

Recitation - project ideas

10/26

Tony, Grace, Alicia, Ced, Jordan, Max

Outline

- Proposal Requirements
- Example Project Proposals
- Example TA Project Ideas
- Project Brainstorming

Proposal Requirements

What do project proposals look like?

Proposals are the same for undergrad and grad projects. Your proposal should be between $\frac{1}{2}$ and 1 page. In a few sentences, with a supporting figure or sketch if necessary, your proposal should include:

- What is the phenomenon or domain you want to study?
- Why is it interesting?
- What is the specific research question that you're hoping to answer with your project?
- What is the relevant background to understand this question?
- What, concretely, are you proposing to do to answer this question (e.g., collect experimental data, implement and/or compare computational models, analyze existing datasets, etc.). You do not need to go into extensive detail - bullet points are fine.
- In a references section at the end of your document, include citations for any work that you cite in your proposal (typically this will contain between 1–3 references).

Example Project Proposals

Our project will be based on the paper *A Generative Model of People's Intuitive Theory of Emotions: Inverse Planning in Rich Social Games* by Houlihan et al., where the authors propose a formal model of people's intuitive theories of others' emotions in the context of a prisoner's dilemma game. The model proposed there is able to (1) infer the players' values and expectations, (2) generate realistic patterns of play, and (3) generate emotion predictions by inverting a generative model of social gameplay. Our aim for this project is to modify the model such that it's able to capture human emotion predictions in the context of ethical dilemmas.

To do that we will (1) identify an appropriate ethical dilemma to base the model on, (2) identify the relevant latent features involved in the ethical dilemma, (3) modify the computational model replacing the latent features, (4) set up web-based experiments using Amazon mTurk to gather human data on what peoples' intuitions for emotion predictions are, and (5) compare the model's emotion predictions to the human ones and scale the model appropriately. For step (4), unlike Houlihan et al. we don't have real-life videos of the scenario we are interested in, so instead will we show to participants written descriptions of the ethical dilemma. Similarly to Houlihan et al. participants will then have them rate the intensity of 20 emotions on a scale from 1 to 7.

Example Project Proposals

For our project, we are seeking to further investigate the topics of intuitive physics and simulation presented in lecture. We were inspired by the work done by Battiglia, Hamrick, and Tenenbaum in Simulation as an engine of physical scene understanding (<http://www.pnas.org/content/110/45/18327.short>) and would like to reproduce parts of this study in a video game setting. As such we design a game in which a scene is presented to the player, and the player must make inferences on the outcome of the scene. A scene might consist of a ball sitting on a sloped table or an array of objects precariously balanced on a dresser. The player must make inferences based on these scenes, such as whether the objects will stay in place or fall, and where objects will land if they do indeed fall. This type of physics simulator can be made using the Unity engine, in which there are a variety of physics engines that can be used. Player interactions can then be quantified by boolean values and Euclidean distance where applicable. Based on the human interactions with certain scenes in the game, we then build a probabilistic WebPPL model defining human inferences about physics. After developing this model, we compare the model based on human experimental data with a model generated from Bayesian probability and laws of physics and highlight similarities and differences. We are interested in seeing which aspects of this inference match with accepted laws of physics, and which aspects a simulation performs better in.

Example Project Proposals

Title: Further Investigations on the Optimality of Everyday Cognition (tentative)

In *Optimal Predictions in Everyday Cognition* [1], Griffiths and Tenenbaum noted the perceived disconnect between human perception/memory and cognitive judgements; whereas the former are seen as optimal inferences based on accurate prior probabilities, the latter are often seen as following error-prone heuristics that are oblivious to priors. The paper examined the actual optimality of human cognition by asking people to make predictions about various numerical values given limited information, and compared them with the outputs of Bayesian prediction functions given various priors (power-law, Gaussian, and Erlang). The results suggested that everyday cognitive judgements in fact follow the same optimal statistical principles as perception and memory.

The core of the project will seek to reimplement and replicate the results from this paper, effectively as an extension to Problem 3 from Problem Set 1. We will gather data in person from other undergraduates, asking them similar questions to those given as examples in the paper.

