Electrophysiology reveals that intuitive physics guides visual tracking and working memory

Halely Balaban[1,2], Kevin A. Smith[1], Joshua B. Tenenbaum[1], Tomer D. Ullman[2]

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[2]Department of Psychology, Harvard University, Cambridge, Massachusetts 02138, USA

Abbreviated title: Intuitive physics guides visual object tracking

Corresponding author: Halely Balaban (halely@mit.edu)

Number of pages: 37; **Number of figures:** 5;

Number of tables: 0; **Number of multimedia:** 0; **Number of 3D models:** 0

Number of words for abstract: 249

Number of words for introduction: 646

Number of words for discussion: 1290

Conflict of interest: The authors declare no competing financial interests.

Acknowledgments: This work was supported by the Defense Advanced Research Projects Agency [DARPA] Machine Common Sense program. We thank Roy Luria for his generosity in providing the experimental site.

Starting in early infancy, our perception and predictions are rooted in strong expectations about the behavior of everyday objects. These intuitive physics expectations have been demonstrated in numerous behavioral experiments, showing that even pre-verbal infants are surprised when something impossible happens (e.g., when objects magically appear or disappear). Yet, the online mental processes that underlie physical expectations remain hidden. In two EEG experiments (N=32 total, male and female), people watched short videos like those used in behavioral studies with adults and infants, and more recently in AI benchmarks. Objects moved on a stage, and were briefly hidden behind an occluder, with the scene either unfolding as expected, or violating object permanence (adding or removing an object). We measured the contralateral delay activity, an electrophysiological marker of online processing, to examine participants' working memory (WM), as well as their ability to continuously track the objects in the scene. We found that both types of object permanence violation disrupted tracking, even though violations involved perceptually non-salient events (magical vanishing) or new objects that weren't previously tracked (magical creation). The physical violation caused WM to reset, i.e., discard the original scene representation before it could recover and represent the updated number of items. Providing a physical explanation for the violations (a hole behind the occluder) restored object tracking, and we found evidence that WM went on representing the item that disappeared 'down the hole'. Our results show how intuitive physical expectations shape online representations and form the basis of dynamic object tracking.

Significance statement

People expect ordinary things to behave in ordinary ways. For example, objects should not appear out of thin air, or suddenly disappear. Decades of research have shown even infants are surprised by physically impossible events. Despite many advances made using behavioral studies, the moment-by-moment neural dynamics of physical expectation violation remain uncharted. Making novel use of electrophysiological markers, we reveal the influence of intuitive physics on online scene processing. Violations of object permanence disrupted object tracking, due to expectations about physical outcomes. Working memory quickly recovered, forming modified scene representations based on physical explanations. This work uncovers a fundamental way in which intuitive physics governs everyday cognition, and provides a new neural method for studying central cognitive processes.

Introduction

People have basic expectations about the physical behavior of everyday objects. Objects do not simply disappear and reappear, and solid bodies do not pass through one another or break apart for no reason. Such expectations, known as 'intuitive physics', help us to efficiently perceive, predict, and interact with the world around us (Spelke, 1990; Battaglia et al., 2013; Kubricht et al., 2017). Core physical expectations are present in very early infancy, and shared by many non-human animals (Baillargeon et al., 1985; Wynn, 1992; Xu and Carey, 1996; Cacchione and Krist, 2004). The existence of these expectations has been established over decades, mostly through measures of overt surprise in the face of physically impossible events. More recent studies have used computational and neuroscientific methods to provide insight into the cognitive mechanisms and brain regions involved in intuitive physics (Fischer et al., 2016; Bear et al., 2021; Piloto et al., 2022). Yet despite the fundamental role of intuitive physics in our understanding of the world, its moment-by-moment neural processing remains largely unknown. In this paper, we use EEG methods and behavioral displays to study the unfolding neural dynamics of physical expectations, from representation, to the detection of anomalies, to resolution.

To uncover the ongoing processing of intuitive physics, we adopted a novel approach of targeting working memory (WM). This mental workspace holds information in an active state, ready to be accessed and manipulated (Baddeley, 1992; Luck and Vogel, 1997). WM is involved in both classic memory paradigms, and whenever material has to be held online (Blaser et al., 2000; Carlisle et al., 2011; Tsubomi et al., 2013). We modified a standard WM paradigm, to include stimuli based on classic developmental studies (Wynn,

1992) and recent AI-benchmarks that probe physical expectations in machines (Piloto et al., 2022).

In two experiments, people watched short animations of one or two objects crossing a stage (Figure 1; example videos: https://youtu.be/w_jaIxFD0HU). The scenes either unfolded as expected, or contradicted object permanence by having an object appear or disappear. Using EEG, we recorded scalp electrical activity as people watched the animations, and tested, for the first time, how violations of physical expectations are processed online.

