(LLM)
Agents
(with Reinforcement Learning)

Session 1, February 17th 2025, Polina Tsvilodub
Slides inspired by this course and this tutorial

Logistics: Schedule

Day 1 (Feb 17th): Introduction, concept of agents

Day 2 (Feb 19th): Cognitive architectures

Day 3 (Feb 18th): Reinforcement learning

Day 4 (Feb 20th): LLM agents & neuro-symbolic systems

- session 4.3: **14:45—16:15** + QA time

Day 5 (Feb 21st): Exam

Days 1—4:

1. 09:00 — 10:30: intro / lecture for topic of the day
2. 11:00 — 12:30: paper discussion session
3. 14:00 — 15:30: discussions, exercises and / or lecture

Day 5:

Exam 09:00 — (12:30)

Logistics: Grading

	3 CP	6 CP
In-person paper expert	10%	5%
Submission of a question for at least two other papers	10%	5%
In-person exam (Feb 21st)	80%	40%
Final group project	✗	50%

Logistics: Grading

Exam:

- one hand written cheat sheet allowed
- based on lectures, papers, exercises — applying discussed concepts

Projects:

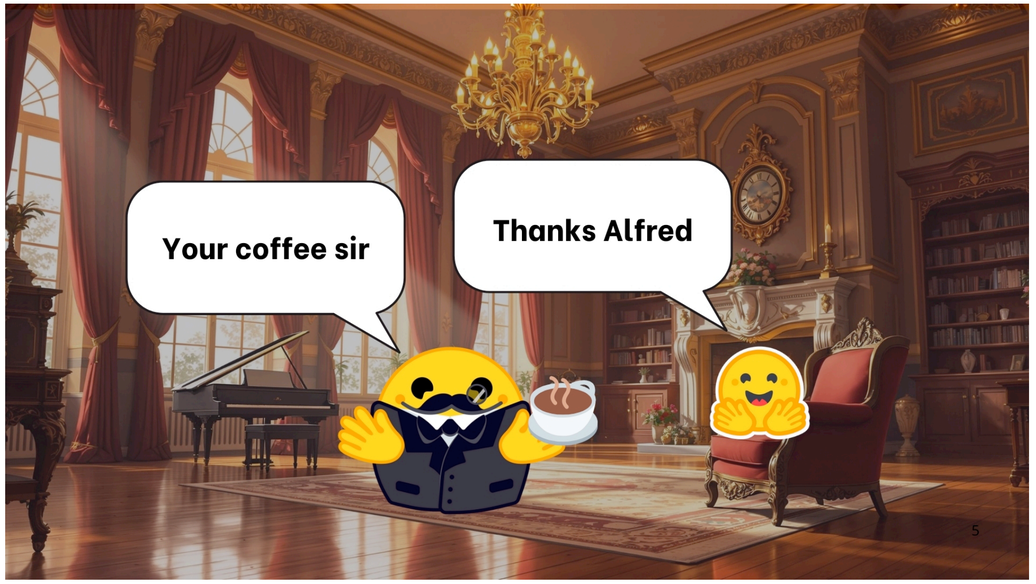
- groups of 3–5
- submission of a conference-paper style report + recorded presentation by March 31st on Moodle (one per group)
  - details & guidelines tba
- project topics will be available in a document on Moodle, tba (“final projects” section)

Introduction: What are agents?


