The fine structure of surprise in intuitive physics: when, why, and how much?

Kevin A. Smith[1,2], **Lingjie Mei**[3], **Shunyu Yao**[4], **Jiajun Wu**[5]
Elizabeth Spelke[2,6], **Joshua B. Tenenbaum**[1,2,3], **Tomer D. Ullman**[2,6]

[1] MIT BCS, [2] Center for Brains, Minds, & Machines, [3] MIT CSAIL,
[4] Princeton University, [5] Stanford University, [6] Harvard Psychology

Abstract

We are surprised when events violate our intuitive physical expectations. Even infants look longer when things seem to magically teleport or vanish. This important surprise signal has been used to probe what infants expect, in order to study the most basic representations of objects. But these studies rely on binary measures – an event is surprising, or not. Here, we study surprise in a more precise, quantitative way, using three distinct measures: we ask adults to judge *how* surprising a scene is, *when* that scene is surprising, and *why* it is surprising. We find good consistency in the level of surprise reported across these experiments, but also crucial differences in the implied explanations of those scenes. Beyond this, we show that the timing and degree of surprise can be explained by an object-based model of intuitive physics.

Keywords: Intuitive physics; Surprise; Violation of expectation; Generative models

Introduction

Imagine going to a magic show, with all its standard tricks: balls levitate, bunnies appear and disappear, assistants are run through with swords but left unharmed. If you believed that these were not tricks but actual magic, you would surely be astounded. This astonishment is not driven by the particular objects involved – you would be just as shocked if a novel object levitated or an animal you had never seen before suddenly materialized. Rather, the surprise is due to the violation of basic intuitive physical principles, such as permanence and solidity.

Surprise is an important measure of our intuitive reasoning. It forms the basis of the Violation of Expectation (VoE) paradigm, which examines what infants know based on how they react to different events (Spelke & Kinzler, 2007; Baillargeon, Spelke, & Wasserman, 1985). This early knowledge is also the foundation of our adult understanding (Spelke, Breinlinger, Macomber, & Jacobson, 1992; Noles, Scholl, & Mitroff, 2005). Examining the 'surprise' signals in computational models has recently been used to assess commonsense physical reasoning in machines (Riochet et al., 2018; Piloto et al., 2018; Smith et al., 2019), and recent studies suggest surprise is also an internal signal for learning in humans (Stahl & Feigenson, 2015; Kidd, Piantadosi, & Aslin, 2012). Here we focus on the way surprise arises from violations of core physical knowledge in order to better understand how we use intuitive physics to make sense of the world.

Despite the impact and importance of VoE and surprise measures, infant studies (and by extension models of infant physical reasoning) often rely on coarse, binary information. These methods can demonstrate that infants find one scene to be more surprising than another, but cannot be used to pick out when or what is surprising, and often cannot compare surprise across scenes (though c.f. Téglás et al., 2011). We present a richer study of surprise in physical reasoning that uses several different measures in adults, with stimuli inspired by the developmental literature. We quantitatively examine the degree (how), timing (when), and explanation (why) of surprising physical events. We assess people's consistency and variation for each of these measures, as well as their overall convergence. We find that regardless of the way they are asked, people consistently find some scenes more surprising than others, with a consistent ordering even among those that all contain physically implausible events. However, the specific surprises that they note in each scene differ, depending on the measure. Our moment-by-moment measures of surprise also let us evaluate physical reasoning models in a new way. We show that one particular model proposed by Smith et al. (2019) for capturing overall surprise can also predict human levels of moment-by-moment surprise.

Methods

To quantify surprise in physical reasoning, we examine the degree to which people find scenes surprising, when they are surprised, and how they explain why the scene is surprising. We study the degree of surprise with data from Smith et al. (2019), and briefly describe their methodology below. All three experiments, including one originally presented in Smith et al. (2019), follow similar procedures and use the same set of stimuli. Participants were all recruited from Amazon's Mechanical Turk, using the psiTurk framework (Gureckis et al., 2016).

Stimuli. Smith et al. (2019) used 8 different scenarios designed to test aspects of core knowledge of physics (Spelke & Kinzler, 2007), including permanence, solidity, and continuity. The stimuli were inspired by classic developmental studies of physical understanding. Each scenario includes one scene that depicts a violation of physics, and one or more control scenes with similar motion patterns that do not depict this violation. For example, a violation might show an object travel behind a screen and disappear, whereas the control would show those events without the object vanishing. Each scenario contained 8 different templates to create violation and control videos. Each template included the same objects, to match the overall visual appearance across videos. The 8 scenarios are shown in Fig 1, and include:

Create (object permanence) Violation: object appears from behind a screen. Control: the object does not appear or already exists (Wynn & Chiang, 1998).