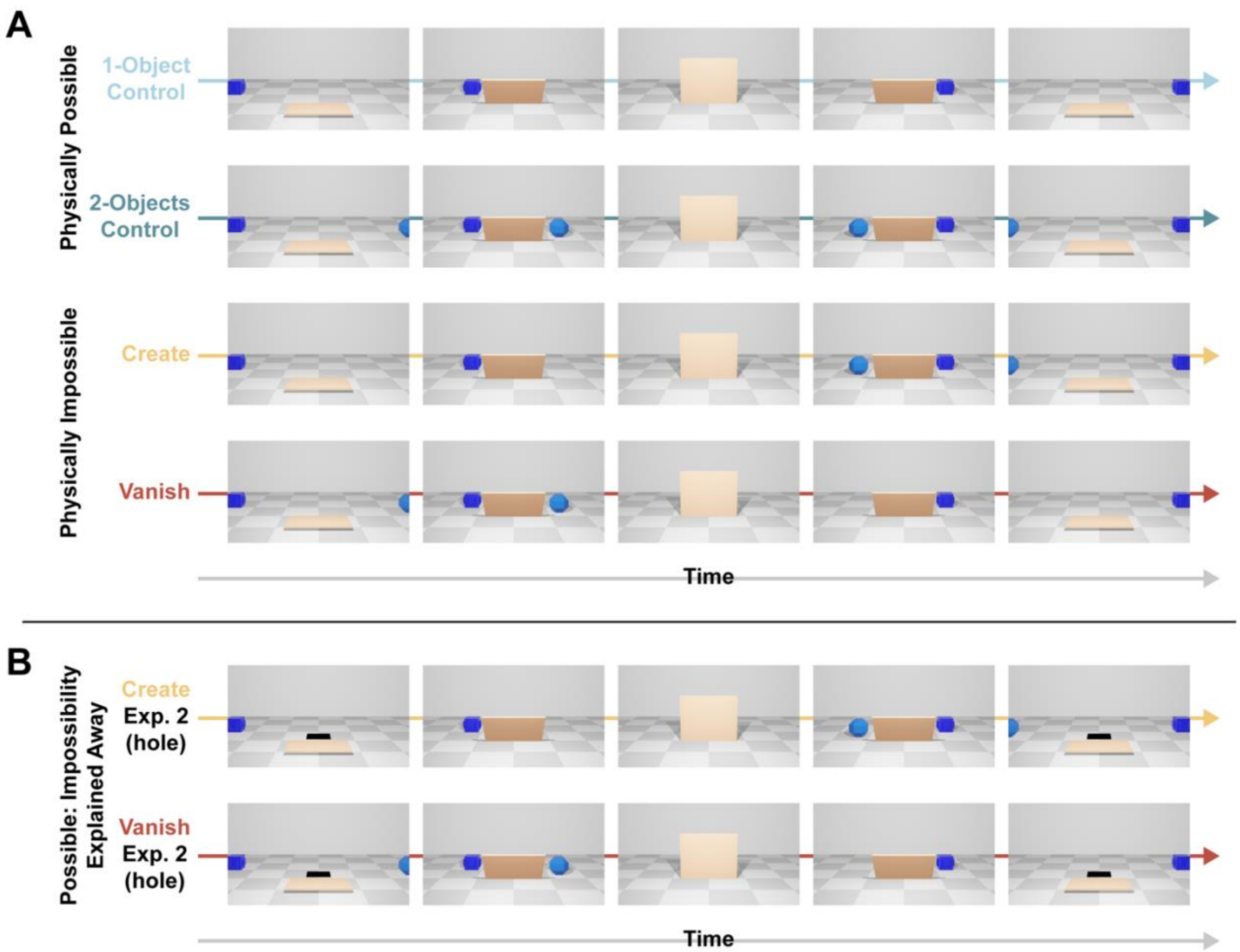

Figure 1. Frames from the animations used as stimuli. (A) Experiment 1's conditions. Top to bottom: 1-Object and 2-Objects Controls, Create, and Vanish. Note that Create starts like 1-Object and ends like 2-Objects, while Vanish is the opposite. (B) Experiment 2's Create and Vanish condition.

To evaluate moment-by-moment active processing, we isolated contralateral delay activity (CDA; see Vogel and Machizawa, 2004; Vogel et al., 2005; Luria et al., 2016), an event-related potential (ERP) index of WM. The CDA's amplitude reflects representational load (while being immune to related factors, for a review see Luria et al., 2016), rising when more items are held online (Figure 2A, top). The CDA also reflects the dynamics of the pointer system, an indexing process connecting online representations with perception. The pointer system implements a one-to-one correspondence through which a specific representation in WM can be accessed and updated when the analogous real-world item changes (Kahneman et al., 1992; Scholl and Pylyshyn, 1999; Pylyshyn, 2000; Balaban and Luria, 2017). If the WM-perception correspondence is disrupted (e.g., if an object splits in two), WM resets (Balaban and Luria, 2017; Balaban et al., 2018a, 2018b, 2019a). Resetting is accompanied by a reliable transient drop in CDA amplitude (Figure 2A, bottom), indicating that an event interrupted the pointer system's tracking ability. The CDA-drop specifically taps into pointer system disruption, and does not reflect related but distinct factors such as general surprise (Balaban and Luria, 2019; see also the Discussion).

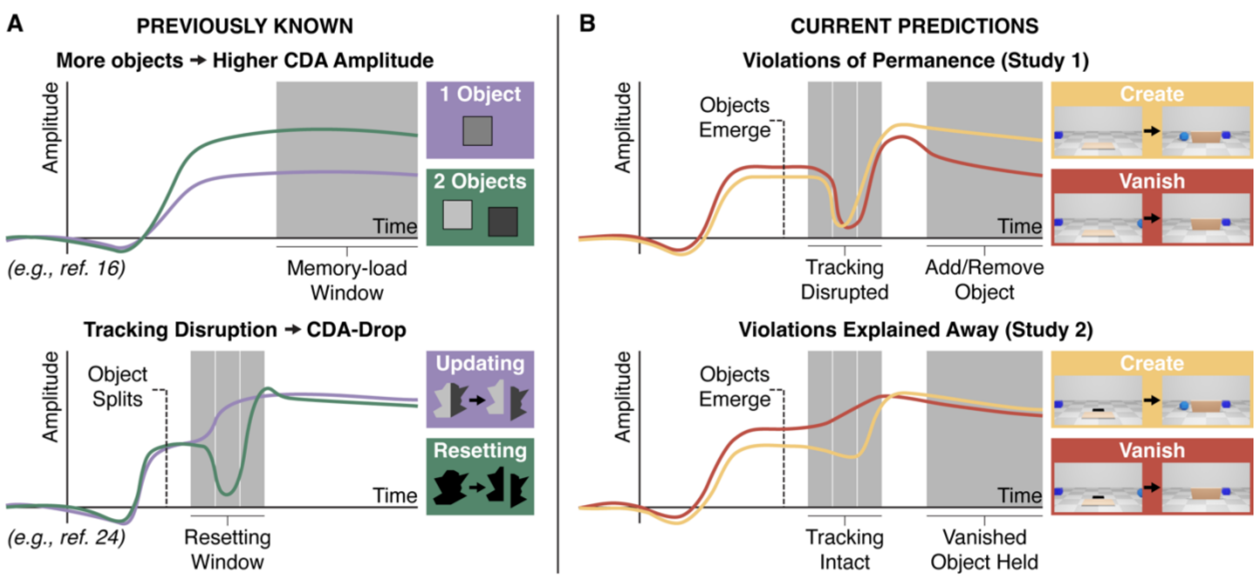

Figure 2. Schematic of past results and current predictions. (A) Previous CDA-based findings regarding online representations (top) and tracking (bottom). (B) Current predictions in Experiment 1 (top) and 2 (bottom).

The excellent temporal resolution of ERPs allowed us to examine different pre-defined time-windows of the CDA amplitude, to uncover evolving mental processing in a fine-grained manner (Figure 2B). Specifically, we tested whether violations of object permanence (Experiment 1) interrupt object tracking, causing a CDA-drop in the previously-established resetting window, followed by the appropriate removal or addition of an object from WM. We also hypothesized that if violations are explained away (Experiment 2), the modified expectations about physical dynamics may prevent disruption, leading to no resetting, and no removal of the vanished object from WM.

Materials and Methods

Data, code, and video examples are available at the Open Science Framework: https://osf.io/csarg.

Participants

Participants were students with normal or corrected-to-normal visual acuity and normal color-vision, who gave informed consent following the procedures of a protocol approved by the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects under protocol 1912000059. Due to the COVID-19 global pandemic, the experiments were run at Tel Aviv University, Israel. Participants were notified of their rights before the experiment, were free to terminate participation at any time, and were compensated monetarily for their time at a rate of $30 an hour.

Each experiment included 16 naïve participants (11 females, mean age 24 in Experiment 1, and 14 females, mean age 24 in Experiment 2). Sample size was determined based on the smallest effect size in the study with the most similar analysis method (Balaban and Luria, 2017), which was d = 0.8, entailing 16 participants for 80% power. Because the experiment required holding fixation and avoiding blinks for a relatively long time (roughly 6 seconds on each trial), participants who could not perform the task under these limitations were released after the first few blocks and replaced (3 in each experiment). Another participant in Experiment 1 was released and replaced for failing to understand the task, and another one was released due to electrode malfunction.

Stimuli and Procedure

In this study, we presented videos that may or may not contain a physical violation, and we measured EEG to examine how the scenes are represented. As is standard, this requires showing two videos – one for each hemisphere – and pre-cuing which side to pay attention to. Using this lateralized display allows defining the contralateral and ipsilateral hemispheres, and subtracting them to get rid of any common perceptual factors. To ensure participants pay attention to the videos and hold objects in WM, after the animations' presentation they were given a task, to pick out one of the objects that they had seen from a set of objects. Figure 3 presents the trial sequence.

