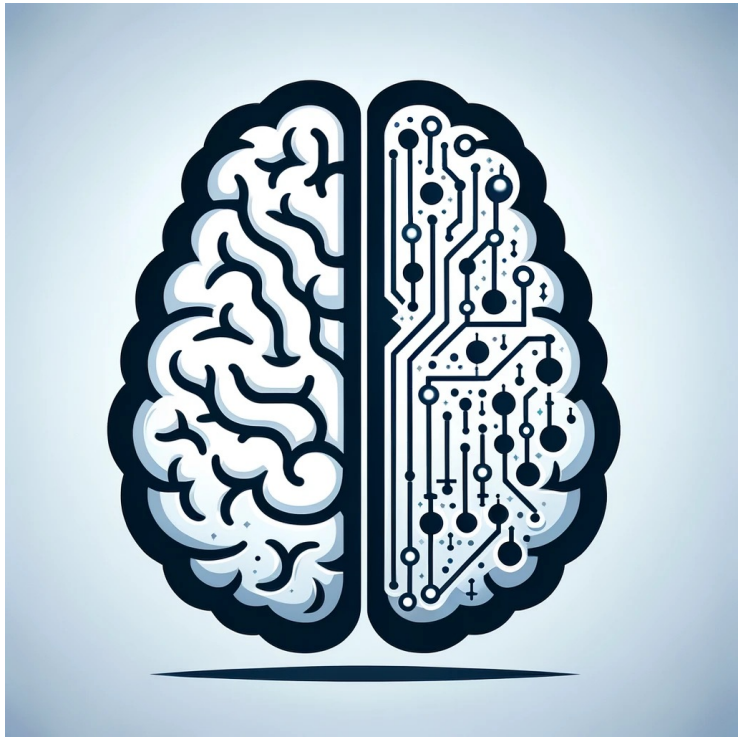
# EN 601.473/601.673: Cognitive Artificial Intelligence (CogAI)



## Lecture 18: Social scene understanding

Tianmin Shu

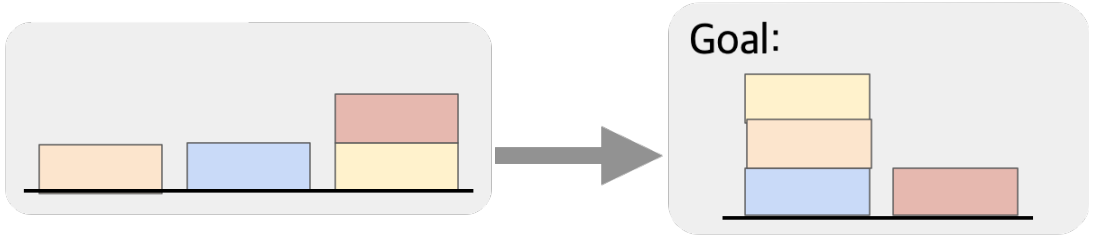
# Physical and social reasoning

- Common sense scene understanding, intuitive theories
- Physical reasoning
- Social reasoning

# Language models as world models

- LLMs fail to plan robustly

Blocksworld: How to move the blocks to the goal state?




GPT-4

**Invalid Action!**
The yellow block is still under the red one.

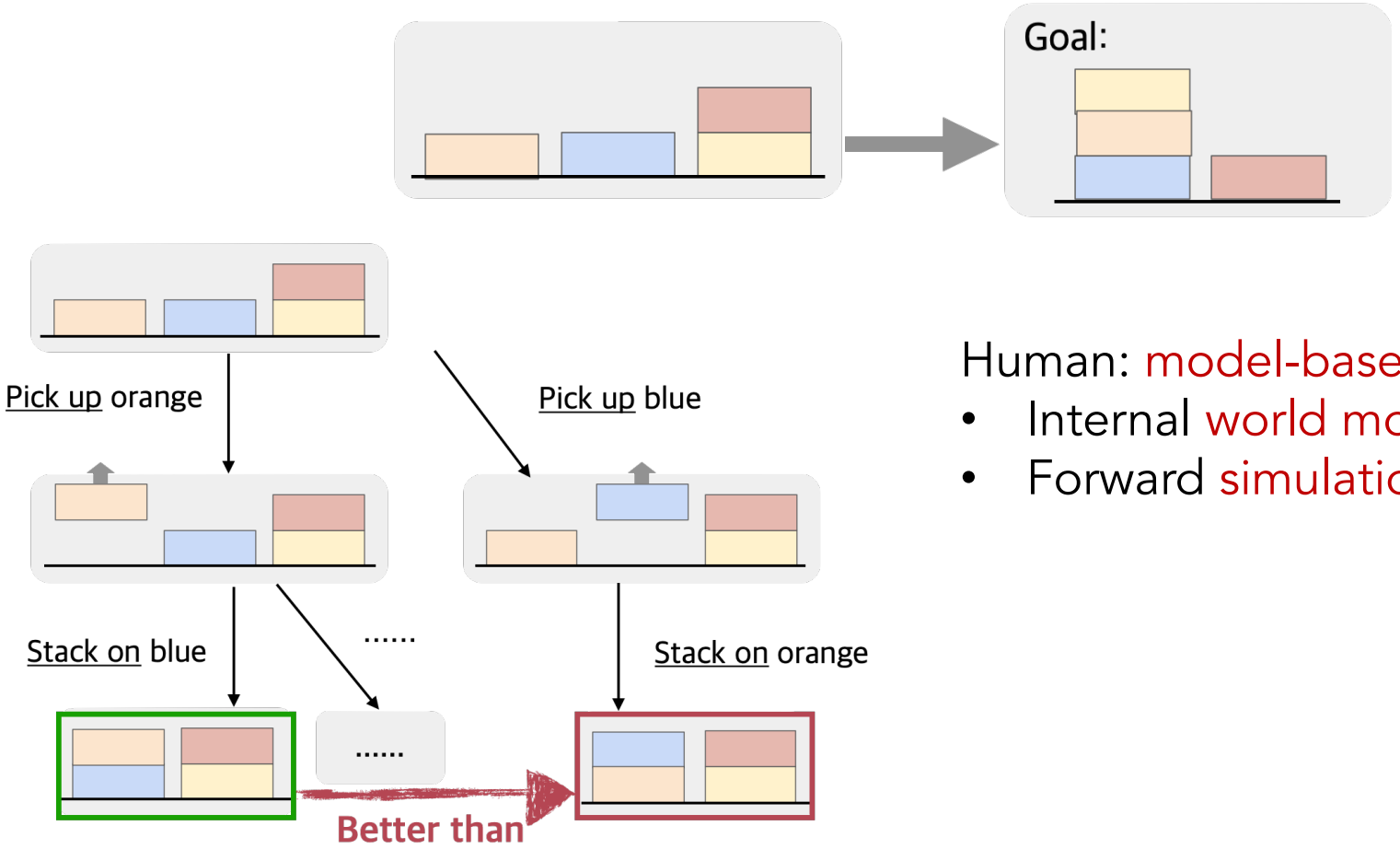
1. Pick up the orange block.
2. Stack it on the blue block.
3. Pick up the yellow block.
4. Stack it on the orange block.
5. Pick up the red block.
6. Put it on the table.

LLMs: Autoregressive plan generation

Valmeekam et al. (2023)
Hao et al. (2023)

# Language models as world models

Blocksworld: How to move the blocks to the goal state?




Human: model-based planning

- Internal world model
- Forward simulation of alternative plans

# Language models as world models

- Reasoning-via-Planning (RAP), Hao et al. (2023)

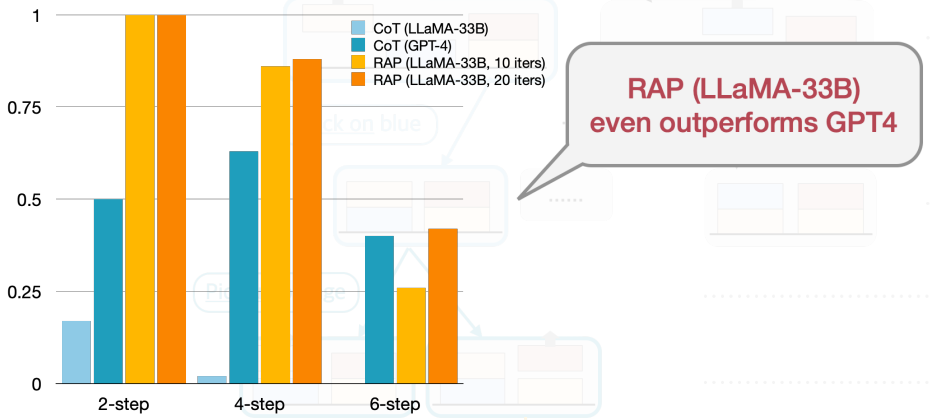
Monte Carlo Tree Search (MCTS):




Model-based planning w/ a small LM (as WM)

>

Autoregressive planning w/ a much larger LM




Using LMs to construct WMs for embodied reasoning

# Language models as world models

- Why can LMs simulate the world?

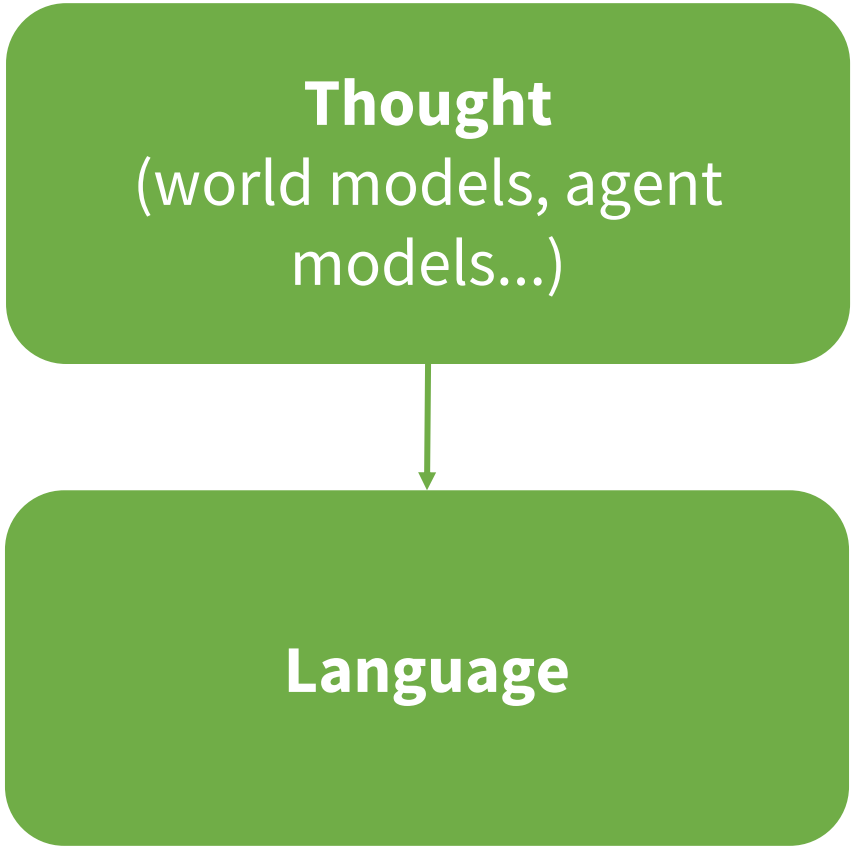
Knowledge of the world encoded in the training text

*"If a wine glass falls onto the ground, it will break."*

*"If a basketball falls onto the ground, it will bounce back."*

# Language models as world models

- Language of thoughts (why text encodes world knowledge?)



# Language models as world models

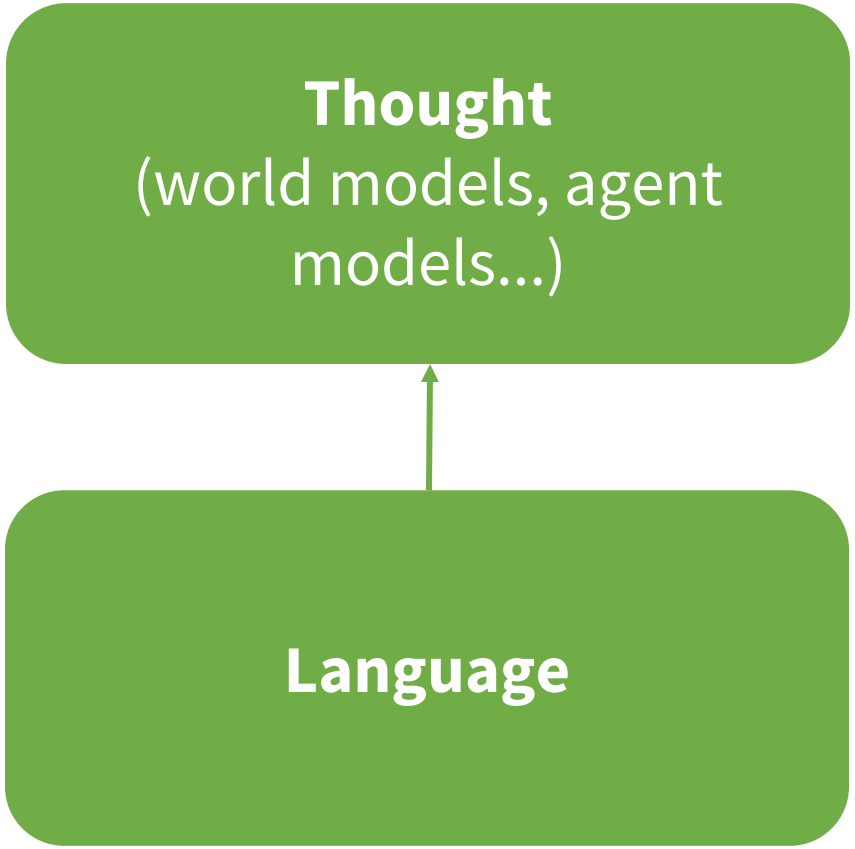
- Probabilistic language of thoughts

**From Word Models to World Models:**
Translating from Natural Language to the Probabilistic Language of Thought

Lionel Wong[1⋆], Gabriel Grand[1⋆], Alexander K. Lew[1], Noah D. Goodman[2], Vikash K. Mansinghka[1], Jacob Andreas[1], Joshua B. Tenenbaum[1]

⋆*Equal contribution.*

[1]MIT, [2]Stanford




Probabilistic programs

# Probabilistic programs with a natural language interface

- Reasoning about the world using language. For example,
- Imagine a table with a red ball placed to the left of a blue ball. We can push the red ball and it hits the blue ball.
- Imagine that the red ball is pretty heavy. And the blue ball is fairly light.
- How fast does the blue ball move after the collision?

# Text to probabilistic programs using LLMs

Text

Imagine a table with a red ball placed to the left of a blue ball. We can push the red ball and it hits the blue ball. Imagine that the red ball is pretty heavy. And the blue ball is fairly light.

Probabilistic program

```
(define choose_shapes...)
(define get_initial_color ...)
(define choose_mass ...)
(define get_initial_x...) ...

(define generate-object
  (mem (lambda (obj-id) (list
(pair 'object-id obj-id) (choose_shape obj-id)
(choose_color obj-id) (choose_mass obj-id)...)))) ...

(define generate-initial-scene-state...) ....

(define simulate-physics (mem (lambda (scene total_t delta_t)
  (let check_collisions ...)
  (let generate_next_scene_state_at_time...) ....))))
```

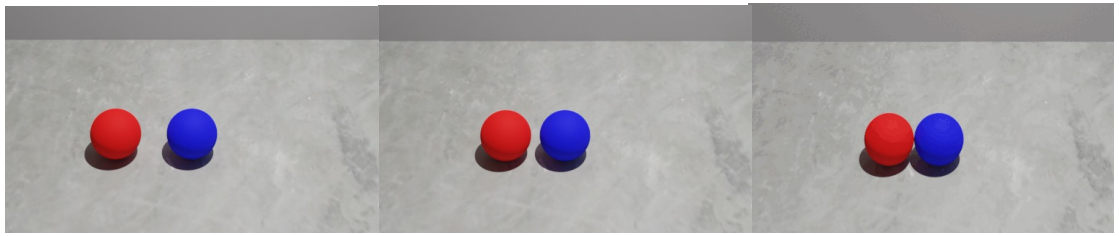
Attributes of objects

```
object-1: { color: red, shape: sphere, mass: 0.2, x: -3, v: 1.0,
            a: -0.05, force: 1.0 ...}
object-2: { color: blue, shape: sphere, mass: 3.0, x: 0, v: 0.0,
            a: 0.0, force: 0.0...}
```

Physics simulation engine

| t=1 | t=2 | ... | t=10 |
|---|---|---|---|
| object-1:<br>{...,<br>x: -2.5,<br>v: 0.95...} | object-1:<br>{...,<br>x: -2.0,<br>v: 0.9...} | | object-1:<br>{...,<br>x: 0.0,<br>v: 0.01...} |

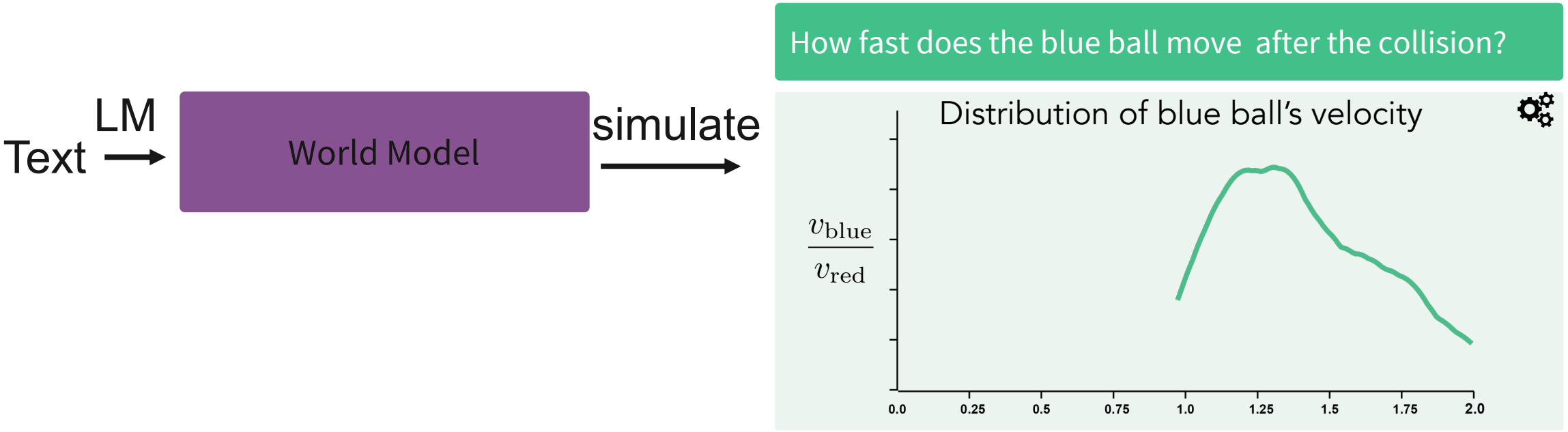
Graphics rendering engine



Gabe Grand, Lio Wong, et al. (2023)

# Text to probabilistic programs using LLMs

Using LMs to construct WMs via *probabilistic programs* for language reasoning




Probabilistic reasoning → faster than the red ball's initial speed

Gabe Grand, Lio Wong, et al. (2023)

# Physical and social reasoning

- Common sense scene understanding, intuitive theories
- Physical reasoning
- Social reasoning

# Social intelligence

**An anonymous reviewer**: … it is unclear why having machines with “human-level social intelligence” is necessary. First, can “social intelligence” be measured or evaluated? … Second, what types of applications will benefit from this skill?

