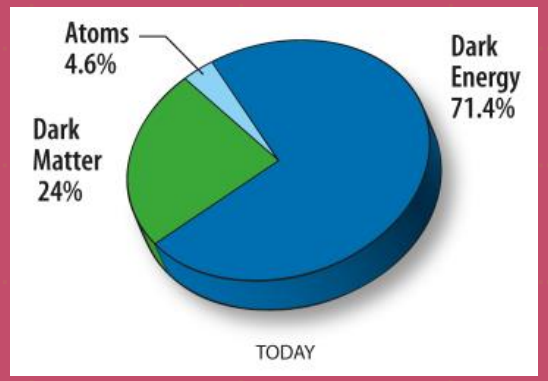
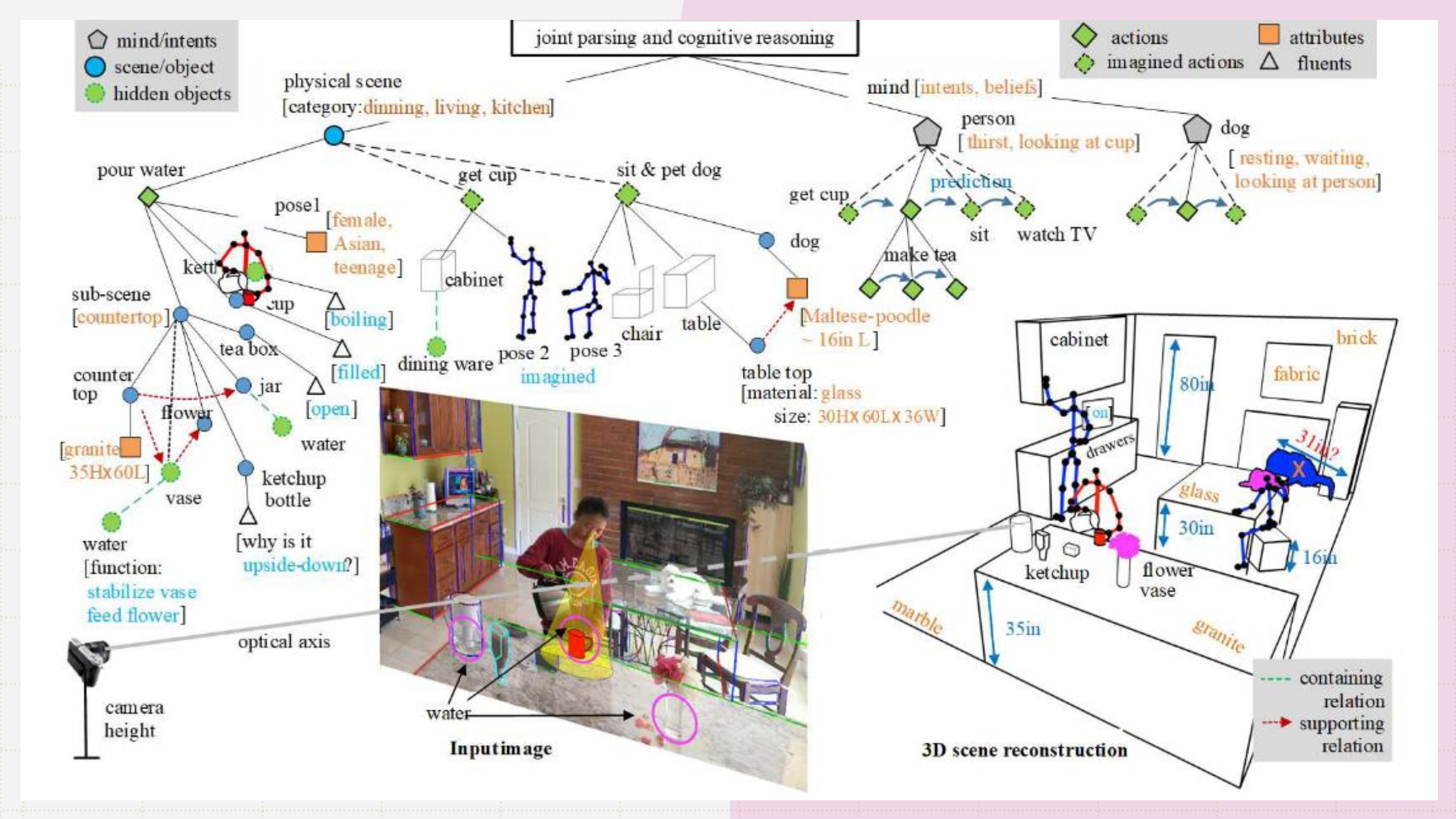
Dark, Beyond Deep: A Paradigm Shift to Cognitive AI with Humanlike Common Sense

Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, Joshua B. Tenenbaum, Song-Chun Zhu




mind/intents
scene/object
hidden objects
joint parsing and cognitive reasoning
actions
imagined actions
attributes
fluents
physical scene
[category: dinning, living, kitchen]
mind [intents, beliefs]
person
[thirst, looking at cup]
dog
[resting, waiting, looking at person]
prediction
get cup
sit
watch TV
make tea
pour water
get cup
sit & pet dog
pose1
[female, Asian, teenage]
kettle
cup
[boiling]
sub-scene
[countertop]
tea box
[filled]
counter top
jar
[open]
flower
water
[granite 35Hx60L]
vase
ketchup bottle
water
[function: stabilize vase feed flower]
[why is it upside-down?]
cabinet
dining ware
pose 2
imagined
pose 3
chair
table
dog
[Maltese-poodle ~ 16in L]
table top
[material: glass
size: 30Hx60Lx36W]
optical axis
camera height
water
Input image
cabinet
brick
fabric
80in
[on]
drawers
31in?
glass
30in
16in
ketchup
flower vase
marble
35in
granite
3D scene reconstruction
containing relation
supporting relation

Paradigm shift

- 1. What, Where?
- Data-driven -> task-driven vision
- 3D reconstruction
- 2. How and Why?
- Visual commonsense:
- **F**unctionality,
- **P**hysics,
- **I**ntentionality,
- **C**ausality,
- **U**tility

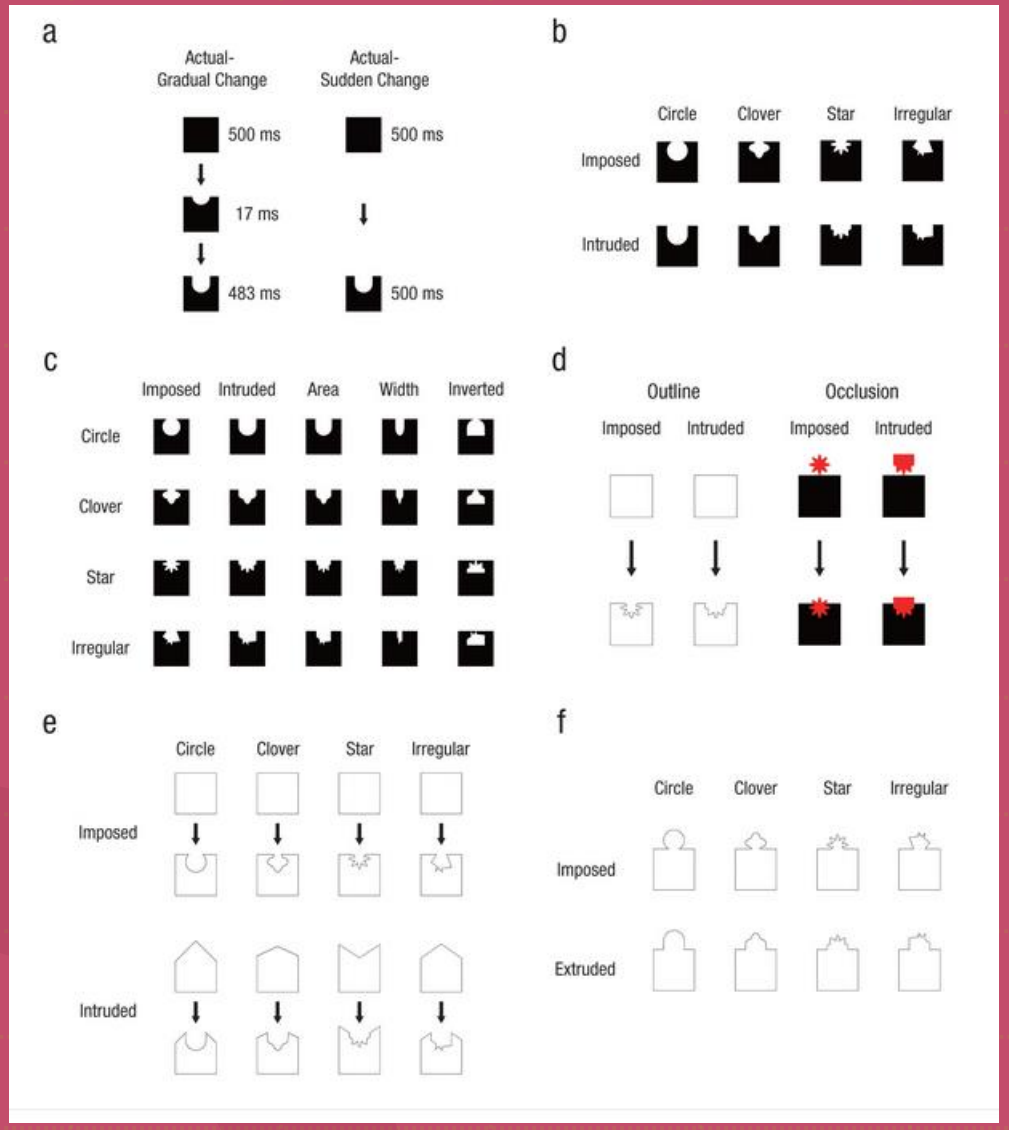
Benefits:
Small sample learning
Generalization through reusable schemas
Bidirectional, continual inference