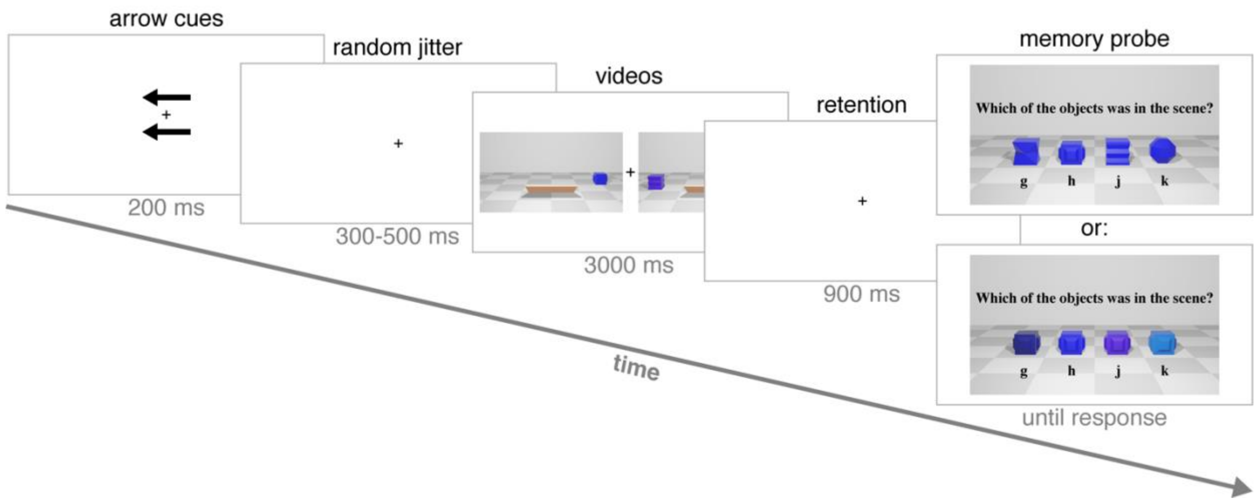

Figure 3. Example of a trial sequence in Experiments 1 and 2. Black arrows indicated to participants which side is relevant for the upcoming trial. The video animations included objects crossing the floor, as an occluder went up and down, briefly hiding them from view. Objects' movement and the occluder were irrelevant to the task. After a retention interval, subjects selected the previously presented object from 4 options, which were either all the same color and varying in shape (a shape test; top), or all the same shape and varying in color (a color test; bottom).

Each trial started with a 1,000 ms display of a fixation black cross 0.7° × 0.7° of visual angle from a viewing distance of approximately 60 cm) in the center of a grey screen (RGB values: 180, 180, 180). Then, two black arrows (3° × 1°) appeared above and below fixation for 200 ms. The arrows pointed either left or right randomly (with an equal probability), and this indicated the side to which participants were asked to attend for the upcoming trial. After another 300-500 ms (randomly jittered) fixation display, a video animation was played on each side of the fixation. Each video spanned about 21° in width and 12° in height, and was placed at a distance of 4.2° from the fixation, and at the middle of the screen's height. The videos on both sides were always from the same condition (see below; from here on, when describing the number of items, we always refer to the relevant side), with all other details drawn randomly (without replacement) independently between the sides. On the main trials, the videos played for 3 second. On catch trials, the videos played for only 300 ms. The goal of these short trials was to ensure participants paid attention not only to the end of the animation. They made up 25% of the trials in the first block, and 10% in all other blocks. Catch trials were not analyzed. After the video animations, a 900 ms retention interval was presented, with only the fixation visible on screen. Then a memory probe appeared, including an image of 4 objects, from which participants chose the one that appeared in the video they saw on the relevant side of the screen. Responses were made via button press (using the "g", "h", "j", and "k" keys on a standard keyboard, based on the objects' location in the image, from left to right). Responses were unspeeded and no feedback was provided. Participants were asked to hold fixation throughout the trial, and to blink only when they press the response key. Participants completed 15 practice trials, followed by 16 experimental blocks of 50 trials each, for a total of 800 trials (including catch trials), and about 180 trials per condition. The experimental session took 2.5-3 hours (including EEG preparation).

Animations were created in Blender, and rendered in the Eavee engine at 24 frames per second. Videos displayed a light-grey and white checkerboard pattern floor, a white back wall, and a rectangular light-orange screen, referred to as the occluder, in the middle of the scene. The occluder started off on the floor, went up until it stood vertically midway through the video, and lowered back down. Additionally, each video included one or two objects crossing the floor from one side to the other, with the occluder hiding them from view for 625 ms. When the animation included two objects, they moved in opposite directions, and one of them was slightly closer to the occluder than the other, so that they would not collide.

There were 4 possible un-namable shapes, and 4 possible colors that were all shades of blue (the exact rendered color changed across the object's surface because of the irregular shapes' shading; RGB values are reported based on one representative area: 45, 45, 142; 60, 60, 230; 105, 60, 220; 45, 130, 210), for a total of 16 possible objects. Each object's movement direction (from right to left or from left to right), color, and shape were determined randomly without replacement in each trial.

There were 4 conditions, varying in the number of objects at the beginning and end of the animation (see Figure 1A). In the two physically possible Control conditions, either one or two objects simply passed behind the occluder (the 1-Object and 2-Objects Controls, respectively). There were two physically impossible conditions: Create, where one object went behind the occluder but two exited, and Vanish, where two objects went behind the occluder but only one exited. Thus, in the Create condition the video's first half was identical to the 1-Object Control and its second half was identical to the 2-Objects Control, and vice versa for the Vanish condition. The Create and Vanish conditions constitute violations of object permanence, which translates to a change in the number of objects, allowing us to leverage the CDA's set-size sensitivity (see below). In the Vanish condition, the item that disappeared was never probed, though participants were not explicitly told of that.

There were only two differences between Experiments 1 and 2. First, in the Create and Vanish conditions of Experiment 2, a small black rectangle was placed right behind the occluder throughout the trial (see Figure 1B). Importantly, this black area was only visible when the occluder was down (i.e., at the beginning and end of the trial). The second difference compared to Experiment 1 was the explanation provided to participants before the experiment started. Participants in Experiment 2 were told that the black area is a hole in the floor, meaning that if the trial starts with two objects, one is going to fall down (the front object), and if the trial starts with one object, one is going to "climb up" from the hole, using a hidden leverage. Participants were also shown example video of the different conditions, and a demonstration of what the Vanish condition would look like if the occluder wasn't there (showing an object falling down the "hole").

EEG Recording and Analysis

EEG was recorded inside a shielded Faraday cage, using a BioSemi ActiveTwo system, from 32 scalp electrodes placed at a subset of the extended 10-20 system's locations, and from two electrodes placed on the mastoids, which served as reference. EOG was recorded from two electrodes placed 1 cm from the external canthi, and from an electrode placed 2 cm beneath the left eye. Data was digitized at 256 Hz.

Offline signal processing was performed using the EEGLAB (Delorme and Makeig, 2004) and ERPLAB (Lopez-Calderon and Luck, 2014) toolboxes, and custom Matlab (The Mathworks, Inc.) scripts. All electrodes were referenced to the average of the left and right mastoid electrodes. Continuous data was segmented into epochs from -200 to +3900 ms from animations' onset (corresponding to the end of the retention interval). Artifact detection was performed on the EOG electrodes using a sliding window peak-to-peak analysis, with a threshold of 80 μV. This procedure resulted in a mean rejection rate of

9.2% in Experiment 1, and 9.5% in Experiment 2 (for evidence that eye movements are not responsible for the CDA or the resetting-drop, see Kang and Woodman, 2014; Balaban and Luria, 2017, 2019; Balaban et al., 2018a). For plotting purposes, the epoched data were low-pass filtered using a noncausal Butterworth filter (12 dB/oct) with a half-amplitude cutoff point at 30 Hz. Statistical analysis was performed on the unfiltered data, to avoid potential effects of filtering on the results.

Epoched data were averaged separately for each condition, and difference waves were calculated by subtracting ipsilateral from contralateral activity, relative to the memorized side on each trial. As was done in previous research (e.g., Balaban and Luria, 2017; Balaban et al., 2018a, 2019a), we focus on the results from the average of 3 parietal-occipital electrode pairs – P7/8, Po3/4, and Po7/8 – but similar patterns of activity were found in each pair separately.