Alfred, I would like a coffee please
2


Alfred's Toolbox
Reason and plan
3

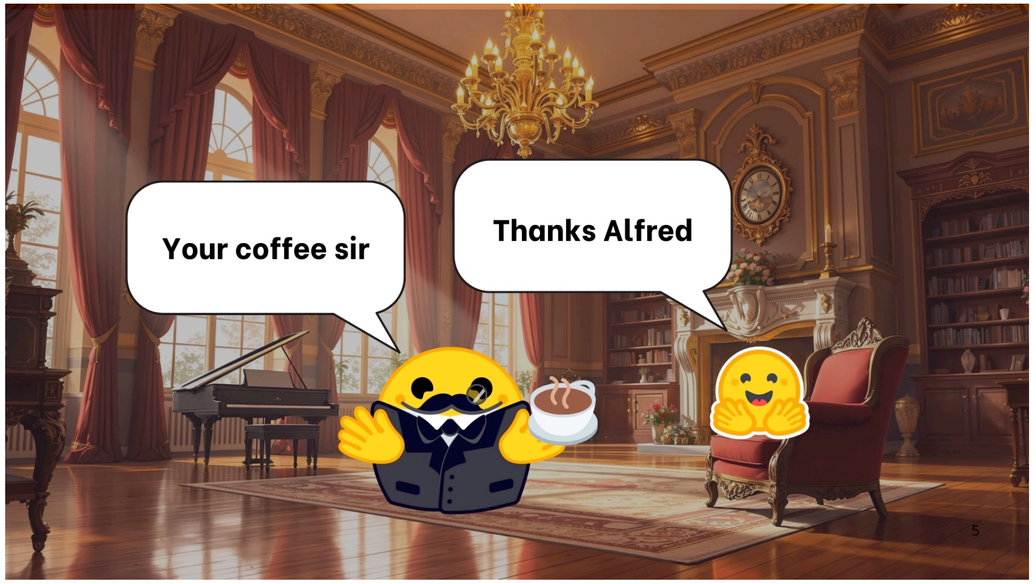

Execute
4


Your coffee sir
Thanks Alfred
5

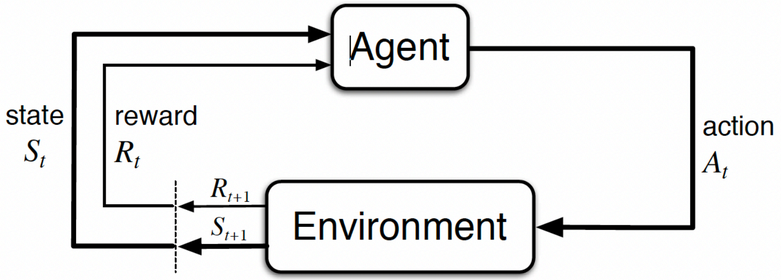

Alfred, I would like a coffee please
2

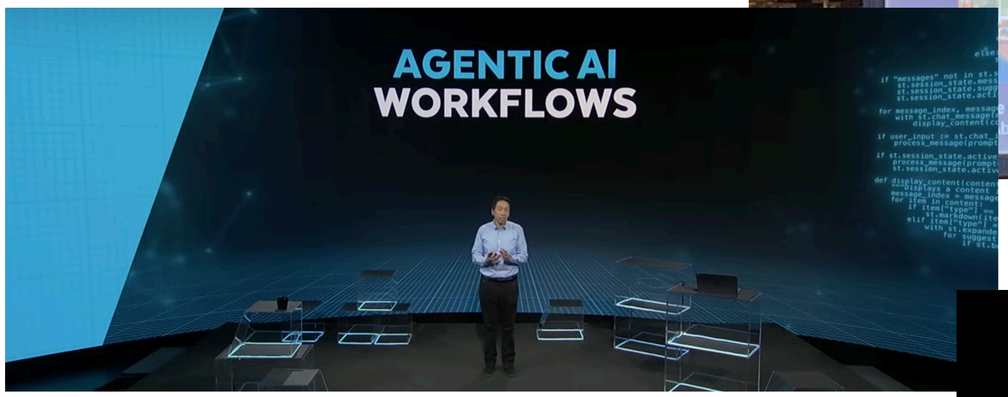

Alfred's Toolbox
Reason and plan
3

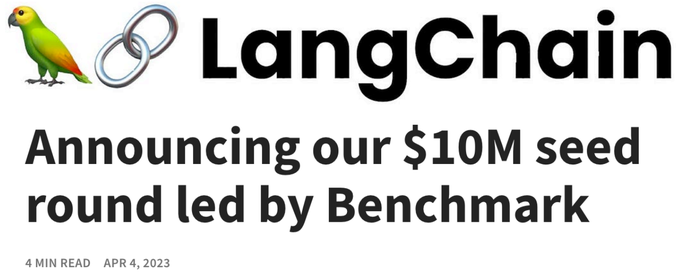

Execute
4


Your coffee sir
Thanks Alfred
5

Course on agents: Why now?

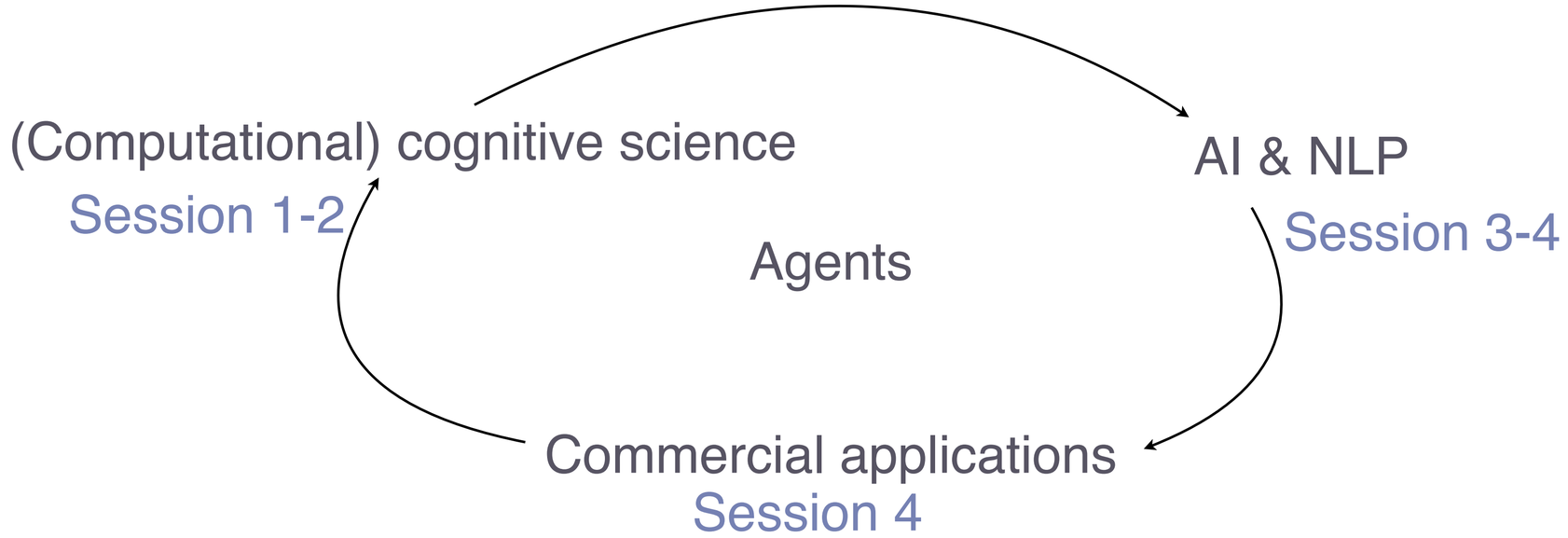






Course on agents: What I'd like you to take from this course

- work towards **understanding of the current AI developments & LLM agents** by **learning about different perspectives on agents**
- develop **system-level, task-decomposition thinking**
  - think: understanding which unit-tests to create for agents
  - MoE models; trend towards specialised, smaller models
- **think critically** and develop your own understanding, critical opinion of current technology
- course structure: mini-deep-dives every day
  - one overarching research question per day that you are invited to think about it



What are agents?

