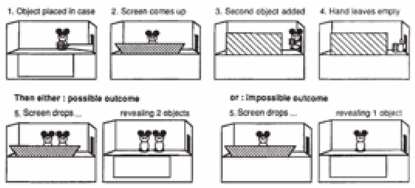
The fine structure of surprise in intuitive physics: when, why, and how much?

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]

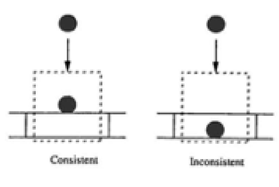
1: MIT BCS, 2: CBMM, 3: MIT CSAIL, 4: Princeton Computer Science, 5: Stanford Computer Science, 6: Harvard Psychology

Introduction

- Surprise when objects/events violate physics is used to understand physical expectations
- Often relies on binary measures: surprising or not
- We study this surprise in a fine-grained manner:
  - *How* surprising is a scene?
  - *When* do people register a surpirse?
  - *Why* do people find a scene surprising?



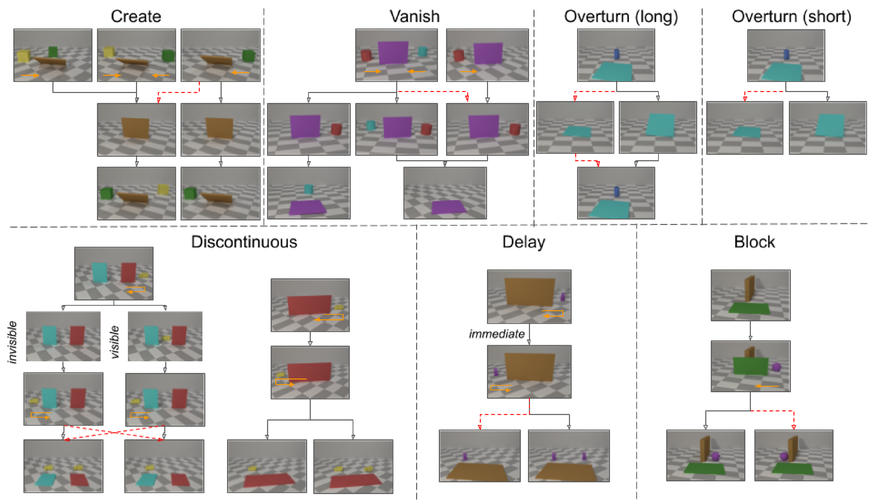
Wynn (1992)



Spelke et al. (1992)

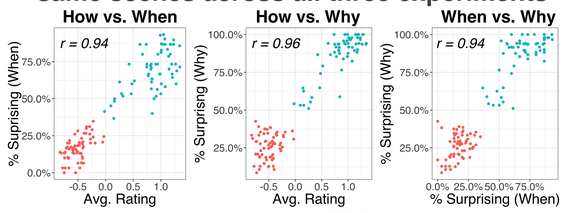
Experiments

- Used Violation of Expectations: register surprise differences between matched scenes with/without physics violation
- Eight types of violations inspired by developmental studies, taken from Smith, Mei, Yao, et al., (2019); measures expectations about object permanence, solidity, & continuity
- Three experiments for three measures of surprise:
  - *How*: rate surprise on a sliding scale (*n=60*)
  - *When*: push a button when surprising event is noticed (*n=60*)
  - *Why*: choose from a list of descriptions of what occured (*n=95*)

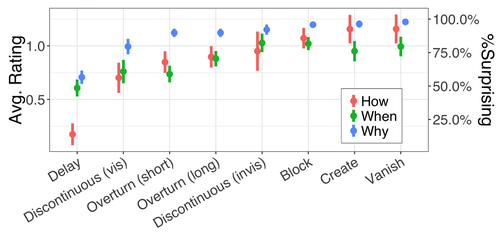


Consistency of ratings

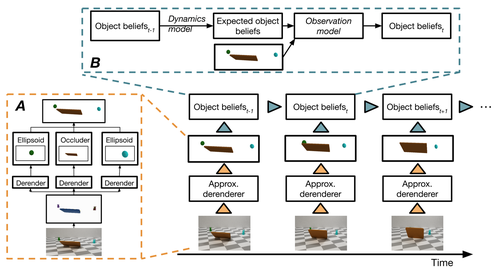
Participants rated surprise at similar rates to the same scenes across all three experiments



