58)*	
Judgment X Agent:Att.	1.01 (0.57)	
Judgment X Agent:Int. (ref. = Acc.)		1.25 (0.58)*
Judgment X Agent:Att. (ref. = Acc.)		2.26 (0.66)***
Judgment X Context	0.45 (0.72)	-1.35 (0.90)
Agent:Acc. X Context	-0.17 (0.57)	
Agent:Att. X Context	0.16 (0.56)	
Agent:Int. (ref. = Acc.) X Context		0.17 (0.57)
Agent: Att. (ref. = Acc.) X Context		0.33 (0.62)
Judgment X Load cond. X Agent:Acc.	2.48 (1.17)*	
Judgment X Load cond. X Agent:Att.	-0.08 (1.10)	
Judgment X Load cond. X Agent:Int. (ref. = Acc.)		-2.48 (1.17)*
Judgment X Load cond. X Agent:Acc (ref. = Acc.)		-2.57 (1.28)*
Judgment X Load cond. X Context	-0.72 (0.96)	-0.72 (0.96)
Judgment X Agent:Acc. X Context	-1.79 (1.15)	
Judgment X Agent:Att. X Context	1.36 (1.14)	
Judgment X Agent:Int. (ref. = Acc.) X Context		1.79 (1.15)
Judgment X Agent:Att. (ref. = Acc.) X Context		3.15 (1.27)*
Load cond. X Agent:Acc. X Context	-0.91 (1.14)	
Load cond. X Agent:Att. X Context	0.73 (1.11)	
Load cond. X Agent:Int. (ref. = Acc.) X Context		0.91 (1.14)
Load cond. X Agent:Att. (ref. = Acc.) X Context		1.64 (1.25)
Constant	1.55 (0.21)***	-0.39 (0.23)
AIC	544.29	544.29
BIC	640.05	640.05

Effect of cognitive load

Log Likelihood	-249.15	-249.15
Number of observations	475	475
Number of subjects	239	239
Variance (subjects)	0.07	0.07

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Effect of cognitive load

We first probe the significant 3-way interactions between judgment, load condition and the agent being assessed further, by exploring the means for the relevant conditions. As noted above and depicted in Fig. 3, for participants not under cognitive load, we find a stronger influence of intentions for judgments of wrongness than punishment. For participants under load, we find a different pattern. Specifically, whereas judgments of punishment appear roughly unchanged (Intentional load: -0.66 [0.10], comparison with no-load Welch $t(113.9) = -1.53$, $p = .13$, Cohen's $d = -0.2$, 95% CI = -0.45 – 0.06; Accidental load: -0.06 [0.18], comparison with no-load Welch $t(57.93) = -0.52$, $p = .61$, Cohen's $d = -0.07$, 95% CI = -0.32 – 0.19; Attempted load: 0.2 [0.18], comparison with no-load Welch $t(58.0) = 0.26$, $p = 0.80$, Cohen's $d = 0.03$, 95% CI = -0.22 – 0.29; Intentional load vs. accidental load Welch $t(46.58) = -2.95$, $p = .005$, Cohen's $d = -0.38$, 95% CI = -0.62 – -0.12; Intentional load vs. attempted load Welch $t(46.01) = -4.20$, $p < .001$, Cohen's $d = -0.54$, 95% CI = -0.8 – -0.28; Accidental load vs. attempted load Welch $t(58.0) = 1.04$, $p = .30$, Cohen's $d = 0.13$, 95% CI = -0.12 – 0.39), judgments of wrongness become more harsh specifically for accidental harms (Intentional load: -0.73 [0.09], comparison with no-load Welch $t(116.1) = -0.75$, $p = .46$, Cohen's $d = -0.1$, 95% CI = -0.35 – 0.16; Accidental load: -0.16 [0.18], comparison with no-load Welch $t(52.41) = -3.92$, $p < .001$, Cohen's $d = -0.51$, 95% CI = -0.76 – -0.25; Attempted load: -0.40 [0.17], comparison with no-load Welch $t(57.63) = 0.58$, $p = .58$, Cohen's $d = 0.07$, 95% CI = -0.18 – 0.33). While we continue to find a difference in judgment of agents intentionally causing harm versus agents accidentally causing harm when under load (Welch $t(43.48) = -2.83$, $p = .007$, Cohen's $d = -0.36$, 95% CI = -0.62 – -0.11), we now no longer observe a difference between agents attempting to harm and

agents who accidentally harm (Welch $t(57.82) = -0.94$, $p = .35$, Cohen's $d = -0.12$, 95% CI = -0.38 – 0.13). And we continue to find no difference between agents causing intentional harm and those attempting to harm (Welch $t(45.18) = -1.74$, $p = .09$, Cohen's $d = -0.22$, 95% CI = -0.48 – 0.03). In sum, the 3-way interaction we observe between judgment, load condition and agent assessed is driven by the fact that agents causing accidental harm are viewed as more wrong under cognitive load. We observe no similar shift for intentional or, importantly, attempted harms.