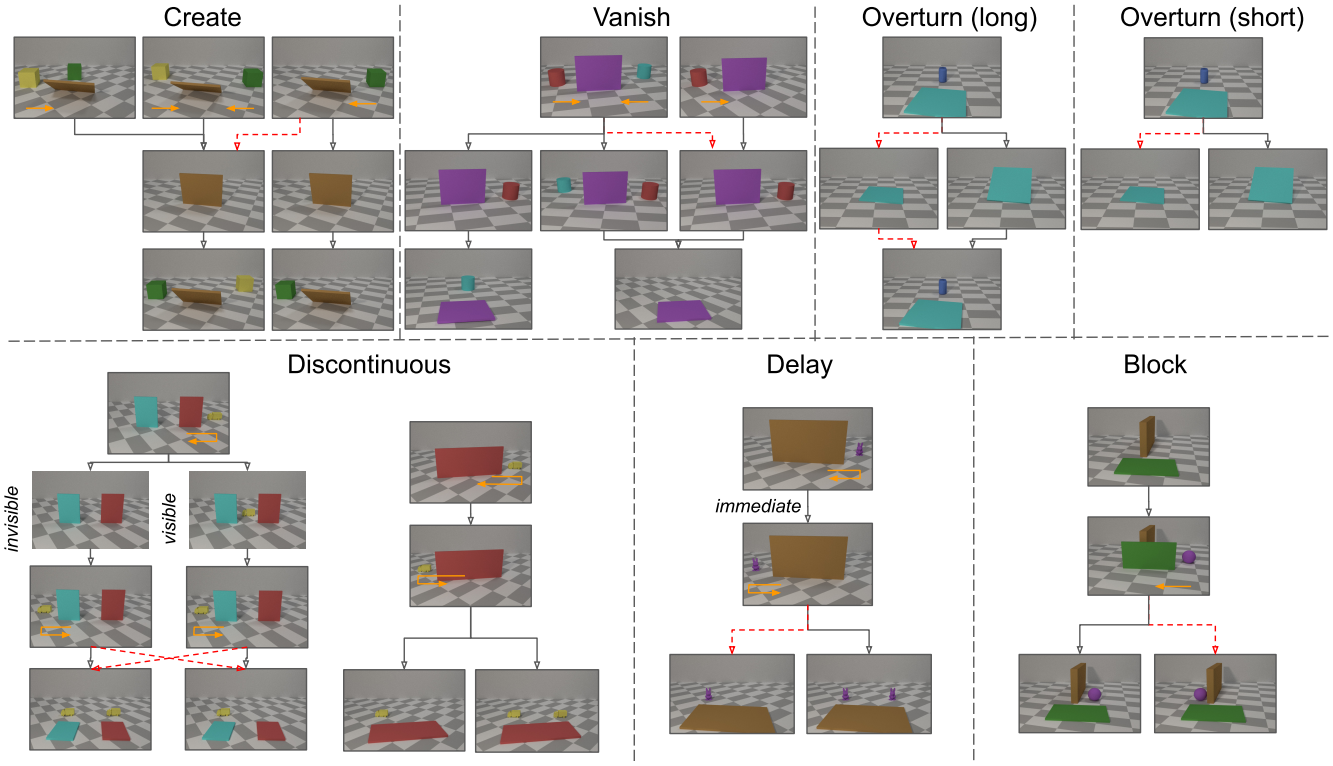

Figure 1: The different scenarios used in all experiments. Black arrows show physically plausible transitions between movie parts. Red dashed arrows show transitions that violate physical expectations. Figure reprinted from Smith et al. (2019).

Vanish (object permanence) Violation: object disappears behind a screen. Control: the object remains or never existed.

Overturn (solidity) Violation: a screen rotates backwards and through an object. Control: the screen stops before it hits the object. In 'short Overturn' , the video ends when the screen stops, in 'long Overturn' the screen rotates back and shows the original object again (Baillargeon et al., 1985).

Discontinuous-invisible (continuity) Violation: two screens with space between them are shown. An object moves out and back from one screen, then an identical object moves out and back behind the other. Screens rotate down to show a single object behind the second screen. Control: two objects remain, or there is one large screen (Spelke, Kestenbaum, Simons, & Wein, 1995).

Discontinuous-visible (solidity) identical to Discontinuous-invisible, except the object is seen moving between two screens, and the violation scene ends with objects behind both screens.

Delay (continuity) Violation: an object moves out and back behind one side of a large screen then immediately moves out and back behind the other side, and the screen lowers to show one object. Control: the screen lowers to show two objects.

Block (solidity/continuity) Violation: a wall is shown behind a screen. An object moves from the side behind the screen, then the screen comes down to show the object on the opposite side of the wall. Control: the object remains on the same side (Spelke et al., 1992).

