Can Adults Revise Their Core Beliefs about Agents?

Rongzhi Liu (rongzhi_liu@berkeley.edu), Fei Xu (fei_xu@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94720 USA

Abstract

A set of fundamental principles governs our reasoning about agents since infancy. Past research has shown that adults are surprised when they observe apparent violations of these principles, which might prime them to learn from the violations and update their beliefs. In the present experiments, we demonstrate that adults can revise their beliefs about these principles in a specific, virtual world when they observe multiple pieces of counterevidence, and generalize their revised beliefs to new agents in the same environment. We discuss these findings together with the findings of a similar study with preschoolers, and we suggest future directions for this line of research.

Keywords: belief revision; core knowledge; intuitive psychology

Introduction

The Core Knowledge view (Spelke, 1988, 2000; Spelke & Kinzler, 2007) argues that human infants are endowed with core knowledge systems – a small number of systems of domain-specific knowledge, each accompanied by a set of principles. Later in development, infants and children construct intuitive theories based on these systems, such as intuitive physics and intuitive psychology (e.g., Carey, 1985; Gopnik & Meltzoff, 1997; Wellman & Gelman, 1992).

One of the core knowledge systems guides how we represent and reason about agents. Between the ages of 6 to 12 months, infants understand that agents' intentional actions are directed to goals (Woodward, 1998), agents choose efficient means to achieve their goals (Gergely & Csibra, 2003), and agents' preferences can be inferred based on violations of random sampling (Wellman et al., 2016). For ease of exposition, we will refer to these as the Goal principle, the Efficiency principle, and the Sampling principle from now on. These principles support further learning in the psychological domain (e.g., Jara-Ettinger et al., 2015; Kushnir et al., 2010; Sodian et al., 2016). They also persist into adulthood – these principles underlie adults' mental state reasoning and action understanding in complex scenes (Baker et al., 2017, Jara-Ettinger et al., 2020).

Thus, these core principles about agents are early emerging, they support learning about the psychological domain throughout development, and they continue to guide reasoning about the psychological world in adulthood. Yet one of the hallmarks of human learning is that beliefs can be revised given new evidence (Chater & Oaksford, 2008; Gopnik & Wellman, 2012; Tenenbaum et al., 2011; Xu, 2019; Ullman & Tenenbaum, 2020). Are these earliest-emerging core principles about agents also subject to revision once we acquire them? If adults are given enough evidence that violates these principles, will they rationally update their beliefs? This is the focus of the current studies.

A set of past studies has shown that preschoolers and adults can revise their beliefs about the core principles of objects given counterevidence (Liu & Xu, 2021, 2022). However, people's beliefs about the principles governing agents' behaviors might be more stochastic than their beliefs about the principles governing object motions. Therefore, it would still be important to investigate whether the core principles of agents are revisable given counterevidence.

Infants and children are sensitive to evidence that violates the core psychological principles and use this evidence to update their beliefs. When an agent violated the Efficiency principle, infants attended to the inefficient agent less compared to an efficient agent, and expected a neutral observer to approach the efficient agent instead of the inefficient agent (Colomer et al., 2020). In addition, infants were less likely to learn novel words from an inefficient agent compared to an efficient agent (Colomer & Woodward, 2023). These findings suggest that infants use agents' inefficient actions to update their beliefs about the knowledgeability of the agents, as well as how other agents would interact with them. Lastly, a past study has shown that preschoolers can revise their beliefs about the Goal and the Sampling principles when given counterevidence (Liu & Xu, 2021). However, the strength of adults' prior beliefs about these principles might be different from that of children's, therefore it is unclear whether a few pieces of counterevidence would also be enough to revise adults' beliefs about these principles.

A growing body of literature shows that adults' reasoning about agents can be captured by Bayesian probabilistic models (Baker et al., 2017, Jara-Ettinger et al., 2020; Shu et al., 2021). They rationally update their beliefs about agents' mental states and the environment given new evidence. In addition, adults are surprised by apparent violations of the core psychological principles. When they observed events that violated the Goal and Efficiency principles, they rated these as more surprising than events that did not violate the principles (Shu et al., 2021). Surprise provides opportunities for learning (Stahl & Feigenson, 2015; 2017). Thus, when adults are surprised by violations of the core psychological principles, they might be primed to learn from the violations and rationally update their beliefs about the principles.

