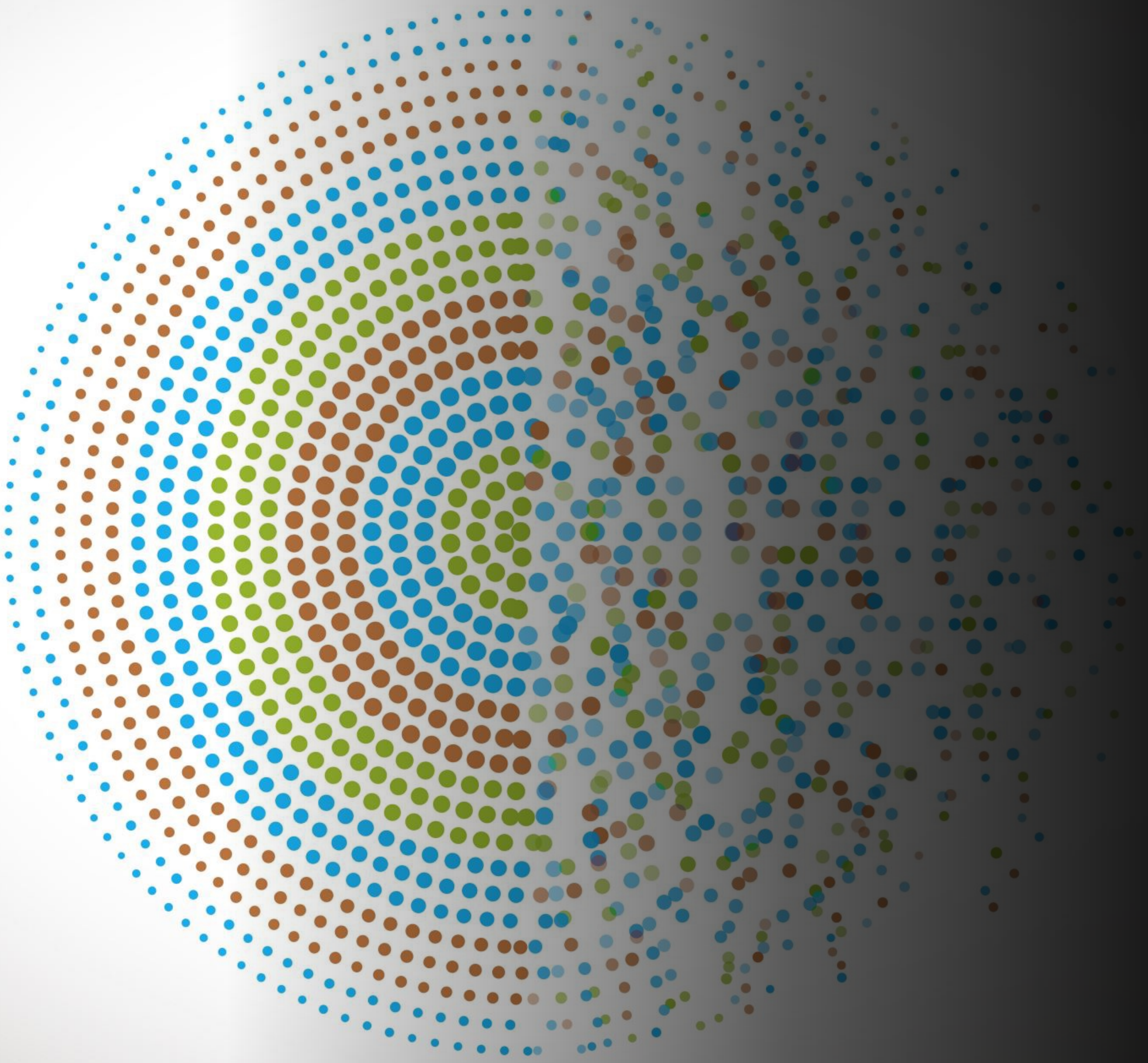


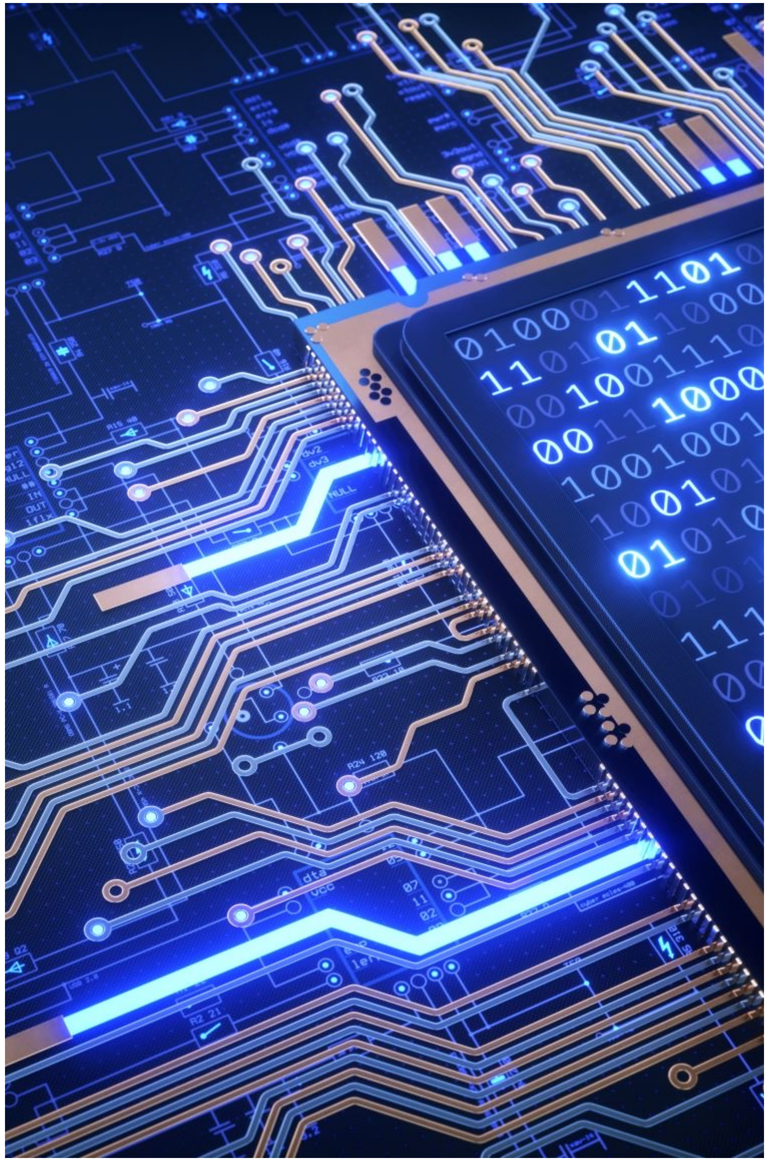
Mitigating False Appropriation in HCXAI

Carson Johnston

cjohn445@uwo.ca

Introduction

- PhD Candidate in Philosophy at Western University in Ontario, Canada.
- Graduate student member at the Rotman Institute of Philosophy.
- My doctoral research consists three interrelated projects.
- (1) Understanding how human-computer-interaction (HCI) and explanation aid the development of novel explainable AI (XAI) techniques
- (2) Use HCI as a learning opportunity for understanding human and machine decision structures
- (3) Questioning insights and consequences of sharing technical vocabularies between disciplines that study human and artificial minds, i.e., “conceptual borrowing.”