Participants always judged one video from each template, and never saw the same object multiple times in the same scenario. Stimuli were counterbalanced to ensure equal numbers of videos from each of the violation and control conditions.

"How" surprising is a scene? Smith et al. (2019) measured people's overall level of surprise, emulating the logic of infant studies where the surprise is considered as a response to the scenario as a whole. Participants (N=60) watched the videos on repeat. After a video had finished playing once, participants were asked "How surprising are the events in this movie?" and responded using a slider that ranged from "Not at all surprising" (0) to "Extremely surprising" (100). Ratings were z-scored within participants, to control for individual uses of the scale.

"When" is a scene surprising? Participants (N=60) were told that i) they will watch several videos, ii) sometimes odd things will happen in the video, iii) not all videos include surprises, and iv) some videos could include multiple surprises. Participants were instructed to push the spacebar any time something surprising happens. Whenever participants pushed the spacebar, the border of the screen flashed, to indicate that the surprise was registered. These button presses should therefore be timed to when participants observe evidence of a surprise – either one that just occurred or evidence that one had happened in the past.

"Why" is a scene surprising? Participants (N=95) watched the videos on repeat, similar to the 'How' experiment, and

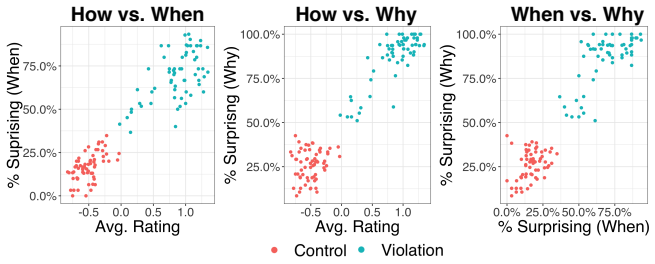

Figure 2: Comparisons of ratings ('How'), and proportion of people who noted a surprising event ('When' and 'Why'). Each point represents a different video. The amount of surprise noted was consistent across all three experiments.

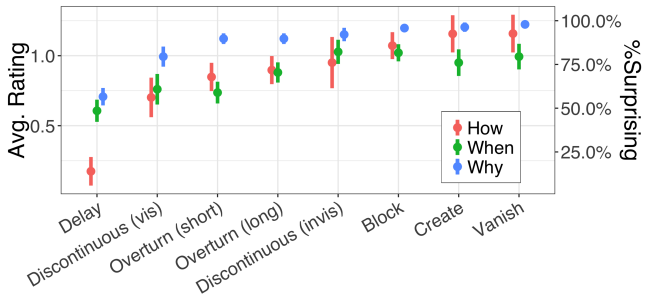

Figure 3: Surprise ratings for violation scenes by scenario foe each experiment. There are differences between scenarios that are consistent across experiments.

were asked "Which of the following surprising events occurred?" Participants were provided with a set of possible answers, including: (1) nothing surprising happened, (2) an object disappeared, (3) an object appeared that wasn't there before, (4) an object teleported, (5) an object moved through another object, (6) an object started moving by itself, (7) an object stopped moving by itself, (8) an object changed shape, (9) an object changed color, and (10) other (with a text box to fill in if checked). Participants could indicate multiple surprises.

Results

Consistency and convergence. People were remarkably consistent in finding the violation scenes to be more surprising than control scenes. Across every one of the 64 scene templates, participants rated the violation video as more surprising than the controls ('How'), were more likely to indicate a surprise ('When') or to pick out a particular surprise ('Why').

In addition, there was high consistency in the *relative* surprise rankings of trials across experiments. Across all trials, there were high correlations between ratings in the 'How' experiment and the proportion of people who marked a trial as surprising in 'When' ($r = 0.94$), between 'How' ratings and the number of people who described a scene with some sort of surprise in 'Why' ($r = 0.96$), and the proportion of people finding a scene surprising in both 'When' and 'Why' ($r = 0.94$; see Fig. 2). This agreement is partly driven by overall differences between the violation and control scenarios, but even within just the surprising scenes there was large agreement across the experiments ('How' vs 'When': $r = 0.61$, 'How' vs 'Why': $r = 0.82$, 'When' vs 'Why': $r = 0.69$). This suggests that responses across all experiments are driven by similar processes, regardless of how surprise is queried.

Within the consistency across experiments, there is variability in how surprising people find different scenes. As can be seen in Fig. 3, there is a large difference in the average ratings for violation scenes ($\chi^2(7) = 104,\ p \approx 0$). Out of the stimuli participants observed, they found scenes where objects are created or destroyed to be the most surprising. Conversely, people consistently find the 'Delay' scenario least surprising, even though there is no way the ball could pass the barrier as quickly as it did without teleporting or an extreme momentary change in velocity.