Experimental Design and Statistical Analyses

In order to examine the maintenance and tracking of online representations in WM during a physics-violation task, we isolated the CDA (Vogel and Machizawa, 2004; Vogel et al., 2005; Luria et al., 2016). The CDA is an ERP component reflecting online processing in visual WM. It was first reported in memory tasks, but can be measured equally well when items are held in WM while being completely visible (Tsubomi et al., 2013), like in search or tracking tasks (Drew and Vogel, 2008; Luria and Vogel, 2011b). Numerous studies have shown that CDA amplitude is not sensitive to processes that are related to, but distinct from, WM, such as spatial attention or task difficulty (Vogel and Machizawa, 2004; McCollough et al., 2007; Ikkai et al., 2010; Feldmann-Wüstefeld et al., 2018, and, for a review, see Luria et al., 2016). Similarly, the resetting-drop in CDA amplitude is extremely specific: When the mapping between WM representations and perceptual input is disrupted there is a characteristic drop, whereas extremely similar situations (even within experiment and participants) that allow this mapping to hold lead to a smooth change in amplitude (Balaban and Luria, 2017; Balaban et al., 2018a, 2019a, and, for a review, see Balaban and Luria, 2019). Last though crucial, the CDA-drop does *not* reflect general surprise. Rather, it is the result of a specific disruption to object tracking. The effect persists after many dozens of exposures to the disrupting events, and can also be observed for events that are completely predictable (Balaban et al., 2019b).

Based on prior work, we analyzed the CDA in several pre-defined time-windows (for a similar approach in different contexts, see Luria and Vogel, 2011a; Drew et al., 2012, 2013; Balaban and Luria, 2015; Peterson et al., 2015; Balaban and Luria, 2016a, 2016b). First, to examine tracking, for each condition we compared mean amplitude across two previously-defined time-windows (Balaban and Luria, 2017): the resetting window, 200-300 ms after the critical event, and the pre-resetting baseline window, immediately preceding it, i.e., 100-200 ms after the critical event. Here, the critical event was the moment when items started to emerge from behind the occluder, or, in the Vanish condition, the time in which an item should have emerge but didn't. This happened 1896 ms after the onset of the video.

We additionally examined the window immediately following the resetting window (300-400 ms after the critical event). In previous studies involving object separation, this delay was enough time for the resetting process to finish (e.g., Balaban and Luria, 2017).

We compared this time-window to the pre-resetting baseline time-window, to test whether in our study WM recovers as quickly as in other contexts.

To establish scene reinterpretation, we compared mean amplitude during the retention interval (3200-3900 ms after trial onset; note that the CDA takes about 200 ms to respond, e.g., to initially rise, see Vogel et al., 2005) across the different conditions. Specifically, we examined whether, after the video ended, participants represented the Create and Vanish conditions similarly to 1 object or to 2 objects (i.e., whether the amplitude in each impossible condition is lower than that of the 2-Objects Control, or higher than that of 1-Object Control).

A resetting-drop was established via a within-subjects Analysis of Variance (ANOVA), with Time (pre-resetting baseline vs. resetting, and pre-resetting vs. prolonged resetting) and Condition as independent factors on mean amplitude as the dependent measure. The final representation was examined with a within-subjects ANOVA, with Condition as an independent factor, on mean amplitude during the retention time-window as the dependent measure. Finally, we analyzed behavioral performance in the task with a one-way within-subjects ANOVA with condition as an independent factor on accuracy as a dependent measure. We followed these ANOVAs with planned comparisons (contrasts), the results of which we focus on, for simplicity. We additionally report effect sizes for all statistical comparisons, and 95% confidence intervals (CIs) for the difference between conditions.

Results

Experiment 1: Active representations rely on intuitive physics

In Experiment 1, we tested how object permanence violations affect online maintenance and tracking. Within a WM task, participants (n = 16) watched animations of objects moving across a stage, behind a rising screen, and back out as the screen lowers (Figure 1A). In the Control conditions, the animations proceeded as expected. In the impossible conditions, while the screen was up an object was added (Create) or removed (Vanish), seeming to appear or disappear magically. It is in principle possible that object tracking is sensitive only to low-level visual properties, rather than physical expectations. In this case, the CDA should rise or fall smoothly. However, if events that contradict intuitive physics disrupt the ongoing function of the pointer system, both Create and Vanish should trigger a CDA-drop (resetting).

To test the possible disruption of the pointer system, we compared the CDA amplitude in the resetting time-window (200-300 ms after the critical event, when an item appeared or failed to appear) to the baseline time-window which immediately precedes any potential CDA-drop (Balaban and Luria, 2017). We also compared the baseline to the immediately subsequent time-window (300-400 ms after the critical event). After establishing a significant interaction of Time and Condition ($F(3,45) = 3.31$, $p = 0.028$), we focused our analysis on the violation conditions. We found a CDA-drop for both types of object permanence violation (Create and Vanish), with lower amplitude in the resetting time-window (Figure 4). This effect was marginally significant in the Create condition ($F(1,15) = 4.47$, $p = 0.05$, $d = 0.5$, 95% CI = [0.0, 0.63] μV), and significant in the Vanish condition (F(1,15) = 5.61, p = 0.03, d = 0.6, 95% CI = [0.02, 0.4]). As detailed above, this CDA-drop does *not* reflect a general surprise from some event, but specifically marks a disruption to the ongoing function of the pointer system. So, the results demonstrate that the pointer system depends on intuitive physics for object tracking.