**We can find answers from social cognition studies!**

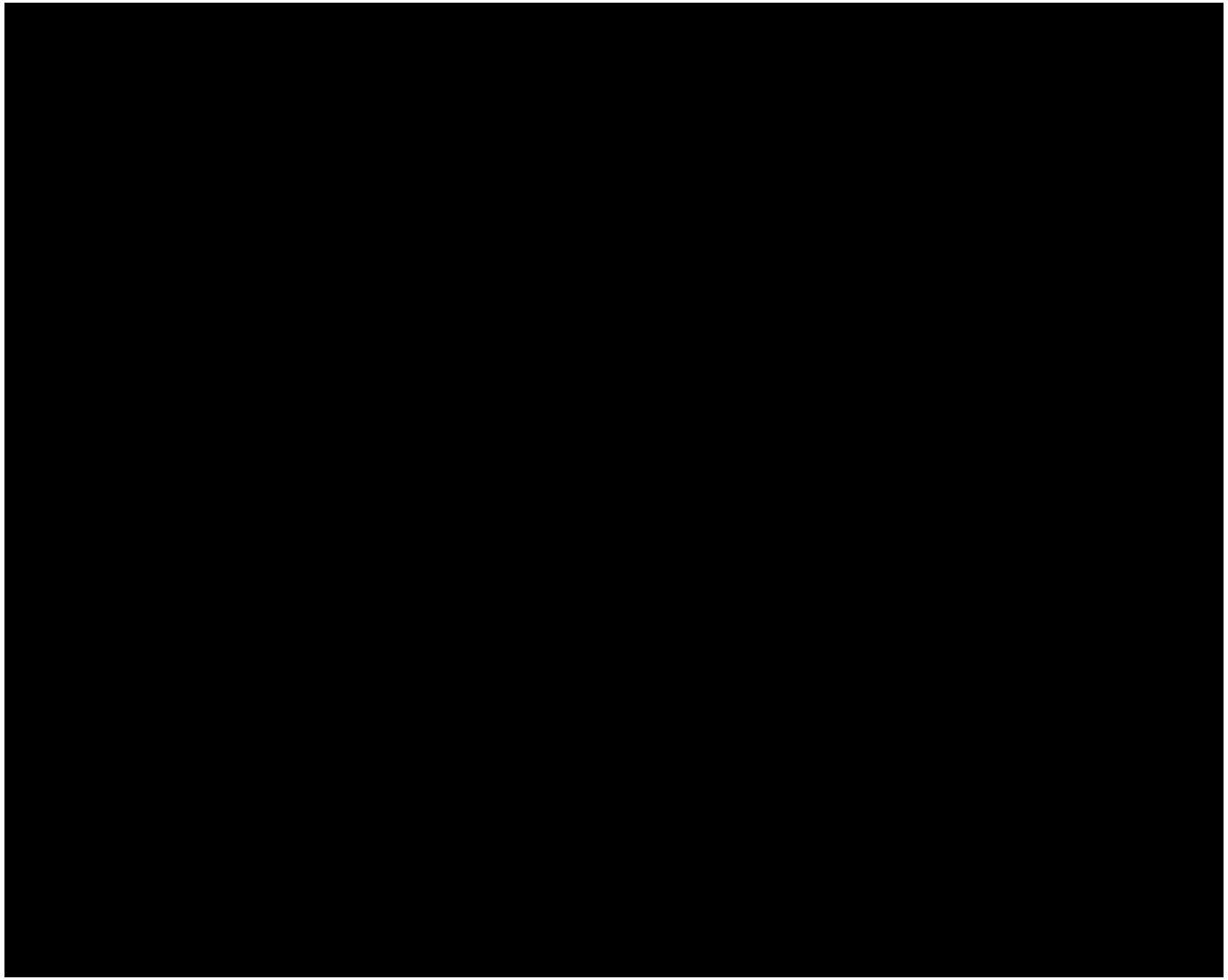
# Social intelligence

Altruistic behavior of an 18-month-old child

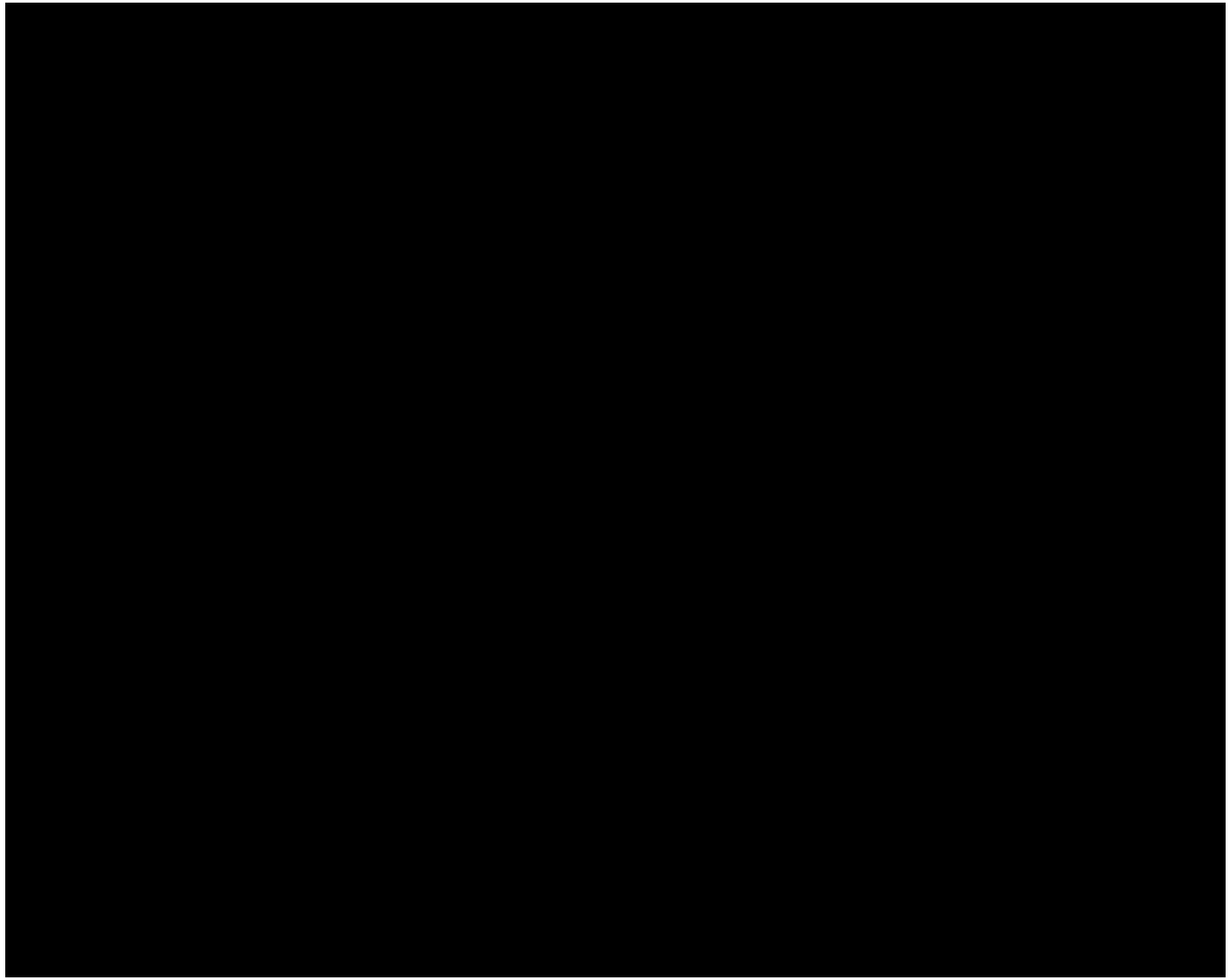
They don’t understand language yet


Warneken & Tomasello (2006)

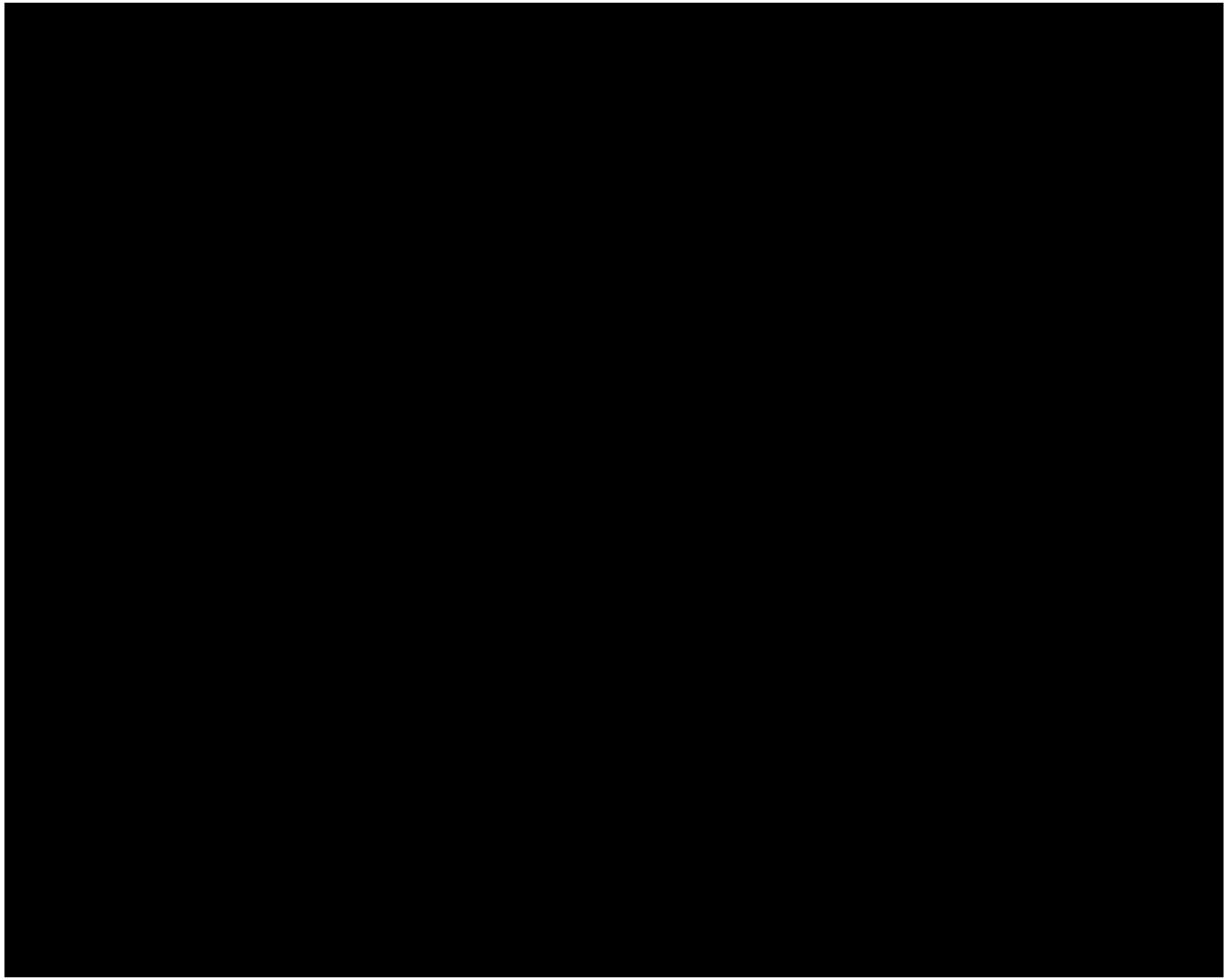
# Flexible and generalizable in different types of tasks



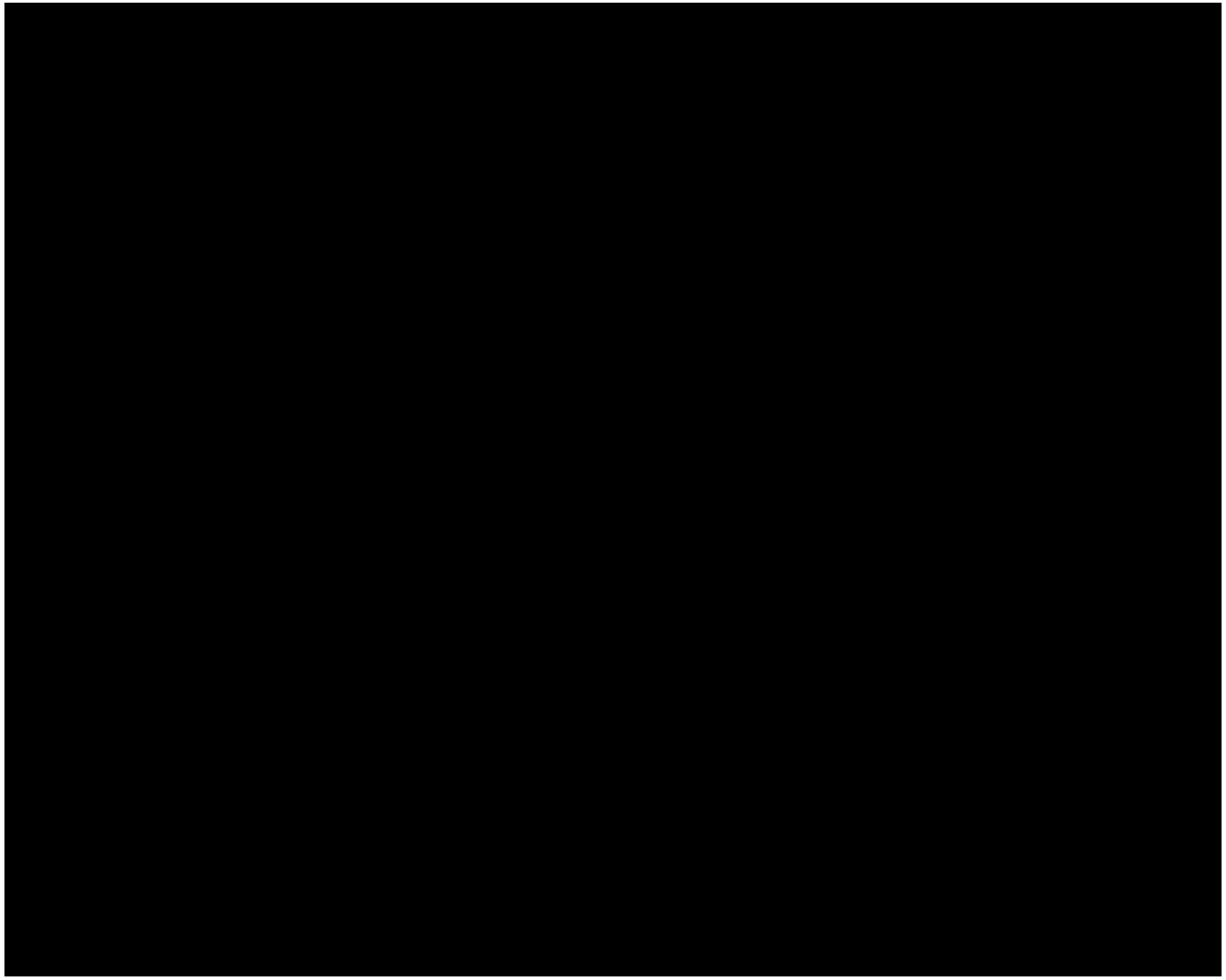
# Flexible and generalizable in different types of tasks



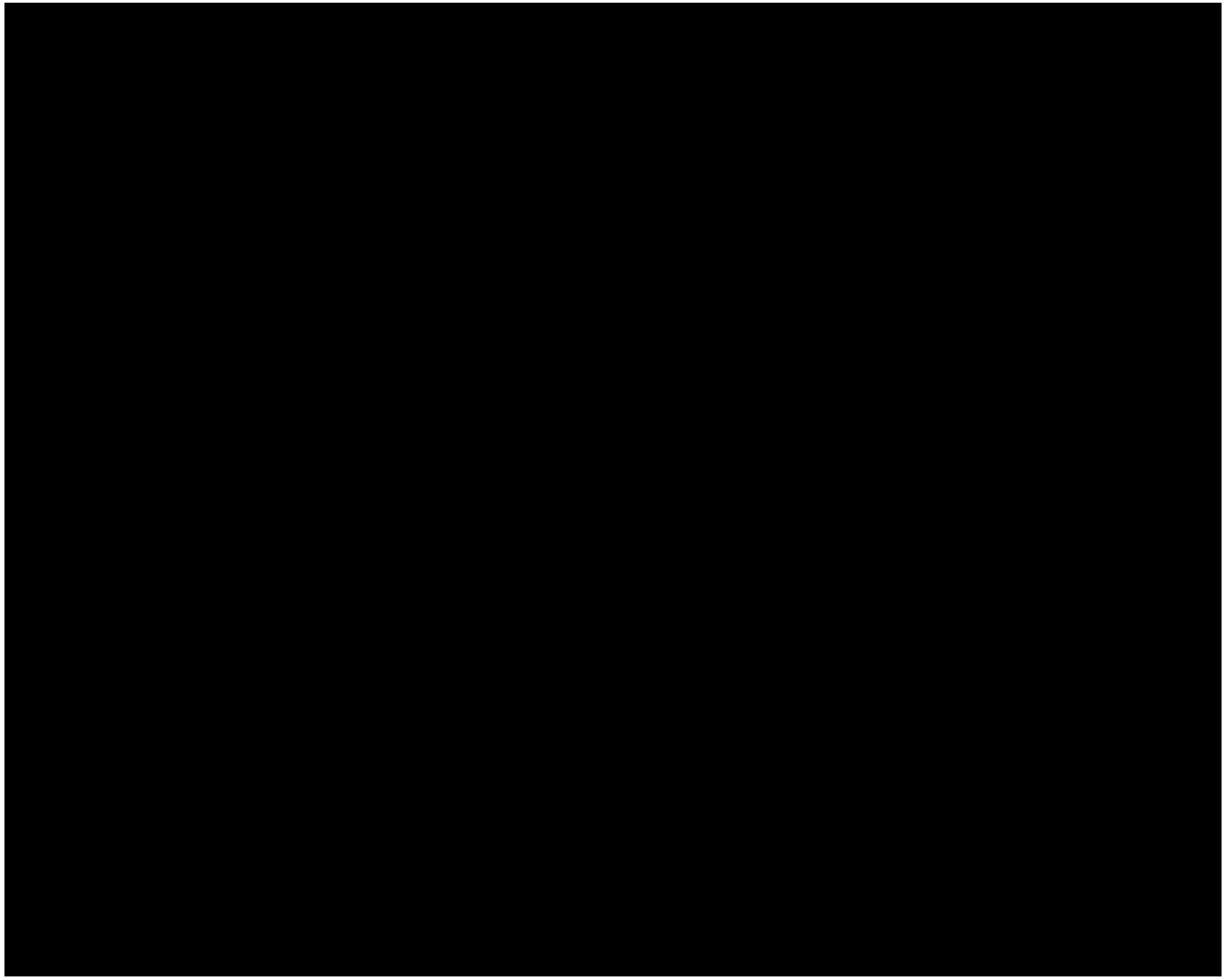
# Flexible and generalizable in different types of tasks