Timing of surprise. We next examine whether moment-by-moment surprise in the 'When' experiment is consistent across participants, or whether there are individual differences. Because all violation trials within a scenario shared the same timing for when objects are created, start moving, change direction, etc., we aggregated data from all trials within a scenario for additional power.

We used Gaussian mixture models to find clusters in the timing response data (Scrucca, Fop, Murphy, & Raftery, 2016). This analysis considers trials with only one or two noted surprises (98% of responses). We separated responses into single and double surprises, and clustered them separately, assuming that people who found two things surprising are responding differently than people who saw one surprise. The number of clusters was selected to minimize BIC.

A distribution of participants' moment-by-moment response for each scenario is shown in the middle rows of Fig. 4, split by number of key-presses, and color-coded according to the clustering algorithm.[1] In half of the scenarios, participants are remarkably consistent in their responses: in 'Create' they note surprise soon after they first see the new object; in 'Overturn (short)' they note surprise soon after the screen passes through the space the object occupies; in 'Discontinuous (visible)' they note surprise as soon as it is revealed there are two objects; and in 'Block' they note surprise as soon as the object is seen on the other side of the block. In the other half of scenarios, participants show interesting variability suggesting they hold different interpretations for these scenes. Here we find that participants are surprised at different times, or that some participants are surprised once, while others are surprised twice. We discuss these patterns in relation to the explanations from 'Why' below.

Explanations of surprise. We first consider the reliability of participants' descriptions of surprising scenes in the 'Why'

[1] In the 'Create,' 'Delay,' and 'Block' scenarios the clustering algorithm breaks what looks like one group of responses into two. This is due to a mismatch between the Gaussian distribution assumed by the algorithm, and the long tails in the response data.

