Running to do evil: Costs incurred by perpetrators affect moral judgment

Julian Jara-Ettinger*, Nathaniel Kim*, Paul Muentener, and Laura E. Schulz
(jjara, nlkim, pmuenten, lschulz@mit.edu)
Department of Brain and Cognitive Sciences, MIT
Cambridge, MA 02139 USA
*These authors contributed equally to the paper.

Abstract

Humans evaluate transgressors focusing on their intentions and the outcome. Here we propose that, in addition to these factors, we also take into account the cost and reward of actions, supported by a fundamental inferential process we call a "naïve utility calculus." Because inferences about costs and rewards trade off, observers can infer that agents who incur higher costs place a higher value on acting. This inference has implications for moral judgments. Our account predicts, somewhat paradoxically, that the higher the costs a perpetrator incurs in transgressing, the more harshly observers will judge him. Less paradoxically, the same principle holds for helpful actions: controlling for intention and outcome, more costly helpful actions will be given more credit. Consistent with our framework, we find that adults and preschoolers make graded social evaluations guided by the costs of the actions.

Keywords: Cognitive Development; Naïve Utility Calculus; Rational Action; Social Cognition; Social Evaluations.

Introduction

Arnold and Bob are identical twins who just bought identical cars. Arnold and Bob drove to offices near each other. Arnold locked his car, set the alarm, and put a club on the steering wheel. Bob left his car unlocked with the keys in the ignition. After some hours, they came out to find their cars had been stolen. Police apprehended the thieves: a guy named Joe stole Arnold's car and one named Phil stole Bob's car.

Arnold and Bob were clearly innocent victims; Joe and Phil were clearly guilty thieves. Nonetheless, we might find that we hold Phil slightly less accountable than Joe. Why?

Research on moral reasoning has investigated many factors that affect moral judgment: the agents' in-group or out-group status; whether the event involves direct or indirect harm; the agents' intentions, and the outcomes of the event (See Baillargeon, Scott, He, Sloane, Setoh, Jin, Wu & Bian, *in press*; Greene, 2003; Hamlin, 2013; Knobe, 2010; Mikhail, 2007 for review). Critically however, in this scenario, none of those contrasts is in play. The agents' social status is left unspecified, the actions are direct, the agents apparently act intentionally, and the outcomes are identical. What then accounts for our graded judgments?

If we hold Joe more accountable than Phil we might invoke the biblical caution against those who "run to do evil" (Isaiah: 5:7). However, although recognized in our ethical canons, the costs a transgressor incurs to commit a wrongdoing have rarely been investigated as a factor in psychological studies of moral reasoning.

Here we propose a formal account of this intuition, suggesting that human beings evaluate others' actions with respect to an intuitive theory of how agents assign costs and rewards to the world, how these cost and rewards combine to produce utilities, and how these utilities inform agents' decisions about what actions to take. We will refer to this as a *naïve utility calculus*. Details of this account are inspired by earlier computational models of theory of mind (Baker, Saxe, & Tenenbaum, 2009, 2011; Ullman, Baker, Macindoe, Evans, Goodman, & Tenenbaum, 2010; Jara-Ettinger, Baker, & Tenenbaum, 2012) and have been developed elsewhere, so here we will discuss the inferences supported by the formalization intuitively (See Jara-Ettinger, Tenenbaum, & Schulz, *in prep*, for a detailed version of the theoretical framework, and Jara-Ettinger, Tenenbaum & Schulz, 2013 and Jara-Ettinger, Gweon, Tenenbaum & Schulz, 2014, for developmental evidence) .

At the core of the naïve utility calculus are a few key claims:

1) Observers have a theory of rational action that resembles classical utility theory in three respects:
    a. Actions generate rewards and incur costs. The rewards minus the costs determine the utility of acting.
    b. Both rewards and costs have an external, agent-independent component and an internal, agent-dependent component.
    c. Rational agents act to maximize the highest expected utility.
2) Observers can use known and observable information about agents and the environment to infer the costs and rewards of actions, enabling predictions about unseen features of the environment, unobserved mental states, and others' future behaviors.
3) These abilities emerge early in development, supporting children's ability to reason about agents' goal-directed behavior.