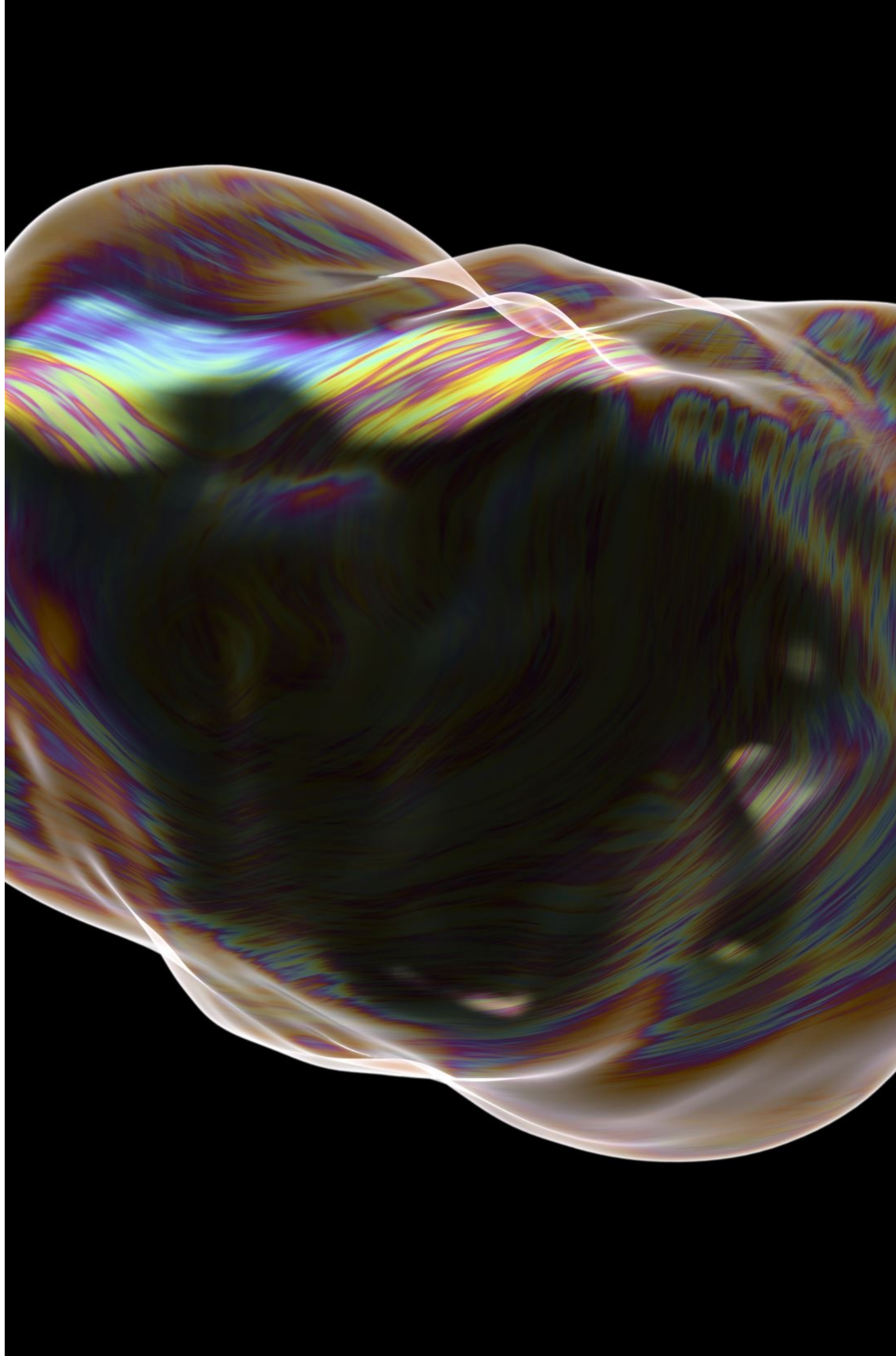
Outline

Today’s presentation concerns project 1: the development of novel XAI techniques to render opaque AI and ML models transparent for various stakeholder categories qua their unique epistemic positions using human-computer–interaction (HCI).

Arg → Cognitively forcing practical inferences is an irreplaceable component of human-centric explainable artificial intelligence (HCXAI).