Brainstorming

Mentimeter code:
8274 2641

Wooldridge & Jennings (1995), Ng (2024)

What are agents?

Definitions

- **AI & Computational modeling**:
  - Russell & Norvig: any entity that perceives its environment through sensors and acts upon it through actuators
  - Wooldridge & Jennings: “entities that exhibit (aspects of) intelligent behavior [with the properties of autonomy, social ability, reactivity, pro-activeness, rationality, representational capacity for knowledge, beliefs, intentions, obligations]”
  - Cybernetics: a system that regulates itself through feedback loops from inputs to adjusting outputs
  - Game theory: rational decision-makers who interact strategically with other agents
- **Reinforcement learning**: decision-making learner that interacts with (an uncertain) environment to learn behavior that optimizes goal achievement (i.e., maximizes a reward)
- **Cognitive science & philosophy**: “intentional stance”: an entity that acts intentionally, and can be attributed beliefs, desires, goals
- **Related**:
  - social entities that understand other agents, e.g., via Bayesian reasoning
  - embodied entities that interact with environment

Wooldridge & Jennings (1995), Sutton & Barto (2018), Russell & Norvig (2020)

What are agents?

Definitions

- **AI & Computational modeling**:
  - Russell & Norvig: any entity that perceives its environment through sensors and acts upon it through actuators
  - Wooldridge & Jennings: “entities that exhibit (aspects of) intelligent behavior [with the properties of autonomy, social ability, reactivity, pro-activeness, rationality, representational capacity for knowledge, beliefs, intentions, obligations]”
  - Cybernetics: a system that regulates itself through feedback loops from inputs to adjusting outputs
  - Game theory: rational decision-makers who interact strategically with other agents
- **Reinforcement learning**: decision-making learner that interacts with (an uncertain) environment to learn behavior that optimizes goal achievement (i.e., maximizes a reward)
- **Cognitive science & philosophy**: “intentional stance”: an entity that acts intentionally, and can be attributed beliefs & desires
- **Related**:
  - social entities that understand other agents, e.g., via Bayesian reasoning
  - embodied entities that interact with environment

Wooldridge & Jennings (1995), Sutton & Barto (2018), Russell & Norvig (2020)

What are agents?
Definitions

- AI & Computational modeling:
  - Russell & Norvig: any entity that perceives its environment through sensors and acts upon it through actuators
  - Wooldridge & Jennings: “entities that exhibit (aspects of) intelligent behavior [with the properties of autonomy, social ability, reactivity, pro-activeness, rationality, representational capacity for knowledge, beliefs, intentions, obligations]”
  - Cybernetics: a system that regulates itself through feedback loops from inputs to adjusting outputs
  - Game theory: rational decision-makers who interact strategically with other agents
- **Reinforcement learning**: decision-making learner that interacts with (an uncertain) environment to learn behavior that optimizes goal achievement (i.e., maximizes a reward)
- **Cognitive science & philosophy**: “intentional stance”: an entity that acts intentionally, and can be attributed beliefs & desires
- Related:
  - social entities that understand other agents, e.g., via Bayesian reasoning
  - embodied entities that interact with environment

Wooldridge & Jennings (1995), Sutton & Barto (2018), Russell & Norvig (2020)

What are properties of agents?

Definitions

- “entities that exhibit (aspects of) goal-directed intelligent behavior with the properties of autonomy, social ability, reactivity, pro-activeness, rationality”

- **Cognitive science & philosophy**: “intentional stance”: an entity that acts intentionally, and can be attributed beliefs & desires

Dennett (1987), Wooldridge & Jennings (1995), Sutton & Barto (2018), Russell & Norvig (2020)

What are properties of agents?

Definitions

- "entities that exhibit (aspects of) goal-directed intelligent behavior with the properties of autonomy, social ability, reactivity, pro-activeness, rationality"

  - autonomy: come up actions on its own based on relevant knowledge
  - social ability: interact with other agents
  - reactivity: perception of / effect on environment
  - pro-activeness: goal & task representation
  - rationality: anticipation of certain subtasks
  - behavior: plan of actions based on reason
  - behavior: execute actions

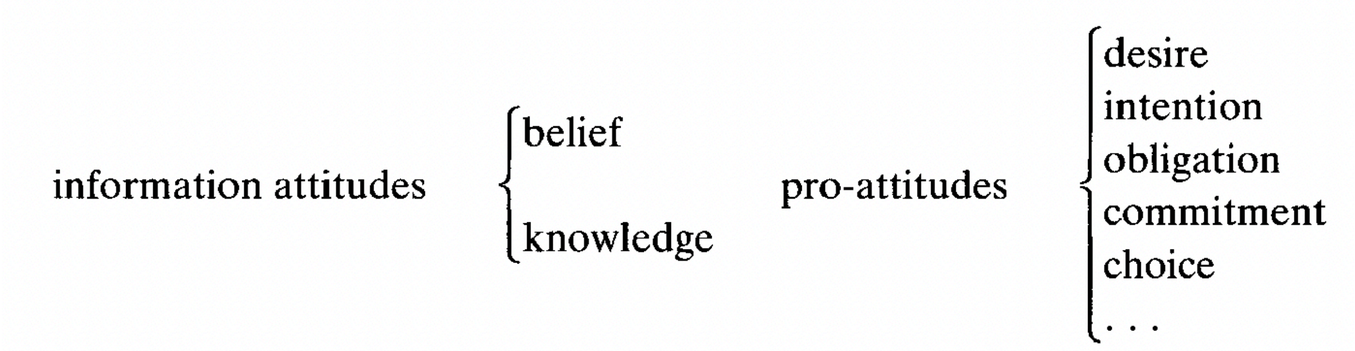
- **Cognitive science & philosophy**: "intentional stance": an entity that acts intentionally, and can be attributed beliefs & desires