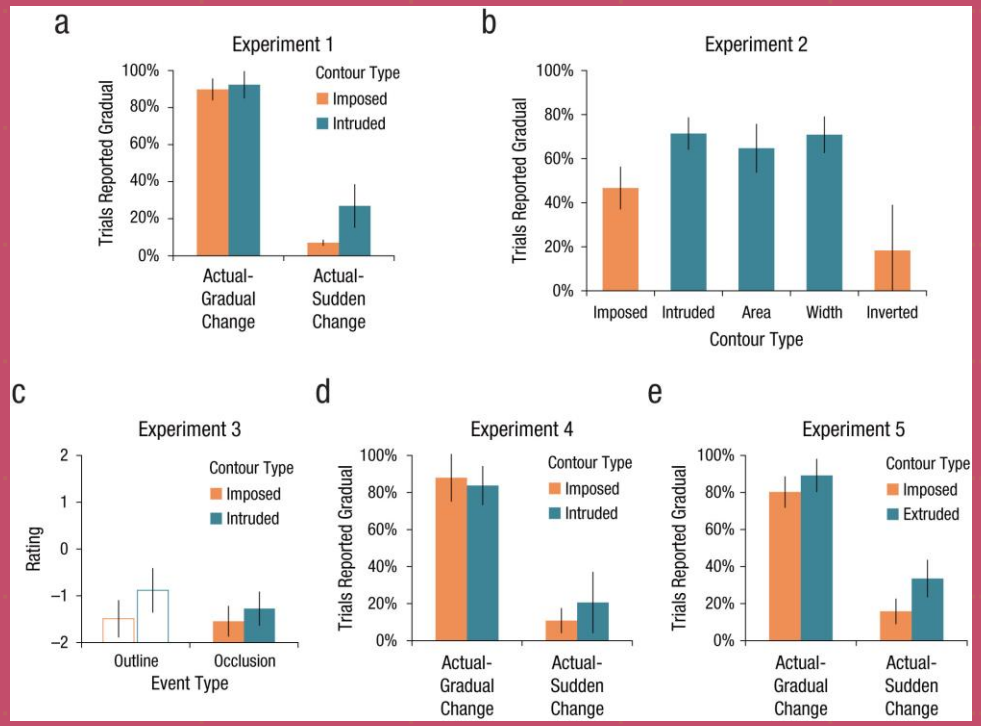
Michotte experiments

Perceptual causality

- We perceive/reconstruct the causal history of input stimuli

Experimental evidence





Chen & Scholl, 2016

Counterfactual perception

- Causal model of the world, or scaffolding?

Gerstenberg et al., 2017

Causal And-Or Graph

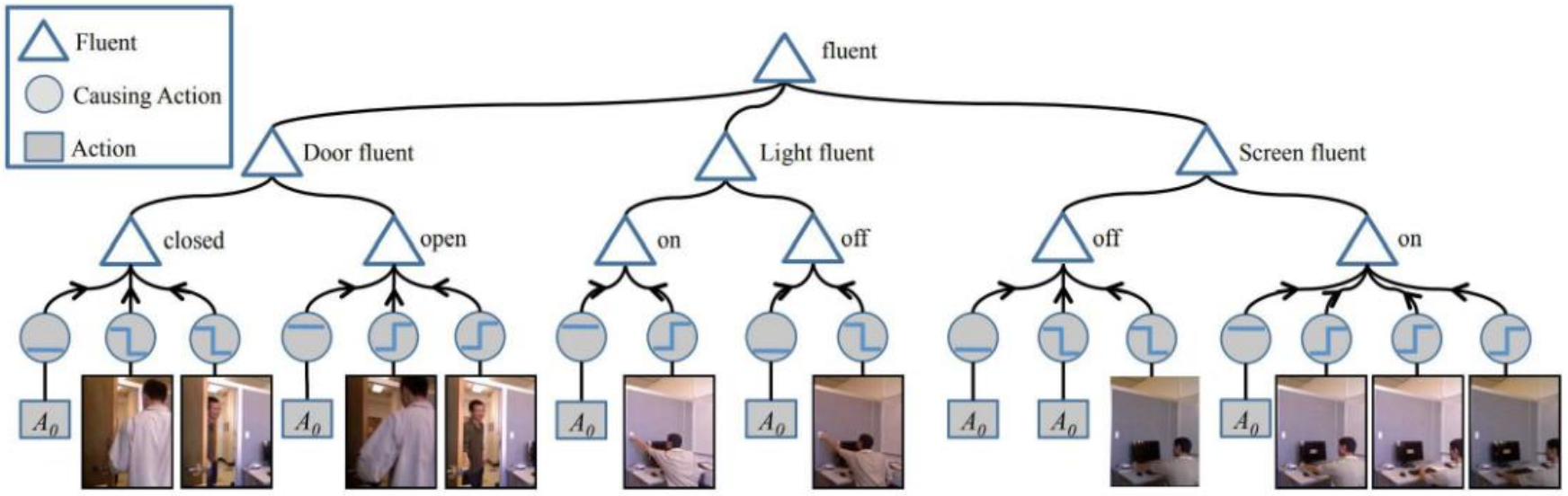
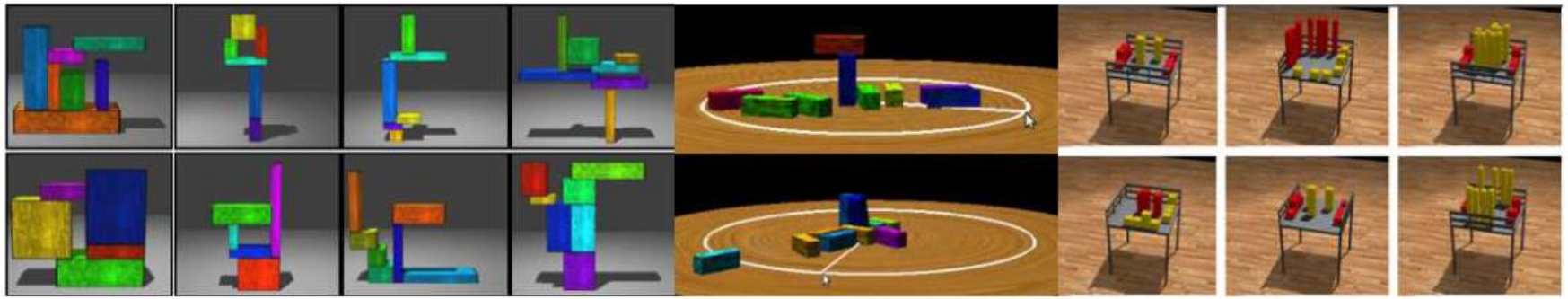