In two experiments, participants observed events that supported or violated the Goal, Efficiency, and Sampling principles, or they did not receive any new evidence about these principles. Then, they made predictions about the outcomes of new events. We compare 3 hypotheses. Our first hypothesis is that participants would not revise their beliefs about the principles given counterevidence. Regardless of the

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

evidence they observe, they would predict outcomes consistent with the principles for new events. Our second hypothesis is that participants who saw the belief-violating evidence would be more likely to predict outcomes inconsistent with the principles, compared to those who saw the belief-consistent evidence and those who did not receive new evidence. However, they would not genuinely accept the counterevidence and genuinely revise their beliefs about the principles. When asked to explain the belief-violating evidence, they would either try to come up with alternative interpretations to explain away the counterevidence, or say that they learned from the pattern in the belief-violating evidence and made predictions accordingly. Our third hypothesis is that participants would genuinely accept the counterevidence and revise their beliefs. Participants would make predictions about new events in the same way as predicted by the second hypothesis. In addition, when asked to explain the belief-violating evidence, they would state that they have accepted the violations of the principles in the evidence.

We also examine how far participants would generalize their revised beliefs about these principles. After observing an agent violate each principle, participants were asked to make predictions about the same agent (easy test trials), new agents of the same kind (hard test trials), and new agents of a different kind (harder test trials).

Experiment 1

Method

Participants Sixty adults (mean age = 33 years; range = 18 to 54; *SD* = 9.41; 35 females) participated in the experiment on Prolific. Participants provided written informed consent prior to participating in the experiment. They completed a 25-minute survey for which they were paid $4.

Stimuli and Procedure Participants were randomly assigned to one of three conditions, the Baseline condition, the Belief Consistent (BC) condition, and the Belief Violation (BV) condition. They were tested on 3 principles, Efficiency, Goal, and Sampling, in counterbalanced orders. For each principle, there were 6 familiarization trials and 4 test trials (2 easy test trials and 2 hard test trials; order counterbalanced). The familiarization trials in the BC condition displayed events that were consistent with the principle, those in the BV condition displayed events that violated the principle. The familiarization trials in the Baseline condition did not display the outcomes of the events, so participants did not receive any new evidence that supported or violated the principles. In test trials, participants chose between the *Belief Consistent (BC) response* and the *Belief Violation* (*BV*) *response*. They never receive feedback about whether their choices were correct or incorrect.

Efficiency principle. In the familiarization trials, a grey wall and 2 agents (i.e., geometric shapes with eyes) appeared. Participants were told that one agent wanted to play with the other agent. The agent went toward the other agent by jumping over the wall. Then, the wall was moved to the side. The agent went toward the other agent by taking a straight path (BC condition) or a jumping path (BV condition), or did not go toward the other agent (Baseline condition) (Figure 1). The goal was a different agent in each familiarization trial.

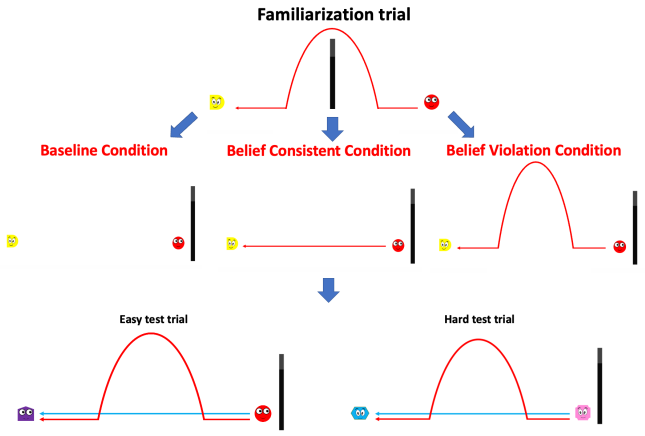

Figure 1: Events shown in the familiarization trials and test trials for the Efficiency principle.