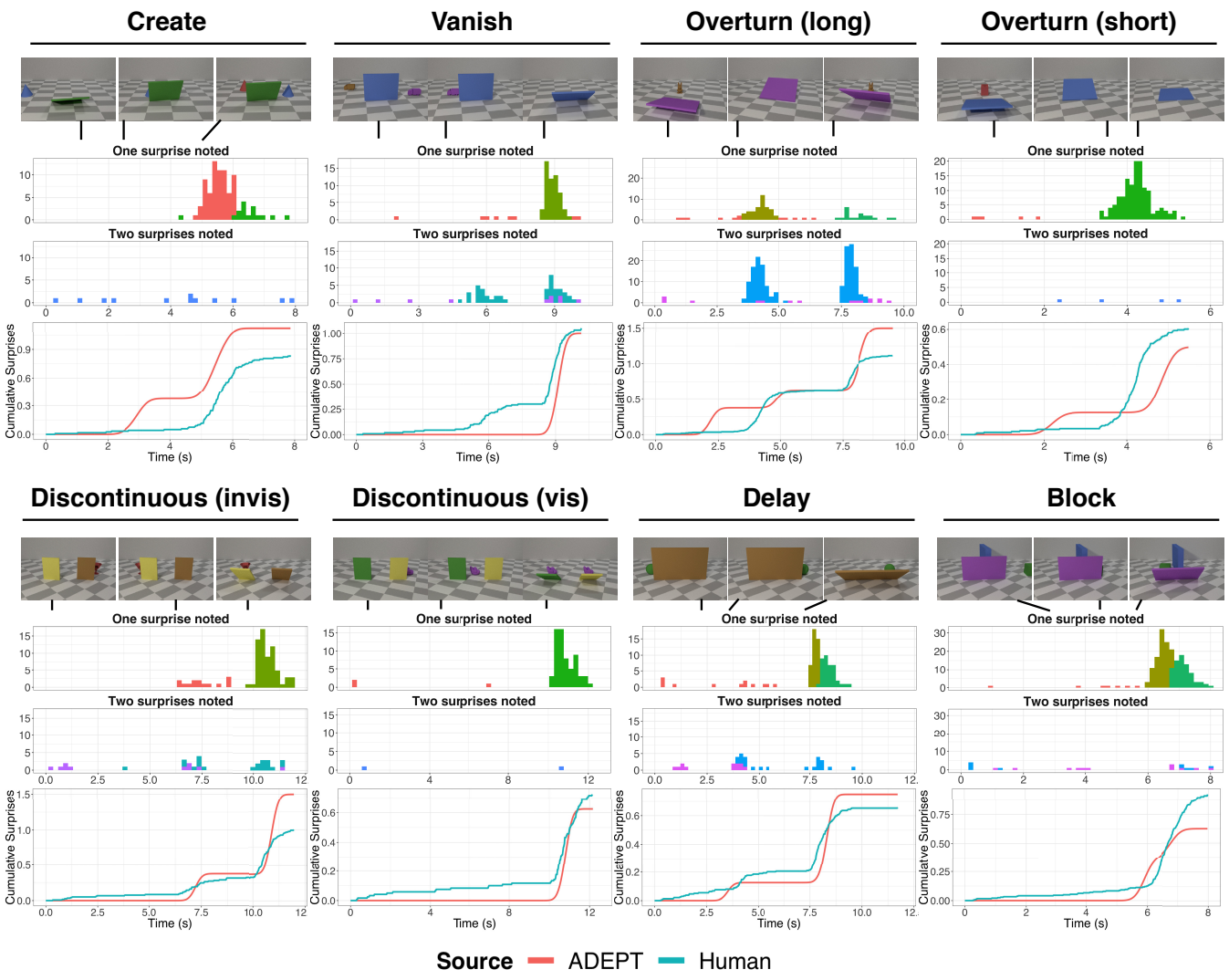

Figure 4: 'When' experiment results. **Top:** frames from violation videos. **Middle:** histogram of timing of surprise, split by participants who noted one surprise (*top*) and those who noted two surprises (*bottom*). Color indicate different clusters of responses. **Bottom:** Average cumulative number of surprises noted by participants (*blue*) and the ADEPT model (*red*).

experiment. Here we again aggregate all violation trials from each scenario, and count the number of responses for which participants marked the exact same set of explanations for that scenario. These explanations are presented in Table 1, including the proportion of surprising trials for which that explanation was endorsed.

Table 1 shows participants were generally consistent in their explanations, with one explanation endorsed in a majority of trials in seven of the eight scenarios. However, there is still some consistent variability across scenarios, suggesting individual differences in interpretation.

Comparison of 'When' and 'Why'. Does the moment-by-moment surprise from 'When' match the ways that those scenarios were explained in 'Why'? In trials that have a single modal pattern of moment-by-moment surprise, explanations typically indicate surprise related to those moments. However, in the scenarios where participants show differences in surprise timings, we find interesting differences between the timing and explanations, suggesting that explanations of surprise are not just based on in-the-moment surprise, but instead require a re-evaluation of the scene.

In the *Overturn (long)* scenario, participants from 'When' were most likely to note surprise twice: first when the screen moved into the object's space, and again when the screen rotated back to show the object again (Fig. 4, blue cluster; 51%). This can only be explained by two distinct surprising events reversing each other: first an object that should be there is not, then the object that was thought to have disappeared is there again. However, the majority of the explanations in 'Why' provide only one source of surprise: either the screen moved through the object, or the object had disappeared (53%) – these participants rarely selected both that an object appeared and an object disappeared (only 10% did, including those that wrote this in with the 'other' op-

Create		**Vanish**		**Overturn (long)**		**Overturn (short)**	
Appears	87%	Disappears	83%	Penetrates	28%	Disappears	60%
Appears & moves	11%			Disappears	25%	Penetrates	16%
				Other	12%	Disappears & penetrates	8%
				Disappears & penetrates	7%		

Discontinuous (invis)		**Discontinuous (vis)**		**Delay**		**Block**	
Teleports	71%	Appears	52%	Teleports	53%	Penetrates	57%
Disappears	13%	Penetrates	10%	Moves	10%	Teleports	16%
		Moves	5%	Other	10%	Teleports & penetrates	12%
				Disappears	8%		

Table 1: Explanations of violation scenes from the 'Why' experiment endorsed in $> 5\%$ of trials, and proportion of time they were selected.

tion). So people were more likely to note multiple surprises in this scenario in 'When' (56%) than in 'Why' (23%; $\chi^2(1) = 53,\ p \approx 0$). This suggests that for many people, what were two in-the-moment surprises could afterwards be re-evaluated more parsimoniously as a single surprising event extended over time.

In the *Discontinuous (invisible)* scenario, people were not more likely to note multiple events in 'When' over 'Why' (18% vs. 13%, $\chi^2(1) = 0.88,\ p = 0.35$), but *how* they interpreted the event differed. Most participants noted a surprise in 'When' only at the very end of the video when the screen came down to reveal that there was no object behind the first screen (Fig. 4 green cluster; 69%), which suggests that they believed that there were two objects through most of the video and were surprised when proven wrong.[2] But participants in 'Why' were most likely to explain this scenario as an object 'teleporting' (71%), which would have had to occur earlier when the object appeared from behind the second screen. So, even though online surprise suggests that people posited two objects and were surprised when one disappeared, participants explained the scene in retrospect as being more likely to be driven by one object teleporting, perhaps because teleportation is seen as easier, or more likely than creating a new object (McCoy & Ullman, 2019). A similar pattern of explanation is seen in the *Delay* scenario, where most people are surprised at the end of the video, but also most people explain the scene as the object teleporting or moving, which would occur earlier than the point indicated by their moment-to-moment surprise.