Example Project Proposals

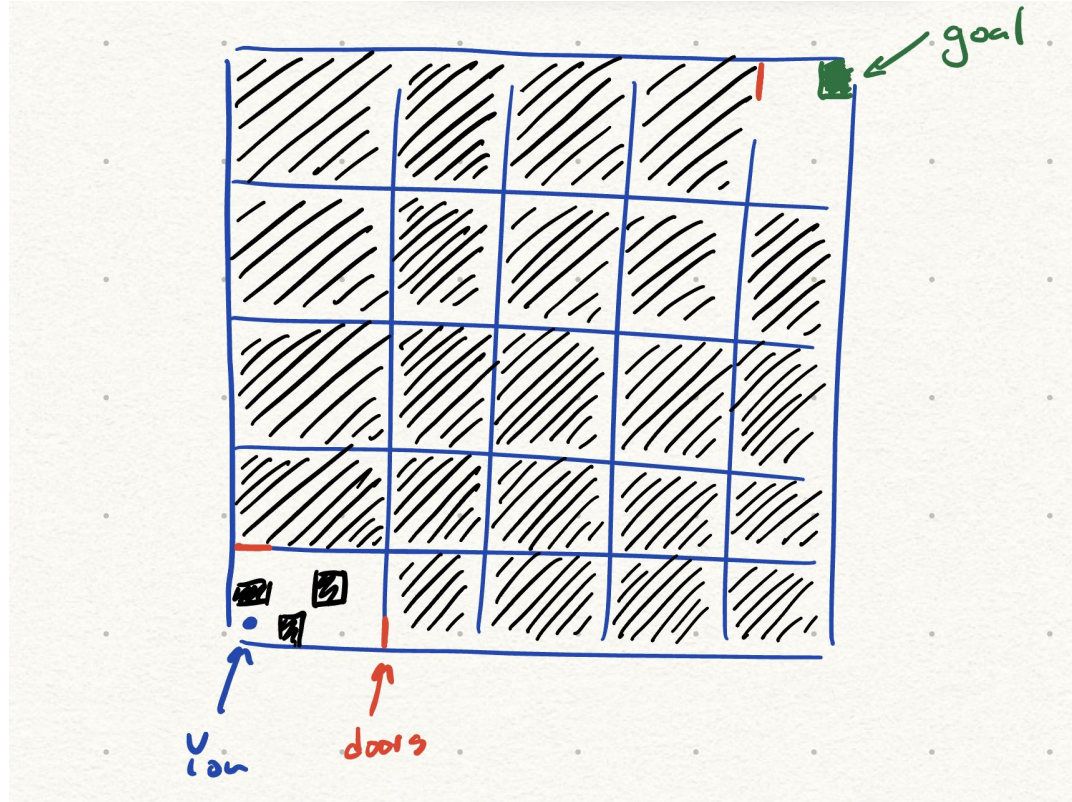
To interpret an acoustic signal, humans must separate the effects of direct sound and reverberation. This poses a computational challenge, as the signal received by the ear combines information from the source and environment. Specifically, the effect of the reflections arriving at an ear can be described by a single linear filter, $h(t)$, and the sound that reaches the ear as the convolution of this filter with the sound of the source: $y(t) = h(t) * s(t)$. How, then, do humans separate the effects of reverberation from the effects of an object?

If humans fail to separate these effects, then subjects may falsely attribute the length of decay to the modes of the object material. In this case, we would predict subjects to systematically misidentify wood as metal or glass in echoic environments (and vice versa in dry environments). On the other hand, if humans can easily separate the effects of the object and the environment, then there are two patterns of behavior we might observe. First, humans may do very well on the task and make near-perfect judgments. Second, longer reverb lengths from the environment may “explain away” longer modes from the material. In this case, then we expect subjects to systematically misidentify metal as wood when they hear longer reverb.