We next probe the 3-way interaction between judgment, context and assessing an accidental vs. attempted harm (see Supplemental Fig. 2). As a reminder, our stimuli include two scenario contexts: one in which the protagonist could swing into the victim and one in which the protagonist could hit the victim with a rock. When stimuli were presented in the swinging context, we observe a similar pattern of no difference between accidental harm and attempted harm for both punishment judgments (Attempted harm: 0.19 [0.18]; Accidental harm: 0.30 [0.17]; Welch $t(59.98) = -0.45$, $p = .65$, Cohen's $d = -0.06$, 95% CI = -0.32 – 0.2) and for wrongness judgments (Attempted harm: -0.11 [0.19]; Accidental harm: 0.21 [0.18]; Welch $t(53.38) = -1.19$, $p = .24$, Cohen's $d = -0.16$, 95% CI = -0.41 – 0.1), as well as no difference between punishment and wrongness judgments (Punishment vs wrongness, accidental harm: $t(56.80) = 0.37$, $p = .71$, Cohen's $d = 0.05$, 95% CI = -0.21 – 0.3; Punishment vs. wrongness, attempted harm: Welch $t(55.06) = 1.14$, $p = .26$, Cohen's $d = 0.15$, 95% CI = -0.11 – 0.4). A different pattern is found in the throwing case. Here, we find no difference between punishment of attempted versus accidental harm, though the effect is marginal (Attempted harm: 0.14 [0.19]; Accidental harm: -0.30 [0.17]; Welch $t(55.11) =$

1.72, p = .09, Cohen's d = 0.22, 95% CI = -0.03 – 0.48). We do find that agents attempting harm are viewed as more wrong than agents accidentally causing harm (Attempted harm: -0.75 [0.12]; Accidental harm: 0.32 [0.17]; Welch $t(53.45) = 5.26$, $p <$ .001, Cohen's d = 0.68, 95% CI = 0.42 – 0.94). And, we find that punishment and wrongness judgments are different for both attempted harm (Welch $t(45.31) = 4.04$, $p <$ .001, Cohen's d = 0.52, 95% CI = 0.26 – 0.78) and accidental harm (Welch $t(58.90) = -2.56$, p = .01, Cohen's d = -0.33, 95% CI = -0.59 – -0.07), though in opposite ways: Whereas accidental harm in the context of throwing is punished more than it is viewed wrong, attempted harm is viewed as more wrong than it is punished. We note that this pattern has no interaction with our load manipulation, and thus does not change our overall interpretation of those results.

4. General discussion

Moral development follows an "outcome-to-intent" shift, in which children's judgments focus less on outcomes caused and more on agent's intentions with age (Armsby, 1971; Baird & Astington, 2004; Costanzo et al., 1973; Cushman et al., 2013; Imamoglu, 1975; Killen et al., 2011; Piaget, 1932; Shultz et al., 1986; Yuill & Perner, 1988; Zelazo et al., 1996; cf Nobes et al., 2016, 2009). But this shift is rather narrow in focus. It occurs especially for judgments of accidental harm (in which these two factors are pitted against each other; Cushman et al., 2013; Killen et al., 2011; Piaget, 1932; Shultz et al., 1986; Zelazo et al., 1996), but less so for judgments of attempted harm. And, it occurs especially for judgments of wrongness, but less so for judgments of punishment (Cushman et al., 2013). Complementary work demonstrates that cognitive

load returns adult moral judgment to a more child-like pattern of response: a greater focus on causation than mental state representations (Buon, Jacob, et al., 2013). However, this prior work did not test whether this effect was selective to judgments of accidental harm, or whether it was selective for wrongness judgments. Here, we provide such a test.