  - representation of state of the world
  - ideal target state of the world

Wooldridge & Jennings (1995), Sutton & Barto (2018), Dennett (1987)

Agents: intentional stance

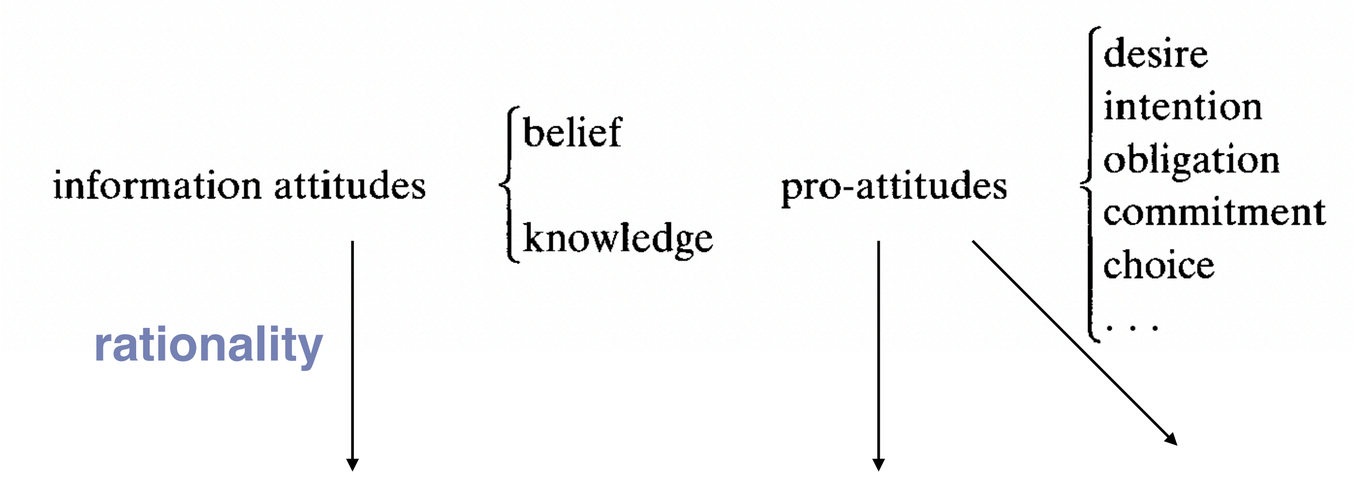
Cognitive science & philosophy: “intentional stance”: an entity that acts intentionally, and can be attributed beliefs & desires



Wooldridge & Jennings (1995)

Agents: intentional stance

Cognitive science & philosophy: “intentional stance”: an entity that acts intentionally, and can be attributed beliefs & desires



Wooldridge & Jennings (1995)

Belief-desire-intention model

Folk psychology based agent model for problem solving

Components:

- **Beliefs**: the beliefs about / model of the world ← how to update?
- **Desires**: desired end state of the world, the objective to accomplish → which ones to achieve first?
- **Intention**: the course of actions currently under execution to achieve the desire of the agent; consistent with beliefs ← how to select actions?

1. Intentions pose problems for agents, who need to determine ways of achieving them.
2. Intentions provide a "filter" for adopting other intentions, which must not conflict.
3. Agents track the success of their intentions, and are inclined to try again if their attempts fail.
4. Agents believe their intentions are possible.
5. Agents do not believe they will not bring about their intentions.
6. Under certain circumstances, agents believe they will bring about their intentions.
7. Agents need not intend all the expected side effects of their intentions.

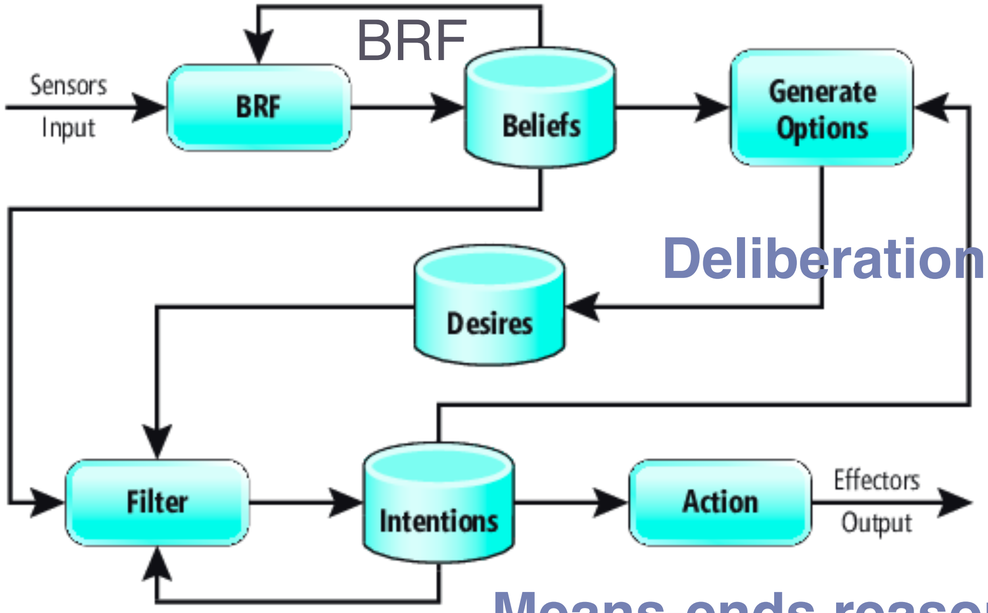
Bratman (1987), Georgeff et al (1999), blogpost

Belief-desire-intention model

Folk psychology based agent model for problem solving

Components:

- **Beliefs**: the beliefs about / model of the world ← how to update?
- **Desires**: desired end state of the world, the objective to accomplish → which ones to achieve first?
- **Intention**: the course of actions currently under execution to achieve the desire of the agent; consistent with beliefs ← how to select actions?