Tony

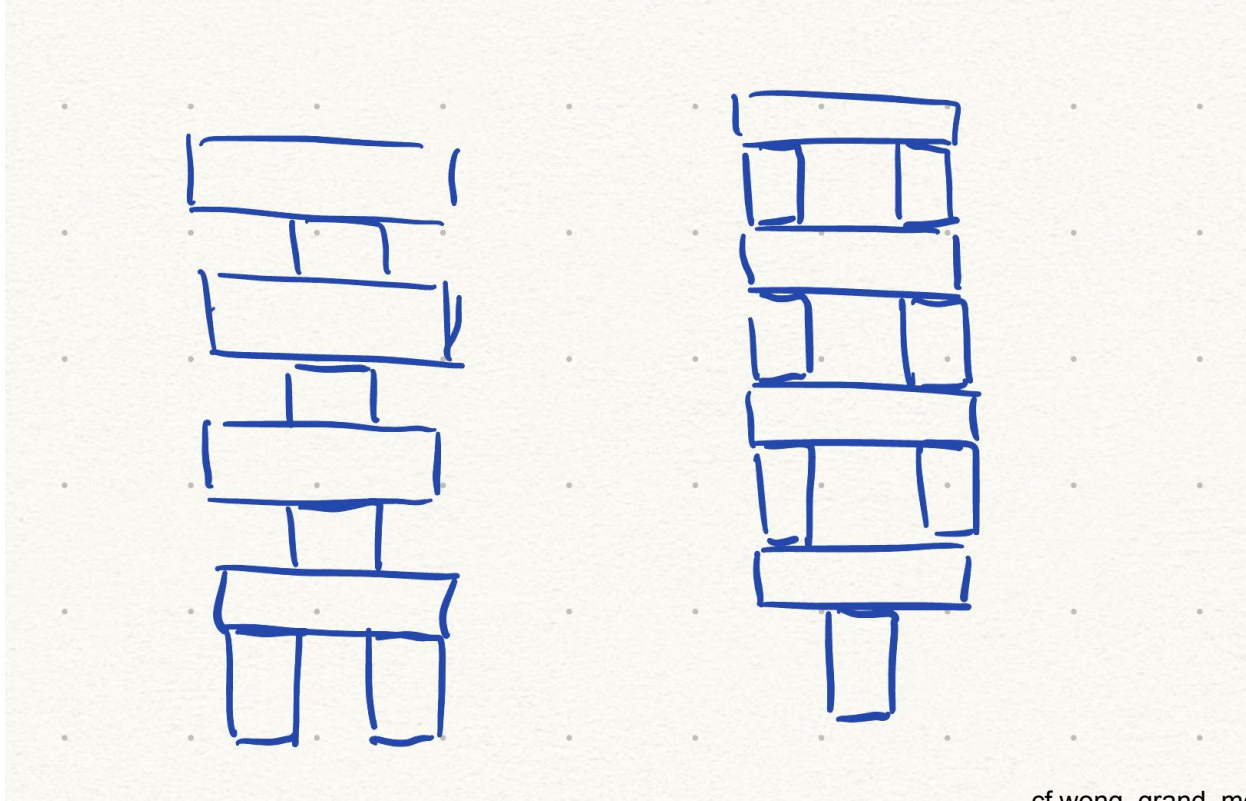
- planning
- inverse planning / theory of mind
- representations and abstractions

Idea 1: Forward vs Backwards planning in navigation

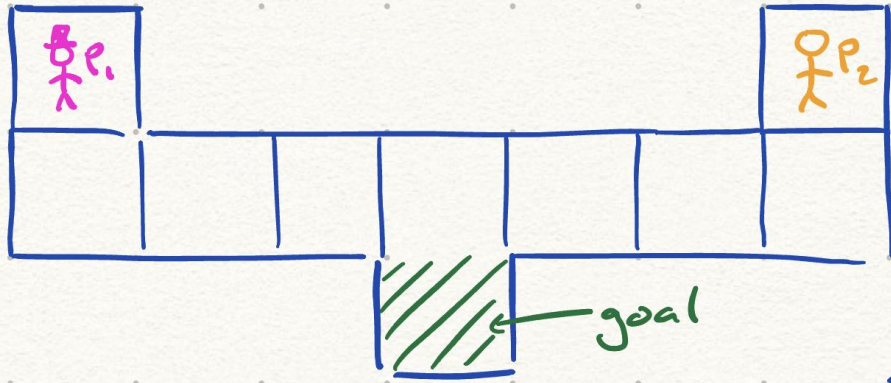


cf afsardeir, karamati 2018
simon, daw 2011
sharp, eldar 2023

Idea 2: Program representations for tower building



Idea : learning norms to break symmetry



$$R((s_1, s_2), (a_1, a_2)) = \begin{cases} 1 & \text{if } s_1 = s_2 = \text{goal} \\ -1 & \text{otherwise} \end{cases}$$

Idea : dog-whistling

A little clunky but could work: <https://aclanthology.org/2020.pam-1.10.pdf>

Jordan+Max:

- 3-8 Players
Ages 13+



Jordan+Max: Spyfall

- Premise of the game is that there is a spy among the players; everyone but the spy knows the location that the players are at
- Players take turns asking each other questions trying to figure out who the spy is but have to make sure the spy doesn't figure out the location
- Win/Lose conditions:
 - Players guess the spy correctly at the end of time:
 - Spy figured out the location: Spy wins
 - Else: Players win
 - Players guess incorrectly: Spy wins whether or not they figured out the location

Jordan+Max: Spyfall

- This is hard to model directly because the question/answer space is intractable (if you have thoughts to make this tractable/doable with LLMs that would be a really interesting project)
- However, we can rephrase the problem as everyone but Y knows X and they want to ask questions that minimize the chance of Y figuring out X while maximizing the chance of catching Y
- In some ways this simplification is similar to a variant of the number game: you want to figure out a number X by using mathematical properties

Jordan+Max: Spyfall

- Given a number X chosen at random in some range, we want to model the best questions to ask depending on which team you're on
- One way we can do this is using loopholes:
 - “Rather than comply or directly refuse, people can subvert an intended request by an intentional misunderstanding.”
- If interested I recommend looking at “Ambivalence by Design: A Computational Account of Loopholes” (Peng et al., 2023)

The (expected) total utility of an action a for L_1 is a combination of the listener's own direct goals, the expected utility of the speaker S , and the social cost of the action:

$$\mathcal{U}_{\text{total}}(a) = \rho \cdot \mathbb{E}_{P(m|u)}[\mathcal{U}_S(a; m)] + \mathcal{U}_{L_1}(a) - \mathcal{C}_{\text{social}}(a), \quad (2)$$

where $\mathcal{U}_{L_1}(a)$ reflects the listener's direct goal (the utility they associate with the action); $\mathcal{U}_S(a; m)$ reflects the speaker's utility, conditioned on a given interpretation m of their intended goals; ρ is a hyper-parameter ranging from -1 to $+1$ that represents the trade-off between the listener's goals and the speaker's goals, similar to the Welfare

Jordan+Max: Spyfall Extension

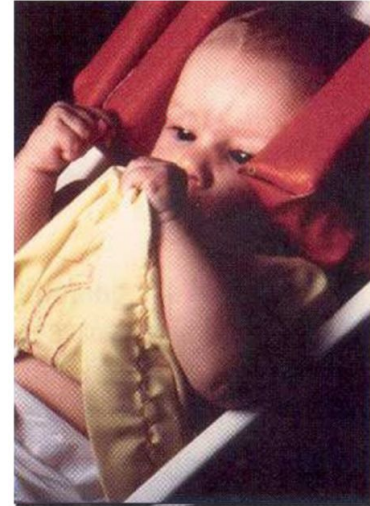
- Probably a bit advanced for undergrads/the simplified number game would be enough of an extension but for grad students/people really interested in this
- Can the simplified game explain the actions/beliefs of people playing Spyfall?
 - In the simplified game, the “question” you ask has a concrete sense of informativeness as the amount of numbers that it rules out
 - But can we relate this to Spyfall where the question/answers are truly intractable and where understanding the informativeness of a question requires understanding natural language?
 - Perhaps ChatGPT might be able to define this? Or perhaps playing multiple games with people and tracking their actions/questions and asking about their beliefs during and after might be insightful

Jordan+Max: Understanding Infant Attention

How can we study infant cognition?

Infants can't speak; can't move very well;
can't understand much language

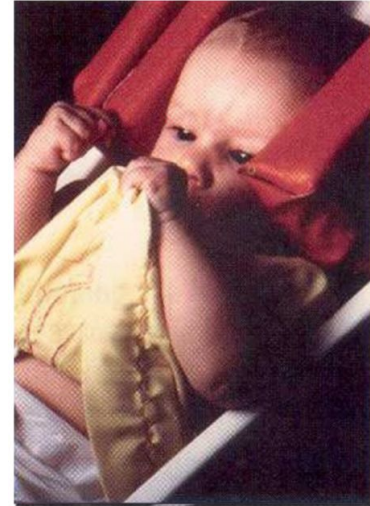
All they really do is stare at things



Jordan+Max: Understanding Infant Attention

This is enough to learn a lot about infants!