# Chimpanzees can only help with out of reach tasks

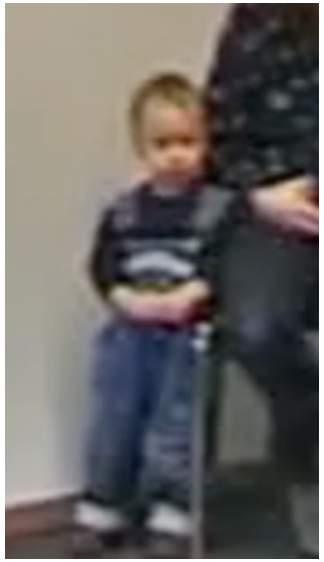


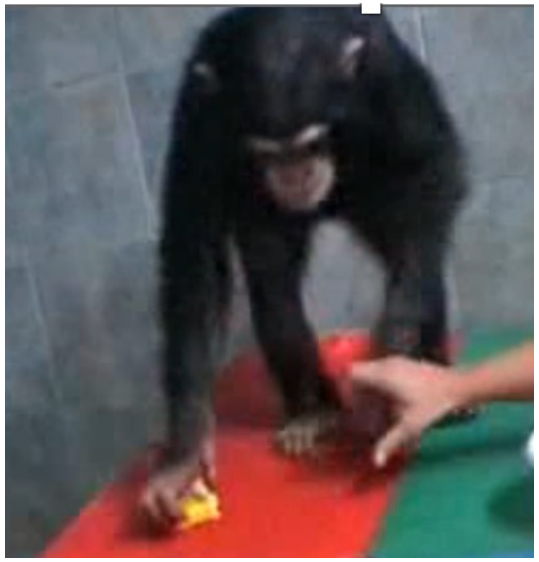
# But not other types of tasks



# Human children vs chimpanzees: what is the gap?

Similar level of intelligence in many areas





# The gap is social

Science

**Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis**

ESTHER HERRMANN , JOSEP CALL, MARÍA VICTORIA HERNÀNDEZ-LLOREDA, BRIAN HARE, AND MICHAEL TOMASELLO Authors Info & Affiliations

- Humans have many cognitive skills not possessed by their nearest primate relatives.
- **The cultural intelligence hypothesis** argues that this is mainly due to a species-specific set of **social-cognitive** skills, emerging early in ontogeny, for participating and exchanging knowledge in cultural groups.
- Two of humans' closest primate relatives, chimpanzees and orangutans vs 2.5-year-old human children on physical and social tasks

# Physical tasks

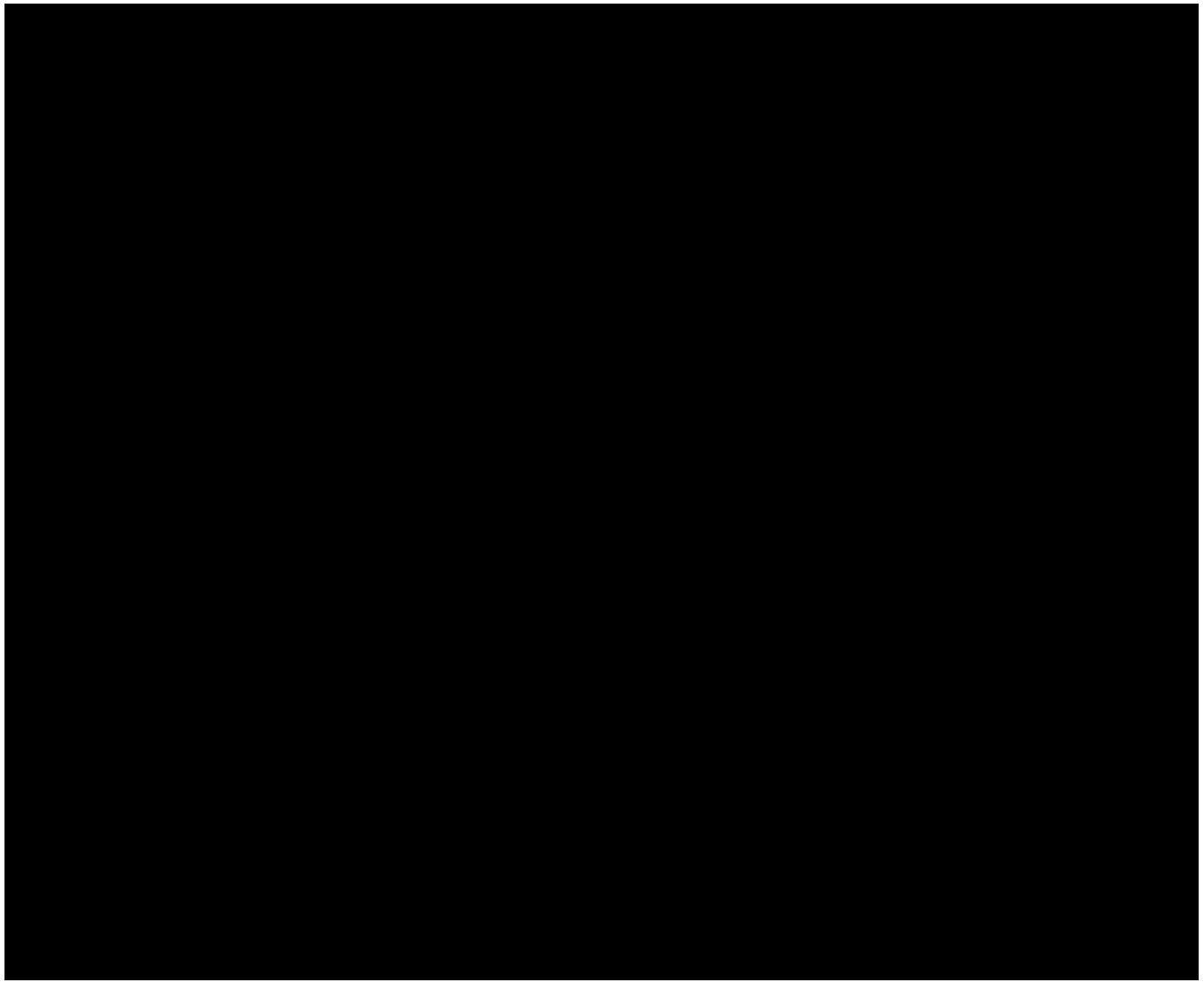
| Domain | Scale | Task | Description |
|---|---|---|---|
| Physical | Space | Spatial memory (1 item, 3 trials) | Locating a reward. |
| | | Object permanence (3 items, 9 trials) | Tracking of a reward after invisible displacement. |
| | | Rotation (3 items, 9 trials) | Tracking of a reward after a rotation manipulation. |
| | | Transposition (3 items, 9 trials) | Tracking of a reward after location changes. |
| | Quantities | Relative numbers (1 item, 13 trials) | Discriminating quantity. |
| | | Addition numbers (1 item, 7 trials) | Discriminating quantity with added quantities. |
| | Causality | Noise (2 items, 6 trials) | Causal understanding of produced noise by hidden rewards. |
| | | Shape (2 items, 6 trials) | Causal understanding of appearance change by hidden rewards. |
| | | Tool use (1 item, 1 trial) | Using a stick in order to retrieve a reward which is out of reach. |
| | | Tool properties (5 items, 15 trials) | Understanding of functional and nonfunctional tool properties. |

# Space: Spatial memory

Locating a reward

# Space: Spatial memory

Locating a reward

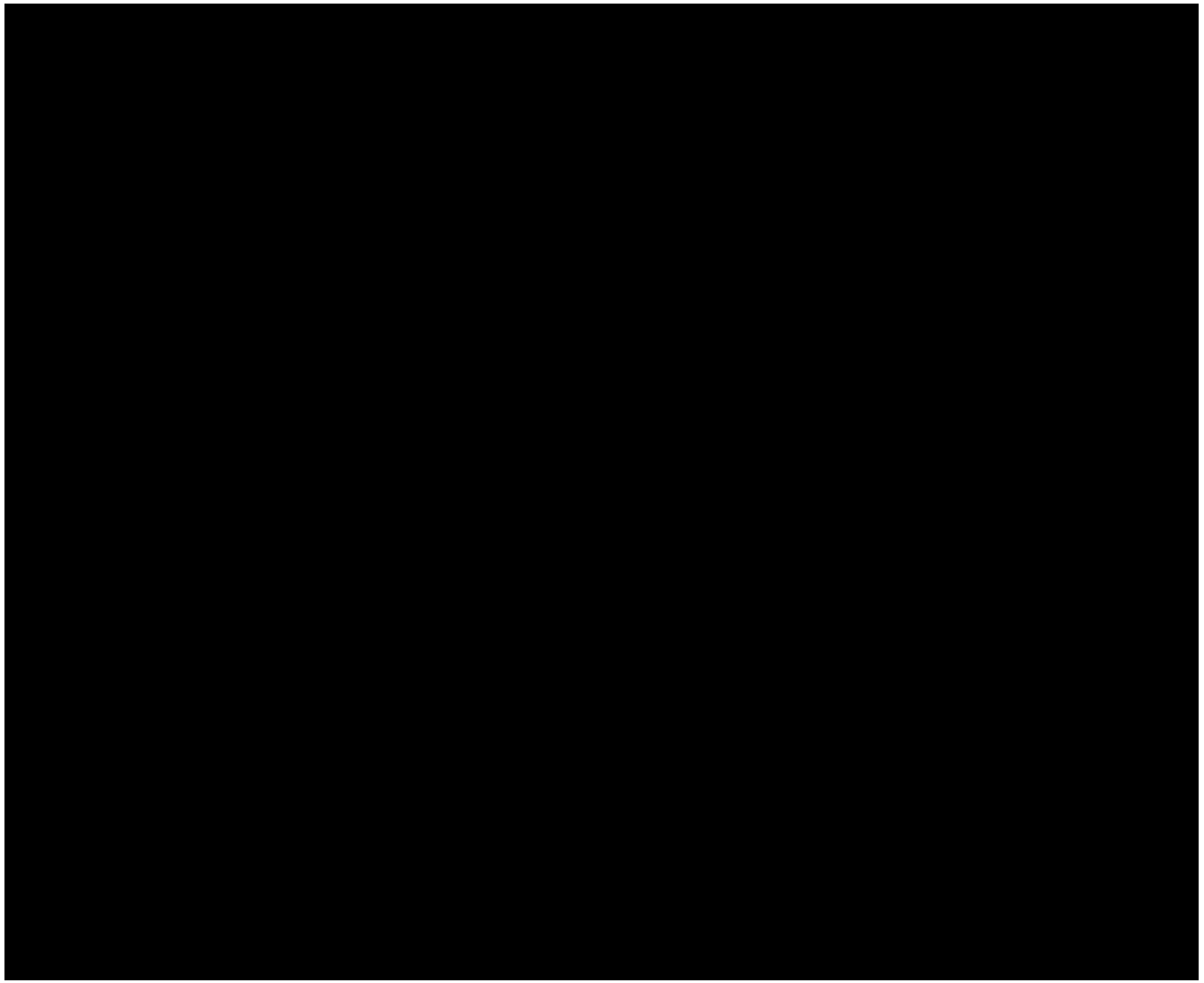


# Quantities: Addition numbers

Discriminating quantity with added quantities

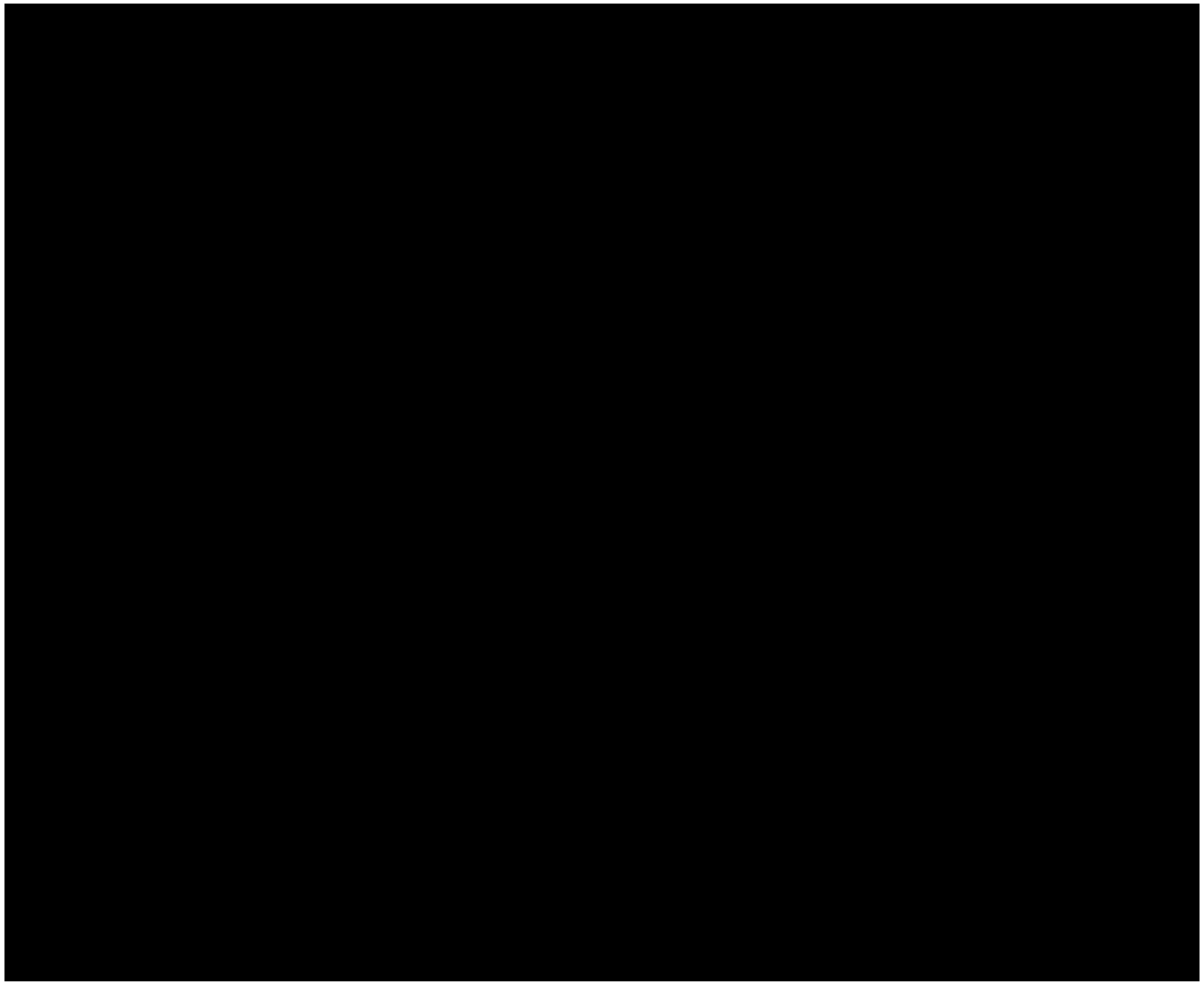
# Quantities: Addition numbers

Discriminating quantity with added quantities



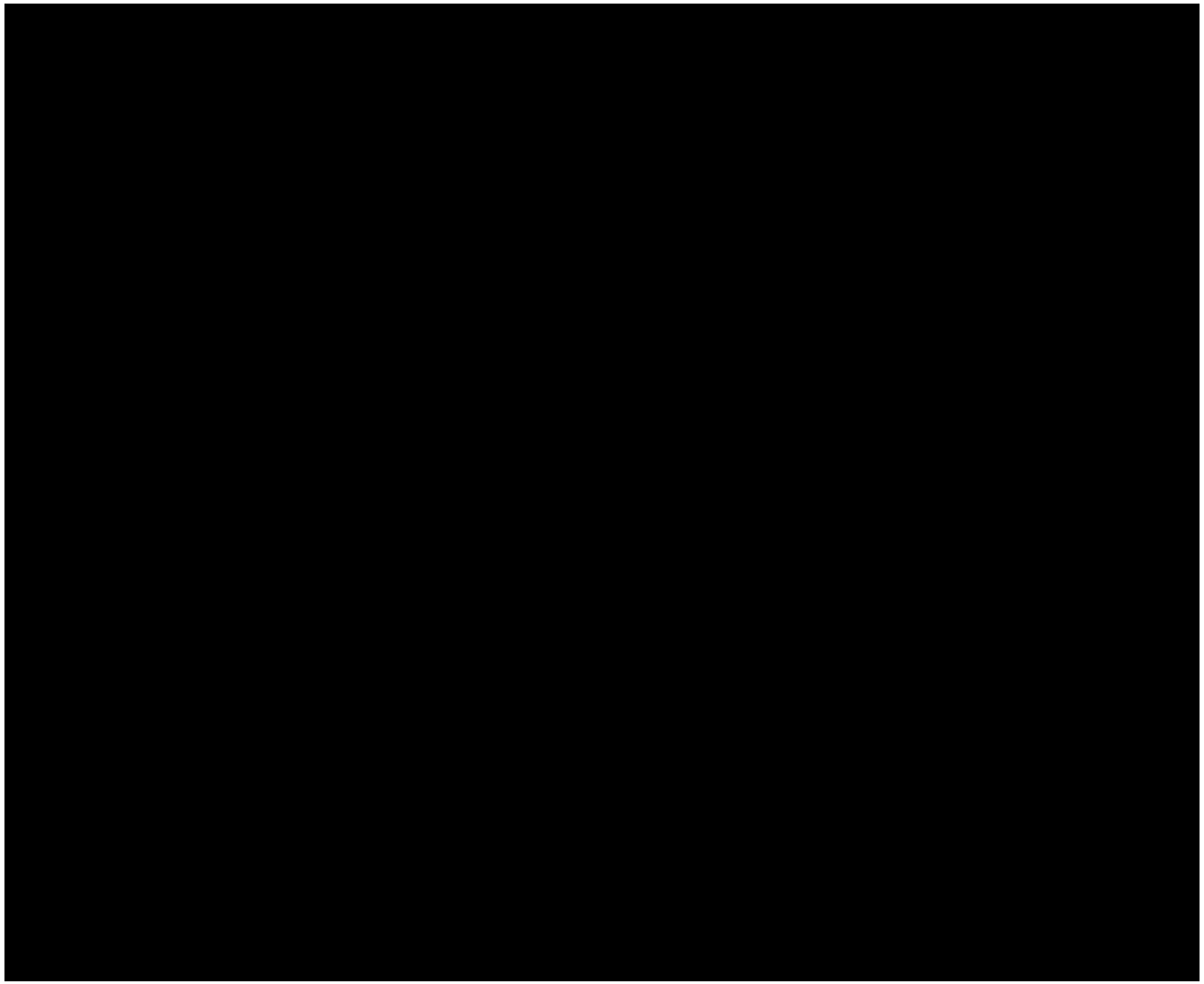
# Causality: shape

Causal understanding of appearance change of hidden rewards.