Figure 14: An example of perceptual causality in computer vision [155], with a causal and-or graph for door status, light status, and screen status. Action A_0 represents non-action (a lack of state-changing agent action). Non-action is also used to explain the change of the monitor status to off when the screensaver activates. Arrows point from causes to effects, and undirected lines show deterministic definition.

- 1. maximize information gain after adding causal relation
- 2. Minimize KL divergence between causal model and observed stats

Intuitive Physics

- Physics engine in the mind



(a) Will it fall?

(b) In which direction?

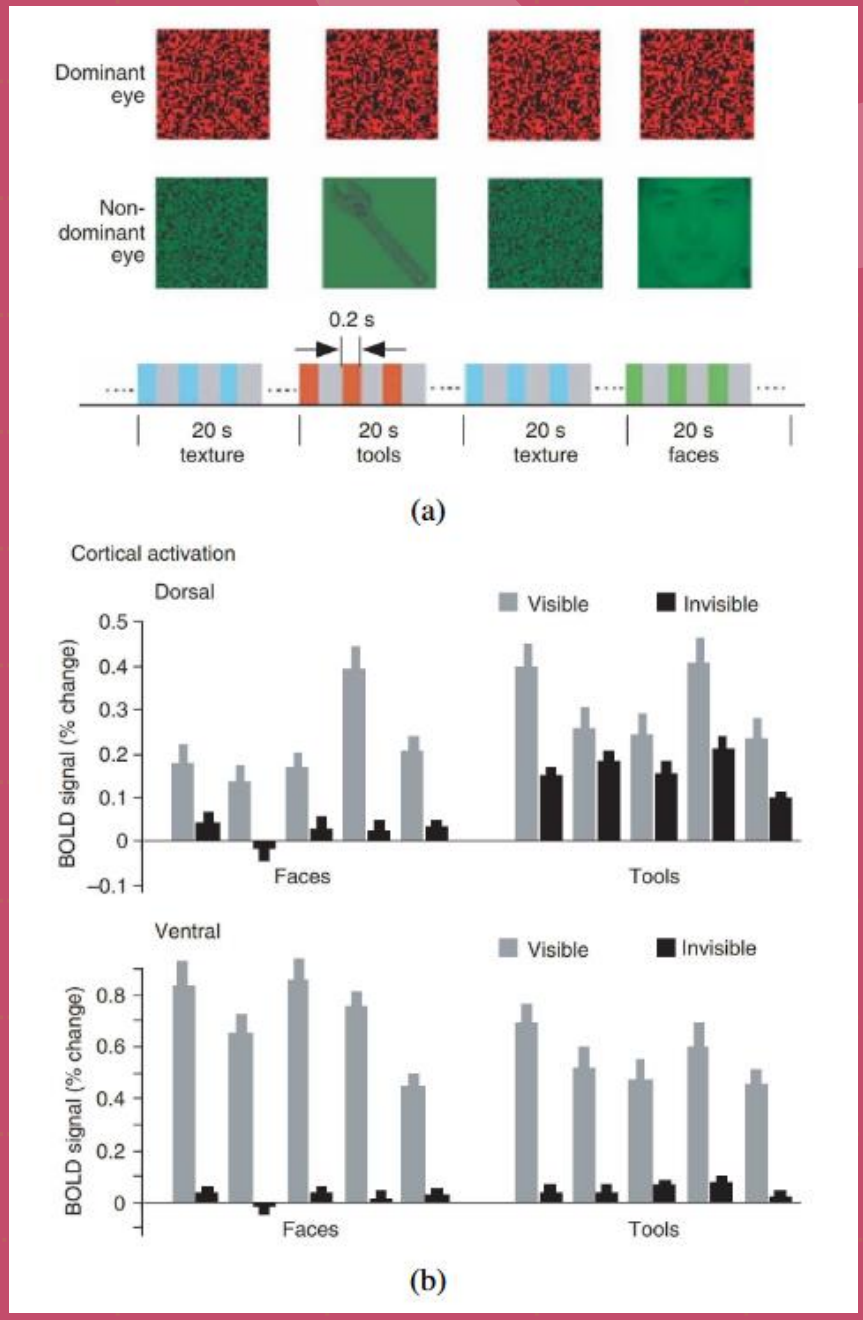
(c) Which is more likely to fall if the table was bumped hard enough, the yellow or the red?