A model of surprise using "extended physics"

Smith et al. (2019) proposed a model that could track and update beliefs about a scene based on videos like the ones that were shown to participants. This model was designed to formalize core knowledge of physics, inspired by infants' ability to reason about arbitrary objects using principles of permanence, continuity, and solidity (Spelke & Kinzler, 2007). They called this the Approximate Derendering, Extended Physics, and Tracking (ADEPT) model.

Smith et al. (2019) found that ADEPT did not just outperform a set of baseline models on differentiating scenes with physical violations from control scenes, but also did so in a more human-like manner. Here we ask whether this same model can also capture the moment-by-moment response of when people indicate they are surprised.

ADEPT model structure. The ADEPT model consists of two components: the Approximate Derenderer, and the Extended Physics + Tracking module (Fig. 5; see Smith et al. (2019) for further details).

The approximate derenderer (Fig. 5A) is a module that can extract symbolic, object-based information from an image (Wu, Tenenbaum, & Kohli, 2017). Inspired by the fact that infants do not keep good track of object shapes (Xu & Carey, 1996; Ullman, Spelke, Battaglia, & Tenenbaum, 2017), yet can still reason about arbitrary objects, this derenderer treats all shapes as simple geometric bodies (elipsoids or cuboids), throwing away information to generalize to new shapes better. This derenderer is used as a visual front end for ADEPT so that there is no need to make assumptions about when objects are visible, allowing this information to be extracted from the videos themselves.

The extended physics + tracking module (Fig. 5B) is used to update beliefs over time by integrating observation in a way that is consistent with physical dynamics. Similar to how people use an Intuitive Physics Engine to make and update predictions over time (Battaglia, Hamrick, & Tenenbaum, 2013; Yildirim, Smith, Belledonne, Wu, & Tenenbaum, 2018), ADEPT predicts what it expects to see at the next moment, then updates its beliefs about the properties of objects based on how well they match the states of objects observed by the approximate derenderer. ADEPT tracks these beliefs using a particle filter in order to retain a probabilistic distribution over world states, and so also about what it expects to see next.

ADEPT "extends" physics by allowing objects and their properties to be resampled from a prior instead of unfolding according to physical dynamics. In this way, ADEPT can change its beliefs about the position, velocity, or even existence of objectives if it observes something that is otherwise

[2] If participants believed that the original object teleported, they should have been surprised when they saw that object appear behind the other barrier, but the rest of the scene would have been consistent with this belief. This would correspond to the red cluster, which contained only 15% of participants.