In the easy test trials, the same agent went toward a new geometric-shaped agent by jumping over a wall. Then, the wall was moved to the side. A red path and a blue path indicated the straight path (the *BC response*) and the jumping path (the *BV response*). Participants chose the path that they believed the agent would take. In the hard test trials, participants made predictions about how a new geometric-shaped agent would behave in the same situation.

In the explanation question, participants were asked to explain why the agent took the straight or the jumping path to get to its goal.

Goal principle. In the familiarization trials, an agent and 2 objects appeared. Participants were told that the agent wanted to play with some toys. The agent went toward one of two objects and took the object 3 times. Then, the two objects switched locations. The agent took the old object at the new location (BC condition) or the new object at the old location (BV condition), or did not take either object (Baseline condition) (Figure 2). A different pair of objects was used in each familiarization trial.

In the easy test trials, a new pair of objects appeared. The same agent took one of the objects 3 times. Then the two objects switched locations. Participants chose the object that they believed the agent would take, either the old object at the new location (*BC response*) or the new object at the old location (*BV response*). In the hard test trials, participants made predictions about a new geometric-shaped agent in the same situation.

In the explanation question, participants were asked to explain why the agent took the respective object after the objects switched locations.

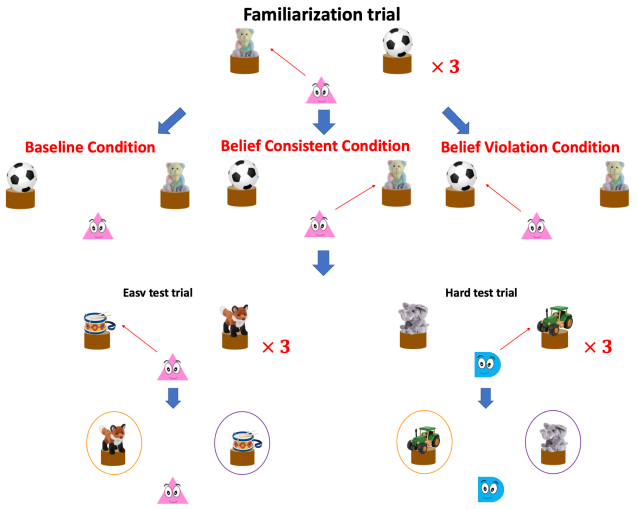

Figure 2: Events shown in the familiarization trials and test trials for the Goal principle.

Sampling principle. In the familiarization trials, an agent and a box of objects appeared. The box contained 7 objects of one type and 31 objects of the other type. Participants were told that the agent wanted to play with some toys. The agent picked out 4 objects of the minority type from the box, and put them into a small box in front of the agent. Then, an object of the minority type and an object of the majority type appeared, equidistant from the agent. The agent went toward the minority type (BC condition) or the majority type (BV condition), or did not go toward either object (Baseline condition) (Figure 3).

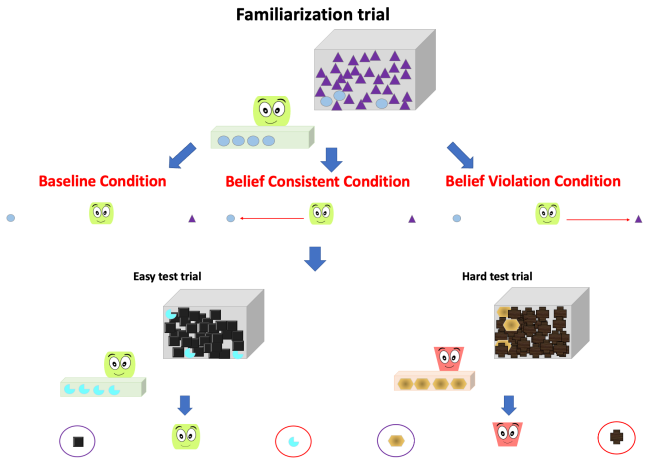

Figure 3: Events shown in the familiarization trials and test trials for the Sampling principle.