# Causality: shape

Causal understanding of appearance change of hidden rewards.



# Social tasks

| Domain | Scale | Task | Description |
|---|---|---|---|
| Social | Social learning | Social learning (3 items, 3 trials) | Solving a simple but not obvious problem by observing a demonstrated solution. |
| | Communication | Comprehension (3 items, 9 trials) | Understanding communicative cues indicating a reward's hidden location. |
| | | Pointing cups (1 item, 4 trials) | Producing communicative gestures in order to retrieve a hidden reward. |
| | | Attentional state (4 items, 4 trials) | Choosing communicative gestures considering the attentional state of the recipient. |
| | Theory of mind | Gaze following (3 items, 9 trials) | Following an actor's gaze direction to a target. |
| | | Intentions (2 items, 6 trials) | Understanding what an actor intended to do (unsuccessfully). |

# Social learning

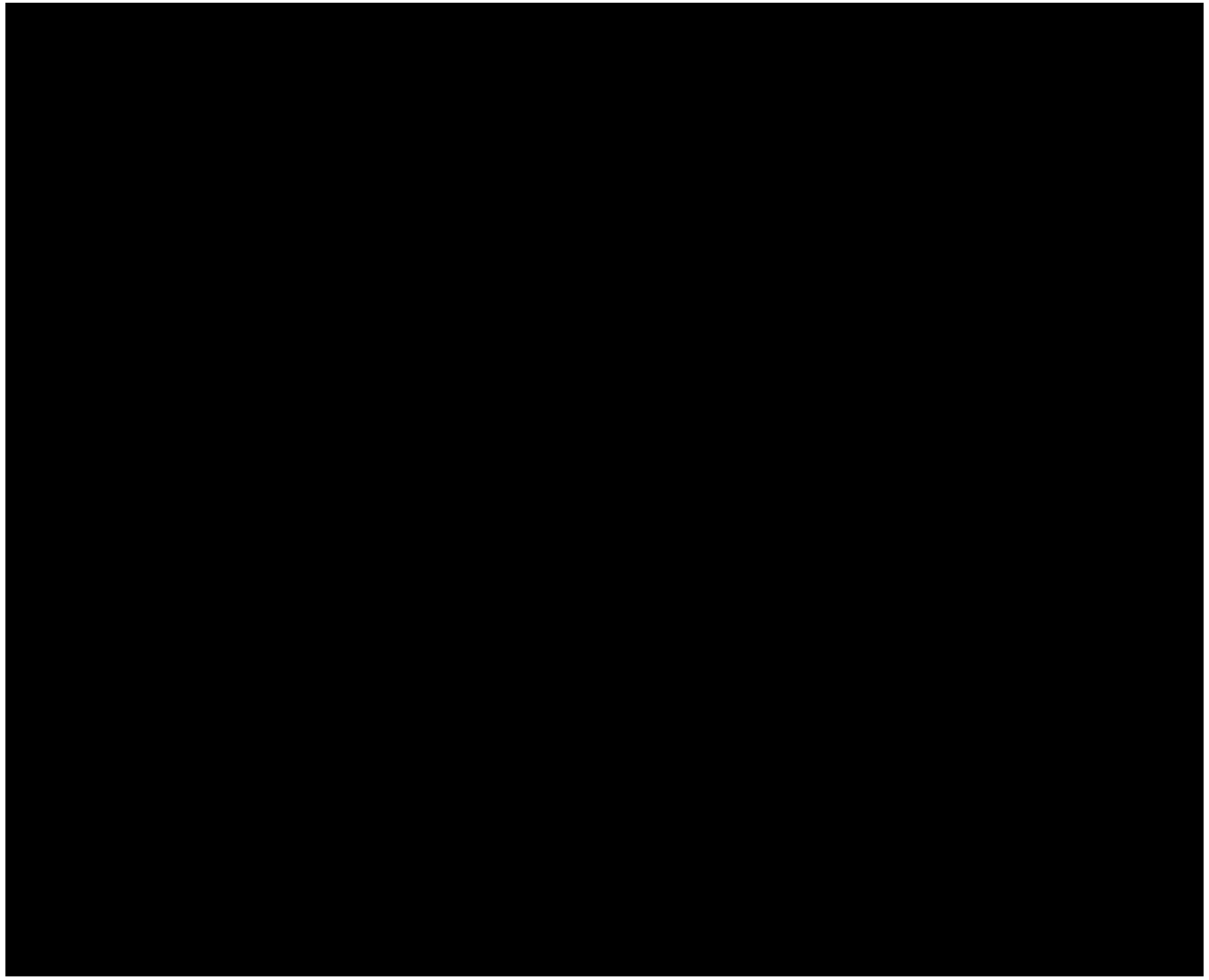
Learning from demonstration (balloon tube)

# Social learning

Learning from demonstration (stick tube)

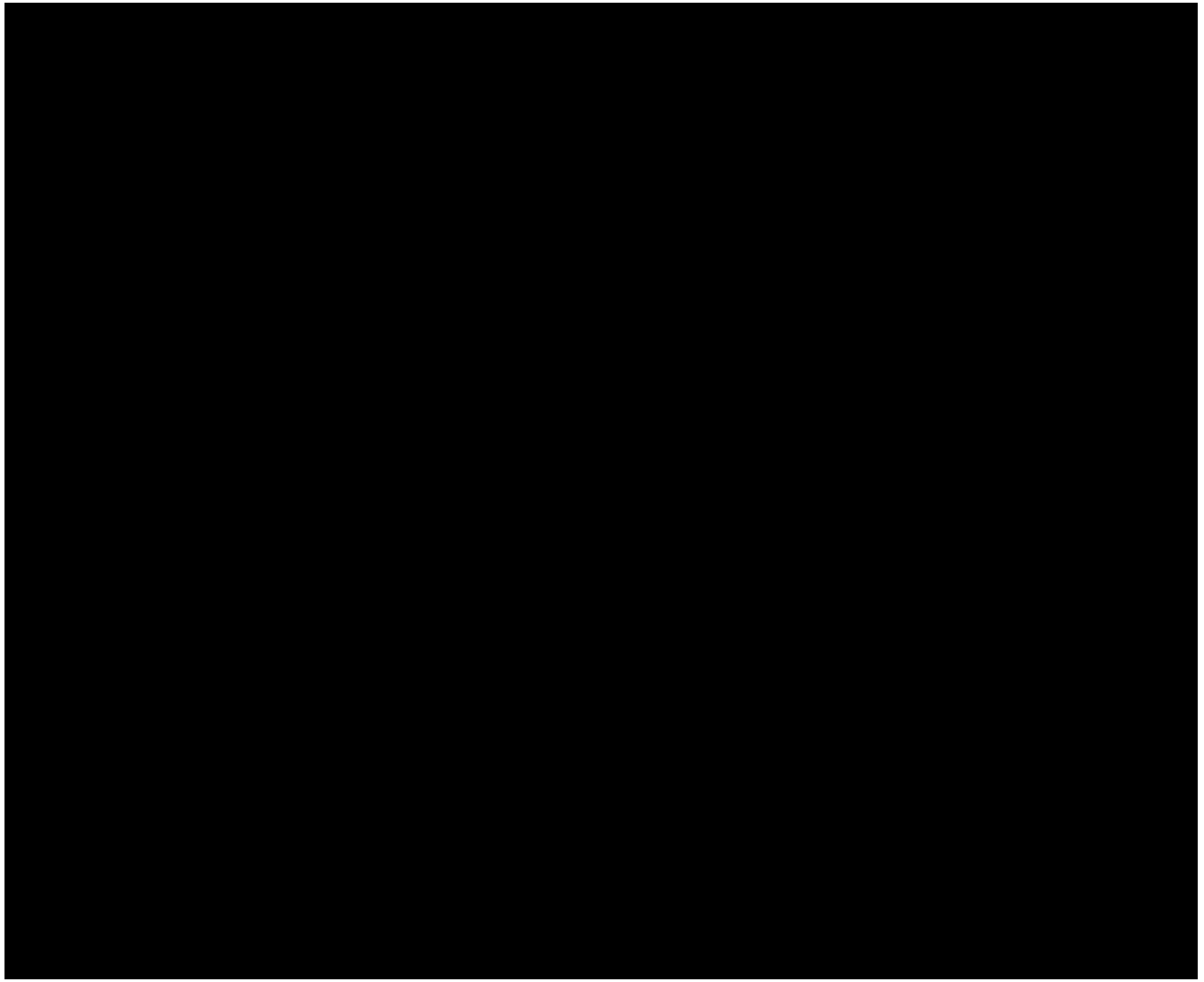
# Communication: comprehension

Understanding communicative cues indicating a reward’s hidden location



# Communication: comprehension

Understanding communicative cues indicating a reward’s hidden location

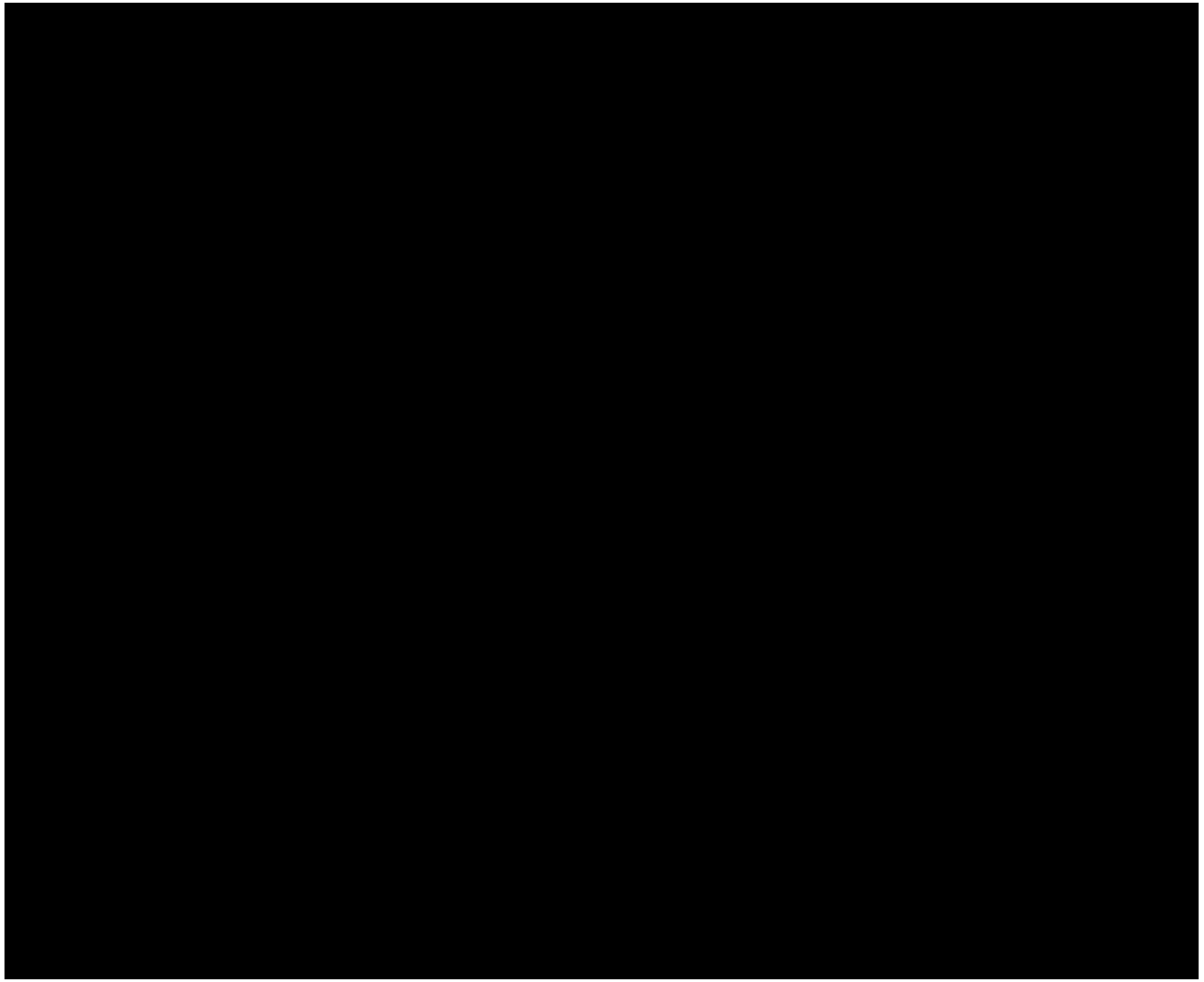


# Theory of Mind: Gaze following

Following an actor’s gaze direction to a target.

# Theory of Mind: Gaze following

Following an actor’s gaze direction to a target.



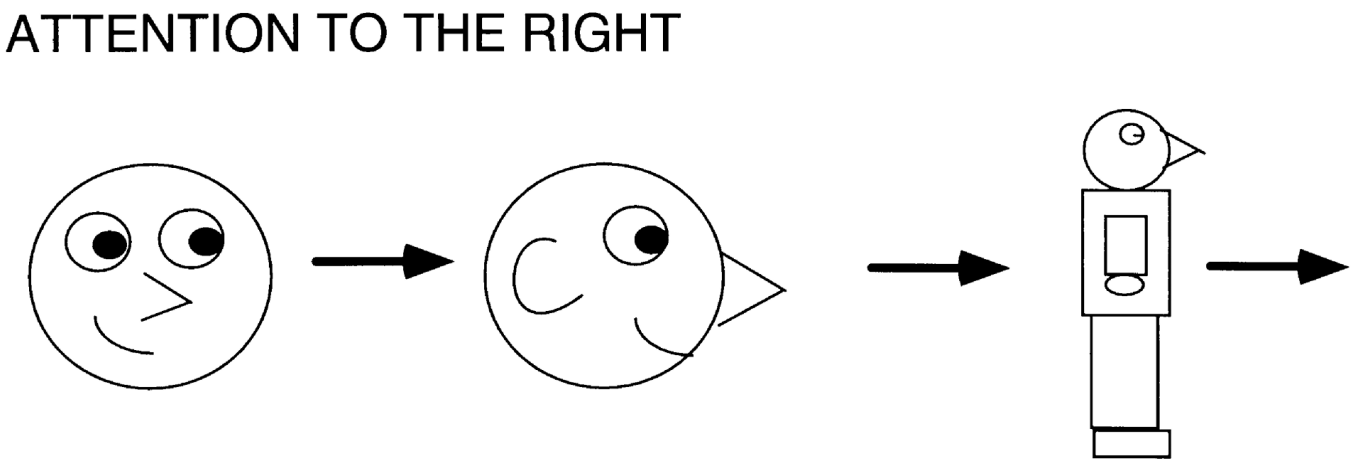
# Gaze is an important component of social interactions

The eyes have it: the neuroethology, function and evolution of social gaze

N.J. Emery*

*Center for Neuroscience, Department of Psychiatry & California Regional Primate Research Center, University of California, Davis, CA 95616, USA.*

Hierarchy of importance of social gaze




Eye Gaze Direction > Head Direction > Body Orientation

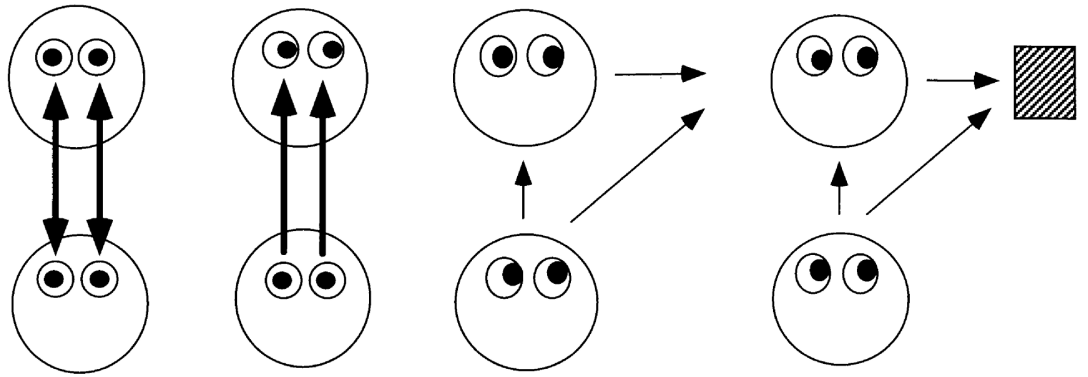
# Humans are extremely sensitive to eye gazes

A scene from the Netflix show “The Queen’s Gambit”

# Facial morphology

- Flatter face
- May have forced a shift in salience from the facing direction to eye gaze