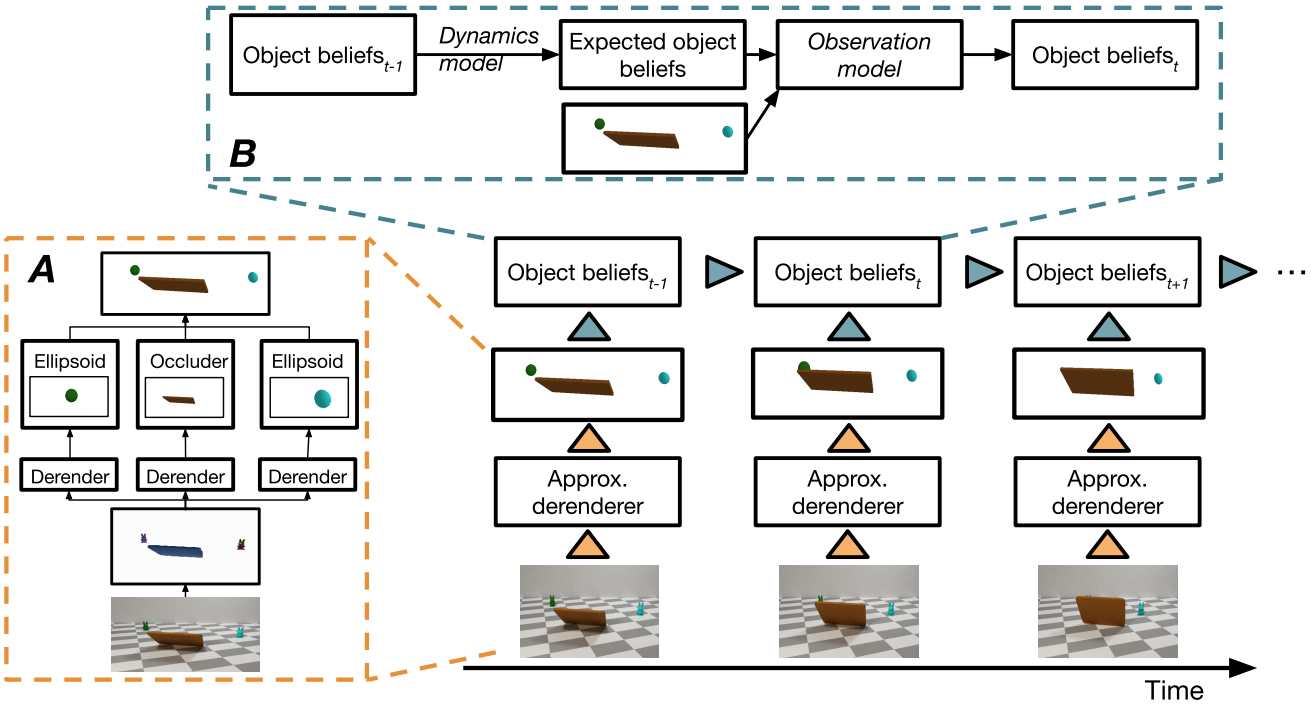

Figure 5: Diagram of the ADEPT model, including the approximate derenderer (*A*), and the extended physics + tracking module (*B*). Figure reprinted from Smith et al. (2019).

inexplicable given its prior beliefs. This resampling only occurs with a very low probability, such that ADEPT does not rely on this component unless there is no other way it could explain the scene.

"Surprise" signals in the ADEPT model. The ADEPT model produces predictions from its dynamics model of what it expects to see, then compares its observations from the approximate derenderer to track and update its beliefs at each point in time. This gives a natural moment-by-moment measure of surprise: the inverse probability of those observations under the dynamics model, including any scene prior resampling required. This surprise signal is formalized as the negative log-likelihood of this probability.

Smith et al. (2019) studied how well ADEPT captured the overall surprise of a scene, and whether it could differentiate scenes with violations from controls. For this purpose they measured the maximum level of the surprise signal over an entire video, and compared the signal generated by violation videos to that from control videos. On this measure, they found that the ADEPT model outperformed baselines that learned dynamics without object representations on overall performance, and was also a closer match to human behavior.

Here we consider whether people and ADEPT use a similar ongoing signal to drive their moment-by-moment surprise. If the signal generated by the ADEPT model is a plausible proxy for the mechanisms underlying human surprise, then we should expect that participants' judgments of *when* they find a scene surprising should be predictable from this surprise signal.

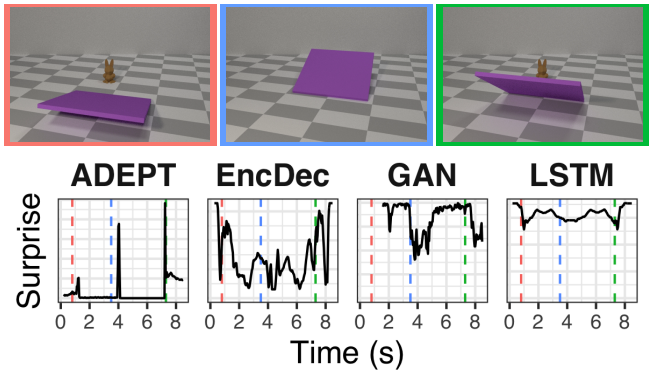

Figure 6: Surprise signal intrinsic to the ADEPT model (left) and baselines over the course of an Overturn (long) trial. Colored dashed lines represent the timing of frames above.

Inspired by how the ADEPT signal produces 'spikes' of surprise signal that are driven by requiring low-probability scene resampling to explain observations (Fig. 6, left), we hypothesize that human surprise might similarly occur when observations suddenly deviate strongly from expectations of an internal dynamics model. We therefore assume that people will note a surprise any time the surprise signal from ADEPT exceeds a set threshold. To get variability in model surprise signals, we ran the ADEPT model multiple times on the same set of violation videos that participants observed, and aggregated signals by scenario in the same way. Finally, because

ADEPT's surprise is produced near-instantaneously upon observing a video frame, whereas people need time to process stimuli and produce a motor response, we shifted these reactions forwards in time (959ms) and smoothed them with a Gaussian kernel. The threshold and time offset were fit to minimize deviations between model predicted surprises and human data.

This model of surprise timing fits participants' behavior well, as can be seen on the bottom rows of Fig. 4. While there are some scenarios in which the ADEPT model misses a human surprise (e.g., the initial stopping in *Vanish*), or others where it is surprised earlier than people (e.g., in *Overturn (long)*), in many cases this model is surprised at roughly the same times and rate as people.

