Durham University

COMP 3647
Human-AI Interaction Design

Topic 6:
Thought Cloning & Human Interestingness

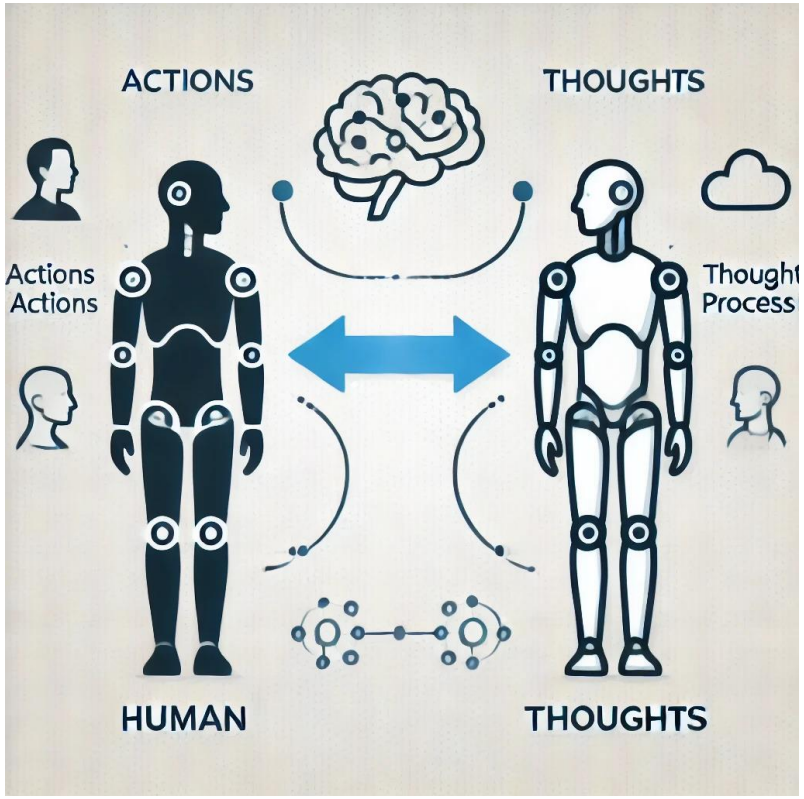
Prof. Effie L-C Law

Introduction to Thought Cloning

Hu, S., & Clune, J. (2024). Thought cloning: Learning to think while acting by imitating human thinking. *Advances in Neural Information Processing Systems (NeurIPS)*, 36. https://www.shengranhu.com/ThoughtCloning/

- **Definition**: Thought Cloning is an AI method where agents learn to imitate not just actions but the thinking processes behind those actions.

- **Goal:** Enhance AI by enabling it to "think while acting," closely mimicking human cognitive processes.