Figure 15: Sample tasks of dynamic scene inferences about physics, stability, and support relationships presented in Ref. [70]: Across a variety of tasks, the intuitive physics engine accounted well for diverse physical judgments in *novel* scenes, even in the presence of varying object properties and unknown external forces that could perturb the environment. This finding supports the hypothesis that human judgment of physics can be viewed as a form of probabilistic inference over the principles of Newtonian mechanics.

Functionality and affordance

- Human-made tools and scenes are designed to serve human functions

- Vision system for facial recognition ≠ vision system for object functionality



(a) (b) (c) (d)

Figure 23: (a) Top three poses in various scenes for affordance (sitting) recognition. The zoom-in shows views of the (b) best, (c) second-best, and (d) third-best choice of sitting poses. The top two rows are canonical scenarios, the middle row is a cluttered scenario, and the bottom two rows are novel scenarios that demonstrated significant generalization and transfer capability. Reproduced from Ref. [233] with permission of the authors, © 2016.

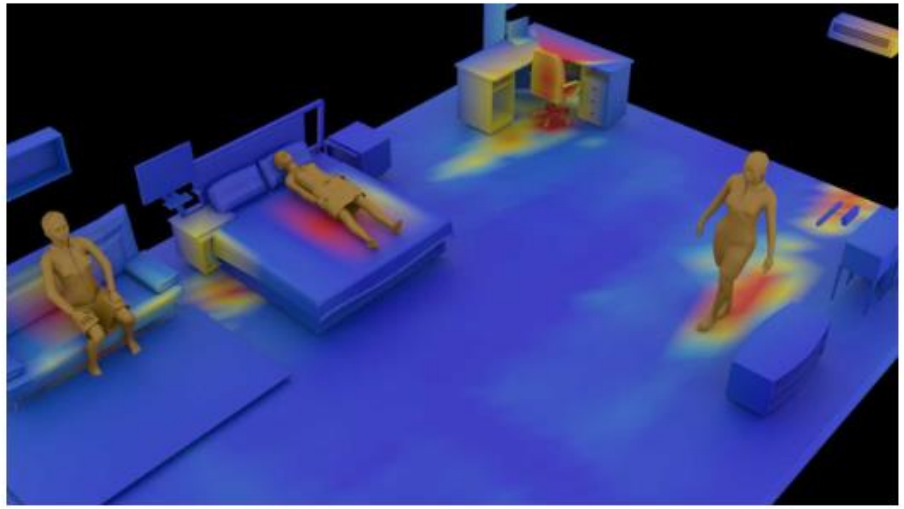
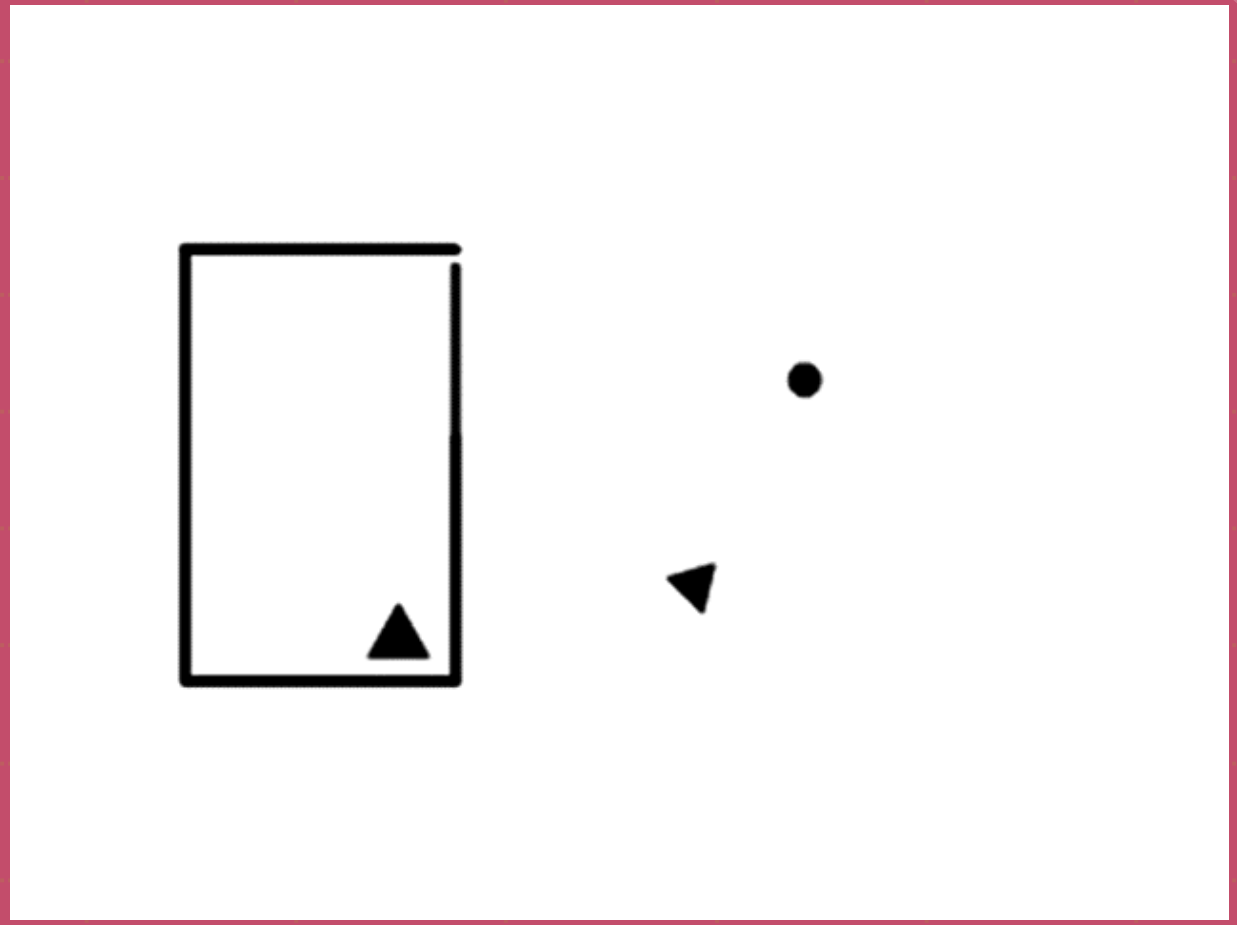