In the easy test trials, the same agent sampled 4 objects of the minority type from a new toy box. Then, an object of the minority type and an object of the majority type appeared, equidistant from the agent. Participants chose the object that they believed the agent liked better, either the minority-type object (*BC response*) or the majority-type object (*BV response*). In the hard test trials, participants made predictions about a new geometric-shaped agent in the same situation.

In the explanation question, participants were asked to explain why the agent went toward the respective object after the sampling behaviors.

Results

The proportion of *BV response* by condition and principle is shown in Figure 4. We used mixed-effect logistic regression to predict participants' binary response (*BV response* = 1, *BC response* = 0) from condition, principle, and test trial type, while controlling for the random effects of individual participants. The best-fitting model included condition and principle as predictors. Participants were more likely to choose the *BV response* in the BV condition than in the Baseline condition ($\beta = 5.39$, $SE = 0.88$, $p < .001$) and the BC condition ($\beta = 7.81$, $SE = 1.10$, $p < .001$), and they were less likely to choose the *BV response* in the BC condition than in the Baseline condition ($\beta = -2.42$, $SE = 0.93$, $p = .009$). The effect of principle showed that, compared to the Goal principle, participants were more likely to choose the *BV response* for the Efficiency principle ($\beta = 1.26$, $SE = 0.34$, $p < .001$), and less likely to choose the *BV response* for the Sampling principle ($\beta = -2.43$, $SE = 0.41$, $p < .001$). Test trial type (easy vs. hard test trials) did not have a significant effect on participants' responses.

To assess the strength of adults' prior beliefs about the principles, we compared the proportion of *BV response* in the Baseline condition against chance. For all 3 principles, adults selected the *BV response* below chance (Exact binomial test: $P_{Efficiency} = .3$ [.20, .41], $p < .001$; $P_{Goal} = .21$ [.13, .32], $p < .001$; $P_{Sampling} = .01$ [.00, .07], $p < .001$).

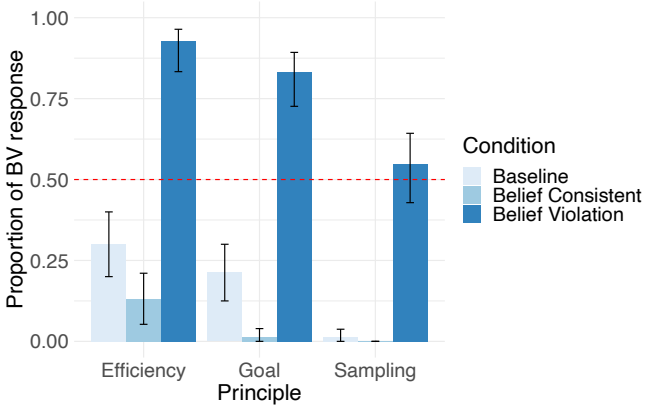

Figure 4: The proportion of trials that participants selected the *BV response* by condition and principle, in Experiment 1. The dashed line indicates chance selection (.5), and the error bars indicate bootstrapped 95% CIs.

For the explanation questions, 2 researchers coded participants' responses into different categories (interrater reliability was excellent, Cohen's Kappa = .89; disagreements were

resolved through discussion). In the BC condition, most responses (89.5%) referred to the principle itself to explain the evidence (other responses were irrelevant to the principle or incomprehensible). In the BV condition, we coded participants' explanations into 4 categories based on the criteria in Table 1.

Table 1: Coding criteria and examples for explanations in the BV condition

Category	Criterion	E.g., Sampling
Accept Evidence	Accepted the violation of the target principle in the counterevidence.	"He likes the purple toy better (the majority type)."
Explain Away	Explained the counterevidence with reasons that would not involve any violations of the target principle.	"He was getting rid of the toy he didn't like."
Pattern	Noted the pattern in the evidence.	"Because that has been the pattern throughout the videos."
Other	Explanations that cannot be categorized into the first three categories.	"I don't know."