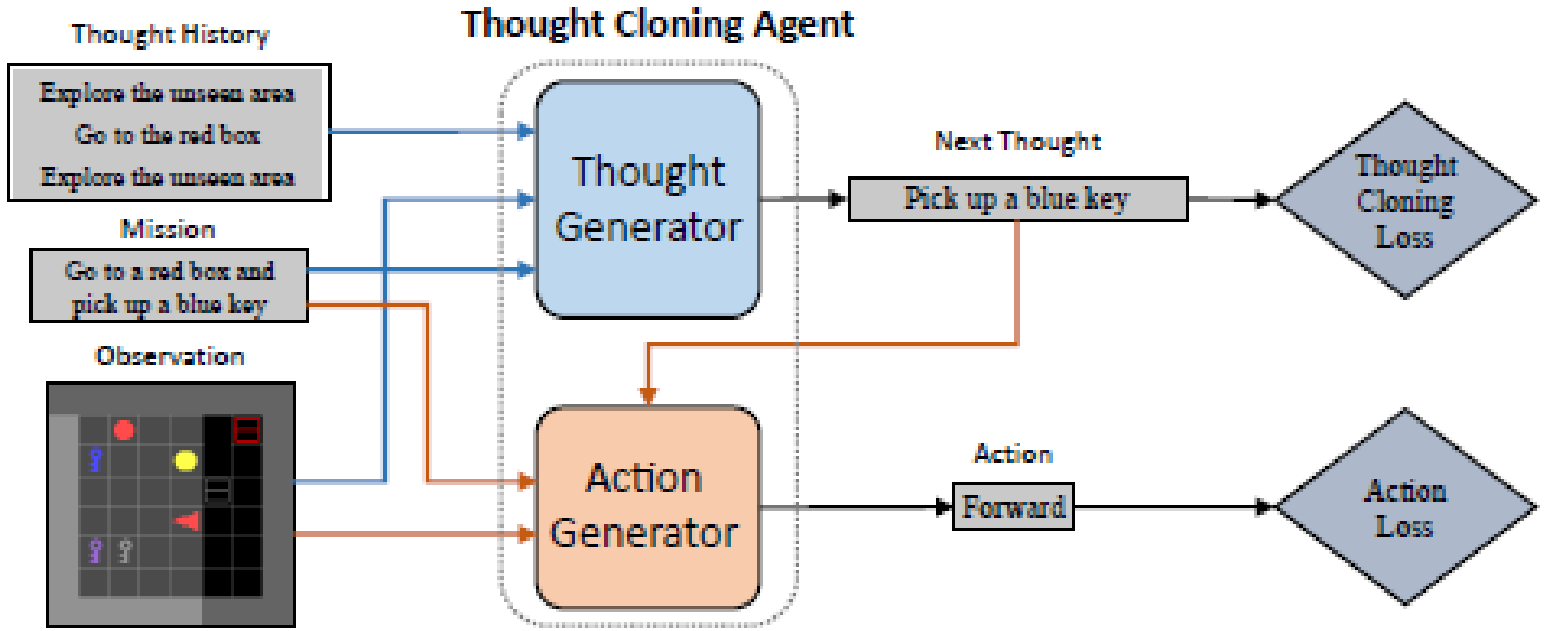
Durham University

The Need for Thought Cloning

- **Current Limitation**: Traditional imitation learning only captures actions, not the underlying thought processes.
- **Complexity in Real-World Tasks**: In many tasks, especially dynamic environments, simply mimicking actions is insufficient.
- **Solution**: Thought Cloning provides a way for AI to capture human-like thinking, reasoning, and decision-making.

Core Concepts of Thought Cloning

- **Thinking Traces**: Recordings of human cognitive processes (e.g., planning, evaluating options).
- **Action Traces**: Sequence of actions taken by a human in response to a situation.
- **Thought Cloning**: Combines both thinking and action traces to train AI.



Durham University

Methodology of Thought Cloning

- **Human Data Collection**: Gather both thinking and action traces by having humans perform tasks while recording their thoughts.

  Example: During gameplay, humans explain their thought process aloud.
- **Training Process**: AI is trained not just to mimic the action but to predict and emulate the thought processes behind actions.
- **Modelling Thought Traces**: Uses techniques such as transformer networks to model thought sequences.

AI Safety in Thought Cloning

- **Interpretability**: One benefit of Thought Cloning is the ability to interpret the AI's reasoning process, making it easier to understand and debug decisions.
- **Safety Concerns**: If AI can "think" in complex ways, there are risks of unpredictability or unintended actions.
- **Mitigation Strategies:**
  - ***Human-in-the-loop***: Ensuring that AI thinking and decision-making are monitored by humans.
  - ***Transparent Models***: Developing models where reasoning steps are clear and auditable.

Interpretability and AI Ethics

- **Challenges in Interpretability**: Understanding how AI "thinks" is crucial for trust and deployment in critical systems.
- **Ethical Considerations**:
  - ***Bias***: Imitating human thinking can reinforce biases if the training data is not carefully curated.
  - ***Decision Accountability***: If AI makes decisions by mimicking human thought processes, who is accountable for those decisions?
- **Solutions**: Explainable AI (XAI): Thought Cloning systems can contribute to the field of explainable AI, as they allow humans to inspect the reasoning process.

Applications of Thought Cloning

- **Robotics**: Robots that can think like humans and make more informed, adaptive decisions.
- **Healthcare**: AI can assist doctors by thinking through diagnostic processes, explaining their reasoning.
- **Gaming**: AI opponents can reason like human players, making them more challenging and unpredictable.

Potential Risks

- **Overfitting to Thought Traces**: The AI might become too tied to specific thought patterns, reducing generalization.
- **Complexity in Thought Traces**: Recording and understanding human thought processes can be much more complex than simply recording actions.
- **Safety Concerns**: If the AI learns faulty reasoning patterns, it could result in unexpected or harmful actions.

Future Research Areas

- **Refinement of Thought Models**: Further research into how to best capture and model human thoughts in a way that is both efficient and accurate.
- **Safety Mechanisms**: Developing robust safeguards for AI that can think autonomously.
- **Scaling**: Figuring out how to scale thought-cloning systems for larger, more complex tasks.

Summary

- Thought Cloning offers a promising way to enhance AI by allowing it to not just mimic actions but think like a human.
- While the potential is vast, issues like safety, interpretability, and ethical concerns remain.
- With ongoing research, Thought Cloning could revolutionize fields like robotics, healthcare, and more.