Figure 25: An example of a synthesized human-centric indoor scene (a bedroom) with an affordance heat map generated by Refs. [99, 288]. The joint sampling of the scene was achieved by alternatively sampling humans and objects according to a joint probability distribution.



Heider-Simmel experiment

Perceptual intentionality

- Assume agent follows “rationality principle”:
- 1) Devote time & resources that change fluents according to intentions
- 2) Achieve intentions optimally given beliefs of the world

Learning utility

- Principle of maximum expected utility

Limitations

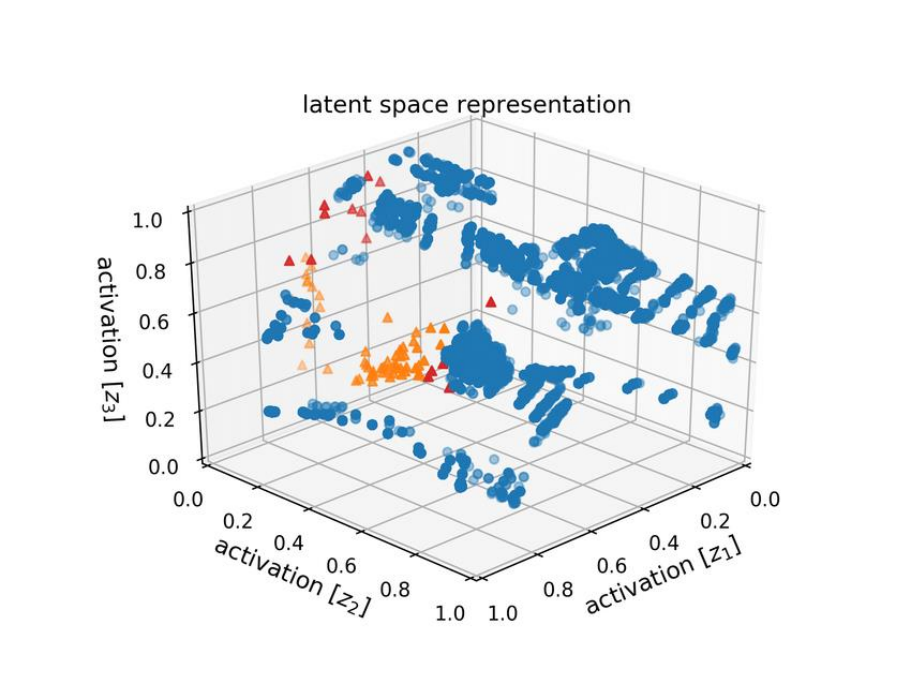
- So far only discussed what a single model should achieve, not a social system of models
- Language, communication and morality