Like adults, infants get bored after staring at the same thing for awhile.



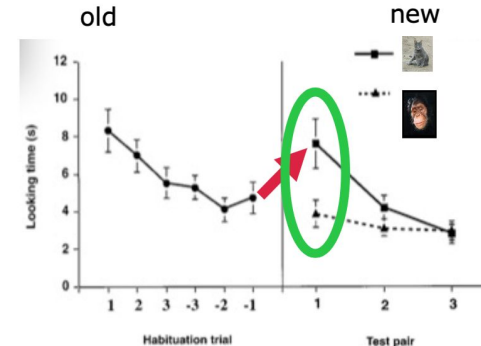
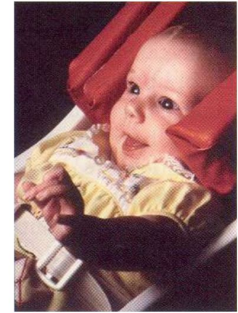
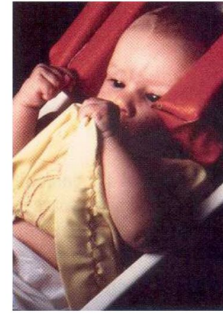
Jordan+Max: Understanding Infant Attention

When infants see something novel, they become interested again.

So, by presenting various stimuli we have a measure of novelty - that the infant detected a difference, when they look away

Many, many studies have used this tool to study the infant mind.

what happens when we change the stimulus?

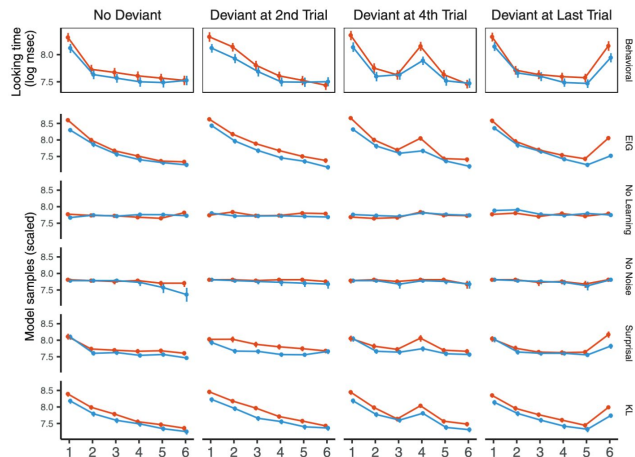


Jordan+Max: Understanding Infant Attention

Why do infants prefer novelty? What factors influence how long they spend looking at a display? Raz et al. (2023) built a computational model which explains looking as a *rational decision to acquire more information*.

Infants look at a display until they decide there is no more information to be gained.

The model works on the same stimuli that infants are shown



Jordan+Max: Understanding Infant Attention

However, it is limited in a few key aspects

- Modeling assumptions limit applicability
 - For example, good for perceptual concepts but not necessarily higher-level concepts
 - Can only handle displays containing one object
- Using probabilistic programming, let's extend the model!

Ced

- Game playing
- Learning world models from natural language
- Vagueness and uncertainty

Game playing 1

Connect 4



Game playing 2

Liar's dice



Learning world models from natural language

A. Prompt, containing unrelated example world model

```
;; We define a probabilistic model in Church of the following scenario.
;; At any given time, about 1% of the population has lung cancer,
;; 20% have a cold, 10% have a stomach flu, and 0.5% have TB.
(define lung-cancer (mem (lambda (person) (flip 0.01))))
(define cold (mem (lambda (person) (flip 0.2))))
(define stomach-flu (mem (lambda (person) (flip 0.1))))
(define TB (mem (lambda (person) (flip 0.005))))

;; If you have a cold, there's a 50% chance you have a cough.
;; 30% of people with lung cancer have a cough, and 70% with TB.
;; There's also a small chance you have a cough even if you're otherwise healthy.
(define cough (mem (lambda (person)
  (or (and (cold person) (flip 0.5))
      (and (lung-cancer person) (flip 0.3))
      (and (TB person) (flip 0.7))
      (flip 0.01))))))

;; Whether a person coughs during a particular visit to the doctor's office
;; depends on whether they have a cough, and a bit of random chance.
;; Note that this will differ each time they go to the doctor's office, so
;; we do not use 'mem' (which memoizes the result).
(define coughs-on-particular-visit (lambda (person) (and (cough person) (flip 0.7))))
```