OMNI: Open-Endedness via Models of Human Notions of Interestingness

Jenny Zhang, Joel Lehman, Kenneth Stanley, Jeff Clune

2023

arXiv preprint arXiv:2306.01711.

What is Open-Ended Learning?

Definition:

- Open-ended algorithms aim to continuously learn *new, interesting* behaviours indefinitely.

Challenges:

- Infinite possible tasks in a vast search space.
- Need to prioritise not only learnable, but also interesting tasks.

Key Problem:

How can we define and quantify "interestingness" for AI to focus on worthwhile tasks?

Human Notions of Interestingness

What is Interestingness?

- "Interestingness" refers to human-like judgments about whether something is **novel, valuable, or worthwhile** to explore or learn.
- It's a **subjective** concept, shaped by personal experiences, curiosity, and expectations of value.

Why is Interestingness Important in AI?

- AI systems that continuously learn must choose from an infinite number of tasks.
- Without a sense of "interestingness," an AI may get stuck on repetitive, trivial, or unimportant tasks.
- Integrating human notions of interestingness enables AI to prioritise tasks that are not only learnable but also **meaningful**.

Challenges in Defining and Quantifying Interestingness

Key Challenge:

- Traditional AI methods focus on measurable, concrete goals like accuracy, novelty, or task success rates.
- Interestingness, however, is abstract and difficult to define or measure in strict quantitative terms.

Previous Approaches:

- Many research papers attempt to optimise novelty, diversity, or exploration in AI learning, but these approaches often fail because:
  - Goodhart's Law: "*When a measure becomes a target, it ceases to be a good measure*."
  - Optimising these metrics leads to trivial or pathological outcomes, such as superficial novelty that doesn't provide genuine value.

Examples of Pathologies:

- AI systems generating many trivial variations of tasks, such as moving objects in slightly different ways, while failing to tackle truly novel or complex challenges.

Human vs. Machine Notions of Interestingness

Human Intuition for Interestingness:

- Humans are adept at identifying tasks or problems that seem promising or interesting, even if the outcome is uncertain.
- This intuition is shaped by personal and societal experiences, history, and values.
- Humans tend to prioritise:
  - Novelty that leads to meaningful discoveries.
  - Tasks that expand knowledge or open up new possibilities.
  - Explorations that promise future rewards or insights (even if not immediate).

Machine's Limitation:

- Without an understanding of what is "interesting," AI tends to focus solely on metrics like learnability or success rates.
- AI lacks the rich, context-aware judgment humans apply to decide which tasks are worthwhile to pursue.

Foundation Models as a Model of Interestingness (MoI)

Foundation Models (FMs) Overview:

- Trained on vast amounts of human-generated data (e.g., texts, images), FMs like GPT-4 have internalised a wealth of human knowledge and can mimic human-like judgments.

Leveraging FMs for Interestingness:

- By prompting FMs to evaluate tasks, OMNI leverages their understanding of human concepts such as novelty, complexity, and usefulness.
- FMs are trained on data where humans naturally write about what they find interesting or boring, giving them a deep reservoir of implicit knowledge about what humans find worthwhile

How FMs Capture Interestingness?

Implicit Learning:

- Through exposure to massive text corpora, FMs have "learned" to understand what types of tasks, ideas, or stories captivate human attention.
- Examples of how FMs capture interestingness:
  - **Narrative Structure**: They understand when a story introduces tension, conflict, and resolution—key drivers of interest in storytelling.
  - **Curiosity-Driven Language**: FMs are familiar with how humans express curiosity or intrigue about new information, patterns, or phenomena.

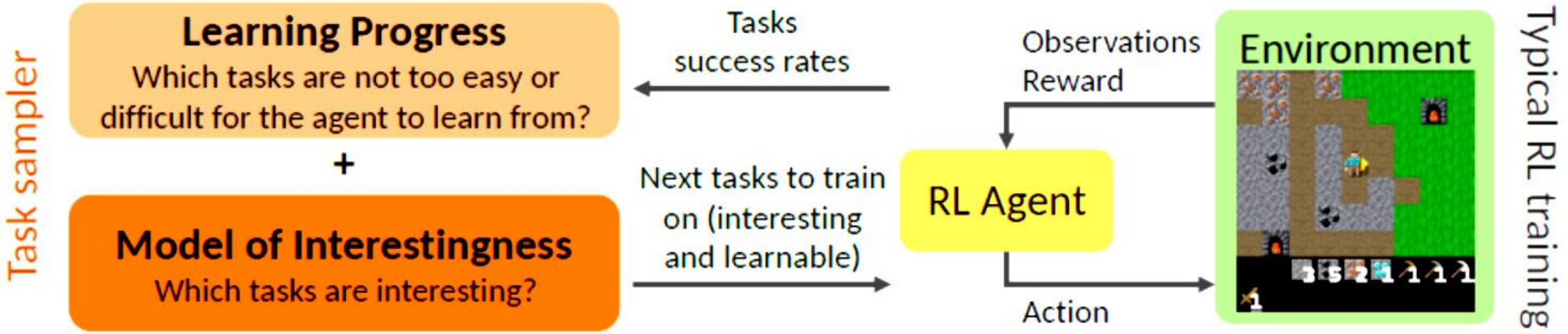
Prompts for Interestingness:

- OMNI uses FMs to select tasks by prompting them with questions like:
- "*What task would be most interesting to learn next*?"
- "*Which tasks would humans typically find valuable at this stage*?"

How OMNI Utilises Human Notions of Interestingness

Step-by-Step Process:

- OMNI prompts an FM (e.g., GPT-4) to assess tasks based on an agent's current progress.
- The FM evaluates the tasks in terms of novelty, complexity, and value—echoing human judgment.
- Interesting tasks are prioritized, while boring, repetitive, or trivial tasks are deprioritised.



Why Human-Like Interestingness Improves AI Learning

Avoiding Triviality:

AI systems often get stuck in loops, focusing on tasks that are easy but provide little value. By incorporating human-like interestingness, OMNI helps the AI avoid such loops.

Encouraging Exploration:

Human notions of interestingness push AI to explore tasks that promise new knowledge or future capabilities, even if they're more complex or difficult.

Balancing Learnability and Novelty:

FMs balance tasks that are at the right level of challenge for the AI with tasks that expand its skills, leading to more efficient and meaningful learning.

OMNI Experiment 1 - Crafter

Setup:

- A 2D Minecraft-like environment where agents complete tasks related to gathering and crafting.
- 15 interesting tasks diluted with 90 "boring" and 1023 "extremely challenging" tasks.

- **Objective:**

  Test whether OMNI can focus on interesting tasks while avoiding distractions. An FM might be prompted to decide between tasks like:

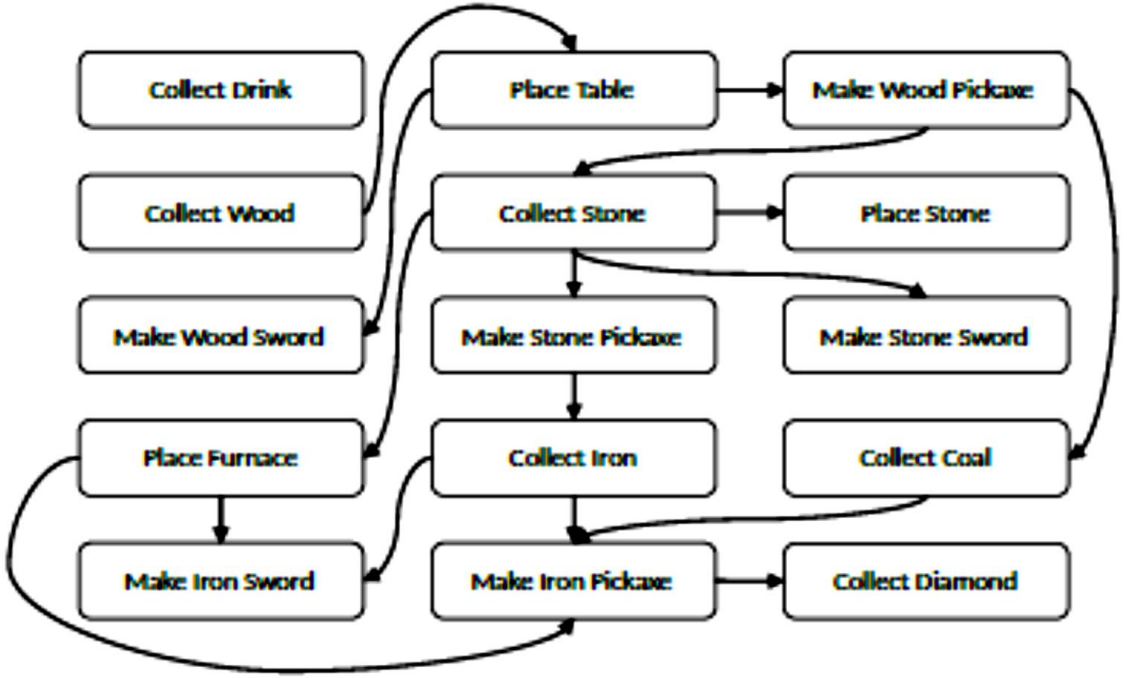
  - Collecting wood again (low interestingness).
  - Crafting a new tool (medium interestingness).
  - Discovering a completely new crafting recipe (high interestingness).