# Gaze provides different kinds of social cues



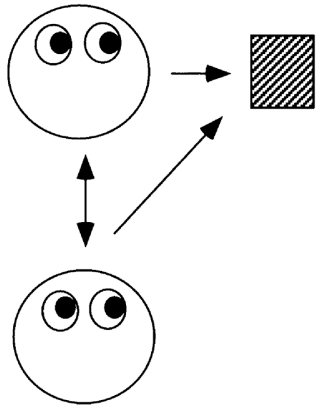
A. Mutual versus Averted Gaze



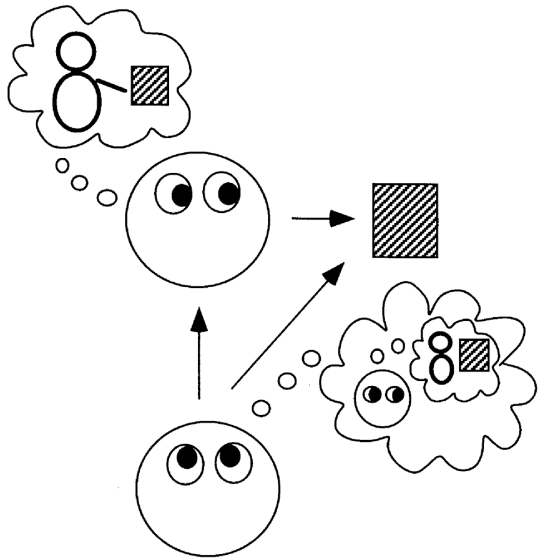
B. Gaze Following



C. Joint Attention



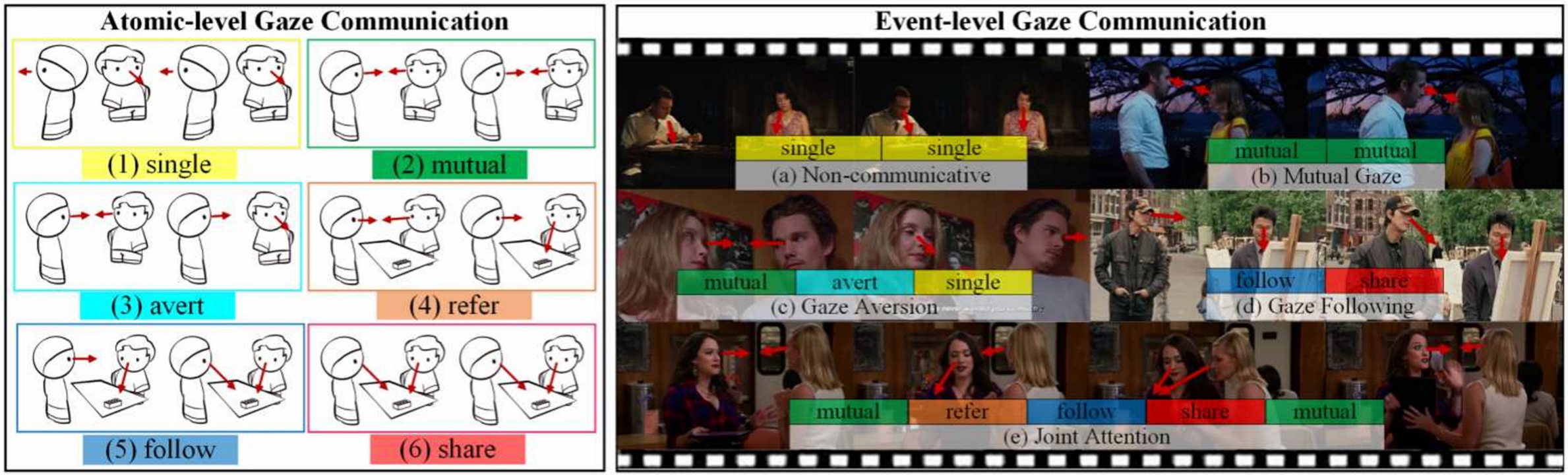
D. Shared Attention



E. "Theory of Mind"

# Understanding Human Gaze Communication by Spatio-Temporal Graph Reasoning

Lifeng Fan[1*], Wenguan Wang[2,1*], Siyuan Huang[1], Xinyu Tang[3], Song-Chun Zhu[1]



# The gap is social

- Human children and chimpanzees share similar cognitive capacities for physical tasks

- However, there is a large gap in their social-cognitive skills

- The social intelligence in human children is fundamental for developing stronger intelligence later in the life
  - Social learning → we can learn from one another, teach one another, and accumulate / transfer knowledge between generations and social groups.

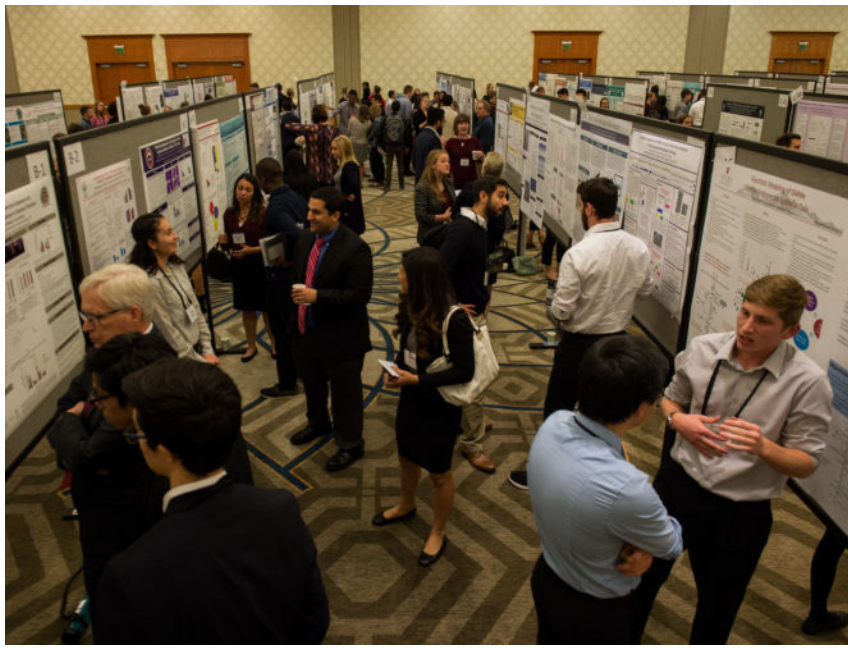
# The gap is social

- Uniquely human cooperation

Play a symphony

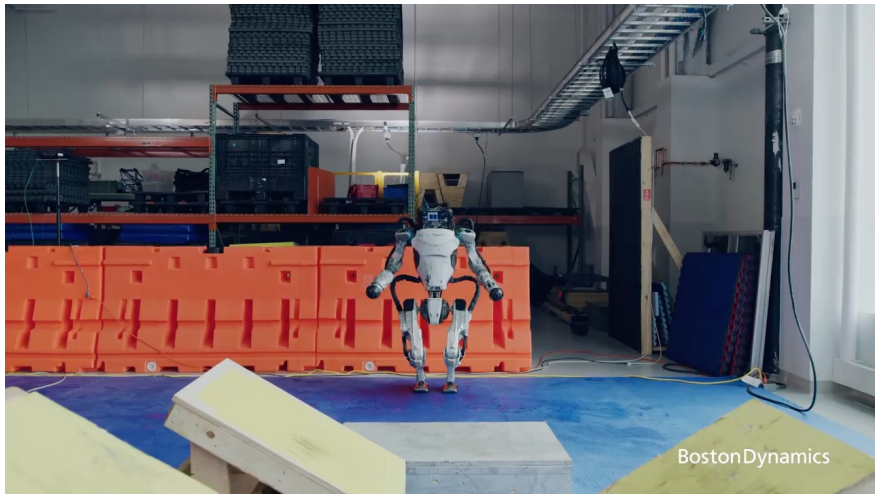
Form a government

Build a scientific enterprise

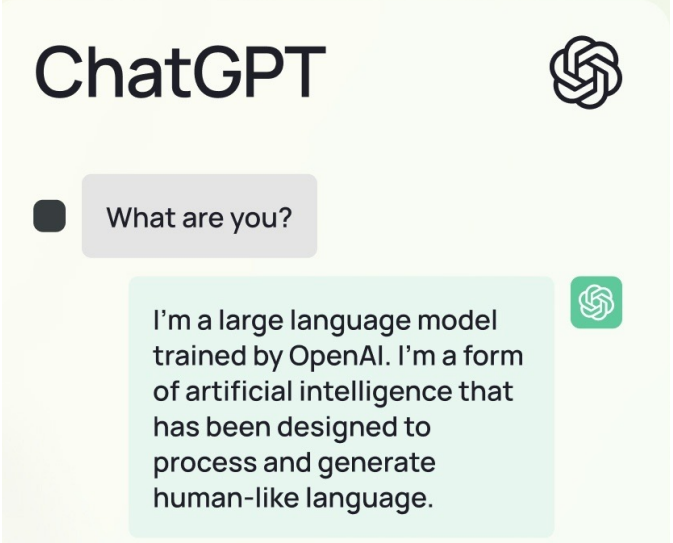


# The gap is social: 18-month-old infant vs machines


Weaker **physical** capacity




Weaker **linguistic** capacity

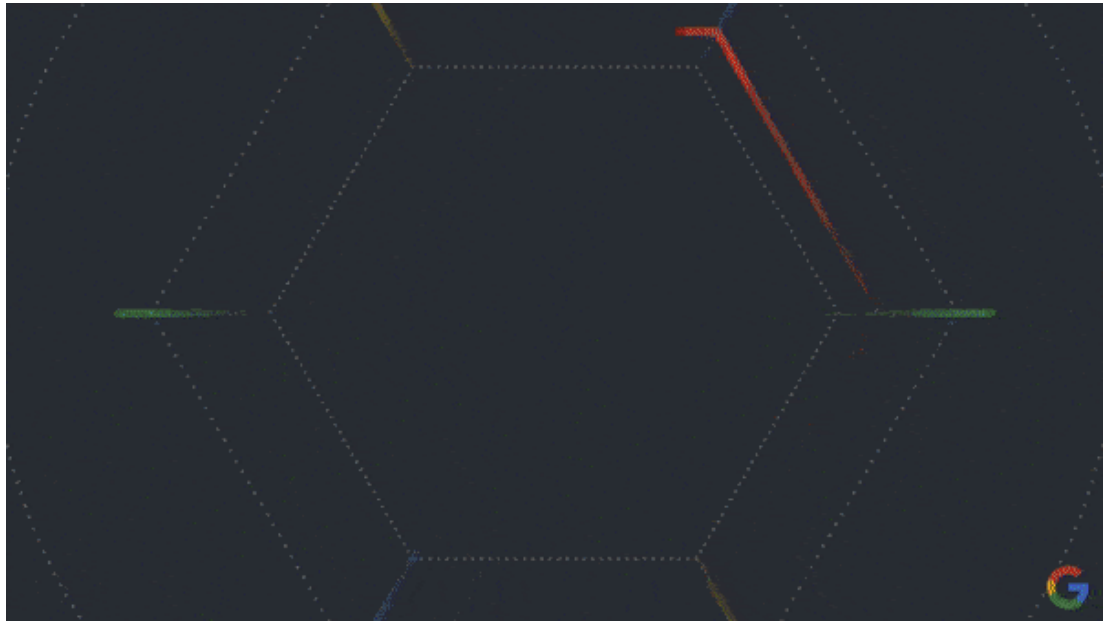


# The gap is social: 18-month-old infant vs machines

Much stronger **social** capacity:
Proactive assistance


**Physical + linguistic** capacity:
Instruction following

**SayCan**



Anh et al. (2022)

# How an 18-month-old child helps another person


Understand other people
1. single-agent behavior
2. multi-agent interactions

Interact with other people

Social Scene Understanding


More advanced topics:
- Recursive reasoning
- Communication
- Moral judgment

Multi-agent Cooperation

# Ultimate goal of social scene understanding: real-world videos



Understanding this video as a 3rd person observer

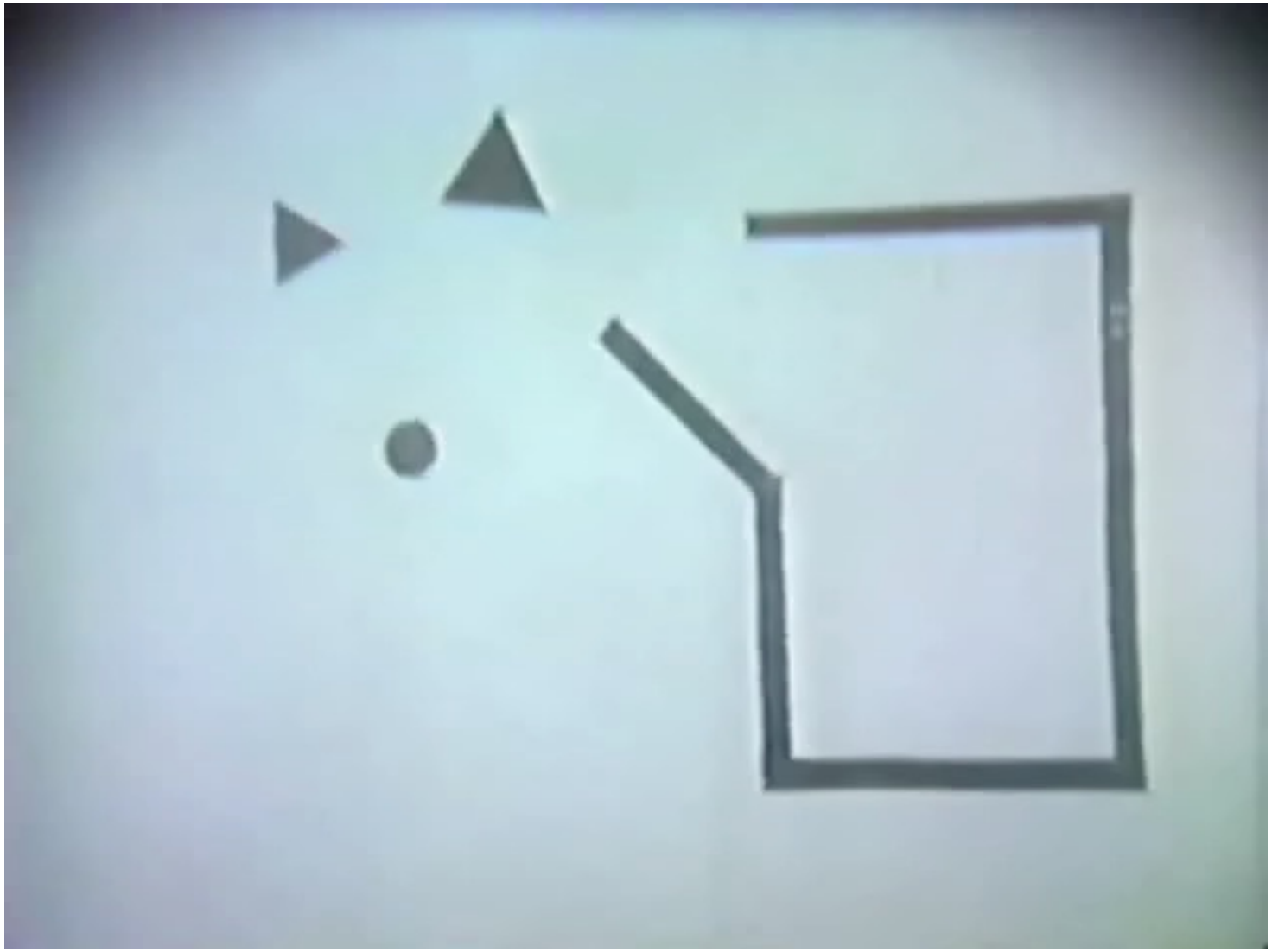
**Dynamics and constraints**: the cabinet is closed; hands are full

**Single agent's goal**: put books to the cabinet

**Multi-agent interaction**: the child is helping the adult

**Relationship**: friendly

# A classic stimulus: Heider-Simmel display



Heider & Simmel (1944)

# A classic stimulus: Heider-Simmel display

**Dynamics and constraints**

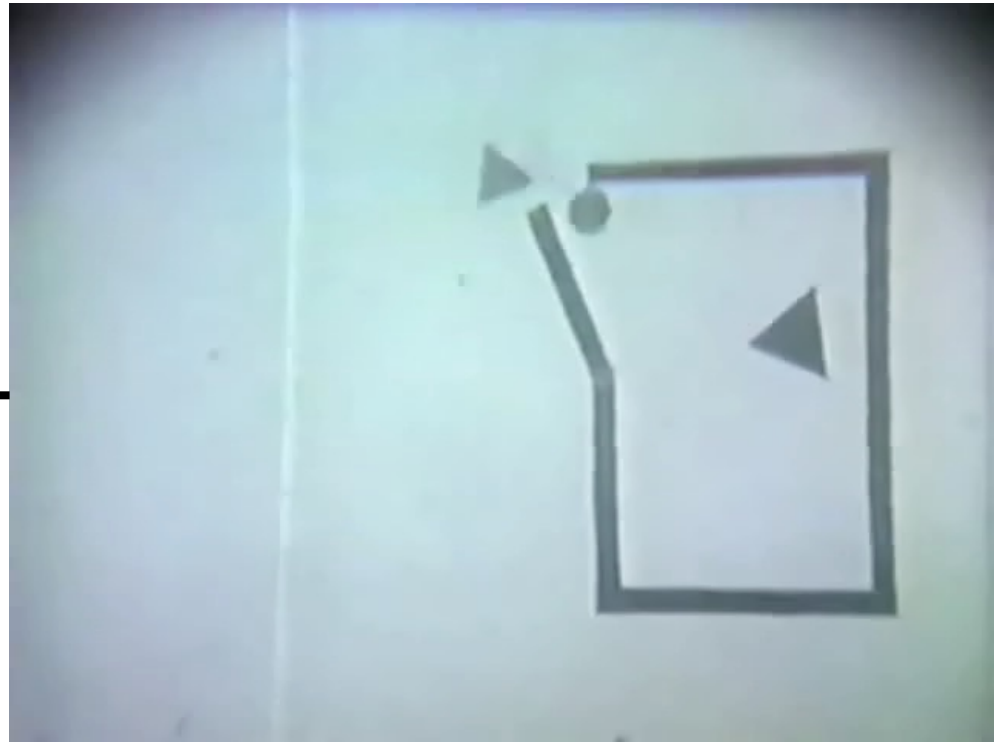
**Strengths**
strong, weak

**Goals**
helping, hurting, escaping

**Relationships**
friends, enemies

**Moral judgment**
good guy, bully

Model
?