B. Defining a new world model from scratch via language-to-code translation

```
;; Now, let's define a different probabilistic model of the following scenario.
;; It is totally unrelated to the previous model and does not reference the functions above.
```

B. Defining a new world model from scratch via language-to-code translation

```
;; Now, let's define a different probabilistic model of the following scenario.
;; It is totally unrelated to the previous model and does not reference the functions above.
```

First, strength levels vary widely from person to person.

```
 (define strength (mem (lambda (person) (normal 100 20))))
```


Furthermore, each person has a percentage of the time that they are lazy.

```
 (define laziness (mem (lambda (person) (uniform 0 1))))
```

The strength of a team is the combined strength of its members, except that in any given match, each player may decide to be lazy, and thus contribute only half of their strength.

```
(define team-strength
  (lambda (members)
    (apply + (map (lambda (member)
      (if (flip (laziness member))
          (/ (strength member) 2)
          (strength member)))
      members)))))
```

Whether one team beats another just depends on which team pulls stronger that match.

```
 (define team-beats-team
  (lambda (team1 team2)
    (> (team-strength team1) (team-strength team2)))))
```

Vagueness and uncertainty

Design experiments to
test human thresholds

“some” “most” “possible”
“heavy”
“often”
“a few” “tall” “short”
“strong”

Alicia

Alicia

Questions I'm interested in:

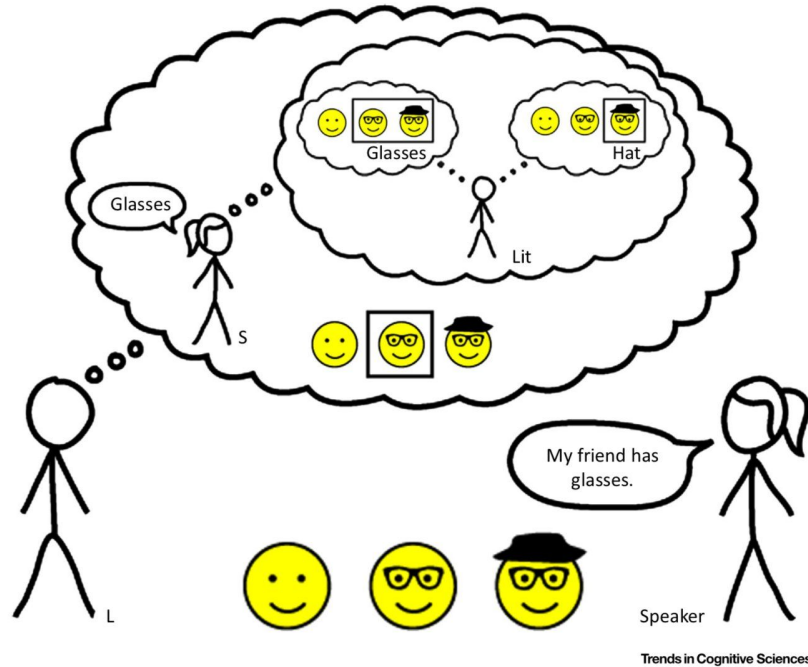
- How do people computationally represent social relationships and how do they use this information to
 - form expectations and norms for how to act
 - evaluate other people's actions
 - communicate about and change the relationship?

Alicia

Questions I'm interested in:

- How do people computationally represent social relationships and how do they use this information to
 - form expectations and norms for how to act
 - evaluate other people's actions
 - **communicate about and change the relationship?**

One approach: rational social inference models



Trends in Cognitive Sciences

Figure 1. Application of Rational Speech Act-Style Reasoning to a Signaling Game. The three faces along the bottom show the signaling game context. Agents are depicted as reasoning recursively about one another's beliefs: listener L reasons about an internal representation of a speaker S, who in turn is modeled as reasoning about a simplified literal listener, Lit. Boxes around targets in the reference game denote interpretations available to a particular agent.