Table 2 shows the number of responses in each category for each principle. We used mixed-effects multinomial logistic regression to predict participants' explanations from principle, while controlling for the random effects of individual participants. We found an effect of principle. When switching from the Sampling principle to the Efficiency principle or the Goal principle, participants became more likely to provide "accept evidence" explanations than the other types of explanations ($ps < .02$).

Table 2: Number of responses by category and principle in the BV condition in Experiment 1

	Efficiency	Goal	Sampling
Accept Evidence	17	15	3
Explain Away	2	1	6
Pattern	1	2	6
Other	1	3	6

Next, we used mixed-effects logistic regression to predict participants' binary choice in the test trials (*BV response* = 1, *BC response* = 0) from the type of explanation they provided, while controlling for the random effects of individual participants. Compared to participants who provided "accept evidence" explanations, participants who provided any other types of explanations were less likely to choose the *BV response* for that principle ("explain away": $\beta = -2.78$, $SE = 0.62$, $p < .001$; "pattern": $\beta = -1.41$, $SE = 0.68$, $p = .040$; "other": $\beta = -1.93$, $SE = 0.60$, $p = .001$).

Discussion

Experiment 1 assessed adults' prior beliefs about the Efficiency, Goal, and Sampling principles, and examined whether adults can revise their prior beliefs about these principles given counterevidence, in a specific, virtual environment. We found that adults had strong prior beliefs about these principles. Most adults who did not receive any new evidence expected agents to behave in ways consistent with these principles. After observing evidence supporting these principles, their prior beliefs were strengthened. Furthermore, we found that adults can revise their beliefs about these principles given counterevidence. After observing evidence violating these principles, they were more likely to predict that agents would behave in ways inconsistent with the principles. Moreover, adults' performance did not differ in the easy and hard test trials, suggesting that they generalized their revised beliefs to new agents.

We also found an interesting effect of principle. Across conditions, adults' likelihood of choosing the inconsistent outcome is higher for the Efficiency principle than the Goal principle, which is in turn higher than the Sampling principle. This suggests that adults had stronger prior beliefs for the Sampling principle than the Goal principle, and stronger prior beliefs for the Goal principle than the Efficiency principle. There was no interaction between principle and condition, suggesting that across the 3 principles, the belief-violating evidence had similar effects on participants' prior beliefs.

Participants' explanations for the evidence showed that they were more likely to accept the belief-violating evidence for the Efficiency and Goal principles than the Sampling principle. Indeed, most participants accepted the counterevidence for the Efficiency and the Goal principles, but most participants did not accept the counterevidence for the Sampling principle. This suggests that the counterevidence for the Sampling principle might not be as compelling as the counterevidence for the other 2 principles. Importantly, across principles, participants who had accepted the counterevidence were indeed more likely to predict outcomes that violated the principles.

In the next experiment, we aim to replicate these findings with more realistic, three-dimensional stimuli, and investigate whether adults can generalize their revised beliefs to agents that are not geometric shapes.

Experiment 2

Method

Participants Eighty-two undergraduate Psychology students (mean age = 20.28 years; range = 18 to 36; $SD = 2.54$; 65 females, 15 males, 2 of unknown gender) participated in the experiment. Participants provided written informed consent prior to participating in the experiment. They completed a 25-minute survey and received course credits.

Stimuli and Procedure The procedure of Experiment 2 was similar to that of Experiment 1, with a few exceptions. First, we used photorealistic, three-dimensional stimuli made with Blender. Second, we added 2 harder test trials for each

principle, where participants were asked to predict how animals would behave in the same situation (Figure 6). Third, participants in the Baseline condition were also asked the explanation questions. Instead of explaining an event in the familiarization trial, they were asked to explain their predictions in an easy test trial.