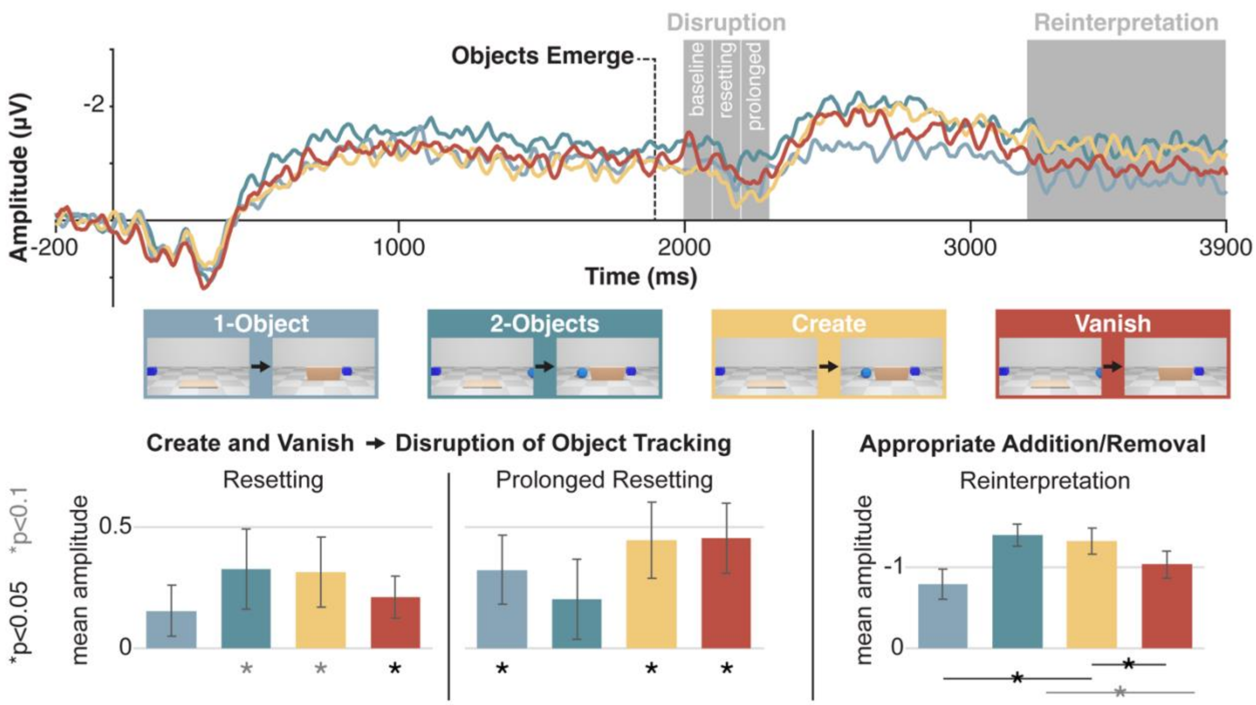

Figure 4. EEG results for Experiment 1. Top: CDA waveforms by condition. Dashed line indicates when objects emerge behind the lowering screen. Analyzed time-windows are in grey. Bottom: Mean amplitude by condition and time-window; error bars show standard error. From left to right: Resetting minus baseline (indicating object tracking disruption), immediately following window minus baseline (indicating prolonged object tracking disruption), and retention interval amplitude (indicating scene reinterpretation, i.e., the number of represented objects at the end of the trial). Asterisks show significant (black)

and marginally significant (grey) contrasts, following the ANOVAs (see Materials and Methods).

Resetting even persisted in the following time-window, with significantly lower amplitudes (as compared with the pre-resetting window) for both the Create ($F(1,15) = 7.67$, $p = 0.01$, $d = 0.7$, 95% CI = [0.1, 0.79]) and Vanish ($F(1,15) = 9.08$, $p = 0.01$, $d = 0.8$, 95% CI = [0.13, 0.78]) conditions. Though we cannot directly compare the current results with experiments that use different stimuli, the prolonged resetting effect suggests that the pointer system took longer to recover from violations of object permanence than what was previously reported for 2D events like separation, in which the same delay was enough for the system to recover (e.g., Balaban et al., 2019a).

In addition to the predicted resetting effects in the violation conditions, we found possible marginal evidence for shorter-lived resetting effects in the Control conditions. In the resetting time-window, there was borderline evidence for a drop in the 2-Objects Control ($F(1, 15) = 3.66$, $p = 0.08$, $d = 0.5$, 95% CI = [-0.04, 0.69]), and no evidence for this effect in the 1-Object Control ($F(1, 15) = 1.97$, $p = 0.18$, $d = 0.4$, 95% CI = [-0.08, 0.39]). In the later time-window, there was no evidence for resetting in the 2-Objects Control ($F(1, 15) = 1.38$, $p = 0.26$, $d = 0.3$, 95% CI = [-0.17, 0.57]), and there was an effect for the 1-Object Control ($F(1, 15) = 4.88$, $p = 0.04$, $d = 0.6$, 95% CI = [-0.01, 0.64]). We believe these effects, if they can be considered that, are artifacts. Beyond the fact that the evidence for them was not strong in this experiment, the results of Experiment 2 (see below) show that a direct replication of the Control conditions result in no effect at all. Still, if one were to try to account for these possible effects, we would suggest that the violations of intuitive physics in the impossible conditions may have led participants to undergo a resetting process in *all* conditions, though more weakly in the Control conditions. In support for this possibility, it has recently been shown that when pointer-disruption events become prevalent, a resetting process will occur in situations that do allow the mapping to hold (Friedman and Luria, 2022).

In addition to tracking disruption, we compared the final CDA amplitude (during the retention interval) of the different conditions, to test how WM representations changed as the scene evolved. We found a significant effect of Condition ($F(3,45) = 4.46$, $p = 0.008$). We expected that by the end of the trial, the CDA amplitude would reflect the updated number of items in the scene. In line with this, a comparison of the different conditions indicated that the CDA amplitude was consistent with an item being added in the Create condition, and removed in the Vanish condition. The CDA amplitude in Create was higher than the 1-Object Control ($F(1,15) = 9.77$, $p = 0.01$, $d = 0.8$. 95% CI = [0.17, 0.9]), and similar to the 2-Objects Control ($F < 1$, $p = 0.76$, $d = 0.1$, 95% CI = [-0.41, 0.55]). The CDA amplitude in the Vanish condition was significantly lower than Create ($F(1,15) = 6.36$, $p = 0.02$, $d = 0.6$, 95% CI = [0.05, 0.58]), marginally lower than the 2-Objects Control ($F(1,15) = 3.94$, $p = 0.07$, $d = 0.5$, 95% CI = [-0.03, 0.8]), and similar to the 1-Object Control ($F(1,15) = 1.81$, $p = 0.2$, $d = 0.3$, 95% CI = [-0.13, 0.57]). These findings indicate that participants successfully and rapidly adjusted their representations after object permanence violations, adequately adding or removing objects from WM.

For completeness, we also examined behavioral performance, and found a significant effect of Condition on accuracy ($F(3,45) = 41.69$, $p < 0.001$). We stress that accuracy in our task does not tap into the ongoing dynamics of representations in WM, given the long presentation time after items exit from behind the occluder. The mean accuracy in the Create condition was lower than in the 1-Object Control ($M = 0.78$ vs. 0.93; $F(1, 15) = 100.8$, $p < 0.001$, $d = 2.5$, 95% CI = [0.12, 0.19]), and similar to the mean accuracy in the 2-Objects Control ($M = 0.79$; $F < 1$, $p = 0.4$, $d = 0.2$, 95% CI = [-0.02, 0.04]). These results match the item load at the end of the trial. The mean accuracy in the Vanish condition was lower than in the 1-Object Control ($M = 0.88$; $F(1, 15) = 31.1$, $p < 0.001$, $d = 1.4$, 95% CI = [0.03, 0.08]), and higher than the 2-Objects Control ($F(1, 15) = 20.8$, $p < 0.001$, $d = 1.1$, 95% CI = [0.05, 0.13]) and Create ($F(1, 15) = 68.4$, $p < 0.001$, $d = 2.1$, 95% CI = [0.08, 0.13]) conditions. So, even though the CDA amplitude indicated a successful removal of the item that disappeared, its presence in the first part of the trial still produced a cost relative to one object.

Experiment 2: Online representations depend on expectations

In Experiment 2, we tested whether changing the interpretation of scenes can affect the moment-by-moment dynamics of object maintenance and tracking. New participants ($n = 16$) watched short animations that were identical to Experiment 1, except that the Create and Vanish conditions included a small black rectangle behind the rising screen (Figure 1B), which was described to participants as a hole that objects could climb up from or fall into. The hole was hidden behind the screen when objects emerged, making the scenes perceptually-identical to Experiment 1 during the critical moments. The two experiments diverged only in terms of participant expectations, as the appearance or disappearance of items could now be explained away (Perez and Feigenson, 2022).

As in the previous experiment, we examined potential disruptions to the pointer system by comparing the resetting time-window, and the subsequent time-window, to the baseline time-window immediately before any CDA-drop, and found a significant interaction of Time and Condition ($F(3,45) = 4.18$, $p = 0.01$). If physical expectations govern online object tracking over and above low-level factors, then making the Create and Vanish scenes physically possible should eliminate the resetting effect. Accordingly, as can be seen in Figure 5, we found that there was no resetting effect for Create nor Vanish (both F's < 1, both p's > 0.4, both d's < 0.2; Create 95% CI = [-0.22, 0.23]; Vanish 95% CI = [-0.11, 0.23]). In the following time-window, there was no resetting effect for Vanish ($F < 1$, $p = 0.9$, $d = 0$, 95% CI = [-0.26, 0.28]), but there was one for Create ($F(1,15) = 4.59$, $p = 0.049$, $d = 0.5$, 95% CI = [0.0, 0.63]). Possibly it was more difficult to explain an appearance, and to predict which exact object will emerge.