(size / velocity / angle…)
A big triangle moves back and forth, while a small triangle and a small circle rotate 360°…

# Two pillars of social scene understandings: Intuitive psychology builds on intuitive physics


Intuitive Psychology

**Goal:**
Put the books inside the cabinet
**Difficulty:**
Cannot open the cabinet door

Intuitive Physics

**Physical constraints:**
Hands are full
The cabinet is closed

# Two pillars of social scene understandings: Intuitive psychology builds on intuitive physics

Intuitive Psychology

**Goal:**
Move boxes
**Difficulty:**
Cannot hold the stack of boxes

Intuitive Physics

**Physical constraints:**
Strengths
Stability

# Outline

## Social Interaction

## Theory of Mind

## Animacy

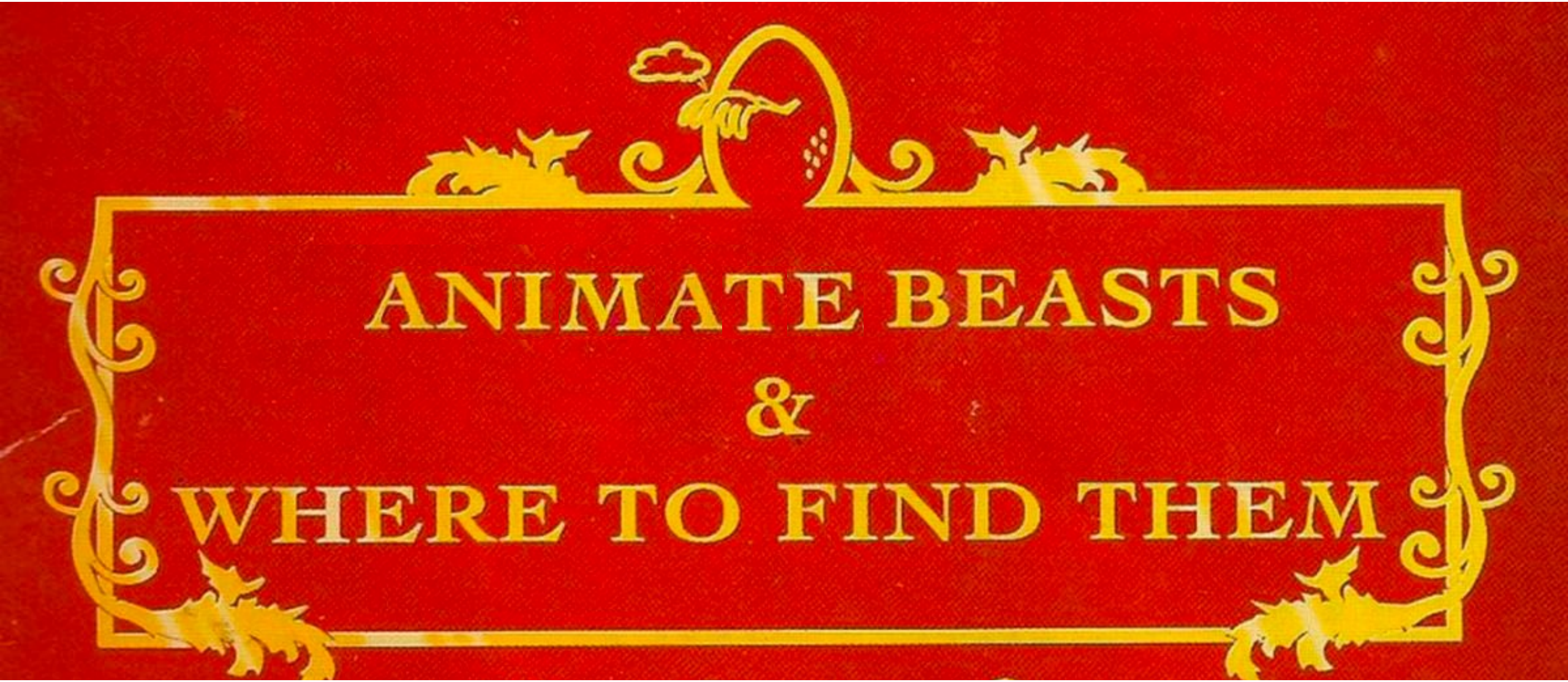
# Outline

Social Interaction

Theory of Mind

**Animacy**

# Animacy Representations & Detection



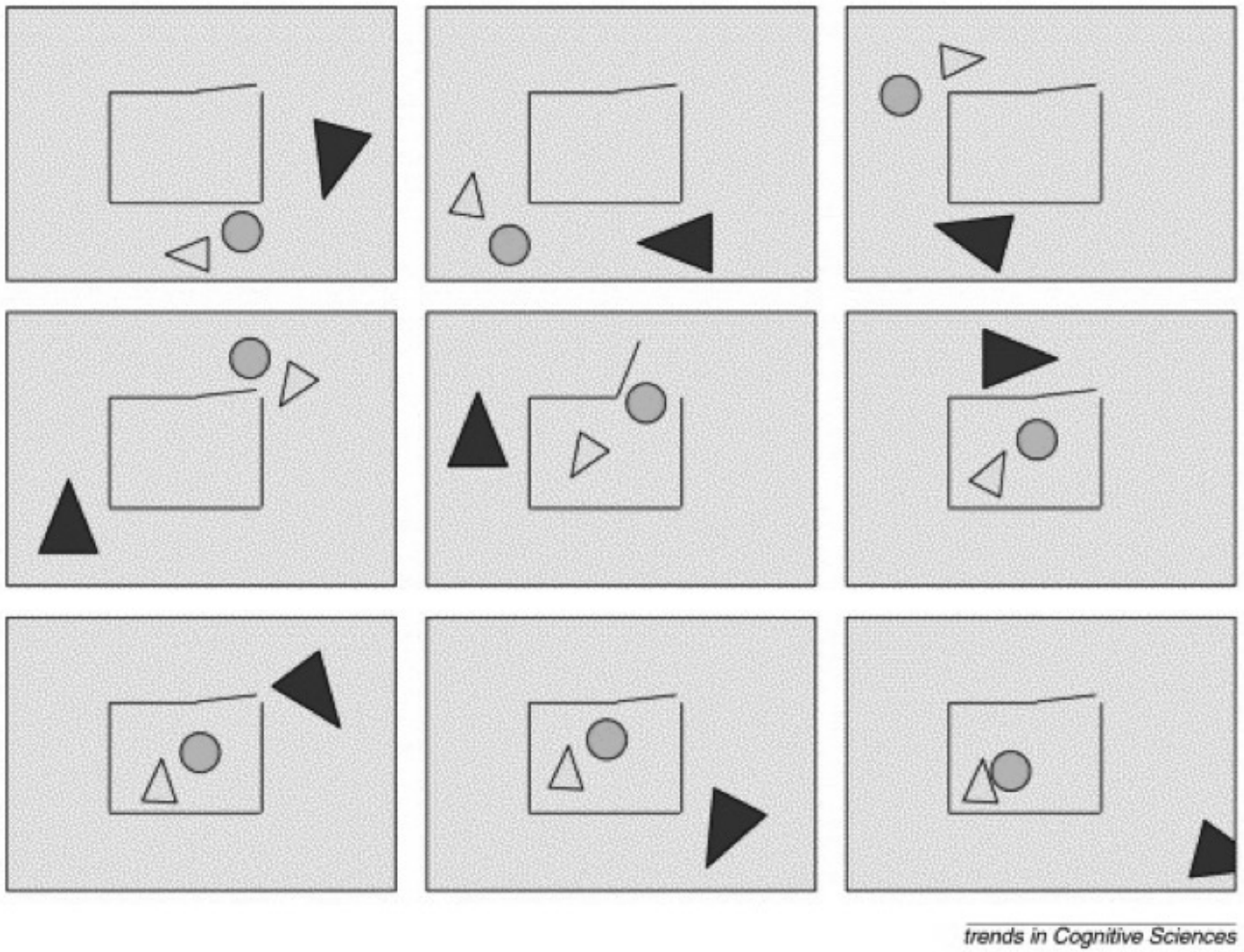
# A hierarchical definition of entities

➡ Objects: entities affected by physical laws of the world

Animates: Entities that *can* 'defy' some physical laws through internal causal power (e.g. force-generator)

➡ Agents: Entities with goals, rewards, costs, etc.

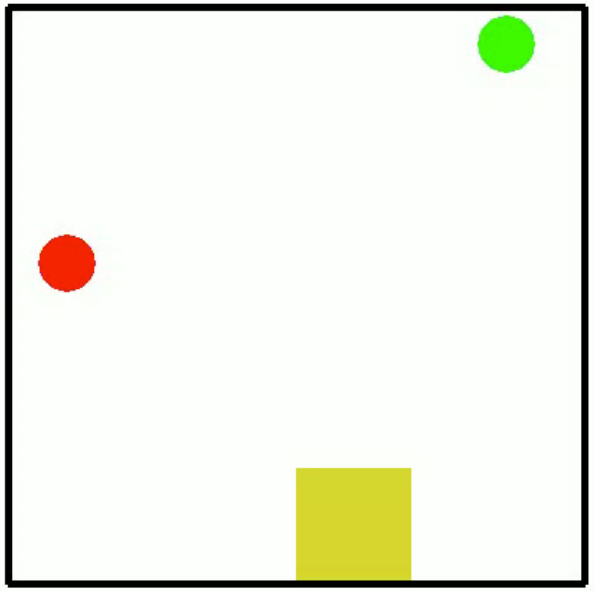
# Basic part of Heider and Simmel (1944)



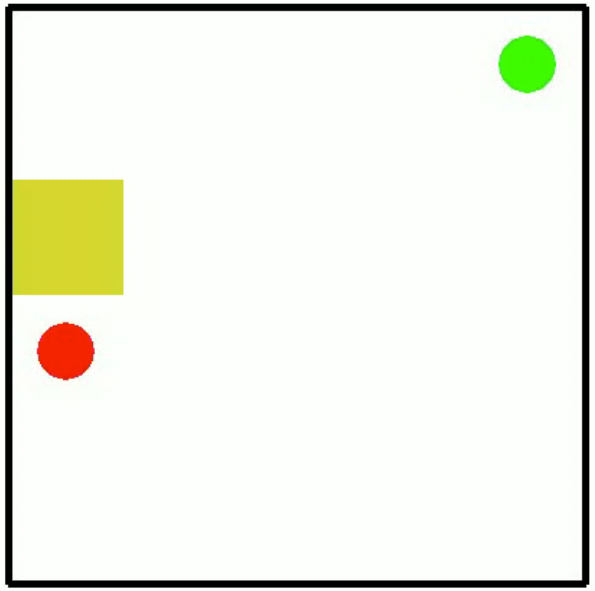
# Animacy detection

Red: Agent or Object
Green: Agent or Object

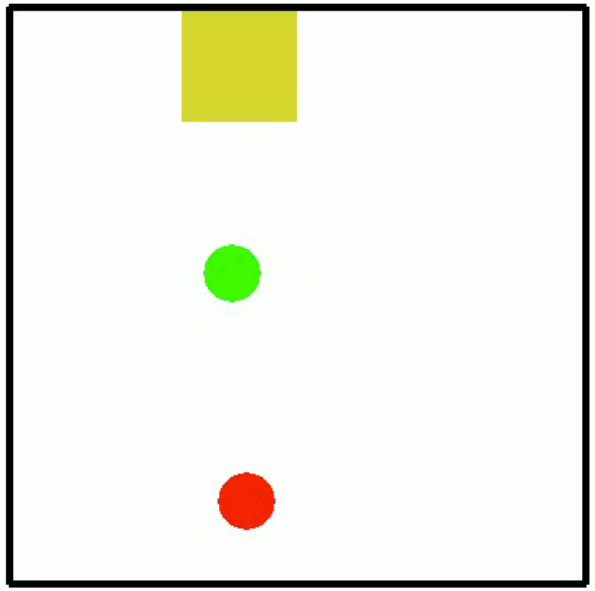
Red: Object
Green: Object



Red: Agent
Green: Object



Red: Agent
Green: Agent



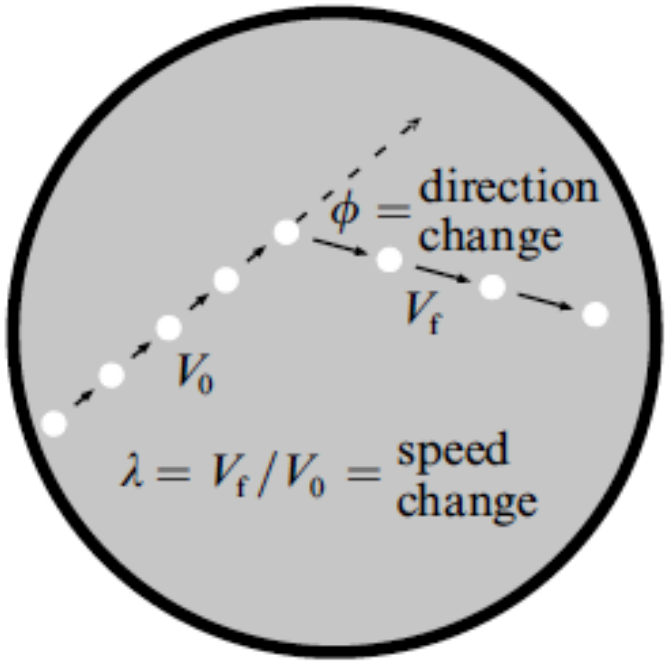
Shu et al. (2021)

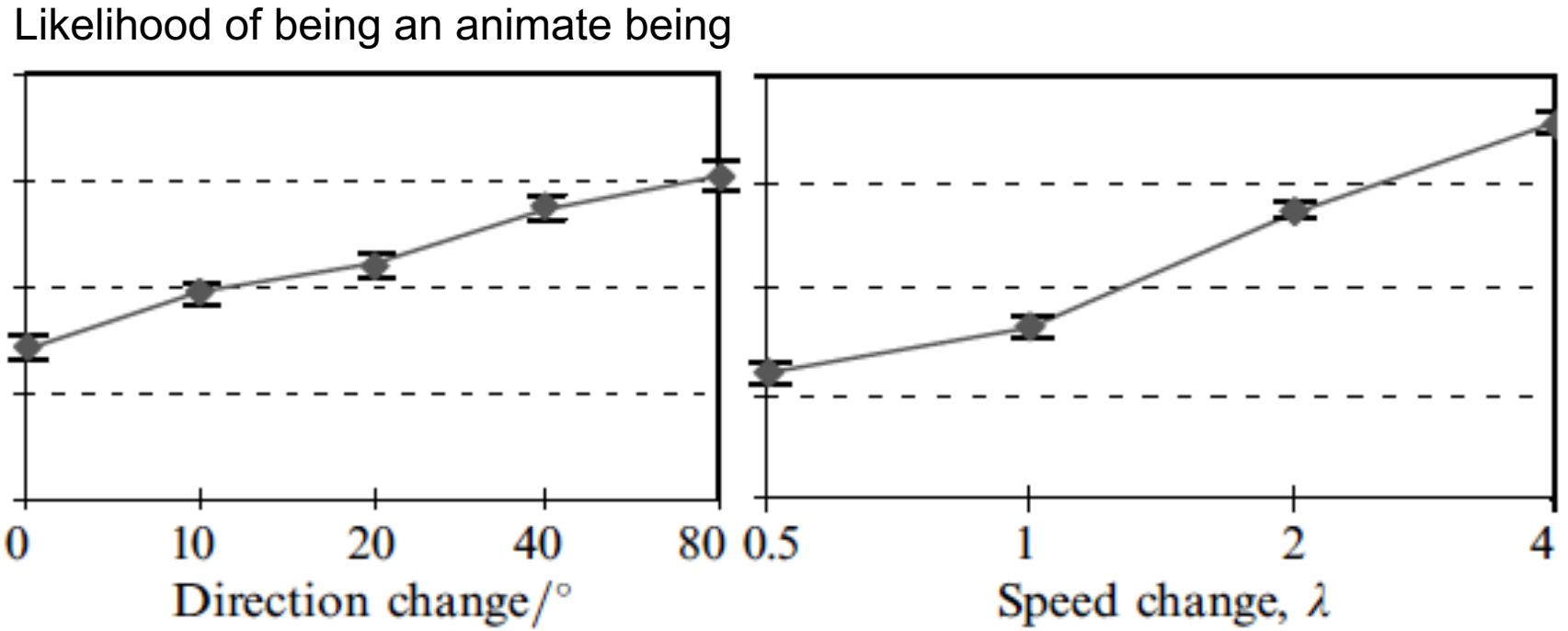
# Classic studies focus on motion cues (features)

- Face/hands/etc.
- Self propulsion / starting from rest: THE grand-daddy of a cue for animacy (starts with Dasser et al 1989? debatable)
- Sudden changes in speed or heading (e.g. Tremoulet and Feldman 2000)
- Patterns of approach and avoidance (Dittrich and Leah 1994, Gao et al 2009)
- Coordinated orientation cues
- "Biological motion": point light displays Dittrich 1993, Johannson 1973, see also evidence from
- chickens)
- "Things moving other things" (Harari et al. PNAS, computational suggestion not empirical study)
- Increases in energy (Bingham et al 1995), hidden source of energy
- Temporal contingency between shapes (e.g. Bassili, J. N. 1976)
- Others (Scholl and Tremoulet 2000)

# Example 1: motion with a single change

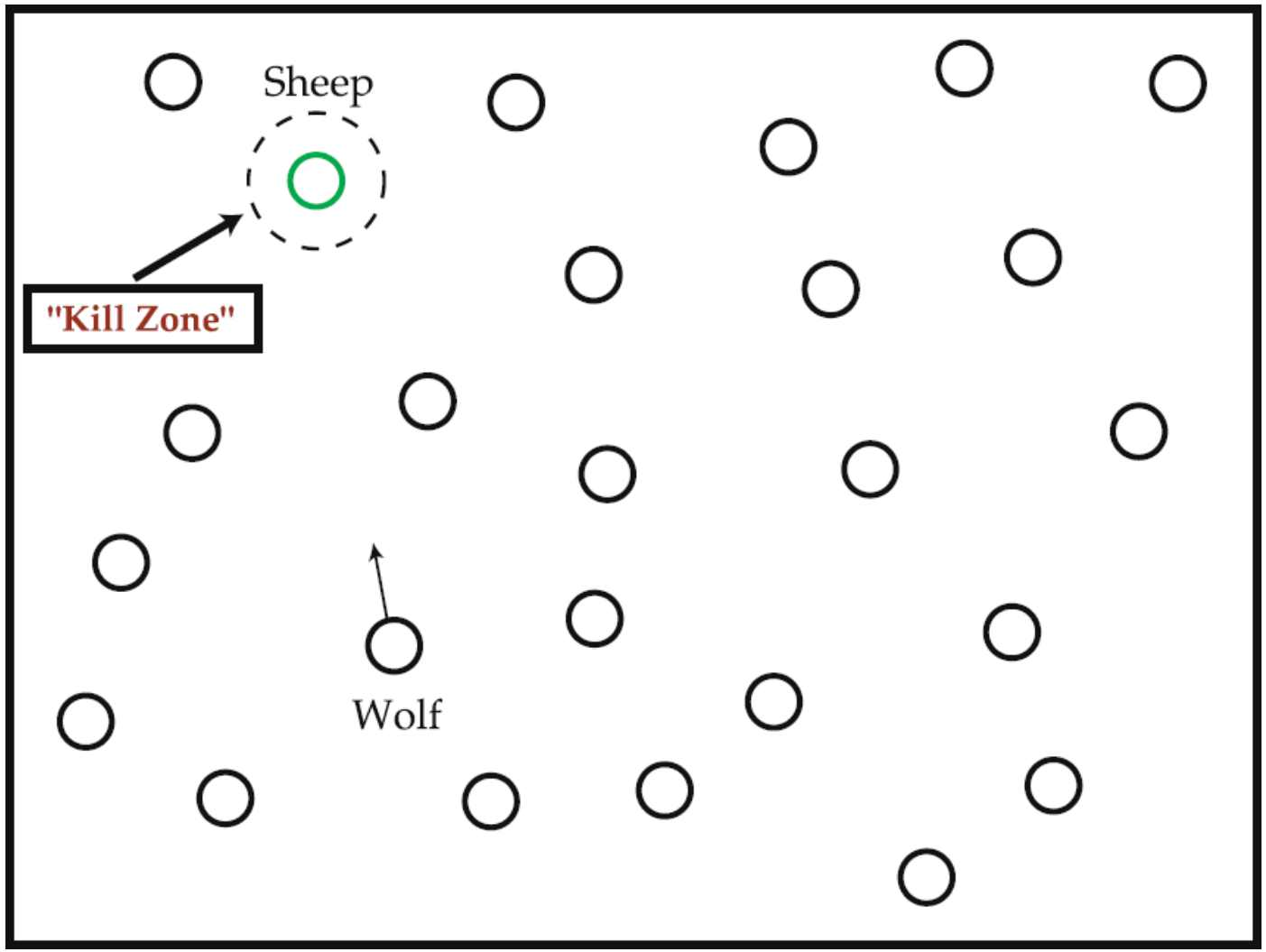
- Tremoulet and Feldman (2000)