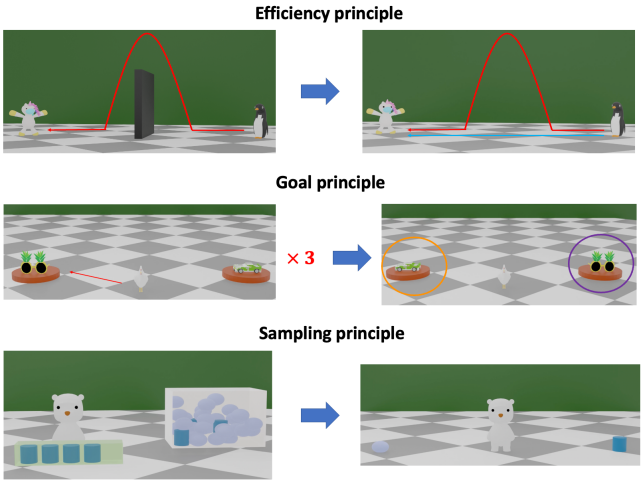

Figure 6: Events shown in the harder test trials for each principle in Experiment 3.

Results

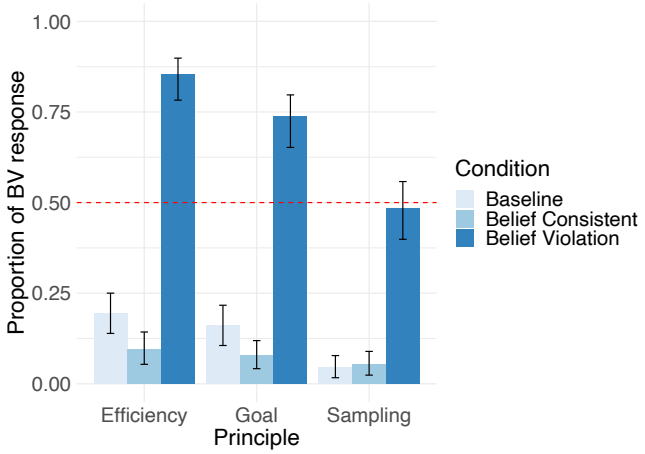

Figure 7: The proportion of trials that participants selected the *BV response* by condition and principle, in Experiment 3. The dashed line indicates chance selection (.5), and the error bars indicate bootstrapped 95% CIs.

The proportion of *BV response* by condition and principle is shown in Figure 7. Mixed-effects logistic regression revealed main effects of condition and principle. Across 3 principles, participants were more likely to choose the *BV response* in the BV condition than in the Baseline condition (β = 4.92, *SE* = 0.73, $p < .001$) and the BC condition (β = 5.82, *SE* = 0.79, $p < .001$); their choices did not differ significantly between the Baseline and the BC conditions (β = -0.90, *SE* = 0.70, $p = .20$). Compared to the goal principle, participants were more likely to choose the *BV response* for the Efficiency principle (β = 0.55, *SE* = 0.21, $p = .009$), and less likely to choose the *BV response* for the Sampling principle (β = -1.52, *SE* = 0.24, $p < .001$). Test trial type (easy vs. hard vs. harder test trials) did not have a significant effect on participants' responses.

To assess the strength of adults' prior beliefs about the principles, we compared the proportion of *BV response* in the Baseline condition against chance. For all 3 principles, adults selected the *BV response* below chance (Exact binomial test: $P_{Efficiency}$ = .20 [.14, .26], $p < .001$; P_{Goal} = .17 [.12, .23], $p < .001$; $P_{Sampling}$ = .06 [.03, .10], $p < .001$).

For the explanation questions, 3 researchers coded participants' responses (interrater reliability was good, Light's Kappa = .72; disagreements were resolved through discussion). In the Baseline and the BC condition, most responses (97.6% and 94.4%) referred to the principle itself to explain the evidence or their predictions (other responses were irrelevant to the principle or incomprehensible).

Table 3 shows the number of responses in each category for each principle in the BV condition (based on the criteria in Table 1). Mixed-effects multinomial logistic regression revealed that when switching from the Efficiency principle to the Goal principle, participants became more likely to provide "accept evidence" explanations than "explain away" explanations (β = 1.79, *SE* = 0.83, $p = .03$); when switching from the Sampling principle to the Goal principle, participants became more likely to provide "accept evidence" explanations than "other" explanations (β = 2.24, *SE* = 1.13, $p = .047$).