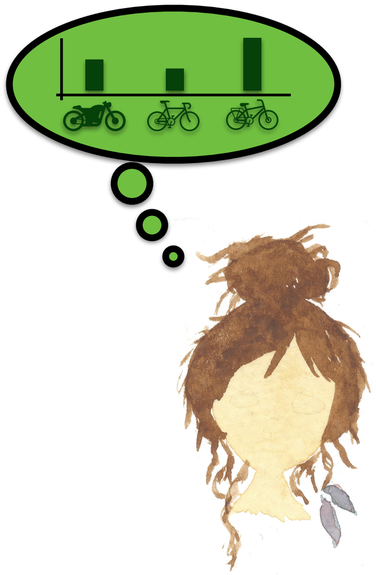
Bratman (1987), Georgeff et al (1999), blogpost

Belief-desire-intention model

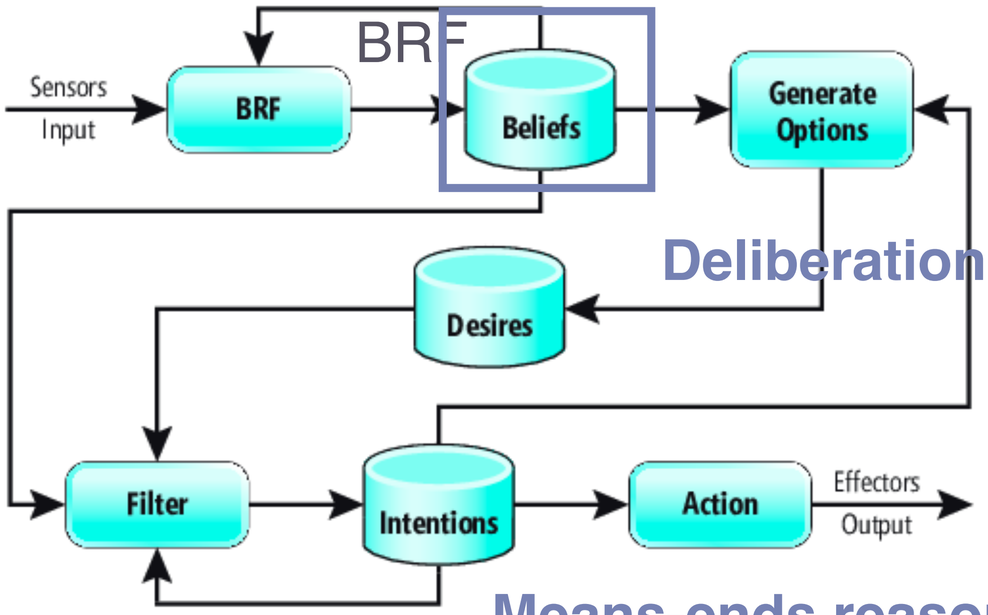
How are beliefs represented?

Components:

- **Beliefs**: the beliefs about / model of the world



First-order beliefs



Means-ends reasoning

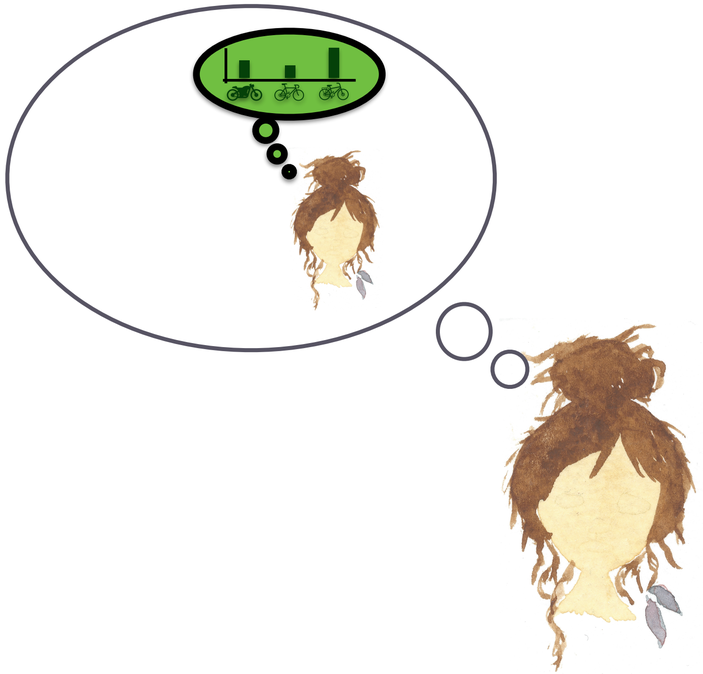
Bratman (1987), Georgeff et al (1999), Wooldridge & Jennings (1995)