- No in-depth discussion of abstract reasoning

- Physically realistic VR/MR Platform: Big Data for Big Tasks

Discussion

- Knowledge representation



https://www.researchgate.net/figure/Latent-space-representation-of-dataset-A-learned-by-autoencoder-architecture-AE-9-after_fig3_319875464

```
:~ initiate(climb).[-1@11].
:~ danger, initiate(observe),
on(agent,platform).[-1@10].
:~ initiate(drop(V1)), more_goals(V1).[-1@9,
V1].
:~ initiate(collect), not lava.[-1@8].
:~ initiate(interact(V1)), not danger, not
on(goal,platform).[-1@7, V1].
:~ initiate(explore(V1)),
occludes_more(V1,V2).[-1@6, V1, V2].
:~ initiate(explore(V1)), occludes(V1).[-1@5,
V1].
:~ initiate(avoid).[-1@4].
:~ initiate(balance).[-1@3]
:~ bigger(V1,V2), initiate(interact(V1)).[-
1@2, V1, V2].
:~ initiate(rotate).[-1@1].
```

https://www.ijcai.org/proceedings/2022/742

Discussion

- Trade-offs:
- Small data -> more structure/inductive bias (overfitting to problem setting?)

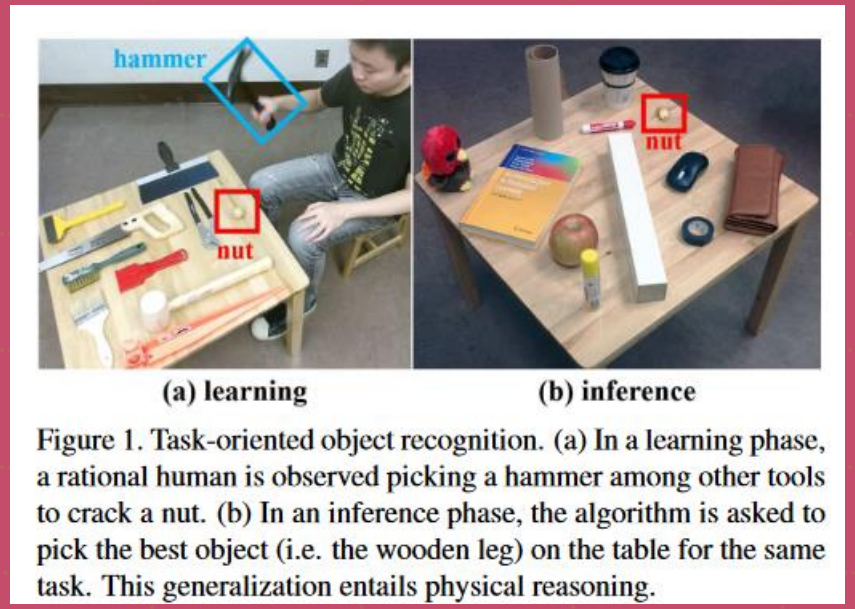


Figure 1. Task-oriented object recognition. (a) In a learning phase, a rational human is observed picking a hammer among other tools to crack a nut. (b) In an inference phase, the algorithm is asked to pick the best object (i.e. the wooden leg) on the table for the same task. This generalization entails physical reasoning.

Discussion

- Human imitation learning
- No technical details (STC-AoGs)
- Bread-first-search literature
- No judgement/ethics
- Embodiment and Scaffolding
- *“Understanding doesn’t emerge in the observer only; it also emerges from the interaction between observer and environment”*
- Representation + set of tools

Bibliography

- Joshua Tenenbaum, Yale University. (2022). What Kind of Computation Is Cognition? In YouTube. https://www.youtube.com/watch?v=NsID1iM8gRw

- Song-Chun , Z. (2017). Qiantan rengongzhineng: xianzhuang, renwu, goujia yu tongyi [AI: The Era of Big Integration Unifying Disciplines within Artificial Intelligence]. Shijiao Qiusuo. English version at: https://dm.ai/ebook/

- Tu, K., Meng, M., Mun Wai Lee, Tae Eun Choe, & Zhu, S.-C. (2013). Joint Video and Text Parsing for Understanding Events and Answering Queries. Arxiv. https://doi.org/10.48550/arxiv.1308.6628

- Zhu, S.-C., & Huang, S. (2021). Computer Vision: Stochastic Grammars for Parsing Objects, Scenes, and Events. Springer Nature.

- Zhu, Y., Gao, T., Fan, L., Huang, S., Edmonds, M., Liu, H., Gao, F., Zhang, C., Qi, S., Wu, Y. N., Tenenbaum, J. B., & Zhu, S.-C. (2020). Dark, Beyond Deep: A Paradigm Shift to Cognitive AI with Humanlike Common Sense. Engineering. https://doi.org/10.1016/j.eng.2020.01.011

- Chen, Y.-C., & Scholl, B. J. (2016). The Perception of History. Psychological Science, 27(6), 923–930. https://doi.org/10.1177/0956797616628525

Technical details

$$E(pg) = \sum_{v \in V^{or}(pg)} E_{or}(v) + \sum_{r \in R(pg)} E_R(r) \qquad (2)$$

$$E_{\text{STC}}(pg) = E_{\text{S}}(pg) + E_{\text{T}}(pg) + E_{\text{C}}(pg) + \sum_{r \in R^*(pg)} E_R(r) \qquad (3)$$

The prior probability of a parse graph pg is then defined as:

$$P(pg) = \frac{1}{Z} e^{-E_{\text{STC}}(pg)} \qquad (4)$$

$$E_v(pg_{vid}) = E_{\text{STC}}(pg_{vid}) - \log p(vid|pg_{vid}) \qquad (5)$$

3.2 Temporal parsing

We perform temporal parsing following the approach proposed in [9], which is based on the Earley parser [54]. The input video is divided into a sequence of frames. The agents, objects and fluents in each frame are identified using the spatial parser and special detectors.