What predictions does this account make for moral reasoning and for our car theft scenario in particular? Understanding how costs and rewards produce utilities, and how utilities guide planning, allows us to partially infer these values from the observable actions. Arnold's car was difficult to steal. It required bypassing the locks, the alarm, the bar, and the ignition. The rewards must have been high

enough to make the overall utility profitable. By contrast, the costs incurred in stealing Bob's car were low; a smaller reward could still result in an overall positive utility. Thus we can be confident that Joe placed a high subjective value on stealing the car because he engaged in costly actions to do so. It is less clear whether Phil placed a high value on stealing the car. Perhaps he would not have done so had the costs been higher. This difference in subjective value might derive from many sources (e.g., perhaps Joe was poorer than Phil). However, in the absence of other information, all we can infer is that Joe had a strong preference for car theft. This has direct implications for moral judgment: we are likely to judge people more harshly to the degree that we believe they are strongly motivated to perform harmful actions.

Our suggestion that a naïve utility calculus underlies moral judgment makes three predictions. First, and perhaps somewhat counter-intuitively, we should hold perpetrators of harmful actions more accountable to the degree that their actions are costly to the perpetrators themselves. We test this prediction in Experiment 1. Second, these inferences should hold for helpful actions as well as harmful ones. (That is, holding intentions and outcomes constant, children should give both more credit and more blame for high cost actions than low cost ones.) Finally, to the degree that a naïve utility calculus underlies moral judgment and is fundamental to our understanding for rational action, we should find evidence for these inferences even in very young children. We test these predictions in Experiment 2.

Experiment 1

In Experiment 1 we look at whether adults take into account the cost of committing a transgression when evaluating an agent. We predict that adults will punish perpetrators more when the perpetrator engages in high cost actions (that is, actions costly to the perpetrator himself) than when the perpetrator engages in low cost actions. This prediction results from a utility calculus in which costly actions license inferences about heightened motivation, in this case, to do harmful acts.

Participants

48 U.S. residents (as determined by their I.P. address) were recruited using Amazon Mechanical Turk. Subjects were randomly assigned to either the long-distance or short-distance condition (24 subjects per condition). Subjects within each condition were assigned to one of three possible theft-value conditions: The low-value theft (stolen iPod), the middle-value theft (stolen iPad), or the high-value theft (stolen Macbook Pro) (8 subjects per condition).

Stimuli

The stimuli consisted of three stories. In all stories a thief stole an object that had been left unattended. Each story had a low-cost version and a high-cost version. In the low-cost versions the thief was very close to the object, thus making the theft low-cost. In the high-cost versions the thief was very far away from the object, making the theft high-cost.

In the first story the owner left the object in a study room. The thief was sitting on an adjacent table in the low-cost version and looking through a window in the high-cost version. In the second story the owner left the object on a park bench. The thief was sitting on the same bench in the low-cost version and looking through a window from the second story of a nearby building in the high-cost version. In the third story the owner left his object on a gym treadmill. The thief was running in the adjacent treadmill in the short-cost version and on the opposite corner of the gym in the high-cost version.

Additionally, we varied the value of the object that was stolen. The lowest value object was an iPod, the middle value object was an iPad, and the highest value object was a Macbook Pro. This generated a total of 18 stories (3 base stories X 2 cost conditions X 3 object value conditions).

Procedure

Participants were randomly assigned to a cost condition (high-cost or low-cost conditions) and to an object condition (low, middle, or high value). This left each participant with three stories to read.

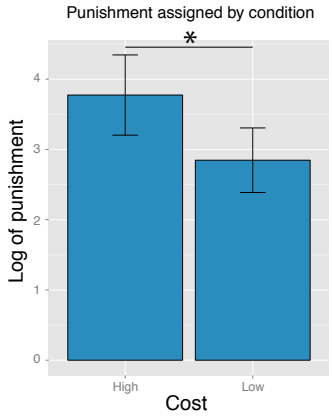

Figure 1: Log punishment given to transgressors across in the high and low cost conditions along with 95% confidence intervals. Thieves in the high cost condition received significantly higher punishments compared to transgressors in the low cost condition. The analysis was performed using the log transformation because, unlike the raw response, it was normally distributed.

Participants were asked to imagine they were jurors, whose job was to decide how long thieves have to spend in a social rehabilitation center. Each participant read the study room, park, and gym stories that corresponded to the condition they were assigned to. Each story contained a control question to ensure participants remembered the thief's starting position. After reading each story participants were asked to decide how many days the thief should spend in a social rehabilitation center.