Belief-desire-intention model

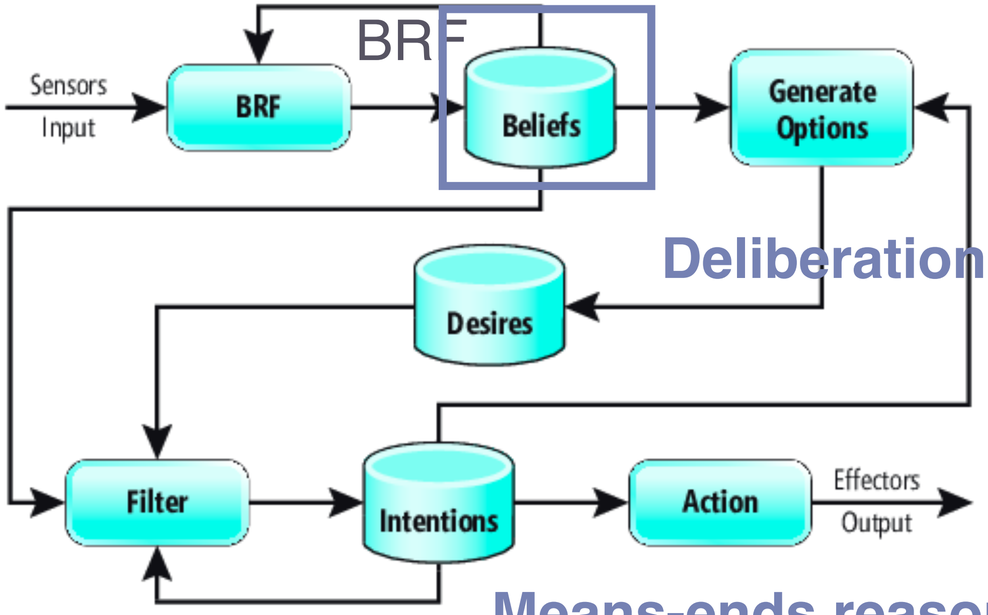
How are beliefs represented?

Components:

- **Beliefs**: the beliefs about / model of the world



Second-order beliefs



Means-ends reasoning

Bratman (1987), Georgeff et al (1999), Wooldridge & Jennings (1995)

Belief-desire-intention model

How are beliefs represented?

Components:

- **Beliefs**: the beliefs about / model of the world
  - fact representation that facilitates *valid automatic reasoning*:
    - e.g., propositions & logic
    - (semantic) networks, hierarchies, scripts
- Common challenges:
  - distinction between knowledge and beliefs (i.e., **uncertainty**)
  - updates based on information from environment
    - conversion between sensory and symbolic representations
    - **frame problem** / relevance
  - granularity
  - flexibility
  - common sense
  - reasoning efficiency
    - **closed world assumption**

Wooldridge & Jennings (1995), McCarthy & Hayes (1969), Schank & Abelson (1977)

From beliefs to actions

How does reasoning work?

Components:

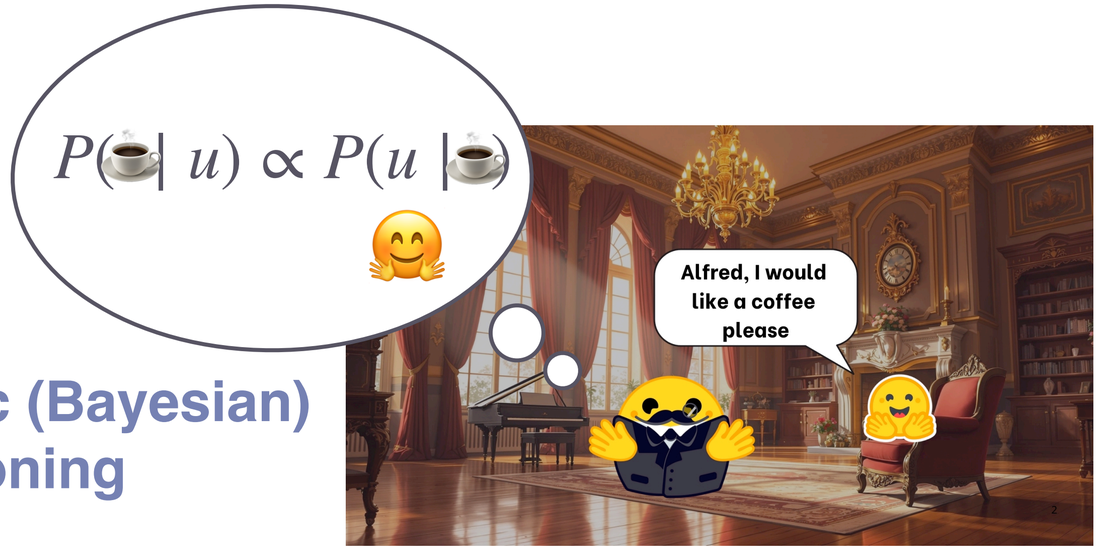
- Intention: the course of actions currently under execution to achieve the desire of the agent; derived via **deliberation / reasoning**
  - Wooldridge & Jennings: modal, intention logic based reasoning
    - side effect problem
  - Newell & Simon: reasoning as search through problem / action space (-> tomorrow!)
- Cognitive perspective:
  - reasoning: drawing conclusions to solve problems and make decisions
    - does not (always) follow classical logic! (e.g., sensitivity to language)
  - many aspects of reasoning are uncertain, i.e., **probabilistic**
    - special role of **social reasoning: theory of mind (ToM)**

Wooldridge & Jennings (1995), Newell & Simon (1956), Tenenbaum et al (2024)