- OMNI uses this guidance to select the next task for the AI to tackle.

OMNI Experiment 1 – Crafter- Results

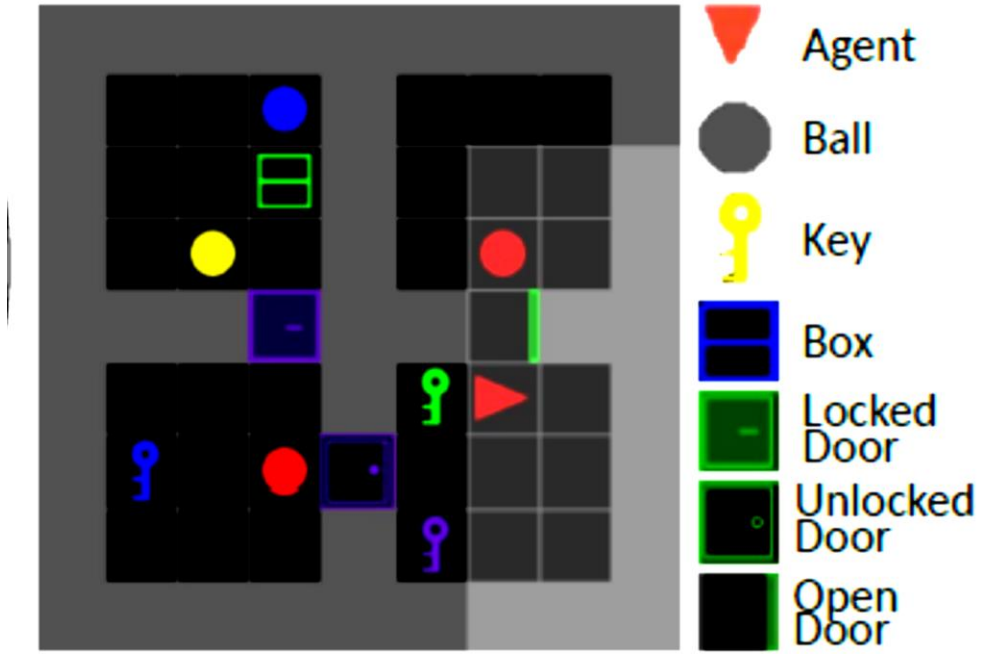
- Without OMNI: The AI might get stuck on slightly varied but ultimately trivial tasks, such as collecting different amounts of the same resource.

- With OMNI: The AI is guided toward tasks like crafting new tools, exploring new terrain, or discovering new resource combinations—more aligned with human ideas of novelty and value.



OMNI Experiment 2 – BabyAI Environment

- **Setup**: A 2D grid world for grounded language learning, with randomly generated room layouts and tasks.
- **Challenge**: 1364 unique tasks, varying in complexity and length.
- **Objective**: Evaluate OMNI's ability to handle a broad range of tasks, including sequential instructions.
- **Result:** OMNI helps the agent focus on complex, multi-step tasks (e.g., navigating obstacles and retrieving objects) rather than simple or repetitive actions.



OMNI Experiment 3 – AI2 THOR (Infinite Space)

Setup:

- A 3D, photo-realistic kitchen environment for embodied robotics tasks.
- Infinite task space generated by Foundation Models (FMs) with task definitions and reward functions.

Objective:

- Assess OMNI in an open-ended, infinite task space.

Result:

OMNI effectively generates and selects interesting tasks, outperforming traditional methods in an infinit task space.

Egocentric View

Bird's Eye View

OMNI's Impact on Open-Ended Learning

Solving Open-Endedness Challenge:

- OMNI's ability to evaluate interestingness solves a key problem in open-ended AI systems: identifying and prioritising tasks that drive continuous learning and growth.

Self-Improving AI:

- With a human-like sense of interestingness, AI systems are better equipped to autonomously select tasks that lead to significant skill advancement and discoveries, mimicking the human pursuit of knowledge and innovation.

Future of Interestingness in AI

Next Steps:

- Expand OMNI's Model of Interestingness with multi-modal models (e.g., vision-language models) to improve the AI's ability to judge interesting tasks in diverse environments.

Beyond Open-Ended Learning:

- Interestingness could be applied in other AI contexts, such as recommendation systems, human-AI collaboration, and creative AI systems.
- Human-like judgments about what's interesting could lead to more meaningful AI applications across industries.