Table 3: Number of responses by category and principle in the BV condition in Experiment 2

	Efficiency	Goal	Sampling
Accept Evidence	21	28	19
Explain Away	9	2	4
Pattern	1	2	4
Other	2	1	6

Mixed-effects logistic regression revealed that participants were more likely to choose the *BV response* if they provided "accept evidence" explanations for that principle, compared to if they provided "explain away" (β = 1.30, *SE* = 0.46, $p = .004$) or "other" (β = 1.54, *SE* = 0.44, $p < .001$) explanations.

Discussion

Experiment 2 replicated the main findings of Experiment 1 in a photorealistic, three-dimensional environment. Adults had strong prior beliefs about the Efficiency, Goal, and Sampling principles. Most adults who did not receive any new evidence expected agents to behave in ways consistent with the principles. Unlike Experiment 1, we did not find statistically significant evidence that observing evidence supporting the principles did not further strengthen their prior beliefs. After

observing evidence violating these principles, adults revised their prior beliefs in this specific context. Their performance did not differ in the easy, hard, and harder test trials, suggesting that they generalized their revised beliefs even to new agents that were not geometric shapes.

We also replicated the effect of principle in Experiment 1. Adults' prior beliefs were stronger for the Sampling principle than the Goal principle, which were in turn stronger than their prior beliefs for the Efficiency principle. We will return to this effect of principle in the general discussion.

Most participants accepted the counterevidence for all 3 principles, including the Sampling principle, suggesting that the photorealistic counterevidence might be more compelling than the stimuli in Experiment 1. There was also some evidence that participants were more likely to accept the counterevidence for the Goal principle than the other 2 principles. It is possible that the counterevidence for the Goal principle was more compelling than the other 2 principles, or that it was less likely to come up with alternative explanations to explain away the counterevidence for the Goal principle. Lastly, participants who had accepted the counterevidence were again more likely to predict outcomes that violated the principles.

General Discussion

The present study reports the first systematic investigation of whether adults can revise their beliefs about the most fundamental principles governing their reasoning about agents. We found that adults had strong prior beliefs about the Efficiency, Goal, and Sampling principles. However, they were able to revise their prior beliefs in a specific, virtual environment with just a few pieces of counterevidence. Moreover, supporting our third hypothesis, a majority of participants genuinely accepted the counterevidence, instead of coming up with alternative interpretations to explain away the counterevidence or simply learning from the statistical pattern observed in the counterevidence.

In addition, adults generalized their revised beliefs to new agents of the same type and new agents of a different type in this environment. After they observed a few agents violate these principles, they expected new agents to also violate the principles, even though they had not observed any behaviors of the new agents. Why? One possibility is that participants interpreted the behaviors that violated the principles as "norms" in this virtual environment. For example, after observing 2 agents jump to get to their goals when there is no obstacle, participants might think that the norm in this virtual environment is to jump to get to goals. Future studies can examine this possibility by asking participants to explain why they expected the new agents to violate the principles. It would also be interesting to examine whether participants would generalize their revised beliefs to new agents entering this world and agents in a completely different virtual environment.

In both experiments, we found that adults had stronger prior beliefs for the Sampling principle than the Goal principle, and stronger prior beliefs for the Goal principle than the Efficiency principle. Interestingly, a previous study (Liu & Xu, 2021) found that preschoolers also showed stronger prior beliefs for the Sampling and the Goal principles than the Efficiency principle, suggesting that the strength of our prior beliefs about these principles might be relatively stable across development. Future studies should systematically investigate the developmental course of beliefs about core psychological principles, and why the strength of our prior beliefs differ across principles. For example, we might have weaker prior beliefs for the principles that are more frequently violated in the real world.