Replicating this prior research (Buon, Jacob, et al., 2013), we find that cognitive load elicits a pattern of moral judgment that is relatively more focused on causation than intentions. We found, however, that the effect of cognitive load is quite specific. It causes people to judge accidental harms more harshly (i.e., cases involving causation, but no malicious intent), but it does not have much effect on their judgments of attempted harms (i.e., cases involving malicious intent, but no causation). Furthermore, this effect is specific to judgments of the moral wrongness of action; it does not arise for judgments of the punishment deserved by the actor. Instead, punishment judgments were mostly unchanged by cognitive load in our study. The correspondence between our results on adult moral judgment under cognitive load and the development of moral judgment in children has several important implications for how we understand the mechanisms underlying moral judgment, their reliance on executive function, and their typical developmental course.

4.1 Implications for theories of moral judgment

Our results help to refine and relate several broad families of hypotheses that account for the psychological basis of intent-based moral judgment and, especially, its development. Our goal is not to perfectly capture every detail of any specific theory,

but rather to draw attention to a set of connected claims that are shared across many prominent examples. As we describe, our work speaks for some of the claims and against others.

4.1.1 Resource-dependent theory of mind

According to one broad family of hypotheses, the outcome-to-intent shift in moral judgment is best explained by development of theory of mind and the capacity to integrate it into moral judgment (Baird & Astington, 2004; Buon et al., 2016; Killen et al., 2011). In other words, it emerges because the child has an increasing ability to represent and reason about the mental states of others, including their benign or harmful intent. A natural extension of this hypothesis posits that representing and reasoning about mental states requires executive function (Buon, Jacob, et al., 2013; Carlson & Moses, 2001; Killen et al., 2011). Presumably, then, introducing cognitive load could "reverse" the developmental attainment of intent-based moral judgment.

This family of hypotheses runs afoul of evidence that, in fact, young children can render negative moral judgments of those who attempt, but fail, to do harm (Costanzo et al., 1973; Cushman et al., 2013; Nobes et al., 2009; Zelazo et al., 1996), and even infants exhibit preferences for agents based not only on the outcomes they cause but also on their mental states (Hamlin, 2013; Woo et al., 2017). Additionally, this approach fails to explain the distinct developmental trajectories for the emergence of intent-based judgments of "naughtiness" versus "punishment" (Cushman et al., 2013). There is no straightforward explanation for why the kind of moral judgment a child is asked to

render would influence their capacity to represent and reason about others' mental states.

For the same reasons, this approach offers no simple account for the pattern we report here, which largely tracks previous findings on the development of intent-based moral judgment. We find that cognitive load influences the judgment of accidents, but not attempts, and only for judgments of wrongness, but not for punishment. Neither of these asymmetries falls naturally out of an account of resource-dependent mental state representation or reasoning.

4.1.2 Resource-dependent conflict resolution

According to a second broad family of theories, the outcome-to-intent shift in moral judgment is best explained by development of the cognitive control necessary to override a prepotent response to condemn actions that cause harm based on a representation of intent (Buon et al., 2016; Margoni, Guglielmetti, et al., 2019; Margoni & Surian, 2016). Thus, the key developmental attainment is not in theory of mind, but instead in executive function. For instance, the ETIC model (Buon et al., 2016) highlights the role of cognitive resources when deploying theory of mind capacities in service of moral judgment but also when inhibiting the automatic negative evaluations arising from the perception of someone harming someone else (see also Margoni et al., 2019).

This approach has several merits. First, it easily accommodates the finding that even very young children are able to make intent-based moral judgments in some contexts. This should occur when there is no conflict between causal- and intent-

based moral judgments—for instance, in cases of attempted harm (Costanzo et al., 1973; Cushman et al., 2013; Nobes et al., 2009; Zelazo et al., 1996). Rather, it is in cases like accidental harms, in which causal and intent-based moral judgments conflict, that this hypothesis predicts the largest developmental change (but see Woo et al., 2017). Young children's judgments will be dominated by the prepotent condemnation of harm, while older children's judgments will increasingly exculpate harmdoers based on their innocent intentions.

For the same reason, this approach provides a clear explanation for our finding that, in adults, cognitive load increases condemnation of accidental harmdoing but not of attempted harmdoing.

Along the same lines, this hypothesis can accommodate distinctive patterns of moral judgment by those with autism. These individuals often show increased condemnation of accidental harms (Buon, Dupoux, et al., 2013; Moran et al., 2011), and recent work has suggested that this may not be due to deficits in the representational ability to attribute intentions but rather limitations in cognitive resources (Margoni, Guglielmetti, et al., 2019). Our results on state-level reductions in cognitive resources accord well with this prediction, as does other work on trait-level reductions associated with those with autism (Buon, Dupoux, et al., 2013; Moran et al., 2011) or through the course of normal aging (Margoni et al., 2018; Margoni, Geipel, et al., 2019).