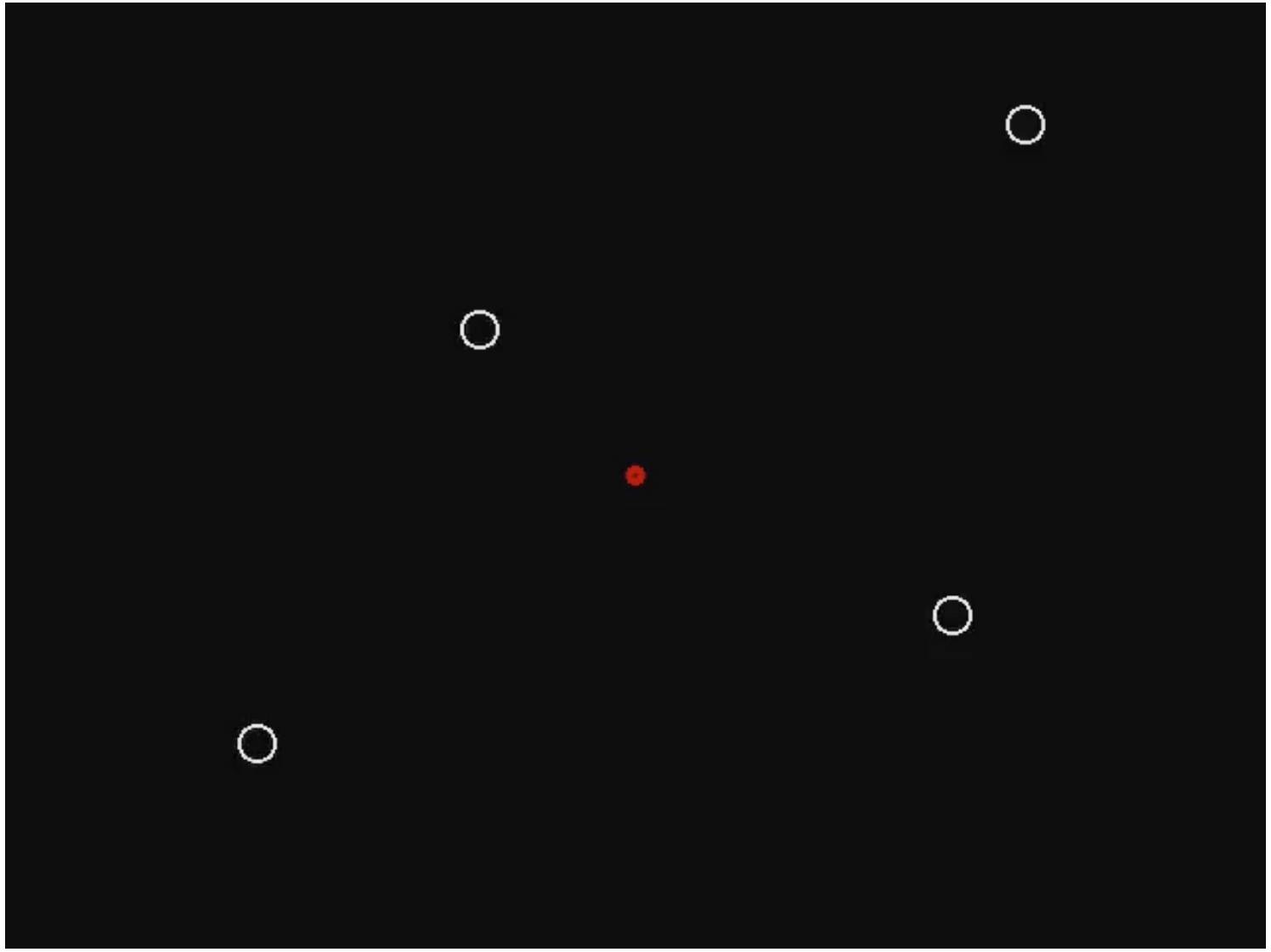
# Example 2: The psychophysics of chasing

- Tao et al. (2009)



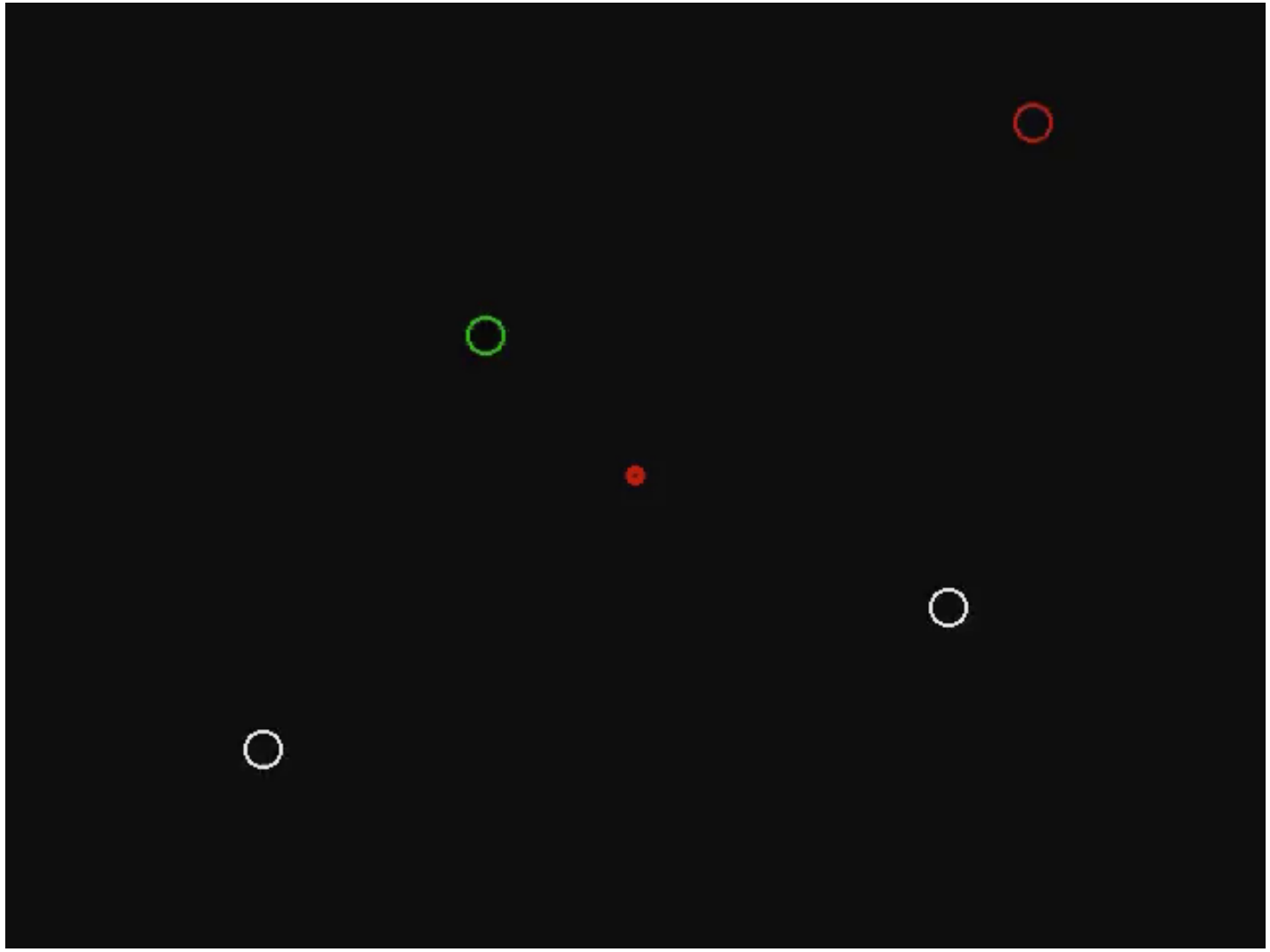
# Example 2: The psychophysics of chasing

- Clap as soon as you spot the wolf, stimulus 1



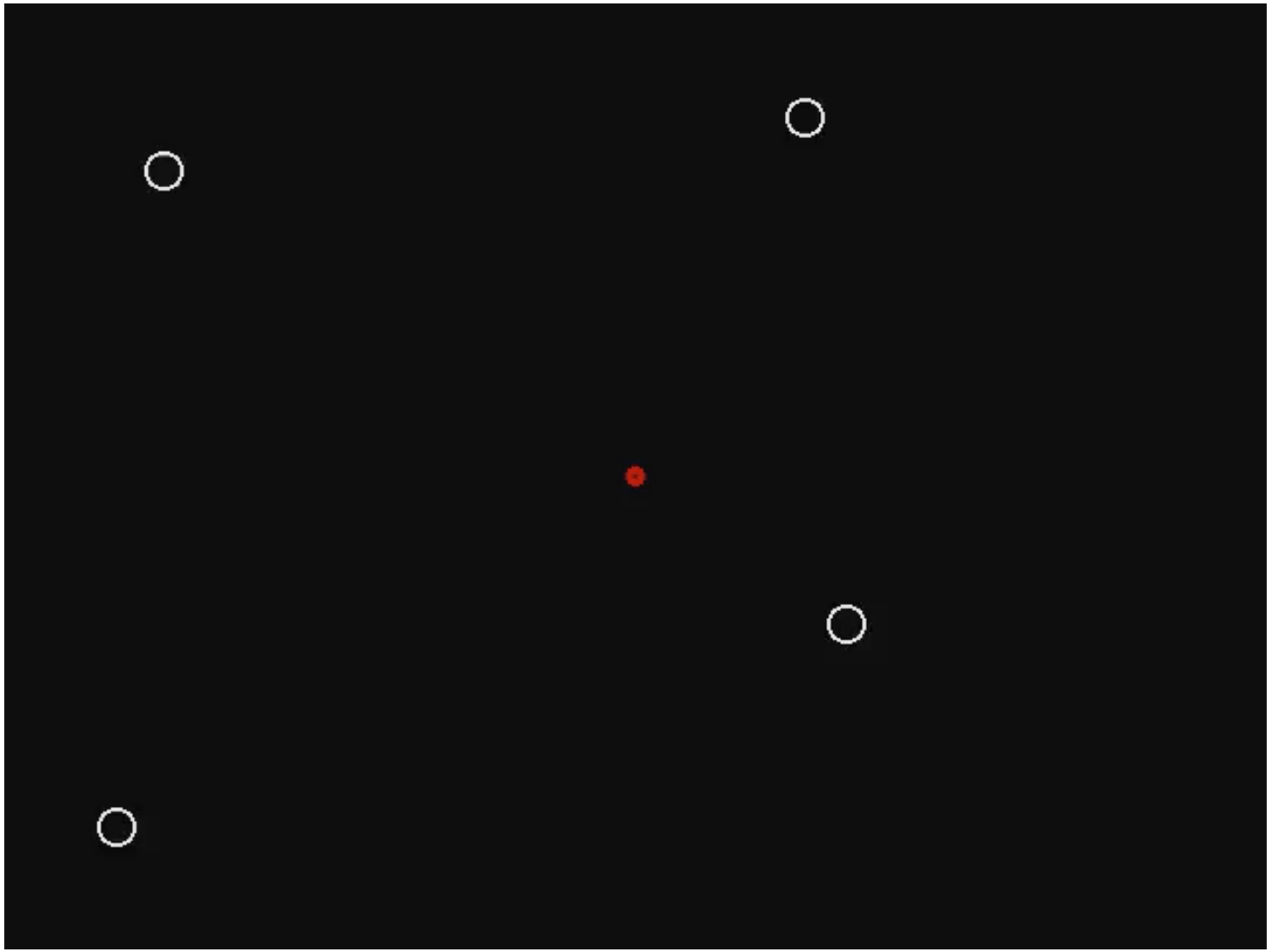
# Example 2: The psychophysics of chasing

- Clap as soon as you spot the wolf, stimulus 1 (identities revealed)



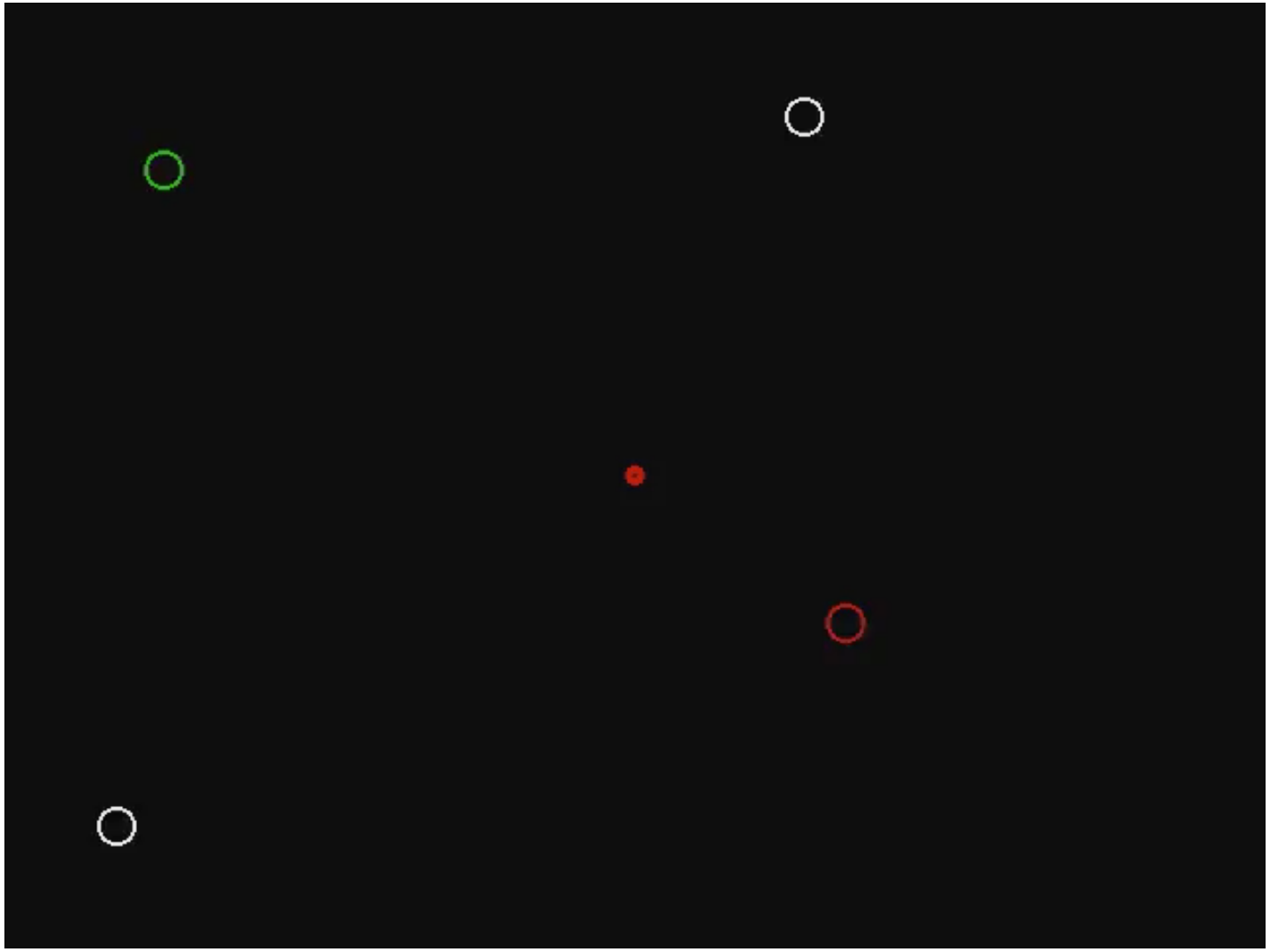
# Example 2: The psychophysics of chasing

- Clap as soon as you spot the wolf, stimulus 2



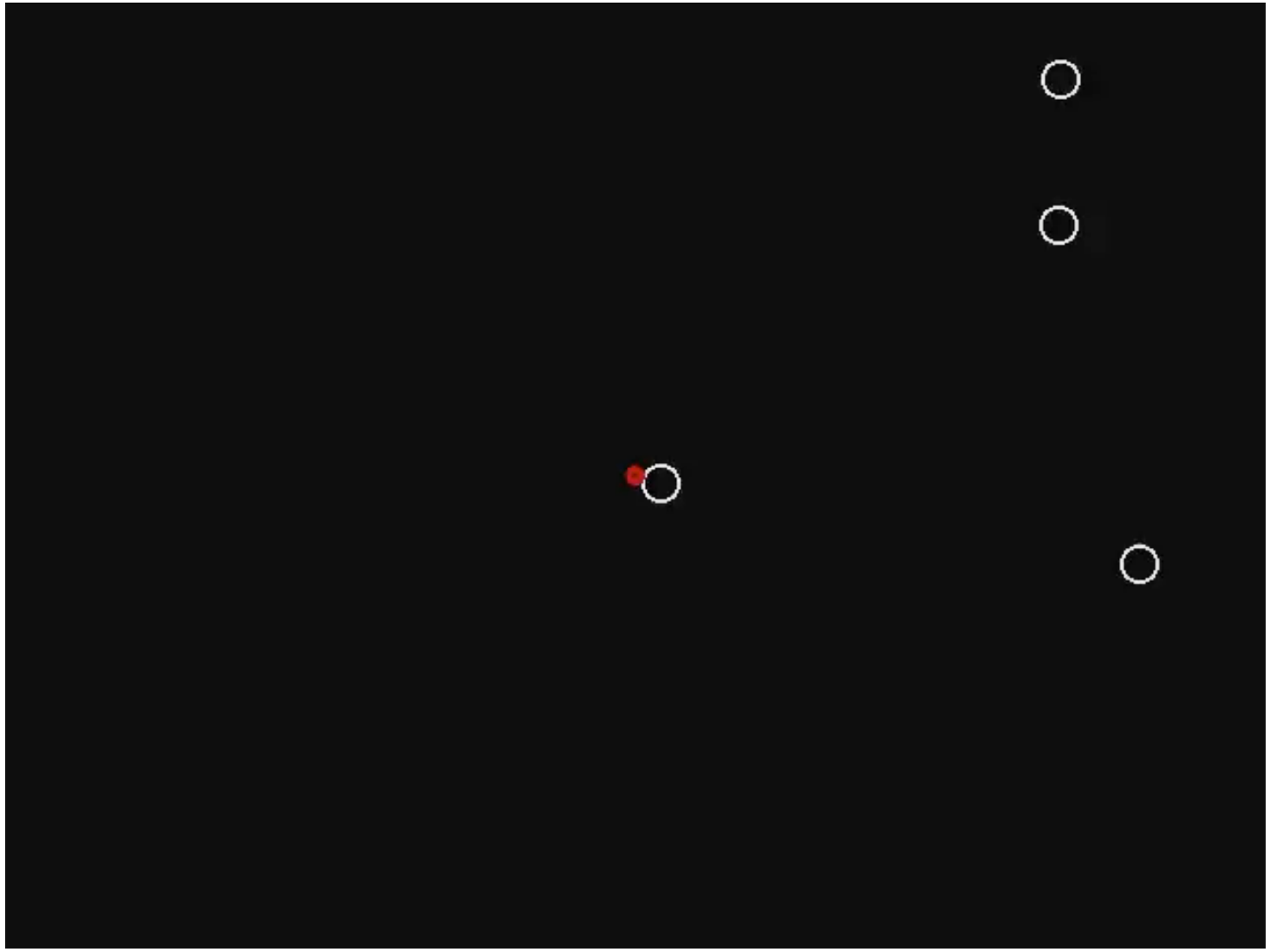
# Example 2: The psychophysics of chasing