Results and Discussion

We calculated each participant's average punishment score across the three stories, excluding those where participants failed to remember the thief's initial location. Figure 1 shows the results from the experiment. Our primary question was whether the cost incurred by the thief affected the participant's judgments. Collapsing across object-value conditions, participants assigned 27.19 days of punishment

in the low-cost condition and 123.01 days in the high-cost condition ($p < 0.012$; Welch two sample t-test on log punishment). Next we analyzed differences in punishment across both the cost and object value conditions using a two-way ANOVA. We found both a significant effect of the cost to the thief and the value of the object ($p < 0.01$ and $p < 0.001$ respectively). Participants assigned higher punishments to thieves who stole more valuable objects.

These results suggest that adult's choice of how much to punish is influenced by the cost the transgressor incurred as well as the stolen object's value. Importantly, each individual participant only saw a single cost condition. Although the control question checked participant's memory for the initial location of the thief relative to the object, participants were given no other information that would enable them to infer that the costs to the thief were relevant to the task. Moreover, the information even about the location of the thief with respect to the object was both very general (i.e., actual distances were never specified) and differed greatly within the stories they heard (i.e., the distance across a gym vs. the distance to a nearby building). Nonetheless, consistent with the predictions of a naïve utility calculus, participants appeared to impute costs automatically, resulting in different judgments across conditions.

However, our data also suggest that participant's judgments were not only affected the costs incurred by the transgressor. Participants also gave longer punishments to transgressors who stole more (objectively) valuable objects. One likely explanation is that participants took into account the loss to the victim and imposed greater punishment for greater losses. Extending the implications of the naïve utility calculus over multi-party interactions remains a rich area for future research.

Experiment 2: Children's Cost Perception

Experiment 1 suggests that adults are sensitive to the cost of actions when evaluating transgressors. In Experiment 2 we extend this study to children, and to both positive and negative social evaluations. Using a somewhat simpler within-subject design, we test the predictions that the cost incurred by an agent affects children's judgments of both credit and blame. Because this aspect of moral judgment had never been previously investigated in children, we chose a broad age range for preliminary investigation. No age trends emerged so here we report all children recruited.

Participants

Twenty-two children (mean age (SD): 5.29 years (0.83 years), range 3.63-6.81 years) were recruited and tested at a local children's museum.

Stimuli

Eight storybooks were used. (See Table 1.) Each story came in two versions, always presented in pairs, one in which the costs of the protagonist's actions were high (as indexed by distance traveled) and one in which they were low, for a total of four trials. The two stories in each pair were identical except for the cost the protagonist incurred to achieve the outcome. In two story pairs, the protagonist performed a pro-social action (Bottle and Pencil stories; in two other pairs, the protagonist performed an anti-social action (Cookie and Gift stories).

Story Name	Type	Narrative	Low-cost	High-cost
Bottle	Nice	Protagonist fetches a baby bottle for her sibling.	Bottle fell right under the protagonist	Bottle rolled into the adjacent room
Pencil	Nice	Protagonist brings pencils to his mother	Pencils are on a nearby table	Pencils are on a table on the second floor
Cookie	Naughty	Protagonist takes a cookie when he was told not to.	The cookie jar is on a low shelf.	The cookie jar is on a high shelf.
Gift	Naughty	Protagonist peeks into a wrapped gift.	Gift is on top of the table next to the protagonist.	Gift is inside the closet.

Table 1: Stories used in Experiment 2.

Procedure

Children were tested individually in a quiet room in the museum. The child was seated in a small table across from the experimenter. The experiment began by placing the high-cost and low-cost version of one of the stories side by side. The experimenter read each story in the pair (See Figure 2 for an example). After both stories in the pair were read, the experimenter asked a control question to ensure the child remembered which protagonist was closer to the goal object (e.g., "Who do you think was closer to the baby bottle at the beginning"). Children were then asked which protagonist was nicer or which protagonist was naughtier in the nicer and naughty story types, respectively (e.g., "Who do you think was nicer?"). The stories were presented in a fixed pseudo-random order so that children never saw two "naughty" or two "nice" stories sequentially. Cookie stories (Naughty) were always followed by the Baby bottle stories (Nice), and Present stories (Naughty) were always followed by Pencil stories (Nice). Otherwise, the order of the stories was counterbalanced across subjects.