What this approach does not easily explain, however, is why the selective effect of cognitive load on judgments of accidental harm (but not attempted harm) is specific to judgments of wrongness, but not judgments of deserved punishment. Similar to

resource-dependent theory of mind hypotheses, this asymmetry does not fall naturally out of an account of resource-dependent conflict resolution.

4.1.2 Theory attainment

A third family of hypotheses attributes intent-based moral judgment to the development of a theory of moral responsibility and wrongness attained in the preschool years (Cushman, 2008; Cushman et al., 2013; Kohlberg, 1969; Piaget, 1932). We therefore refer to this as the "theory attainment hypothesis".

According to one variety of theory attainment hypothesis, moral judgment is initially characterized by two relatively automatic processes: one process detects causal responsibility for harmful outcomes (a "bad outcome detector"), while another process detects certain categories of morally proscribed action (a "bad acts" detector). Both of these processes emerge early in development and operate in a relatively automatic fashion in adulthood. Then, around 5-7 years of age, children acquire a novel theory of moral wrongness (Cushman et al., 2013). This theory focuses on the moral status of acts, and is principally concerned with the mental states that caused the act. If the act was performed with the belief or intention that it would cause harm, the act was wrong; otherwise, the act was permissible. This theory of wrongness is the principle means by which others' mental states bear on moral judgment, both in late-childhood and throughout adulthood. This theory plays some role in constraining punishment judgments, but incompletely. As a result, adult's judgments of moral wrongness show very little influence of the detector-type systems (and thus of causal responsibility), whereas adult's judgments of deserved punishment still retain a

substantial influence of mere causal responsibility (Cushman, 2008; Martin & Cushman, 2016). In sum, according to this hypothesis, early moral judgment (<5 years of age) focuses less on intentions not because these children lack a resource-dependent capacity, but rather because the concept of wrongness has not yet emerged.

At the most basic level, if cognitive load returns adult moral judgment to a more child-like state, this theory attainment hypothesis suggests that cognitive load should undo the changes associated with the emergence of a concept of wrongness. From this perspective, two features of this theory attainment hypothesis which depend upon the concept of wrongness are critical for present purposes. First, this hypothesis centers moral development on changes in judgment of accidental harm and not judgment of attempted harm. The early moral theory of the child condemns accidental harms because the "bad outcome detector" delivers a negative moral verdict, while the only system capable of responding to mental states—the "bad act detector"—is simply silent. In contrast, the late moral theory of the child–which includes the concept of wrongness–can override the condemnation of accidental harm by the "bad outcome detector" because it identifies moral wrongness with malicious intent. In the absence of such intent it is not silent; rather, it exculpates. There is no similar shift of judgments of attempts: the agent is condemned due to the operation of the "bad act detector" early in life and due to the identification of malicious intent through the concept of wrongness later in life.

Second, this hypothesis centers moral development principally on changes in wrongness judgments (especially for accidents) and only later through changes in punishment judgments. This is because the exculpation of accidents is driven by a

theory of *morally wrong acts*, which subsequently places an incomplete constraint on judgments of punishment. In childhood, the emergence of this theory immediately impacts judgments of wrongness (as it is a theory of wrong acts). However, this theory of wrong acts only begins constraining punishment judgments with development, explaining the developmental asymmetry in judgments of accidents for wrongness and punishment. And in adults, judgments of wrongness mostly reflect the later-emerging theory, while adult judgments of punishment largely continue to reflect the early-emerging "detector"-based system.

Thus, the theory attainment hypothesis can accommodate the quite specific pattern of judgments rendered under cognitive load in our study, just as it explains the corresponding pattern of developmental change in young children. This implies that, whatever conceptual reorganization occurs in middle childhood (e.g., through the emergence of a concept of wrongness), this reorganization does not fully replace the early-emerging processes that automatically assign punishment on the basis of causal responsibility and negative character attributions on the basis of malicious intent. Rather, these automatic processes remain intact into adulthood and reassert themselves under the presence of severe cognitive load. Future research should explore this possibility in greater detail.