Ordering of surprise to different violations is consistent



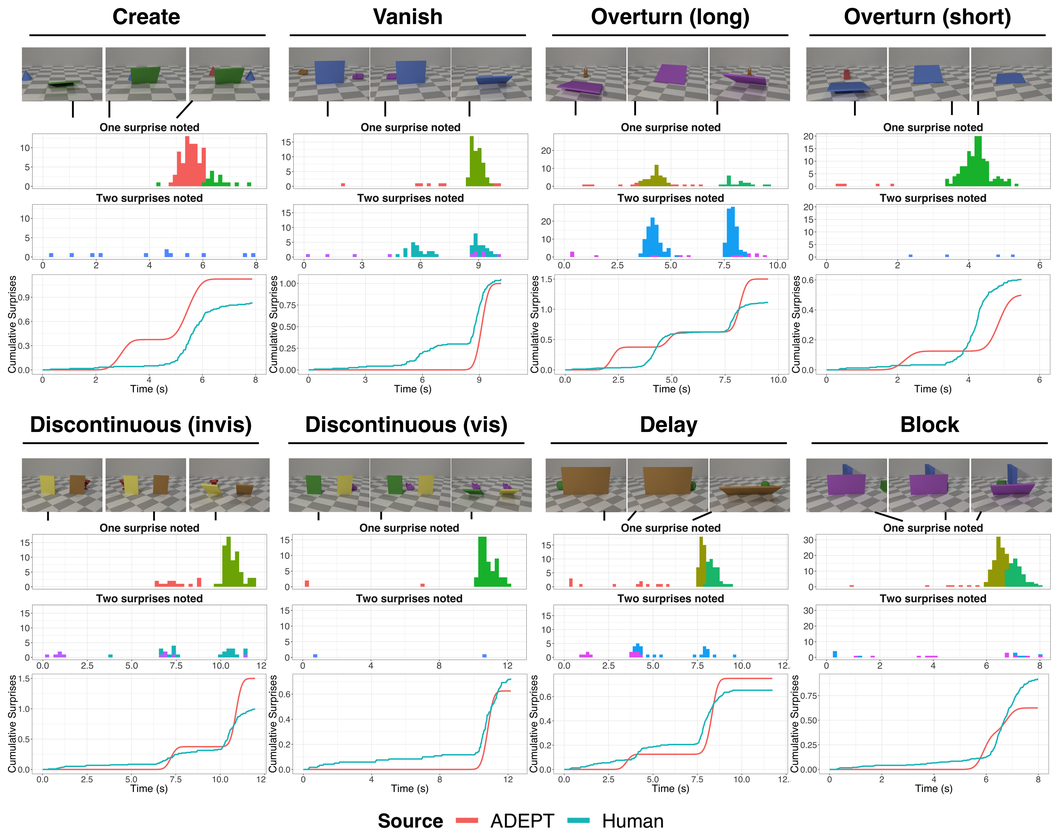
ADEPT model of surprise



Smith, Mei, Yao, et al. (2019)

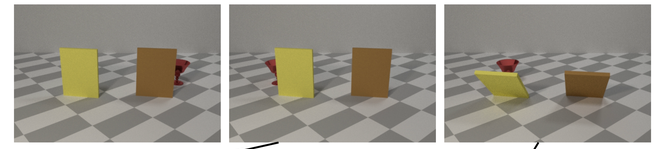
Measuring *when*

- Consistency across subjects; some variability suggests different interpretations
- Timing responses are explained well by the ADEPT model designed to interpret scenes



When vs. *why*

- Explanations endorsed in *why*, implied by timing in *when* often differ
- E.g., in Discontinuous (invisible):



Possible surprise: the object teleported to the other screen. Surprise endorsed by 71% in *why*, but only 15% of participants in *when* (red cluster)

Possible surprise: initial belief that there are two identical objects, surprised when one is seen. Surprise endorsed by only 13% in *why*, but 69% of participants in *when* (green cluster)

- Suggests re-evaluation of explanations

Discussion

- Surprise for physical violations is mostly consistent no matter how measured... but differences in how surprise is explained
- Consistent with a model of physics understanding relying on probabilistic reasoning
- Future directions: reconciling explanations with behavioral / cog-neuro methods