Results and Discussion

One child was dropped from analyses on all but the gift story because s/he refused to choose between the protagonists. An additional child was excluded from analysis on the pencil story due to experimenter error. Finally, we excluded children from analyses on any story where they failed to answer the control question correctly, resulting in n = 18, 19, 19, and 20 children in the baby bottle, pencil, gift, and cookie stories, respectively. Examining each story individually we found that children chose the high-cost protagonist significantly above chance ($p < 0.05$ in each story by binomial test). A total of 14 children completed all four storybooks. 12 of these 14 children (85.71%) selected the high-cost protagonist at least 3 (out of 4) times and 6 children (42.85%) performed at ceiling ($p < 0.0001$; binomial test).

Consistent with our predictions, children were sensitive to the costs the protagonists incurred, and they were equally sensitive for attributions of credit and blame. The naughty stories (cookie and gift stories) replicated the qualitative pattern seen in adults in Experiment 1. Additionally, note that while previous research has established that children can make categorical social judgments (e.g., in distinguishing helpers, hinderers, and bystanders; Kuhlmeier, Wynn, & Bloom, 2003; Hamlin, Wynn, & Bloom, 2007), the current results show that children can also make graded judgments within social categories. These results are consistent with children's ability to use a naïve utility calculus to evaluate agents' actions.

In our experiment, children were given a two-alternative forced-choice question. As a result, we were able to establish that children use costs to make social evaluations. However, it is an open question whether children, like adults in Experiment 1, spontaneously use cost information in these tasks. Future experiments might address this.

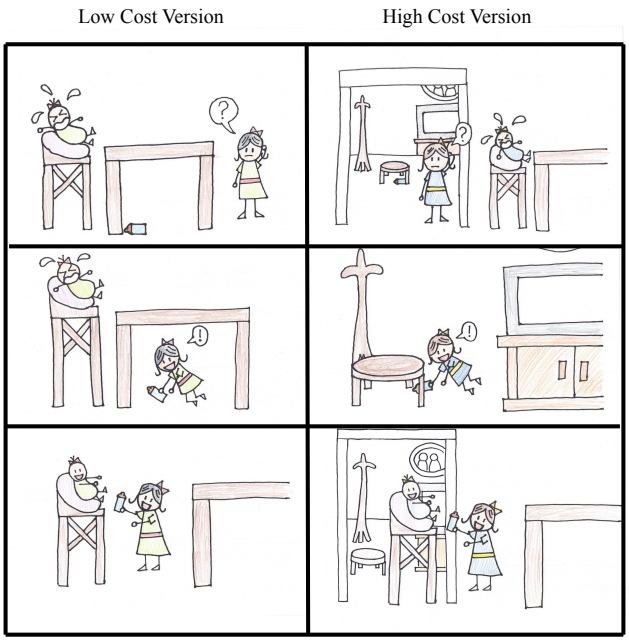

Figure 2**:** Example figures from the baby bottle story. The left column shows the pictures from the low cost version and the right column shows the pictures from the high cost version. The bottom row shows the two pictures children saw when they were asked the control and test question.

General Discussion

Here we proposed that a naïve utility calculus -an intuitive theory of the costs and rewards of decisions- underlies our ability to make graded social evaluations from relatively sparse data about agents' actions and environmental constraints. In the context of moral judgments, the current results suggest that the costs an agent incurs to perform a helpful or harmful action are critical to our moral evaluations. This reasoning sometimes produces seemingly paradoxical results such as attributing more blame to perpetrators who themselves incur greater costs in committing a transgression.

In our study, we looked at graded judgments of actions that were not distinguishable on other grounds relevant to moral reasoning (e.g., social status, directness, intention, or outcome). By this we do not mean to say that judgments of intentionality played no role here. The generalizability of an agent's actions may well be related to judgments of how intentional the action was. As we noted in the Introduction, previous work has manipulated cues to intentionality very directly (by contrasting knowledgeable, volitional agents to those who act in ignorance, under duress or accidentally.) Here none of those contrasts obtain. Nonetheless it is possible that to the degree that we can make graded judgments of how intentional an action is, information about costs bears on our intentionality judgments. In graded judgments we may treat even an action by a knowledgeable, volitional agent as "less intentional" when the action is low cost than when it is high cost. Future research might look at how the naïve utility calculus bears on judgments of intentionality.