The theory attainment hypothesis faces three limitations, however, which must be addressed in future research. First, although it posits a developmentally early-emerging "bad act" detector (Cushman et al., 2013) that can account for some evidence for intent-based moral judgment in infants and toddlers (e.g., Hamlin, 2013; Margoni & Surian, 2020), this element of the theory remains underdeveloped. A

particular challenge is to explain what features of an agent's actions or mental states this detector responds to, and how these are differentiated from the more complete theory of intent-based moral wrongness that emerges between 4 and 6 years old. Second, as originally formulated, the theory does not depend on cognitive control or executive function, but instead upon a conceptual attainment. Thus, in order to explain the present results, it is necessary to propose that this "theory" of moral wrongness is inhibited under cognitive load. Finally, some key aspects of the basic developmental trends remain inconsistent across studies. For instance, while several studies find that the outcome-to-intent shift in the judgment of harmdoing occurs principally for judgments of accidents, but not for attempts (Costanzo et al., 1973; Cushman et al., 2013; Nobes et al., 2009; Zelazo et al., 1996), others find the opposite pattern (Margoni & Surian, 2017). And, while two studies show that this shift is more pronounced for judgments of badness/naughtiness than for judgments of punishment (Cushman et al., 2013; Margoni & Surian, 2017), these studies provide conflicting evidence on whether the former type of judgment constrains the latter. Whether the young child acquires a new concept of moral wrongness in the preschool years, broadly consistent with the seminal theories of Piaget (1932) and Kohlberg (1969), remains an important area for further study.

4.2 Limitations and future directions

We present important evidence that cognitive load has a specific influence on moral judgment, but more work exploring the relationship between cognitive resources and moral judgment is needed. For instance, our manipulation of cognitive load,

borrowed from prior research (Buon, Jacob, et al., 2013), was a powerful one and occupied attention, working memory, language processing as well as cognitive inhibition, among other capacities. Future research should determine which elements of this broad manipulation are necessary.

Relatedly, only limited work has begun exploring how moral judgment may be changed through the process of normal aging and the cognitive changes associated with it. This work suggests that causal processes, which are down-regulated through development in childhood, may reassert themselves in older adulthood. Two recent studies have indeed found that older adults (>63 years of age) condemn accidental harms to a greater degree than younger adults (Margoni et al., 2018; Margoni, Geipel, et al., 2019). Moreover, in one of these studies, the effect of age was larger for judgments of wrongness than for judgment of punishment (Margoni, Geipel, et al., 2019). Notably, no mediation by working memory (Margoni, Geipel, et al., 2019) or cognitive flexibility (Margoni et al., 2018) was detected, as might have been predicted given our current results. Executive function is a varied construct (Miyake et al., 2000), however, and possibly the inhibition of information about intentions on wrongness judgments posited in our theory attainment hypothesis occurs less through working memory or cognitive flexibility and more through other subtypes of executive function, such as cognitive inhibition. Regardless, the general pattern of increased condemnation of accidents in older adults is consistent with the idea that latent automatic processes begin to carry more weight in the process of normal aging, similar to their increased role early in development. It remains an important topic for future

research to investigate other signatures of these processes in adults, which might also include reaction time effects or re-emergence due to cognitive impairments.

It also remains to be seen whether the theory attainment hypothesis can explain atypical patterns of moral judgment in individuals with autism. Recent studies find that individuals with autism show an abnormal (and selective) increased condemnation of accidental harm when judging permissibility (Moran et al., 2011) and punishment (Buon, Dupoux, et al., 2013) relative to controls. Recent work has added to these findings by showing that young children with autism judge cases of accidental harm no different from controls when the processing demands of the task are reduced (Margoni, Guglielmetti, et al., 2019), suggesting that differences in judgment in those with autism may stem from capacity limitations. Further research should determine if this pattern is more pronounced for judgment of wrongness compared to judgments of punishment as the theory attainment hypothesis would predict.

In addition, future work should explore the process by which intentions transition from being recognized and tracked by infants and children to being incorporated into moral judgment. The results of Buon and colleagues (2013) demonstrated that adults under cognitive load could identify and process whether an agent intended to cause harm, but were not able to incorporate that information into their judgment of that agent. An interesting question is whether such a stage is also present in development; that is, a stage during which children can identify agent's negative or positive intentions but cannot use them in their moral judgments. Some evidence in favor of this possibility is the ability of children in some studies to pass comprehension questions requiring them to correctly report agents' intentions despite not using that information

in their subsequent moral judgments (Cushman et al., 2013; Zelazo et al., 1996). Relatedly, as we discuss above, recent work has shown that even infants make sociomoral evaluations of agents based upon their intentions (Hamlin, 2013; Woo et al., 2017). Future work should explore the processes that mediate this transition from a stage in which other's intentions are correctly identified but not incorporated into moral judgment to a stage in which intentions do influence moral judgment.

