- **Introduction**
  - Hi im Lukas you may not know me
  - Going to present my research and if people want to stick around we can try the task which I'm about to send to murk
  - If anyone is confused or has a question please feel free to interrupt
  - This is my first time running through this stuff so would really appreciate any and all feedback.
- **Heider and Simmel**
  - Heider and Simmel found that with a combination of moving shapes on a screen people often see a story unfold. They notice that one triangle is mad at the other or if two shapes are having fun.
  - In those simple shapes, people can easily find agents with emotions, intentions, an perceptions.
  - In doing this, we use a Theory of mind to infer the mental states of other agents.
- **Inverse Planning**
  - In the traditional conceptual model of planning, desires and beliefs create intentions which then influence actions.
  - In RL a reward function and World model (or state transition) create a policy which tells the agent which actions to choose
  - In the header and simmel video we are doing the reverse of this. We use the observed actions of an agent to infer the agent's beliefs and desires, or in RL terms Reward function and World Model (transition function)
- **this approach works**
  - There is definitely something to this idea
  - You can infer the reward function or goals — where the agent is going or what it wants
  - Inferring the beliefs in a partially observable MDP framework - to infer where the agent thinks some object may be
  - Can infer past behavior or trajectories from indirect evidence and low amounts of data
  - Can learn preferences and whether someone else is helping/hindering
  - An MDP model of an agent can clearly model real agents well.
  - **But are there limitations?**
- **Some agents plan differently than other 1**
  - limitations - you treat all agents in the same way, but intuitively out there in the world that is not the case… depending on the creature you are watching you infer different things
  - When using inference against a turtle you can infer where they're going but also that they can't remember many things or see above them
  - With babies they not have object permanence or pass theory of mind false belief tasks
  - You know they have beliefs and desires but they compose them differently
- **Some agents plan differently than other 2**
  - There is some evidence that if you categorize peoples minds therses different dimensions
  - People think of minds as different
  - Participants compared the mental capacities of various human and nonhuman characters via online surveys.
  - PCA Factor analysis revealed two dimensions of mind perception, Experience (for example, capacity for hunger) and Agency (for example, capacity for self-control).
  - A robot has low experience in real life but a decent amount of agency
  - Adults have high experience and agency
- **Some agents plan differently than other 3**
  - We clearly categorize different minds as having different capacities
  - But how do we do this? Where does the last figure come from?
  - There are two possibilities for this
    - Takes a while to learn what an agent can or can't do - maybe you need to spend a week or year with the new agent

    - Or it happens very quickly and naturally (just like mental state inference/inverse planning)
- **Some agents plan differently than other 4**
  - With an example like this we can see the necessity for Mental inference in conjunction with inverse planning
  - If we only did inverse planning in this example we would assume the turtle wants to fall down the hole
  - But instead we assume something like the turtle cannot perceive the floor in front of it
  - Or that the turtle cannot plan at all
  - These are distinctions which cannot be made by finding a reward function and need a new sort of inference
- **Transition slide**
  - To do this sort of inference we look at the input channels to the planner and the sort of state or representations that the agent can use in its world model.
- **Questions slide**
  - This presents two big questions
    - What is the space in which minds can vary?
      - We've seen experience and agency as possible dimensions but surely there are more
    - And how do we make this sort of inference? With what sort of model? With how much data?
  - The first question is rather complex — how do we quantitatively define how a crows mind differs from a baboon
  - So we will focus on the second question for this and assume a subset of possible minds.
- **Conceptual Model Intro**
  - Here is our broad conceptual model of the mind.
  - Desires and beliefs feed into actions and are influenced by perceptions from actions
- **Desires**
  - We assume all agents have desires because if they don't have desires they don't take any actions and if they don't take actions you can't infer their minds
  - For our subset of minds we assume all agents have the same desire
- **Beliefs**
  - Beliefs can vary widely from creature to creature
  - Can they have a memory at all
    - How big is that memory
  - Can they imagine hypothetical experiences
- **Perceptions and Actions**
  - Some minds are optimized to fly or swim
  - Some can hear extremely well or see in the dark
  - Some can smell things at far distances and make difficult classifications (drug dogs)
- **Changing the model**
  - So we will bring this variety into the model
- Starting with **perception**, the agents in our model can all see
  - But some of them cannot hear
  - In the grid world environment hearing means detecting movement behind their field of view
- The next thing we change is **beliefs**
- **No beliefs**
  - Imagine a fly buzzing around the room with no real model or memory, just moving toward its immediate goals
  - We will call this no object permanence
- **Variation within beliefs**
  - We decided to divide beliefs up into a variety of parameters

    - Capacity
      - How many places or objects can the agent remember
      - This could involve waypoints a long a path
      - Or enemies the agent is trying to avoid
    - Decay
      - Once the agent has remembered something
      - Is this memory permanent or does it fade overtime until they no longer hold a representation of the object/place
    - Some agent **can hold beliefs about objects as agents**
      - This involves conceptualizing the other agent as some sort of creature with a planner going towards a goal
- **Task stimulus**
  - Introduce the problem
    - Blue circle is the thief
    - Red triangle the guard
    - Gold star the treasure
  - First task - here you can infer this agent cannot hear and not much else
  - Second
    - Agent has high memory capacity (patrolling)
  - Third
    - Agent seems to be able to represent the thief as an agent heading toward the treasure
    - This can be seen in how it intercepts
  - Fourth
    - Here you can see obj perm and memory decay
- **Inference approach**
  - To infer the minds we use a generative model to create new minds along 5 different capacities
  - We measure the likelihood of the mental properties by seeing how well the generated minds' behavior match the observed behavior
  - Currently because we are only using a subset of minds we can get the full distribution using an exhaustive approach
  - Sometimes for things like searching which uses probability things get tricky
    - generative model may not be best for this.
- **Closing notes**
  - It feels intuitively easy to make these mental inferences
  - We can use a generative model to make these inferences and it requires little data (also intuitively easy)
  - One drawback is that the model usually makes very confident inferences
    - Hard to come up with cases in which graded inferences arise because the tasks are created using an optimal guard planner
    - But cool to see how quickly they can make confident inferences
    - This could say something about the lack of ambiguity in mental inference