- Clap as soon as you spot the wolf, stimulus 2 (identities revealed)



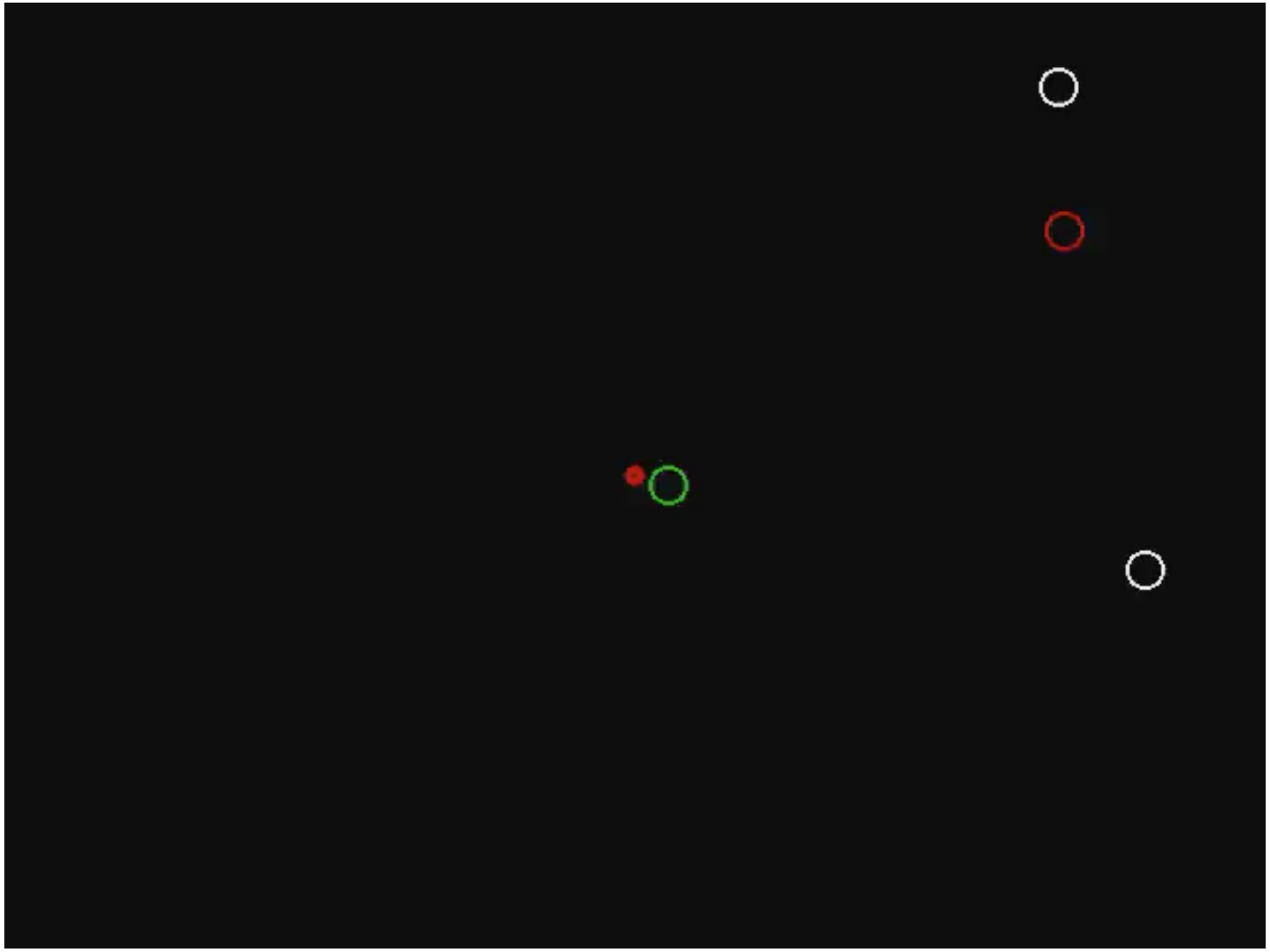
# Example 2: The psychophysics of chasing

- Clap as soon as you spot the wolf, stimulus 3



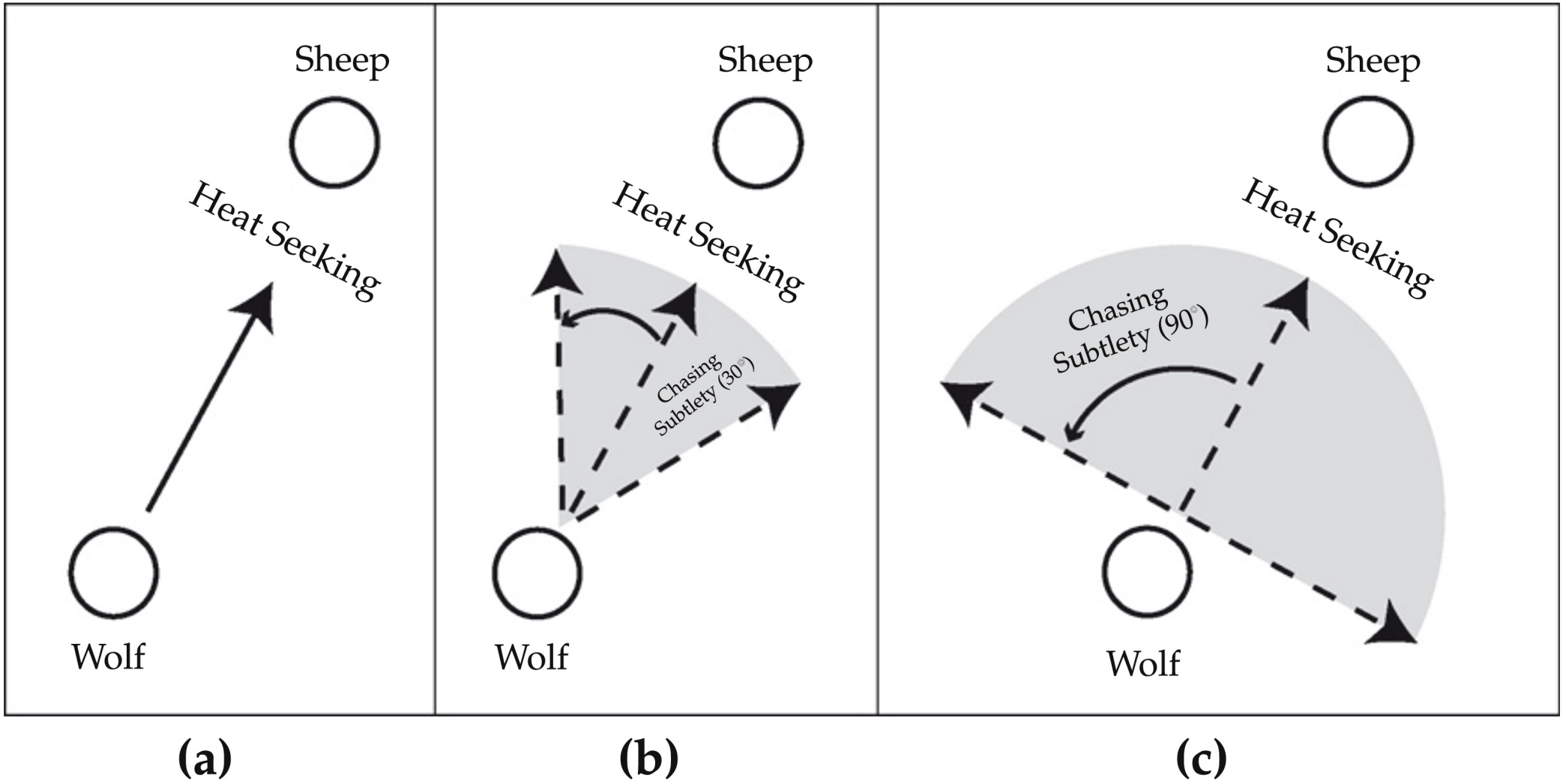
# Example 2: The psychophysics of chasing

- Clap as soon as you spot the wolf, stimulus 3 (identities revealed)



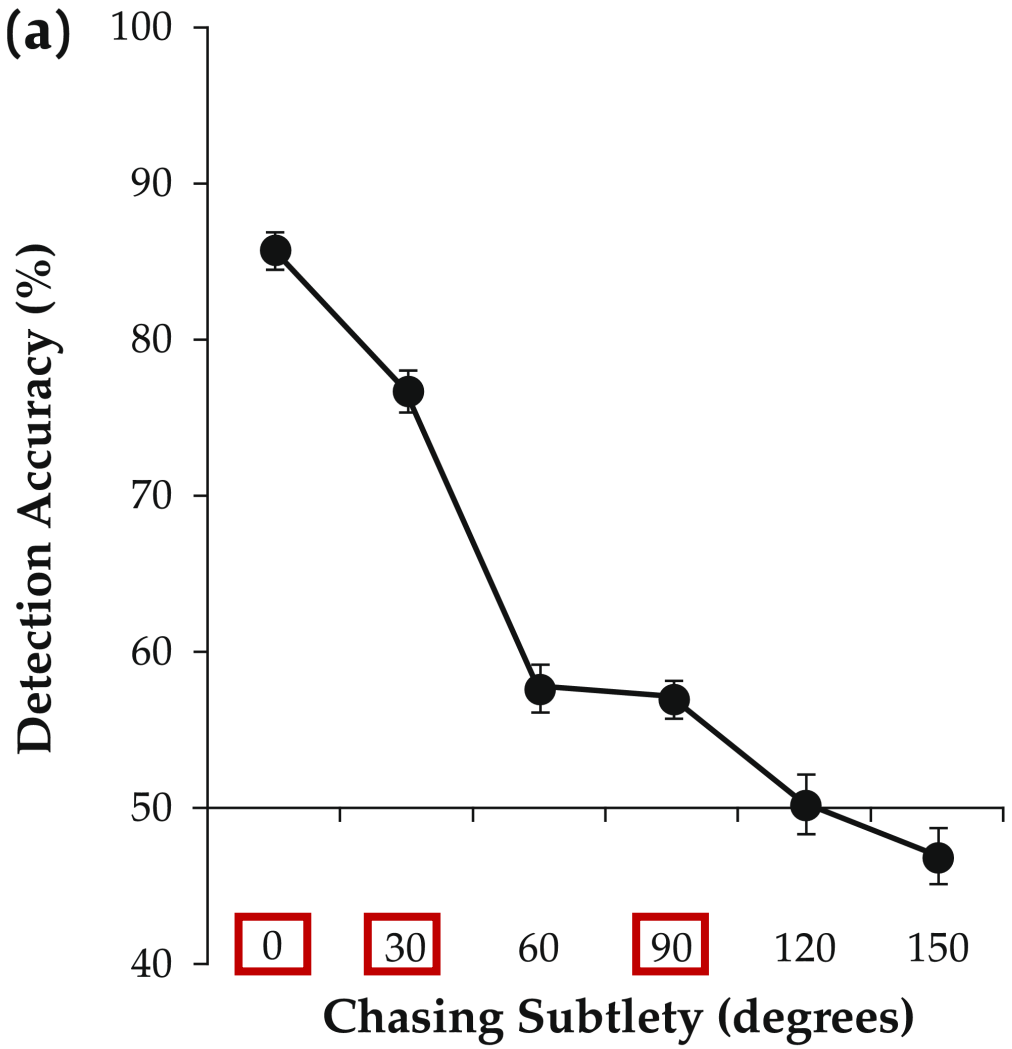
# Example 2: The psychophysics of chasing

Varying degree of chasing subtlety

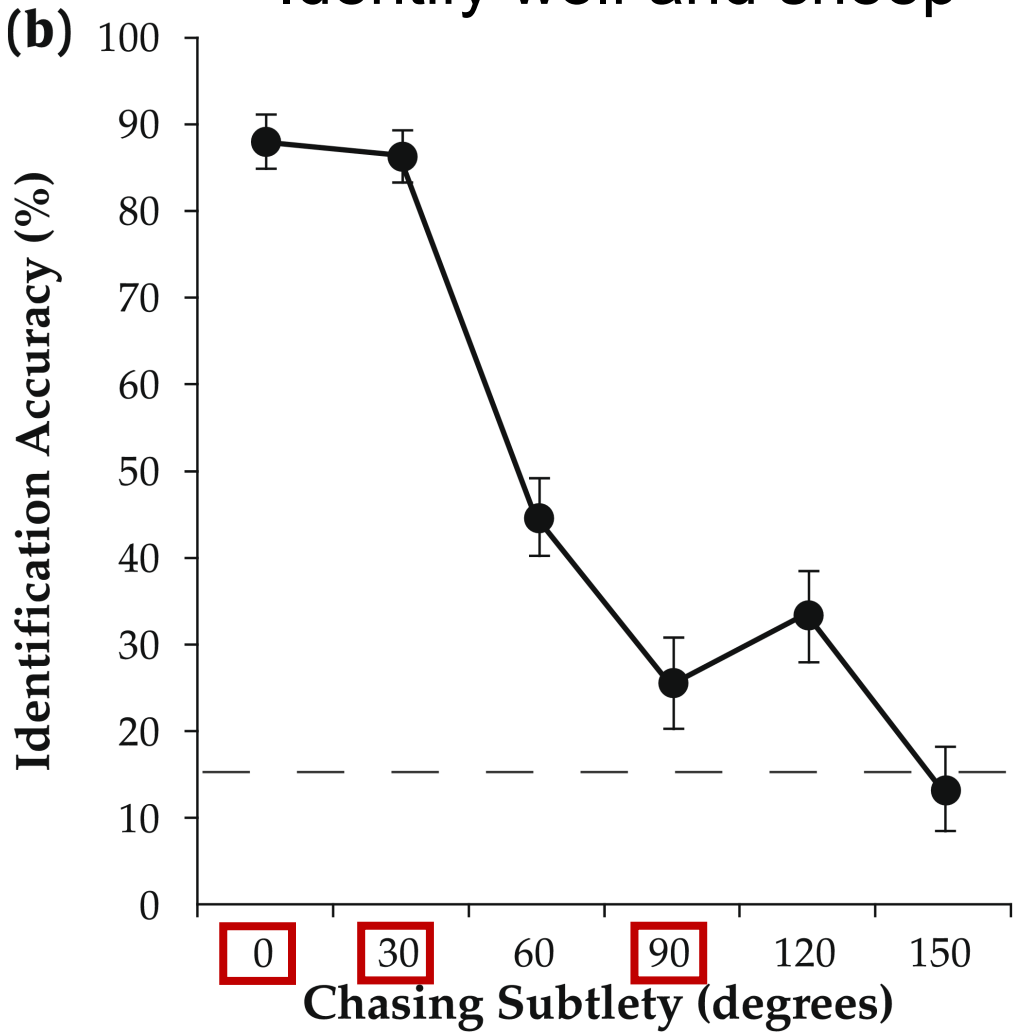


# Animacy detection vs chasing subtlety

Detect chasing




Identify wolf and sheep

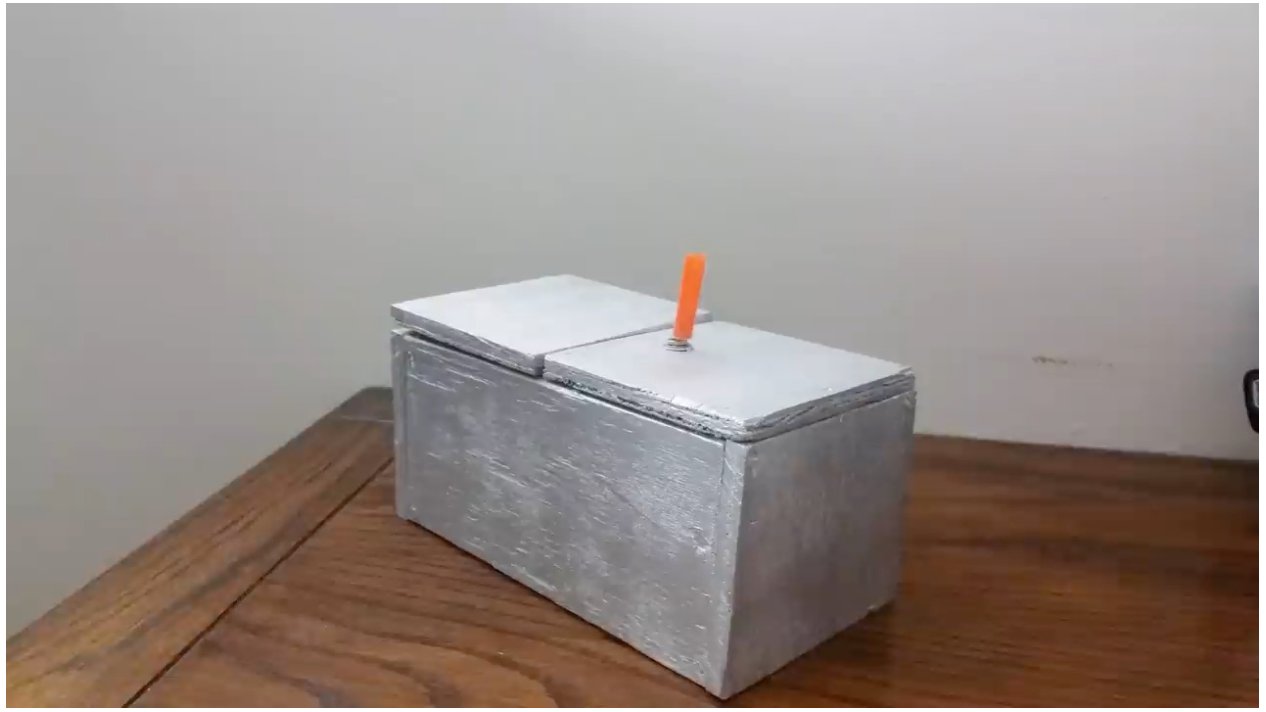


# “Useless machines”

Which machine is more like an agent?

**Useless machine A**

**Useless machine B**



Need a mentalistic representation!