XAI & HCXAI

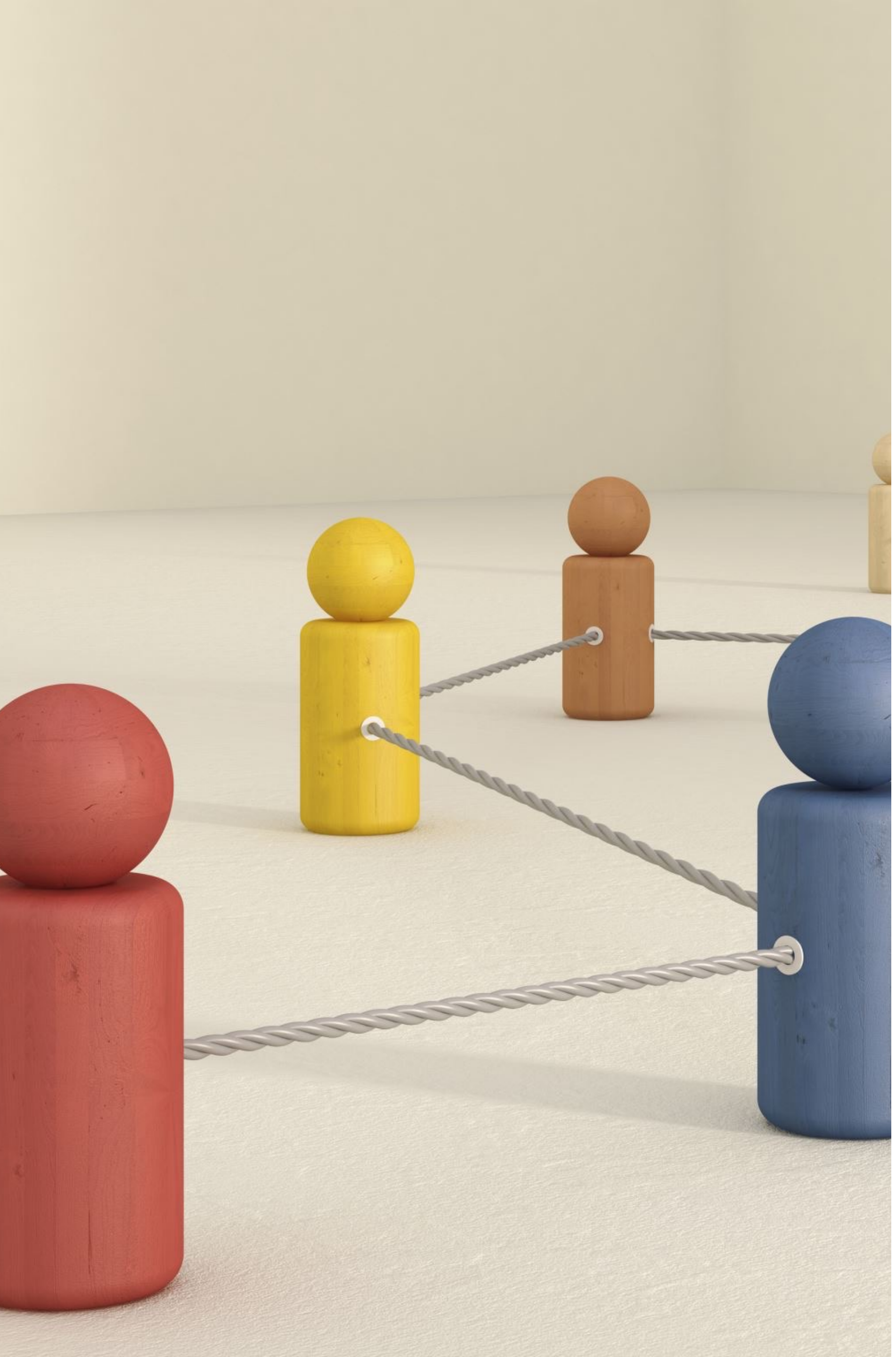
- eXplainable AI (XAI) → rendering opaque models transparent for domain experts *qua* their enlightened epistemic positions.
- Human-centric eXplainable AI (HCXAI) → rendering opaque models transparent for laypersons *qua* their uninformed and differential epistemic positions.
- Rhetoric aimed at circumvention not full transparency: *I therefore, O.*



Transparency

✓ Instrumentally valuable for domain experts.

✓ Intrinsically valuable for lay persons.



Sister Domain – Faithful CoT

- Faithful Chain-of-Thought (CoT) reasoning.
- Tackles the eXplainability of large-language-models (LLMs).
- LLMs have also been introduced in XAI and HCXAI to personalize XAI models to various stakeholder groups, e.g., by training LLMs in accordance with dominant personality or archetypes.

The Virtues of Explanation

- Interactive → conversational, iterative, cyclic, additive…
- Personalized → customizable, adaptable to different persons and model…
- Intelligible → uses non-technical language, is simple, readable, understandable, interpretable, scrutable…

- Vice → the false appropriation of explanatory information.

Roadmap

- Summary of emerging trends in XAI and HCXAI
- Unify intuitions under "folk psychology"
- Diagnose the problem of false appropriation
- Propose a hypothetical remedy

Emerging Trends in XAI

- Model agnostic > model specific
- Local explanation > global explanation
- Development of surrogate models for black-box models
- Post-hoc → *I therefore, O.*

Emerging Trends in XAI cont'd.