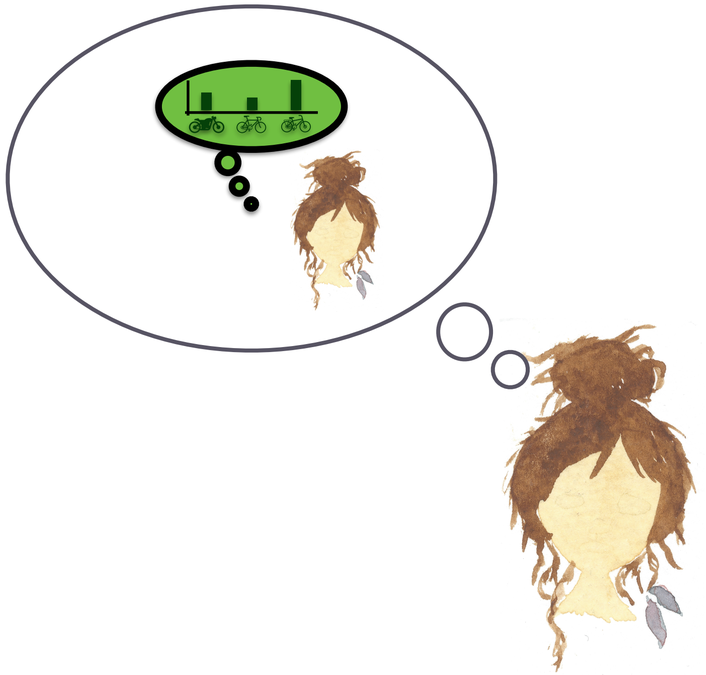
From beliefs to actions

How does reasoning work?

- Cognitive perspective:
  - reasoning: drawing conclusions to solve problems and make decisions
    - does not (always) follow classical logic! (e.g., sensitivity to language)
  - many aspects of reasoning are uncertain, i.e., **probabilistic**
    - special role of **social reasoning: theory of mind (ToM)**



Probabilistic (Bayesian) reasoning



Second-order beliefs

From beliefs to actions

How does reasoning work?

Components:

- Intention: the course of actions currently under execution to achieve the desire of the agent; derived via **deliberation / reasoning**
  - Wooldridge & Jennings: modal, intention logic based reasoning
    - side effect problem
  - Simon: reasoning as search through problem / action space (-> tomorrow!)
- Cognitive perspective:
  - reasoning: drawing conclusions to solve problems and make decisions
    - does not (always) follow classical logic! (e.g., sensitivity to language)
  - many aspects of reasoning are uncertain, i.e., **probabilistic**
    - special role of **social reasoning: theory of mind (ToM)**
  - Evans, Kahneman & Tversky: decision makers are **boundedly rational** and **have different reasoning processes: dual process theory**

Stanovich; Evans; Kahneman & Tversky

Dual process theory

How does reasoning work?

System 1 (Intuition)	**System 2** (Reasoning)
Fast	Slow
Parallel	Sequential
Automatic / default	Controlled
Effortless	Effortful
Associative	Rule-governed
Slow-learning	Flexible
Contextualised (subject to priors, context effects)	Abstract
can be evoked by language operates on conceptual represnetations	

- although the overall distinction is widely supported, there are debates around:
  - is the distinction at the level of modes of thinking, or cognitive mechanisms?
  - mechanisms involved in S1 are also involved in S2
  - degree of modularity of the mind (Sperber, Fodor)

Intermediate summary

What are agents?

- agents combine many complex properties like the ability to achieve goals, autonomously solve novel problems, interact with an environment based on available information, interact with other agents
- goal-directed behavior can be decomposed into different sub-components
- engineering approaches have strived to develop agents for different tasks, but flexible general systems have remained elusive
- cognitive science has provided insights into how humans navigate the complexities of real world

Outstanding questions:

- How can we combine insights about the human mind with engineering demands to build flexible and efficient artificial systems?
- How are intentions and goals formed in open-ended agents?
- Do these systems need to be *fully agentic*?

Disclaimer: we are leaving emotions, affect, consciousness and other aspects out of the picture

What are agents?

Definitions

- AI & Computational modeling:
  - Russell & Norvig: any entity that perceives its environment through sensors and acts upon it through actuators
  - Wooldridge & Jennings: “entities that exhibit (aspects of) intelligent behavior [with the properties of autonomy, social ability, reactivity, pro-activeness, rationality, representational capacity for knowledge, beliefs, intentions, obligations]”
  - Cybernetics: a system that regulates itself through feedback loops from inputs to adjusting outputs
  - Game theory: rational decision-makers who interact strategically with other agents
- **Reinforcement learning**: decision-making learner that interacts with (an uncertain) environment to learn behavior that optimizes **goal** achievement (i.e., maximizes a reward)
- **Cognitive science & philosophy**: “intentional stance”: an entity that acts intentionally, and can be attributed beliefs & **desires**
- Related:
  - social entities that understand other agents, e.g., via Bayesian reasoning
  - embodied entities that interact with environment

Wooldridge & Jennings (1995), Sutton & Barto (2018), Russell & Norvig (2020)

Paper discussion logistics

Chu, Tenenbaum & Schulz (2024). In praise of folly: human flexible goals and human cognition

- How / What intentions and goals are formed in open-ended agents?
- Do these systems need to be *fully agentic*?
- How can we combine insights about the human mind with engineering demands to build flexible and efficient artificial systems?

1. split in two groups, half of the experts in each group
2. discuss the paper (e.g., start with questions of non-expert participants)
3. experts: responsible for adding key points, insights, new questions of the group to shared Google slides: https://docs.google.com/presentation/d/1BH53A2ipfzrix9C0gR39Cd1uUIHZaIZ2_CseZmhe81U/edit?usp=sharing
   a. maximally 3 slides!
   b. make slides such that they will be helpful for exam!
4. joint discussion

Can vanilla LLMs be seen as approximating agents?

LLMs: Rudimentary models of agents / beliefs / actions?