Goodman & Frank, 2016

Has been extended to

- communication about closeness about relationships, in domain of food sharing (Hung et al., CogSci 2022)
- modeling people's goals when they punish (Radkani et al., CogSci 2022)

Project ideas

- Extend these models to other areas of social cognition and communication.
- Possibilities:
 - How do people communicate about the horizon of a relationship (i.e. how long they expect the relationship to last)?
 - How do people communicate about themselves in a relationship, and how does this interact with how long they expect the relationship
 - Different goals that people may have in social relationship (reputational goals, how much do you want to spend time with each other, etc)
 - Many things to do / brainstorm here
- Things not related to social relationships:
 - Modeling social affiliation and group affiliation
 -

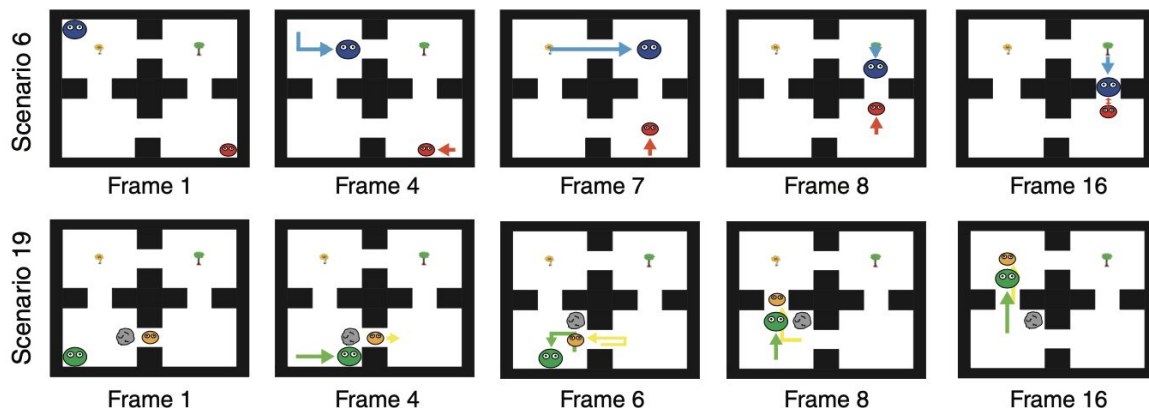
Grace

1. Social goal inference

- Two types of goals: **object-oriented** (flower, tree) and **social** (helping, hindering)
- Task: When paused at a certain “frame”, infer the **goal** of each agent (helping, hindering, flower, tree)

Project: Implementing the inverse planning model is sufficient with some slight to significant modification to better model more complex goals (undergrads: formalize this; grads: implement this)

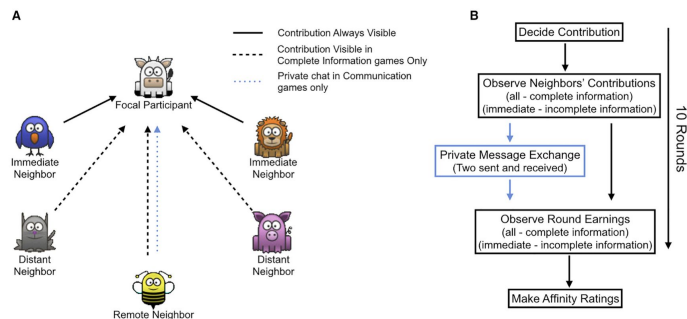
Example: incorporate more interactions (allow for small agents to also react to what they think the other agent's goal is)



Grace

2. Gossip/Information sharing

- **Idea 1:** Formalize an information-sharing game (Jolly & Chang: public-goods game where ability to **share information** and **see others' actions** were varied)
 - Discussion was dominated by talking about other participants when information was *incomplete*; information-sharing also encouraged participants to mirror actions of distant neighbors
- **Idea 2:** Restrict domain to something like fake news about x , and model how likely an untrustworthy source is to be shared based on things like domain knowledge, who it was shared by



Grace

3. Modeling influence of numerical news headlines

- Consider news topic like 'COVID-19 vaccination,' or 'World Series' that has numerical headlines (vaccination rate, exponential decay, game/player stats)
- Consider opinion one might have about the topic, e.g. severity of COVID-19 in a city, or prediction of who will win the World Series
- **Idea:** model how can representations of the number in the headlines influence your belief
 - “more than a million people in LA county have tested positive for Covid” has a different level of urgency if you incorrectly think LA county has **100 million people** instead of **10 million**, and that may influence your social distancing practices
- Instead of numbers, can also focus on some other linguistic aspect of headlines

Brainstorming a Project Idea