We also consider whether the momentary predictions of the non-object-based baseline models studied in Smith et al. (2019) can explain the timing of human surprise. These baselines include a set of architectures that make predictions directly from pixels, and have been previously used as possible explanations of plausible and implausible physical events (Riochet et al., 2018). These models therefore are tests of how systems that do not have object representations might form predictions of these physical events. However, as can be seen in Fig. 6, these baseline surprise traces do not map cleanly onto the moments people intuitively find surprising; because of this, we were unable to find models using these baselines that did not degenerate into, e.g., never noticing any surprises.

Discussion

Across three experiments, we quantitatively studied how people find scenes with physical violations surprising. We compared overall judgments of surprise, moment-by-moment surprise, and explanations of what is surprising, and found consistency across these three measures. However, digging deeper we found that *what* people find surprising differs based on whether they are reporting moment-by-moment surprise, or explaining a scene in retrospect. Finally, we showed that the ADEPT model – designed to measure overall surprise levels – also explains peoples' moment-by-moment surprise.

The differences between momentary and retrospective surprise suggest that there might be two different types of surprise: one that requires an immediate repair to our beliefs about the world, and one that captures whether we are able to form an explainable set of beliefs about past states – just as we might be shocked when a magician makes a rabbit appear from a hat, but later revise our feelings when we learn that the rabbit was in a hidden compartment all along. The stimuli used here did not differ in whether a scene would be surprising in the moment or in retrospect, just in how it might be explained. However, future work will focus on disentangling these two types of surprise.

Ultimately, we would like to use this work to inform our understanding of the foundations of object knowledge. We asked adults to make judgments about surprise because we are able to measure their responses in more precise ways than we can for infants. However, we hope to use this work to make quantitative predictions about the timing and amount of surprise we expect infants will show for various physical violations, and measure the correlates of this surprise using online measures such as pupil dilation or blink suppression. Finally, we hope to use neuroimaging techniques with high temporal frequency, such as EEG or MEG, to measure surprise in both infants and adults, so that we might test for shared neural representations and foundational representations of objects from infancy to adulthood.

References

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, *20*(3), 191–208.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks Effect: Human Infants Allocate Attention to Visual Sequences That Are Neither Too Simple Nor Too Complex. *PLoS ONE*, *7*(5).

McCoy, J., & Ullman, T. (2019). Judgments of effort for magical violations of intuitive physics. *PLOS ONE*, *14*(5), e0217513.

Noles, N. S., Scholl, B. J., & Mitroff, S. R. (2005). The persistence of object file representations. *Perception & Psychophysics*, *67*(2), 11.

Piloto, L., Weinstein, A., TB, D., Ahuja, A., Mirza, M., Wayne, G., … Botvinick, M. (2018). Probing Physics Knowledge Using Tools from Developmental Psychology. *arXiv:1804.01128 [cs]*. (arXiv: 1804.01128)

Riochet, R., Castro, M. Y., Bernard, M., Lerer, A., Fergus, R., Izard, V., & Dupoux, E. (2018). IntPhys: A Framework and Benchmark for Visual Intuitive Physics Reasoning. *arXiv:1803.07616 [cs]*. (arXiv: 1803.07616)

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R journal*, *8*(1), 289–317.

Smith, K. A., Mei, L., Yao, S., Wu, J., Spelke, E. S., Tenenbaum, J. B., & Ullman, T. D. (2019). Modeling Expectation Violation in Intuitive Physics with Coarse Probabilistic Object Representations. In *33rd Conference on Neural Information Processing Systems*. Vancouver, Canada.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*(4), 605–632.

Spelke, E. S., Kestenbaum, R., Simons, D. J., & Wein, D. (1995). Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British Journal of Developmental Psychology*, *13*(2), 113–142.

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.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure Reasoning in 12-Month-Old Infants as Probabilistic Inference. *Science*, *332*(6033), 1054–1059.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind Games: Game Engines as an Architecture for Intuitive Physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Wu, J., Tenenbaum, J. B., & Kohli, P. (2017). Neural Scene De-rendering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7035–7043). Honolulu, HI: IEEE.

Wynn, K., & Chiang, W.-C. (1998). Limits to Infants' Knowledge of Objects: The Case of Magical Appearance. *Psychological Science*, *9*(6), 448–455.

Xu, F., & Carey, S. (1996). Infants' Metaphysics: The Case of Numerical Identity. *Cognitive Psychology*, *30*(2), 111–153.

Yildirim, I., Smith, K. A., Belledonne, M., Wu, J., & Tenenbaum, J. B. (2018). Neurocomputational Modeling of Human Physical Scene Understanding. In *2018 Conference on Cognitive Computational Neuroscience*. Philadelphia, Pennsylvania, USA: Cognitive Computational Neuroscience.