Why are the costs a transgressor incurs so important to our judgments of agents' motivations, and thus to our moral judgment? We suggest that the cost an agent incurs provides valuable predictive information about the agent's future actions. If a perpetrator commits a low-cost transgression, we do not know if they would transgress if the costs were higher. By contrast, if a perpetrator commits a high-cost transgression, there is every reason to suspect that they would also commit transgressions at lower costs. Joe might not steal Arnold's car, but Phil would almost certainly steal Bob's. To the degree that punishment and moral judgment act as a deterrent against future transgressions, scaling these to the costs a perpetrator incurs may allow us to most effectively deter those most likely to offend in a broad range of contexts.

More subtly, one key component of the cost of an action is often the time it takes to perform the action: costly actions are typically time-consuming. Given that in principle, transgressors can change their minds and reform their ways at any point in time, costly transgressions may indicate not

merely the intensity but also the stability of perpetrators' bad intentions. Thus, less punishment should be assigned to agents who act more impulsively.

In Experiment 1 we found that adult intuitions of punishment were influenced not only by the cost to the transgressor but also by the value of the object. In future work, we hope to extend the predictions of the naïve utility calculus to multi-party interactions. When agents choose to act we expect them to be empathetic towards others. That is, we expect the utilities of moral agents to be recursive: if I am a prosocial agent, your utilities affect my utilities (See Ullman, et. al., 2010 for a similar approach). Therefore, the value an agent assigns to their belongings should be included in the transgressor's costs (e.g., you need more motivation to steal an object if the victim assigns great personal value to it). Given no other information, the market value of the object is the best indicator of how much the victim valued his/her belonging. To fully test this we require a more fleshed out account of the naïve utility calculus that goes beyond the scope of this paper but remains a promising area for future work.

In our experiments, the costs agents incurred were mainly indexed by the distance and time they had to travel. However, costs can be influenced by many things, including non-obvious properties internal to the agent (e.g., the agent's strength or competence). Additionally, some actions probabilistically incur extrinsic costs due to potential negative consequences (e.g., getting caught stealing). In our experiment we established that children and adults were sensitive to some kinds of costs. However we do not know yet how costs are represented and integrated, whether some types of costs are more salient than others, and how sensitivity to different kinds of costs changes throughout development. These remain rich areas for future work.

Returning to the current work however, our findings suggest that a naïve utility calculus may be a fundamental component of our general moral calculus, evident even in the judgments of young children. By using the costs an agent incurs to infer the value the agent places on acting, we can move beyond merely deciding whether actions are intentional and direct, and make graded judgments of an agent's motivation and the range of contexts under which the agent is likely to act again. In adding costs to the moral calculus we can formalize many of our social intuitions and gain insight into the principles that support our understanding of others' behavior.

Acknowledgments

We thank the Boston Children's Museum and the families who volunteered to participate. We thank Josh Tenenbaum for useful comments and discussions. This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF STC award CCF-1231216, and by the Simons Center for the Social Brain (SCSB) award 6926004.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the thirty-second annual conference of the cognitive science society* (pp. 2469-2474).

Baillargeon, R., Scott, R. M., He, Z., Slona, S., Setoh, P., Jin, K., Wu, D., & Bian, L. (*in press*). Psychological and sociomoral reasoning in infancy. *APA Handbook of Personality and Social Psychology: Vol . 1. Attitudes and Social Cognition.*

Greene, J. (2003). From neural 'is' to moral 'ought': what are the moral implications of neuroscientific moral psychology?. *Nature Reviews Neuroscience*, *4*(10), 846-850.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557-559.

Hamlin, J. K., 2013: Does the infant posses a moral concept? *Concepts: Core Readings.Volume II.*

Jara-Ettinger, J., Baker, C. L., & Tenenbaum, J. B. (2012). Learning what is where from social observations. In *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society* (pp. 515-520).

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2013). Not so innocent: Reasoning about costs, competence, and culpability in very early childhood. In *Proceedings of the thirty-fourth annual conference of the cognitive science society* (pp. 663-668).

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. *in prep*. Naïve Utility Calculus.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. I'd do anything for a cookie (but I won't do that): Children's understanding of the costs and rewards underlying rational action. In *Proceedings of the thirty-fifth annual conference of the cognitive science society*.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*(4), 315.

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological Science*, *14*(5), 402-408.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future.*Trends in cognitive sciences*, *11*(4), 143-152.

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O. R., Goodman, N. D., & Tenenbaum, J. B. (2010). *Help or hinder: Bayesian models of social goal inference*. Neural Information Processing Systems Foundation.