Finally, our stimuli included only harm violations and not other kinds of moral wrongs, including violations of norms about purity, loyalty or other concerns (Graham et al., 2009). Recent work suggests that the influence of intentions on moral judgment is strongest for harm violations and less strong for other kinds of violations, including purity violations (Chakroff et al., 2016; Chakroff, Dungan, & Young, 2013; Chakroff, Russell, Piazza, & Young, 2017; Chakroff & Young, 2015; Dungan, Chakroff, & Young, 2017; cf. Giner-Sorolla & Chapman, 2017). From this perspective, harm violations are an ideal set of stimuli with which to test our hypotheses. Nevertheless, future work should explore whether cognitive load has a similar influence on other kinds of violations, including purity violations.

5. Conclusion

We find that cognitive load produces a pattern of moral judgment strikingly similar to that observed in early childhood: People are more likely to judge accidental harms as morally wrong when under load. They are not, however, less likely to condemn attempted harms, nor are their judgments of deserved punishment affected by

cognitive load. Taken together with past research, these results support a two-process model of moral judgment.

References

Armsby, R. (1971). A reexamination of the development of moral judgments in children. *Child Development*, *42*(4), 1241–1248. http://www.ncbi.nlm.nih.gov/pubmed/5157112

Baird, J. A., & Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. *New Directions for Child and Adolescent Development*, *2004*(103), 37–49. https://doi.org/10.1002/cd.96

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M. T., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M., Kushnick, G., Pisor, A., Scelza, B. A., Stich, S., von Rueden, C., Zhao, W., & Laurence, S. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, *113*(17), 4688–4693. https://doi.org/10.1073/pnas.1522070113

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7*.

Buon, M., Dupoux, E., Jacob, P., Chaste, P., Leboyer, M., & Zalla, T. (2013). The Role of Causal and Intentional Judgments in Moral Reasoning in Individuals with High Functioning Autism. *Journal of Autism and Developmental Disorders*, *43*(2), 458–470. https://doi.org/10.1007/s10803-012-1588-7

Buon, M., Jacob, P., Loissel, E., & Dupoux, E. (2013). A non-mentalistic cause-based heuristic in human social evaluations. *Cognition*, *126*(2), 149–155. https://doi.org/10.1016/j.cognition.2012.09.006

Buon, M., Seara-Cardoso, A., & Viding, E. (2016). Why (and how) should we study the

interplay between emotional arousal, Theory of Mind, and inhibitory control to understand moral cognition? *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-016-1042-5

Carlson, S. M., & Moses, L. J. (2001). Individual Differences in Inhibitory Control and Children's Theory of Mind. *Child Development*, *72*(4), 1032–1053.

Chakroff, A., Dungan, J., Koster-Hale, J., Brown, A., Saxe, R., & Young, L. (2016). When minds matter for moral judgment: Intent information is neurally encoded for harmful but not impure acts. *Social Cognitive and Affective Neuroscience*, *11*(3), 476–484. https://doi.org/10.1093/scan/nsv131

Chakroff, A., Dungan, J., & Young, L. (2013). Harming Ourselves and Defiling Others: What Determines a Moral Domain? *PLoS ONE*, *8*(9), e74434. https://doi.org/10.1371/journal.pone.0074434

Chakroff, A., Russell, P. S., Piazza, J., & Young, L. (2017). From impure to harmful: Asymmetric expectations about immoral agents. *Journal of Experimental Social Psychology*, *69*, 201–209. https://doi.org/10.1016/j.jesp.2016.08.001

Chakroff, A., & Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition*, *136*, 30–37. https://doi.org/10.1016/j.cognition.2014.11.034

Costanzo, P. R., Coie, J. D., Grumet, J. F., & Farnill, D. (1973). A reexamination of the effects of intent and consequence on children's moral judgments. *Child Development*, *44*(1), 154–161. https://doi.org/10.1111/j.1467-8624.1973.tb02127.x

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

Effect of cognitive load https://doi.org/10.1016/j.cognition.2008.03.006

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21. https://doi.org/10.1016/j.cognition.2012.11.008

Dungan, J. A., Chakroff, A., & Young, L. (2017). The relevance of moral norms in distinct relational contexts: Purity versus harm norms regulate self-directed actions. *PLoS ONE*, *12*(3), 1–16. https://doi.org/10.1371/journal.pone.0173405

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Forgeot d'Arc, B., & Ramus, F. (2011). Belief attribution despite verbal interference. *Quarterly Journal of Experimental Psychology (2006)*, *64*(5), 975–990. https://doi.org/10.1080/17470218.2010.524413