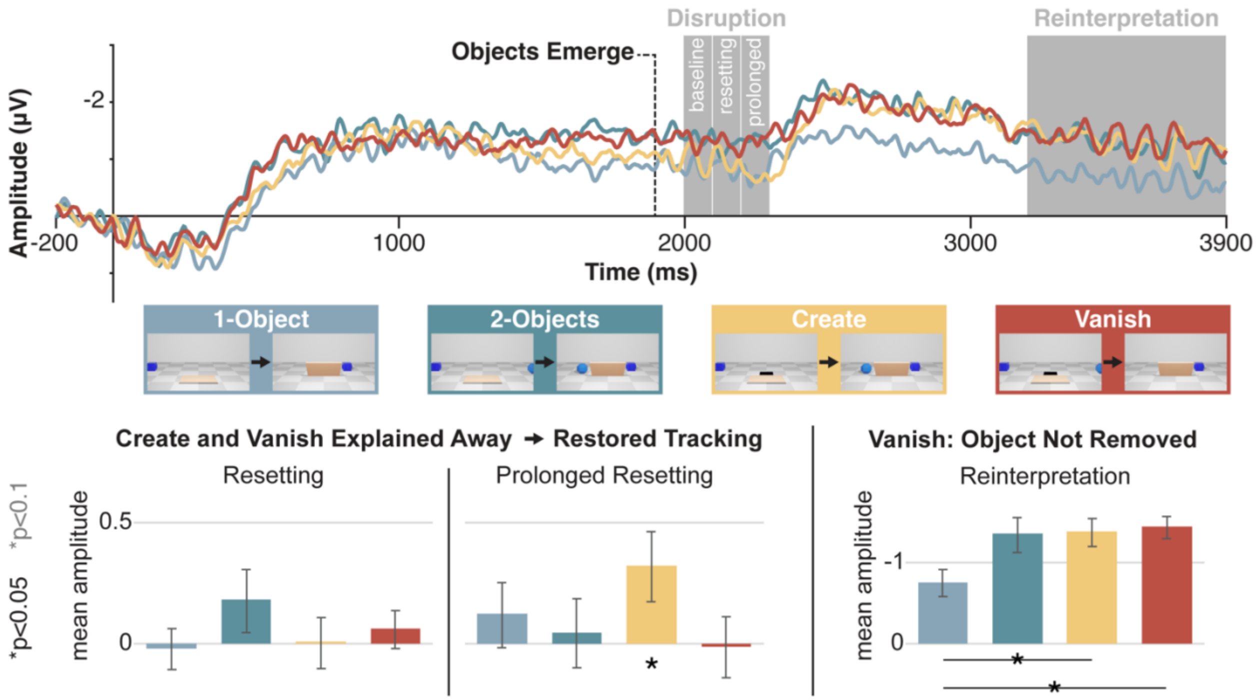

Figure 5. EEG results for Experiment 2. Top: CDA waveforms by condition. Dashed line indicates when objects emerge behind the lowering screen. Analyzed time-windows are in grey. Bottom: Mean amplitude by condition and time-window; error bars show standard error. From left to right: Resetting minus baseline (indicating object tracking disruption), immediately following window minus baseline (indicating prolonged object tracking disruption), and retention interval amplitude (indicating scene reinterpretation, i.e., the number of represented objects at the end of the trial). Asterisks show significant (black) and marginally significant (grey) contrasts.

We next examined whether there was a resetting effect in the Control conditions, which were identical to the Control conditions of Experiment 1. As a reminder, we found some weak evidence for this effect in the Control conditions of Experiment 1, and posited it is likely an artifact. In Experiment 2, we found no drop effect in the Control conditions, in either the resetting time-window (1-Object Control: $F(1, 15) = 1.76$, $p = 0.2$, $d = 0.3$, 95% CI = [-0.11, 0.45]; 2-Objects Control: $F < 1$, $p = 0.8$, $d = 0.1$, 95% CI = [-0.16, 0.2]) or the prolonged resetting window (both F's < 1, both p's > 0.3, both d's < 0.3; 1-Object Control 95% CI = [-0.17, 0.41]; 2-Objects Control 95% CI = [-0.26, 0.35]). Because the Control conditions were identical in Experiment 1 and 2, the non-effect in Experiment 2 suggests that indeed the marginal evidence for resetting in Experiment 1's Control conditions had something to do with the context produced by the impossible conditions, in line with recent work (Friedman and Luria, 2022). This further shows that the resetting results of Experiment 1 do not reflect a condition-general effect (e.g., as the need to track items through occlusion, or an overall reduction in CDA amplitude over time).

As with Experiment 1, we established how the scenes are interpreted after all the events took place by comparing the final CDA amplitude across the different conditions, which resulted in a significant effect of Condition ($F(3,45) = 4.46$, $p = 0.008$). We found that in the Create condition, the final CDA amplitude followed the updated number of objects in the scene, in that it was higher than the 1-Object Control ($F(1,15) = 11.4$, $p = 0.004$, $d = 0.8$, 95% CI = [0.23, 1.02]), and similar to the 2-Objects Control ($F < 1$, $p = 0.9$, $d = 0.0$, 95% CI = [-0.41, 0.47]). Interestingly, in the Vanish condition the amplitude was also significantly above the 1-Object Control ($F(1,15) = 11.57$, $p = 0.004$, $d = 0.9$, 95% CI = [0.26, 1.13]), and similar to the 2-Objects Control ($F < 1$, $p = 0.7$, $d = 0.1$, 95% CI = [-0.4, 0.6]), as well as the Create condition ($F < 1$, $p = 0.7$, $d = 0.1$, 95% CI = [-0.4, 0.6]). This result suggests that participants continued to hold the vanished object in WM, even though, as in Experiment 1, the object disappeared from view, and was never probed during the memory test. The item that 'fell down the hole' is out of view, but still part of the scene. This suggests that online representations are not determined simply by predictability (which would lead participants to never represent the to-be-vanished item), but also by physical explanations.

Again for completeness, we examined people's behavioral performance on the memory task, finding a significant effect of Condition on accuracy ($F(3,45) = 33.89$, $p < 0.001$). We found the same pattern as in Experiment 1. The mean accuracy in the Create condition was lower than in the 1-Object Control ($M = 0.79$ vs. 0.93; $F(1, 15) = 61.47$, $p < 0.001$, $d = 2.0$, 95% CI = [0.09, 0.16]), and similar to the 2-Objects Control ($M = 0.8$; $F(1, 15) = 1.71$, $p = 0.2$, $d = 0.3$, 95% CI = [-0.01, 0.04]). The mean accuracy in the Vanish condition was lower than in the 1-Object Control ($M = 0.88$; $F(1, 15) = 11.58$, $p = 0.004$, $d = 0.9$, 95% CI = [0.02, 0.09]), and higher than the 2-Objects Control ($F(1, 15) = 16.52$, $p = 0.001$, $d = 1.0$, 95% CI = [0.03, 0.1]) and Create ($F(1, 15) = 21.92$, $p < 0.001$, $d = 1.2$, 95% CI = [0.05, 0.12]) conditions. So, while people's behavioral memory performance was the same across Experiment 1 and 2, their online dynamics diverged. This again shows the importance of tools like the CDA in uncovering hidden representational dynamics.