3.3 Causal parsing

After all the events are detected in temporal parsing, we then perform causal parsing. For each fluent change detected in the video using special detectors, we collect the set of events that occur within a temporal window right before the fluent change and run the Earley parser again based on the sub-graph of the C-AOG rooted at the detected fluent change.


EXIT

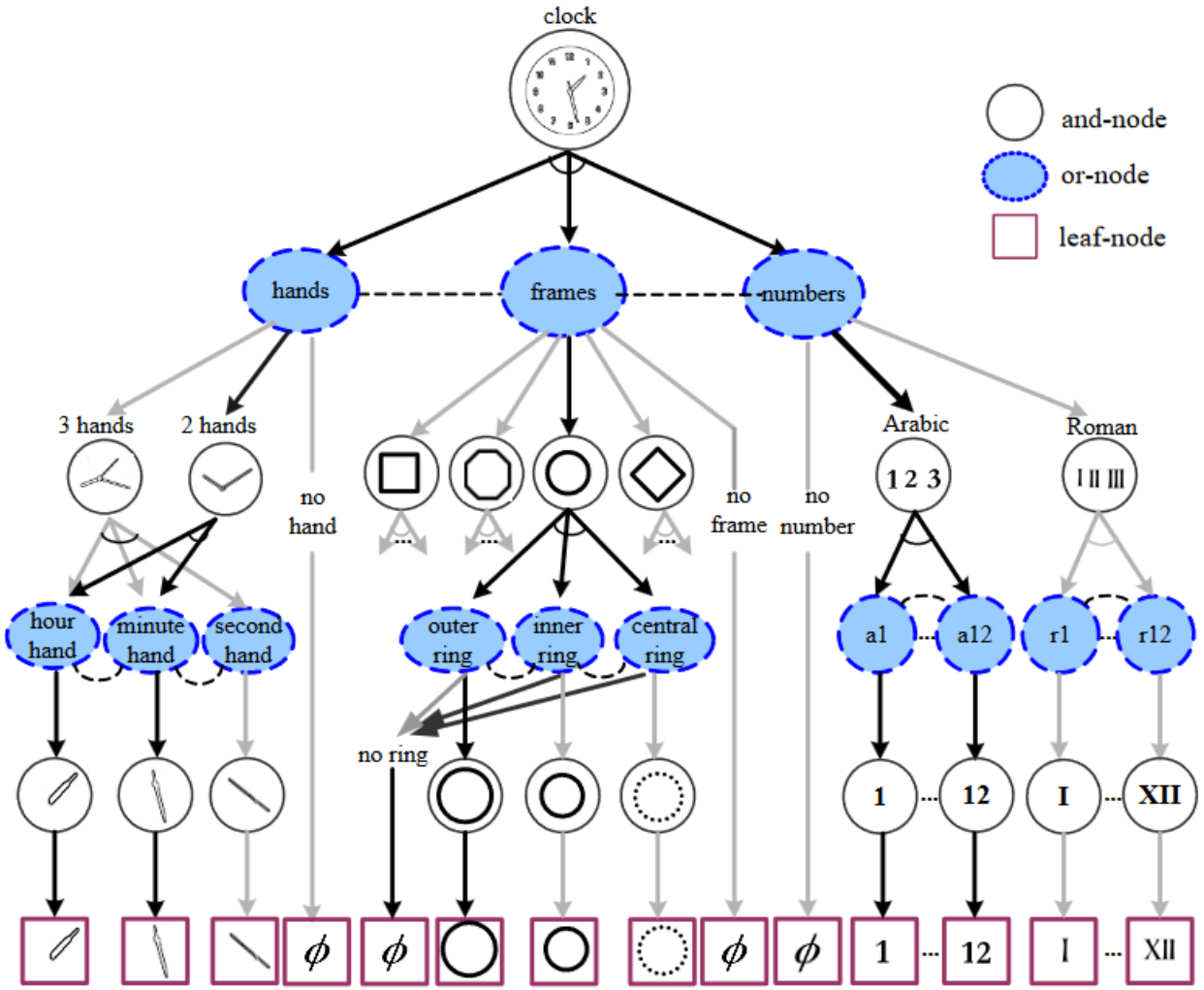


Figure 3.18: An And-Or graph example for the object category – clock. It has two parse graphs shown in Figure 3.16, one of which is illustrated in dark arrows. Some leaf nodes are omitted from the graph for clarity. From [87].