Giner-Sorolla, R., & Chapman, H. A. (2017). Beyond Purity. *Psychological Science*, *28*(1), 80–91. https://doi.org/10.1177/0956797616673193

Graham, J., Haidt, J., & Nosek, B. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Green, P., & Macleod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*,

7(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *The Neuroscience of morality: Emotion, brain disorders, and development* (pp. 35–80). MIT Press. http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2007-14534-005&site=ehost-live&scope=cite

Gummerum, M., & Chu, M. T. (2014). Outcomes and intentions in children's, adolescents', and adults' second- and third-party punishment behavior. *Cognition*, *133*(1), 97–103. https://doi.org/10.1016/j.cognition.2014.06.001

Halekoh, U., & Højsgaard, S. (2014). *A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed Models —The R package pbkrtest* (pp. 1–32). Journal of Statistical Software. https://doi.org/10.18637/jss.v059.i09

Hamlin, J. K. (2013). Failed attempts to help and harm: intention versus outcome in preverbal infants' social evaluations. *Cognition*, *128*(3), 451–474. https://doi.org/10.1016/j.cognition.2013.04.004

Hamlin, J. K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not like me = bad: infants prefer those who harm dissimilar others. *Psychological Science*, *24*(4), 589–594. https://doi.org/10.1177/0956797612457785

Hamlin, J. K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. *Developmental Science*, *16*(2), 209–226. https://doi.org/10.1111/desc.12017

Hebble, P. W. (1971). The Development of Elementary School Children's Judgment of

Intent. *Child Development*, *42*(4), 1203. https://doi.org/10.2307/1127804

Hermer-Vazquez, L., Spelke, E. S., & Katsnelson, a S. (1999). Sources of flexibility in human cognition: dual-task studies of space and language. *Cognitive Psychology*, *39*(1), 3–36. https://doi.org/10.1006/cogp.1998.0713

Imamoglu, E. O. (1975). Children's Awareness and Usage of Intention Cues. *Child Development*, *46*(1), 39–45. https://doi.org/10.2307/1128831

Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition*, *119*(2), 197–215. https://doi.org/10.1016/j.cognition.2011.01.006

Kohlberg, L. (1969). *Stage And Squence: The Cognitive-Developmental Approach To Socialization*. 347–480.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences*, *110*(14), 5648–5653. https://doi.org/10.1073/pnas.1207992110

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

Margoni, F., Geipel, J., Hadjichristidis, C., & Surian, L. (2018). Moral judgment in old age: Evidence for an intent-to-outcome shift. In *Experimental Psychology* (Vol. 65, Issue 2). https://doi.org/10.1027/1618-3169/a000395

Margoni, F., Geipel, J., Hadjichristidis, C., & Surian, L. (2019). The influence of agents' negligence in shaping younger and older adults' moral judgment. *Cognitive Development*, *49*(January), 116–126. https://doi.org/10.1016/j.cogdev.2018.12.002

Margoni, F., Guglielmetti, G., & Surian, L. (2019). Brief Report: Young Children with

Autism Can Generate Intent-Based Moral Judgments. *Journal of Autism and Developmental Disorders*, *49*(12), 5078–5085. https://doi.org/10.1007/s10803-019-04212-9

Margoni, F., & Surian, L. (2016). Explaining the U-Shaped Development of Intent-Based Moral Judgments. *Frontiers in Psychology*, *7*(February), 1–6. https://doi.org/10.3389/fpsyg.2016.00219

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

Margoni, F., & Surian, L. (2020). Conceptual continuity in the development of intent-based moral judgment. *Journal of Experimental Child Psychology*, *194*, 104812. https://doi.org/10.1016/j.jecp.2020.104812

Martin, J. W., & Cushman, F. (2015). To Punish or to Leave: Distinct Cognitive Processes Underlie Partner Control and Partner Choice Behaviors. *PLOS ONE*, *10*(4), e0125193. https://doi.org/10.1371/journal.pone.0125193

Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition*, *147*, 133–143. https://doi.org/10.1016/j.cognition.2015.11.008

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Moran, J. M., Young, L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism.