Discussion

Our findings show for the first time how violations of intuitive physics are processed moment-by-moment to support object tracking and WM updating. We presented scenes inspired by developmental studies and AI benchmarks in a WM task, and monitored the neural dynamics of processing physically surprising events. Specifically, we examined the

CDA (Vogel and Machizawa, 2004; McCollough et al., 2007; Luria et al., 2016), an ERP index of online processing, in two different ways. First, the presence or absence of a CDA-drop after a given event indicated whether this event prevented objects in the scene to continue being tracked (Balaban and Luria, 2017, 2019; Balaban et al., 2018a, 2019). Second, comparing the CDA amplitude between different conditions showed how the scenes were represented in WM after each event took place. The novel use of a well-established electrophysiological marker revealed that violations of object permanence disrupt the pointer system's ability to track objects (Experiment 1), due to expectations about physical outcomes (Experiment 2).

The disruption of tracking during violations of physical expectations sheds light on the principles governing the normal function of the pointer system. First, the presence of a resetting effect in the Vanish condition shows that an event does not have to be perceptually salient to disrupt object tracking. Second, the resetting effect in the Create condition suggests that the pointer system is sensitive to the physical aspects of the scene in tracking objects: The new object's appearance is only "impossible" in the sense that it should not have appeared in a place that was previously empty (note that the mere appearance of a new object in a scene does not in itself trigger resetting; (Balaban and Luria, 2017). Taken together, these two effects suggest that in maintaining a continuous correspondence between the perceptual input and the active representations in WM, the pointer system is not solely driven by low-level factors, as is usually assumed in discussions of object tracking (for a review of this subject, see Holcombe, 2023). While the pointer system obviously relies on visual input, the present results show that it also incorporates the physical interpretation of visual events (Lau and Brady, 2020).

Our findings suggest that the brain's tracking system predicts an object's future location based on intuitive physics, perhaps via noisy quantitative physical simulation (Battaglia et al., 2013; Ullman et al., 2017; Smith et al., 2019). Our results help explain numerous separate past findings, which showed pointer system resetting following violations of object separation (Balaban and Luria, 2017; Balaban et al., 2018a, 2018b, 2019a), object replacement (Balaban and Luria, 2017; Friedman and Luria, 2022), and feature switching (Park et al., 2020). While these situations were not originally explained in such a way, they violate cohesion, object permanence, and kind-identity, respectively.

The CDA further allowed us to decipher people's flexible interpretation of events, and how it changes to fit the inferred meaning of the unfolding scene. We found that WM recovers after resetting within the time frame examined, and that scene representations are correctly adjusted to add or remove objects (Experiment 1), based not only on what is available in perception, but also on the physical explanations of events (Experiment 2). This highlights another way in which common sense physical understanding shapes WM's online representations, over and above low-level visual properties.

The results of our experiments do not reflect surprise itself, but rather a disruption of the pointer system. We found this system to be sensitive to physical violations, which could feed downstream to a surprise signal. The experiments also show that the disruption of the pointer system can be mitigated by expectations, but that this mitigation depends on the format of the expectations. In Experiment 1, participants saw the stimuli dozens of times, which forms the overall statistical expectation that in principle objects can sometimes disappear or appear in the videos. This statistical expectation did not prevent a robust and repeated disruption of tracking. By contrast, the expectations in Experiment 2 were successful in eliminating the disruption. But, these expectations had a perceptual and causal format tied to a specific stimulus: the hole in the ground explained causally why on this trial an object might disappear or appear, and did so in the same visual format as the rest of the stimuli. Past studies of resetting in the pointer system support the distinction between expectation formats: Numerous exposures to a disruptive event did not eliminate the disruption, and even making a disruptive event perfectly predictable failed to do so (Balaban et al., 2019b), but subtle visual cues that altered the way the scene was carved into distinct objects, thereby changing the targets for tracking, were successful right away (e.g., Balaban et al., 2019a). Future work can adopt the present approach to further examine how the pointer system communicates with other processes, by looking at what other physical situations and expectations disrupt or preserve object tracking, including perceptually weak but causally specific expectations such as telling participants the rising screen acts as a magic box on specific trials. While more work is needed, our conclusion is not just that the pointer system's ability to track objects is based on physical reasoning, but that the format of this physical reasoning matters.

Our findings introduce a novel way of tackling both new and outstanding open questions. Identifying different cognitive processes from task performance measures alone (e.g., accuracy or reaction time) can be challenging, as they may reflect the sum of all processing stages. While behavioral performance is useful and important, apparent task effects in online processing tasks might not reflect WM or object tracking, but sub-stages that precede or follow them, such as perception or decision making (Awh et al., 2007). This may be why measures of performance on WM or tracking tasks can show mixed results with regards to intuitive physics (e.g., vanMarle and Scholl, 2003; Mitroff et al., 2004;

Franconeri et al., 2012). In contrast, the CDA factors out contributions of perceptual facets (Ikkai et al., 2010; Luria et al., 2010; Ye et al., 2014), by using a bilateral paradigm. It also factors out response-related facets, by focusing the analysis on the period preceding the test phase, meaning before participants can initiate any response. The step-by-step neural analysis of online processing, tracking, and representations allows researchers to address the different sub-stages of intuitive physics that underlie overt surprise, and can tease apart generic surprise from the violation of intuitive physics more specifically. A developmentally appropriate version of measuring a CDA-like pattern in infants can also potentially establish that infant overt surprise in the face of physically impossible events reflects disrupted physics-based object tracking. This can be expanded to other cases that robustly show behavioral surprise based on violations of expectations. In the physical domain, some events are surprising without interfering with tracking, such as a mug falling off a table and bouncing, which could reveal that it is made out of rubber, which is physically surprising but does not interfere with tracking. In the psychological domain, both adults and young children are surprised when agents act in an inefficient way (e.g., Baillargeon et al., 2016). Nevertheless, we would predict such surprise would not lead to resetting, as this violation does not interfere with object tracking.

More generally, our method can shed light on how people perceive, track, understand, and remember physical events, by bridging traditional cognitive science, developmental psychology, and neuroscience perspectives. We see great promise in hybrid approaches that, as we do here, use simplified but ecologically foundational stimuli based on infant studies, while measuring well-characterized neural markers resting on years of rigorous quantitative validation from carefully controlled psychological experiments with adults. Examining the CDA in this novel manner allowed us to reveal previously-hidden influences of intuitive physics on everyday representation, understanding, and reasoning.

References

Awh E, Barton B, Vogel EK (2007) Visual Working Memory Represents a Fixed Number of Items Regardless of Complexity. Psychol Sci 18:622–628.

Baddeley A (1992) Working Memory. Science 255:556–559.

Baillargeon R, Scott RM, Bian L (2016) Psychological Reasoning in Infancy. Annu Rev Psychol 67:159–186.

Baillargeon R, Spelke ES, Wasserman S (1985) Object permanence in five-month-old infants. Cognition 20:191–208.

Balaban H, Drew T, Luria R (2018a) Delineating resetting and updating in visual working memory based on the object-to-representation correspondence. Neuropsychologia 113:85–94.

Balaban H, Drew T, Luria R (2018b) Visual working memory can selectively reset a subset of its representations. Psychon Bull Rev 25:1877–1883.

Balaban H, Drew T, Luria R (2019a) Neural evidence for an object-based pointer system underlying working memory. Cortex 119:362–372.

Balaban H, Drew T, Luria R (2019b) Neural evidence for a dissociation between the pointer system and the representations of visual working memory. Journal of Vision 19:82c.

Balaban H, Luria R (2015) The number of objects determines visual working memory capacity allocation for complex items. NeuroImage 119:54–62.