AI agents in the age of LLMs

Language can be used as a vehicle for knowledge representation, reasoning, communication

> ***Pat watches a demonstration of a bowling ball and a leaf being dropped at the same time in a vacuum chamber. Pat, who is a physicist, predicts that*** *the bowling ball and the leaf will fall at the same rate.*

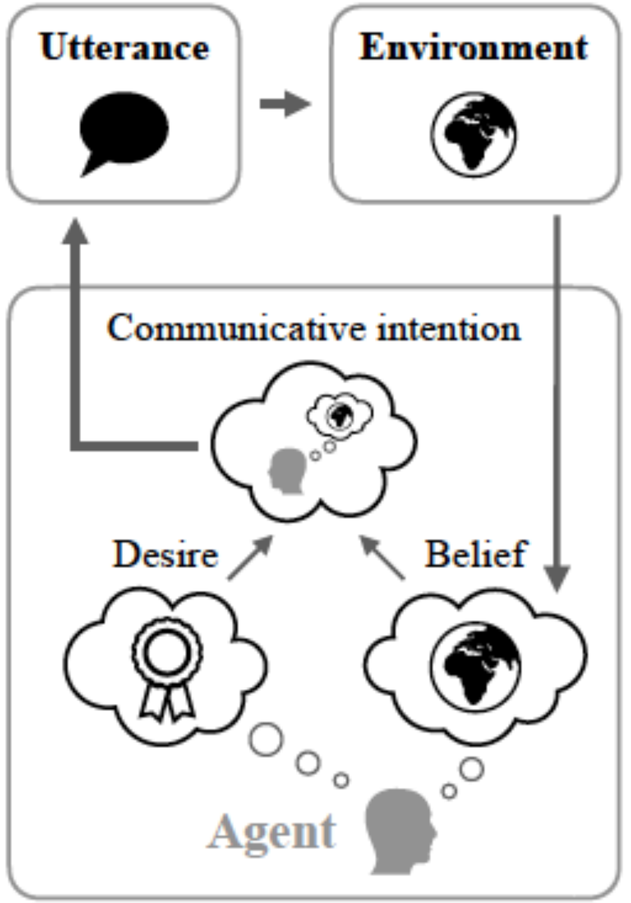
Changing the last sentence of the prompt, we get:

> ***. . . Pat, who has never seen this demonstration before, predicts that*** *the bowling ball will fall to the ground first. This is incorrect. In a vacuum chamber, there is no air*

Language Models as Agent Models [Andreas, 2022]

From text to communicative intentions

- (vanilla) LLMs are trained only on text, without access to the mental states (beliefs, desires, intentions) of the agents that generated that text
  - an LLMs nonetheless learn to approximate intentions of the authors?
- communicative intention (Speech Act theory from linguistics):
  - S intends the audience to produce a certain response to their message and recognise S's intention of doing so
  - literal meaning + results of the message, different types (e.f., assertions vs. directions …)
- maybe LLMs can learn underlying communicative intentions similarly to learning latent grammatical structure of language?



Andreas (2022)

From text to communicative intentions

(C1) In the course of performing next-word prediction in context, current LMs sometimes infer approximate, partial representations of the *beliefs, desires* and *intentions* possessed by the agent that produced the context, and other agents mentioned within it.

(C2) Once these representations are inferred, they are causally linked to LM prediction, and thus bear the same relation to generated text that an intentional agent's state bears to its communciative actions.

1. Agents with beliefs B and desires D are sampled from a population:

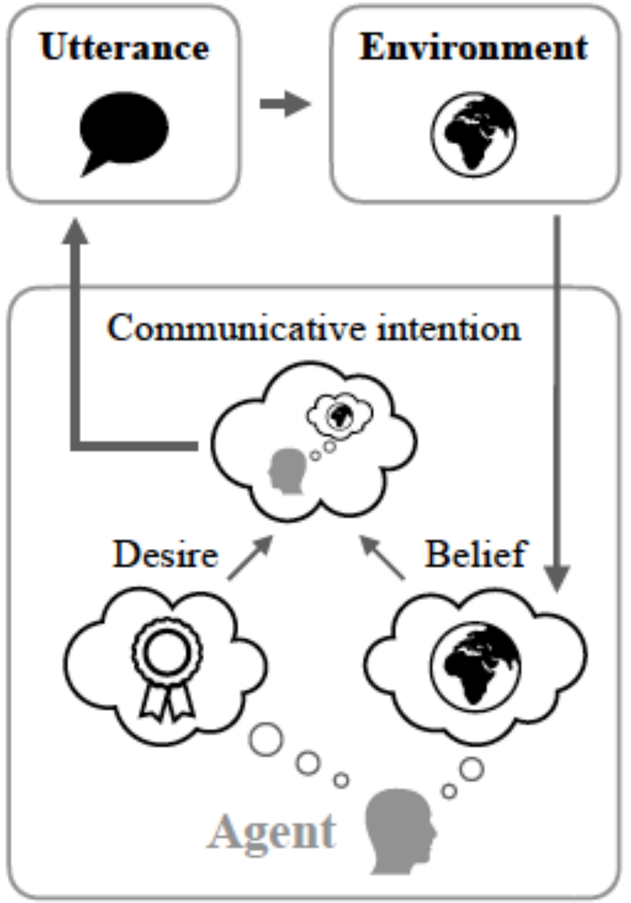
$$(B, D) \sim p_{\text{agent}}(\cdot, \cdot) \quad (2)$$

2. Each agent forms a communicative intention consistent with its current beliefs and desires:

$$I \sim p_{\text{intention}}(\cdot \mid B, D) \quad (3)$$

3. This communicative intention is realized as an utterance:

$$U \sim p_{\text{utterance}}(\cdot \mid I) \quad (4)$$



- the intention arises as the text is sampled by the LM itself!
  - what information is available in texts?
- controlled experiments with small models have found some evidence for these claims (e.g., a causal sentiment neuron for review generation)
  - how well can LLM architectures infer these latent variables in principle?
- let's try to test this!

Andreas (2022)