Proceedings of the National Academy of Sciences, *108*(7), 2688–2692. https://doi.org/10.1073/pnas.1011734108

Mulvey, K. L., Gönültaş, S., & Richardson, C. B. (2020). Who Is to Blame? Children's and Adults' Moral Judgments Regarding Victim and Transgressor Negligence. *Cognitive Science*, *44*(4). https://doi.org/10.1111/cogs.12833

Newton, A., & Villiers, J. de. (2007). Thinking While Talking: Adults Fail Nonverbal False-Belief Reasoning. *Psychological Science*, *18*(7), 574–579. https://doi.org/10.1111/j.1467-9280.2007.01942.x

Nobes, G., Panagiotaki, G., & Bartholomew, K. J. (2016). The influence of intention, outcome and question-wording on children's and adults' moral judgments. *Cognition*, *157*(December), 190–204. https://doi.org/10.1016/j.cognition.2016.08.019

Nobes, G., Panagiotaki, G., & Pawson, C. (2009). The influence of negligence, intention, and outcome on children's moral judgments. *Journal of Experimental Child Psychology*, *104*(4), 382–397. https://doi.org/10.1016/j.jecp.2009.08.001

Orwig, G. (1979). Effects of Verbal Shadowing on the Recognition of Visually Presented Verbal and Nonverbal Information. *ECTJ*, *27*(1), 25–30. http://link.springer.com/article/10.1007/BF02765312

Peschke, C., Ziegler, W., Kappes, J., & Baumgaertner, A. (2009). Auditory–motor integration during fast repetition: The neuronal correlates of shadowing. *Neuroimage*, *47*(1), 392–402. https://doi.org/10.1016/j.neuroimage.2009.03.061

Piaget, J. (1932). *The moral judgment of the child*. Kegan Paul. http://library.brown.edu/find/Record/b1155458

Ratliff, K. R., & Newcombe, N. S. (2008). Is language necessary for human spatial reorientation? Reconsidering evidence from dual task paradigms. *Cognitive Psychology*, *56*(2), 142–163. https://doi.org/10.1016/j.cogpsych.2007.06.002

Shultz, T., Wright, K., & Schleifer, M. (1986). Assignment of Moral Responsibility and Punishment. *Child Development*, *392*(X), 177–184. http://www.jstor.org/stable/10.2307/1130649

Sloman, S. (2014). *Two Systems of Reasoning, An Update*. 69–79.

Sommerville, J. A., Schmidt, M. F. H., Yun, J. E., & Burns, M. (2013). The Development of Fairness Expectations and Prosocial Behavior in the Second Year of Life. *Infancy*, *18*(1), 40–66. https://doi.org/10.1111/j.1532-7078.2012.00129.x

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *The Behavioral and Brain Sciences*, *23*(5), 645–665; discussion 665-726. http://www.ncbi.nlm.nih.gov/pubmed/11301544

Trémolière, B., & Djeriouat, H. (2016). The sadistic trait predicts minimization of intention and causal responsibility in moral judgment. *Cognition*, *146*, 158–171. https://doi.org/10.1016/j.cognition.2015.09.014

Wellman, H. M., Cross, D., & Bartsch, K. (1986). Infant Search and Object Permanence: A Meta-Analysis of the A-Not-B Error. *Monographs of the Society for Research in Child Development*, *51*(3), 1–67.

Woo, B. M., Steckler, C. M., Le, D. T., & Hamlin, J. K. (2017). Social evaluation of intentional, truly accidental, and negligently accidental helpers and harmers by 10-month-old infants. *Cognition*, *168*, 154–163. https://doi.org/10.1016/j.cognition.2017.06.029

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, *107*(15), 6753–6758. https://doi.org/10.1073/pnas.0914826107

Young, L., Cushman, F., Adolphs, R., Traniel, D., & Hauser, M. (2006). Does emotion mediate the relationship between an action's moral status and its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture*, *6*(1), 1–2. https://doi.org/10.1163/156853706776931312

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235–8240. https://doi.org/10.1073/pnas.0701408104

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*, *1*(3), 333–349. https://doi.org/10.1007/s13164-010-0027-y

Young, L., & Saxe, R. (2009a). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, *21*(7), 1396–1405. https://doi.org/10.1162/jocn.2009.21137

Young, L., & Saxe, R. (2009b). Innocent intentions: a correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *47*(10), 2065–2072. https://doi.org/10.1016/j.neuropsychologia.2009.03.020

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent

across moral domains. *Cognition*, *120*(2), 202–214. https://doi.org/10.1016/j.cognition.2011.04.005

Yuill, N., & Perner, J. (1988). Intentionality and knowledge in children's judgments of actor's responsibility and recipient's emotional reaction. *Developmental Psychology*, *24*(3), 358–365. https://doi.org/10.1037/0012-1649.24.3.358

Zelazo, P. D., Helwig, C. C., & Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development*, *67*(5), 2478–2492. https://doi.org/10.2307/1131635