Balaban H, Luria R (2016a) Integration of Distinct Objects in Visual Working Memory Depends on Strong Objecthood Cues Even for Different-Dimension Conjunctions. Cereb Cortex 26:2093–2104.

Balaban H, Luria R (2016b) Object representations in visual working memory change according to the task context. Cortex 81:1–13.

Balaban H, Luria R (2017) Neural and Behavioral Evidence for an Online Resetting Process in Visual Working Memory. J Neurosci 37:1225–1239.

Balaban H, Luria R (2019) Using the Contralateral Delay Activity to Study Online Processing of Items Still Within View. In: Spatial Learning and Attention Guidance (Pollmann S, ed), pp 107–128 Neuromethods. New York, NY: Springer US.

Battaglia PW, Hamrick JB, Tenenbaum JB (2013) Simulation as an engine of physical scene understanding. Proc Natl Acad Sci USA 110:18327–18332.

Bear DM, Wang E, Mrowca D, Binder FJ, Tung H-YF, Pramod RT, Holdaway C, Tao S, Smith K, Sun F-Y, Fei-Fei L, Kanwisher N, Tenenbaum JB, Yamins DLK, Fan JE (2021) Physion: Evaluating Physical Prediction from Vision in Humans and Machines. Available at: https://arxiv.org/abs/2106.08261.

Blaser E, Pylyshyn ZW, Holcombe AO (2000) Tracking an object through feature space. Nature 408:196–199.

Cacchione T, Krist H (2004) Recognizing Impossible Object Relations: Intuitions About Support in Chimpanzees (Pan troglodytes). Journal of Comparative Psychology 118:140–148.

Carlisle NB, Arita JT, Pardo D, Woodman GF (2011) Attentional Templates in Visual Working Memory. J Neurosci 31:9315–9322.

Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods 134:9–21.

Drew T, Horowitz TS, Vogel EK (2013) Swapping or dropping? Electrophysiological measures of difficulty during multiple object tracking. Cognition 126:213–223.

Drew T, Horowitz TS, Wolfe JM, Vogel EK (2012) Neural Measures of Dynamic Changes in Attentive Tracking Load. Journal of Cognitive Neuroscience 24:440–450.

Drew T, Vogel EK (2008) Neural Measures of Individual Differences in Selecting and Tracking Multiple Moving Objects. J Neurosci 28:4183–4191.

Feldmann-Wüstefeld T, Vogel EK, Awh E (2018) Contralateral Delay Activity Indexes Working Memory Storage, Not the Current Focus of Spatial Attention. Journal of Cognitive Neuroscience 30:1185–1196.

Fischer J, Mikhael JG, Tenenbaum JB, Kanwisher N (2016) Functional neuroanatomy of intuitive physical inference. Proc Natl Acad Sci USA 113.

Franconeri SL, Pylyshyn ZW, Scholl BJ (2012) A simple proximity heuristic allows tracking of multiple objects through occlusion. Atten Percept Psychophys 74:691–702.

Friedman S, Luria R (2022) Visual working memory adaptability: What more can we learn about updating and resetting of visual working memory representations? Journal of Vision 22:3465–3465.

Holcombe A (2023) Attending to Moving Objects, 1st ed. Cambridge University Press.

Ikkai A, McCollough AW, Vogel EK (2010) Contralateral Delay Activity Provides a Neural Measure of the Number of Representations in Visual Working Memory. Journal of Neurophysiology 103:1963–1968.

Kahneman D, Treisman A, Gibbs BJ (1992) The reviewing of object files: Object-specific integration of information. Cognitive Psychology 24:175–219.

Kang M-S, Woodman GF (2014) The neurophysiological index of visual working memory maintenance is not due to load dependent eye movements. Neuropsychologia 56:63–72.

Kubricht JR, Holyoak KJ, Lu H (2017) Intuitive Physics: Current Research and Controversies. Trends in Cognitive Sciences 21:749–759.

Lau JS-H, Brady TF (2020) Noisy perceptual expectations: Multiple object tracking benefits when objects obey features of realistic physics. Journal of Experimental Psychology: Human Perception and Performance 46:1280–1300.

Lopez-Calderon J, Luck SJ (2014) ERPLAB: an open-source toolbox for the analysis of event-related potentials. Frontiers in Human Neuroscience 8.

Luck SJ, Vogel EK (1997) The capacity of visual working memory for features and conjunctions. Nature 390:279–281.

Luria R, Balaban H, Awh E, Vogel EK (2016) The contralateral delay activity as a neural measure of visual working memory. Neuroscience & Biobehavioral Reviews 62:100–108.

Luria R, Sessa P, Gotler A, Jolicœur P, Dell'Acqua R (2010) Visual Short-term Memory Capacity for Simple and Complex Objects. Journal of Cognitive Neuroscience 22:496–512.

Luria R, Vogel EK (2011a) Visual Search Demands Dictate Reliance on Working Memory Storage. Journal of Neuroscience 31:6199–6207.

Luria R, Vogel EK (2011b) Shape and color conjunction stimuli are represented as bound objects in visual working memory. Neuropsychologia 49:1632–1639.

McCollough AW, Machizawa MG, Vogel EK (2007) Electrophysiological Measures of Maintaining Representations in Visual Working Memory. Cortex 43:77–94.

Mitroff SR, Scholl BJ, Wynn K (2004) Divide and conquer: How object files adapt when a persisting object splits into two. Psychological Science 15:420–425.

Park B, B. Walther D, Fukuda K (2020) Dynamic Representations in Visual Working Memory. Journal of Vision 20:900.

Perez J, Feigenson L (2022) Violations of expectation trigger infants to search for explanations. Cognition 218:104942.

Peterson DJ, Gözenman F, Arciniega H, Berryhill ME (2015) Contralateral delay activity tracks the influence of Gestalt grouping principles on active visual working memory representations. Atten Percept Psychophys 77:2270–2283.

Piloto LS, Weinstein A, Battaglia P, Botvinick M (2022) Intuitive physics learning in a deep-learning model inspired by developmental psychology. Nat Hum Behav 6:1257–1267.

Pylyshyn ZW (2000) Situating vision in the world. Trends in Cognitive Sciences 4:197–207.

Scholl BJ, Pylyshyn ZW (1999) Tracking Multiple Items Through Occlusion: Clues to Visual Objecthood. Cognitive Psychology 38:259–290.

Smith K, Mei L, Yao S, Wu J, Spelke E, Tenenbaum J, Ullman T (2019) Modeling Expectation Violation in Intuitive Physics with Coarse Probabilistic Object Representations. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Spelke ES (1990) Principles of object perception. Cognitive Science 14:29–56.

Tsubomi H, Fukuda K, Watanabe K, Vogel EK (2013) Neural Limits to Representing Objects Still within View. J Neurosci 33:8257–8263.

Ullman TD, Spelke E, Battaglia P, Tenenbaum JB (2017) Mind Games: Game Engines as an Architecture for Intuitive Physics. Trends in Cognitive Sciences 21:649–665.

vanMarle K, Scholl BJ (2003) Attentive Tracking of Objects Versus Substances. Psychol Sci 14:498–504.

Vogel EK, Machizawa MG (2004) Neural activity predicts individual differences in visual working memory capacity. Nature 428:748–751.

Vogel EK, McCollough AW, Machizawa MG (2005) Neural measures reveal individual differences in controlling access to working memory. Nature 438:500–503.

Wynn K (1992) Addition and subtraction by human infants. Nature 358:749–750.

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

Ye C, Zhang L, Liu T, Li H, Liu Q (2014) Visual Working Memory Capacity for Color Is Independent of Representation Resolution Martinez LM, ed. PLoS ONE 9:e91681.