While adults revised their beliefs about all 3 core psychological principles, preschoolers in the previous study (Liu & Xu, 2021) only revised their beliefs about the Goal and the Sampling principles, but not the Efficiency principle. One possibility is that, unlike adults, children do not consider jumping as a violation of the Efficiency principle (e.g., they might find jumping more rewarding than taking the straight path, and therefore worth the additional time and energy). In future research, we could show children different types of violations (e.g., an agent taking a detour when walking on the ground) to see if those would revise children's beliefs about the Efficiency principle. More generally, we could show participants multiple types of violations for each principle, and ask them to make predictions about different types of events, to further rule out the possibility that participants were simply learning from the statistical pattern in the evidence.

The present study found that adults can revise their beliefs about the core psychological principles when they observe geometric-shaped agents violate these principles in a virtual world. Past research suggests that adults expect geometric-shaped agents to behave similarly to humans in the real world in terms of the core psychological principles (Baker et al., 2017, Jara-Ettinger et al., 2020; Shu et al., 2021). Thus, the present findings imply that adults have the potential to (1) generalize their revised beliefs about these principles to humans in the real world, and (2) revise their beliefs when they observe humans violate these principles in the real world. These are important directions for future research.

References

Amsterlaw, J., & Wellman, H. M. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development*, *7*(2), 139–172.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press/Bradford Books.

Chater, N., & Oaksford, M. (Eds.). (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford University Press, USA.

Colomer, M., Bas, J., & Sebastian-Galles, N. (2020). Efficiency as a principle for social preferences in infancy. *Journal of Experimental Child Psychology*, *194*, 104823.

Colomer, M., & Woodward, A. (2023). Should I learn from you? Seeing expectancy violations about action efficiency hinders social learning in infancy. *Cognition*, *230*, 105293.

Doan, T., Denison, S., Lucas, C., & Gopnik, A. (2015). Learning to reason about desires: An infant training study In D. C. Noelle et al. (Eds.), *Proceedings for the Annual Meeting of the Cognitive Science Society* (pp. 578–583). Austin, TX: Cognitive Science Society.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gopnik, A., & Meltzoff, A. (1997). *Words, thoughts and theories.* Cambridge, MA: MIT Press.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–1108.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The Naïve Utility Calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, *21*(8), 1134–1140.

Liu, R. & Xu, F. (2021). Revising core beliefs in young children. In T. Fitch, C. Lamm, H. Leder, & K. Tessmar (Eds.), *Proceedings of the 43rd Annual Conference of the Cognitive Science Society.* Cognitive Science Society.

Liu, R. & Xu, F. (2022). Can adults revise their core beliefs about objects? In J. Culberston, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., et al. (2014). The child as econometrician: A rational model of preference understanding in children. PLoS One, 9(3), e92160.

Ma, L., & Xu, F. (2011). Young children's use of statistical sampling evidence to infer the subjectivity of preferences. *Cognition*, *120*(3), 403–411.

Shu, T., Bhandwaldar, A., Gan, C., Smith, K. A., Liu, S., Gutfreund, D., Spelke, E., Tenenbaum, J. B., & Ullman, T. D. (2021). AGENT: A benchmark for core psychological reasoning. *ArXiv:2102.12321 [Cs]*.

Sodian, B., Licata, M., Kristen-Antonow, S., Paulus, M., Killen, M., & Woodward, A. (2016). Understanding of goals, beliefs, and desires predicts morally relevant theory of mind: A longitudinal investigation. *Child Development*, *87*(4), 1221–1232.

Spelke, E. S. (1988). The origins of physical knowledge. In L. Weiskrantz (Ed.), *Thought without language*. Clarendon Press/Oxford University Press.

Spelke, E. S. (2000). Core knowledge. *American Psychologist, 55*(11), 1233–1243.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Stahl, A. E., & Feigenson, L. (2017). Expectancy violations promote learning in young children. *Cognition*, *163*, 1–14.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, *2*, 533-558.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual review of psychology*, *43*(1), 337-375.

Wellman, H. M., Kushnir, T., Xu, F., & Brink, K. A. (2016). Infants use statistical sampling to understand the psychological world. *Infancy*, *21*(5), 668-676.

Woodward, A. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.

Xu, F. (2019). Towards a rational constructivist theory of cognitive development. *Psychological Review*, *126*(6), 841–864.