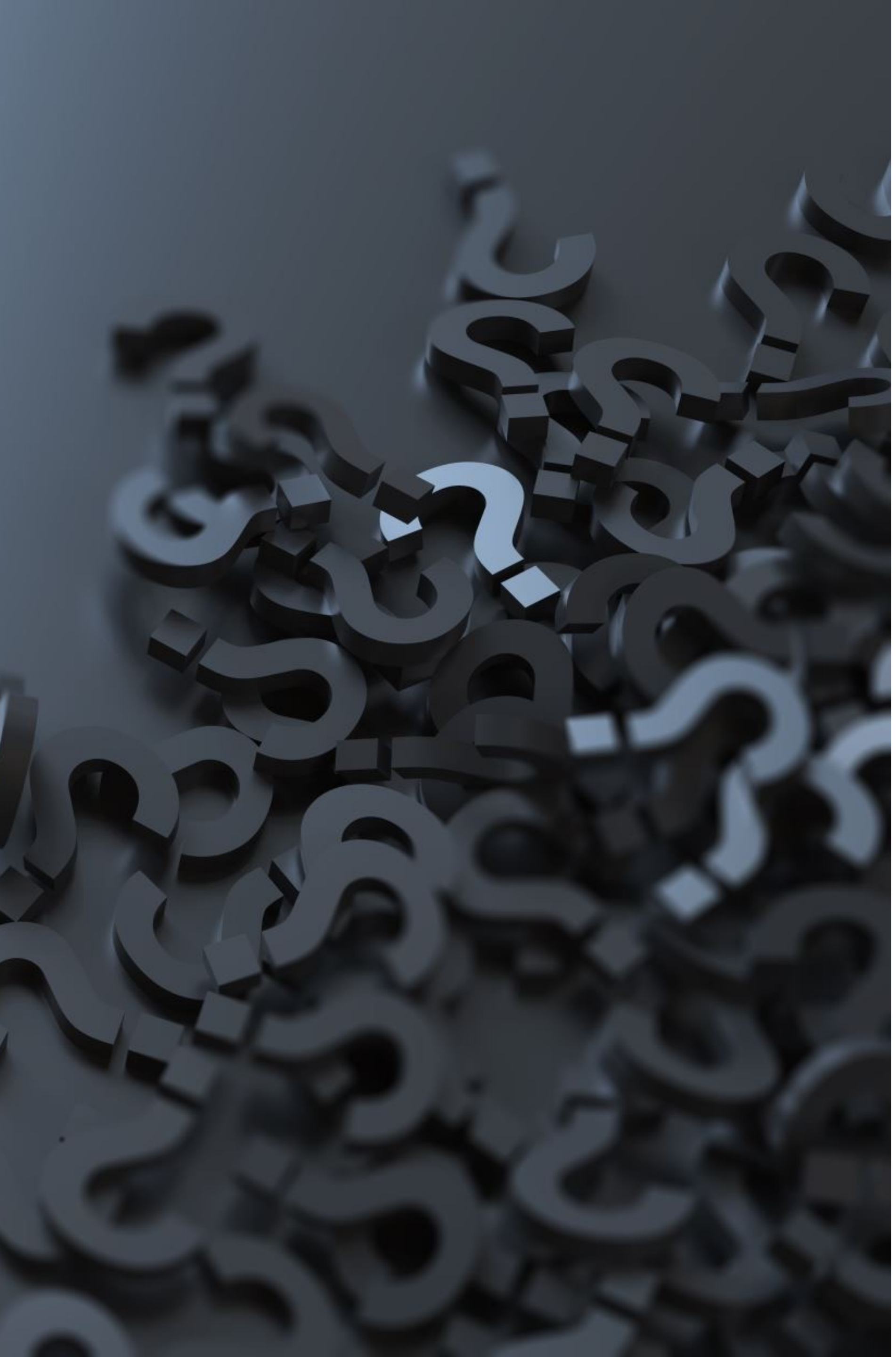
- Rationale generation → language explanations
- Rationale generation manifests as intentional-explanations – explanations reflect a models "propositional attitudes"
- Intentional explanations furnish a valuable explanatory language
- Greater success with interactive explanations
- Interactive + intentional = incredibly human-centric.

Emerging Trends “mental models”

- From interactive explanations we get insights about an explainee’s mental model and how to improve it, we also get a hypothetical mental model representing the decision structure of an XAI model.
- Essentially, via interactive-intentional explanations there is a mirroring-effect. Explainee’s expect intentional explanations from a model and by way of a dialogue, the decision structure of a model can be uncovered….(can it?).

Emerging Trends “dual-process theory”

- Avoid system 1 cognition: heuristic reasoning.
- Encourage system 2 cognition: analytic reasoning.
- How → use cognitive forcing functions (CFFs).
- When an explanation does not falsely appropriate explanatory information, the explanation is a true test of a human-centric explanation that renders an opaque model transparent.

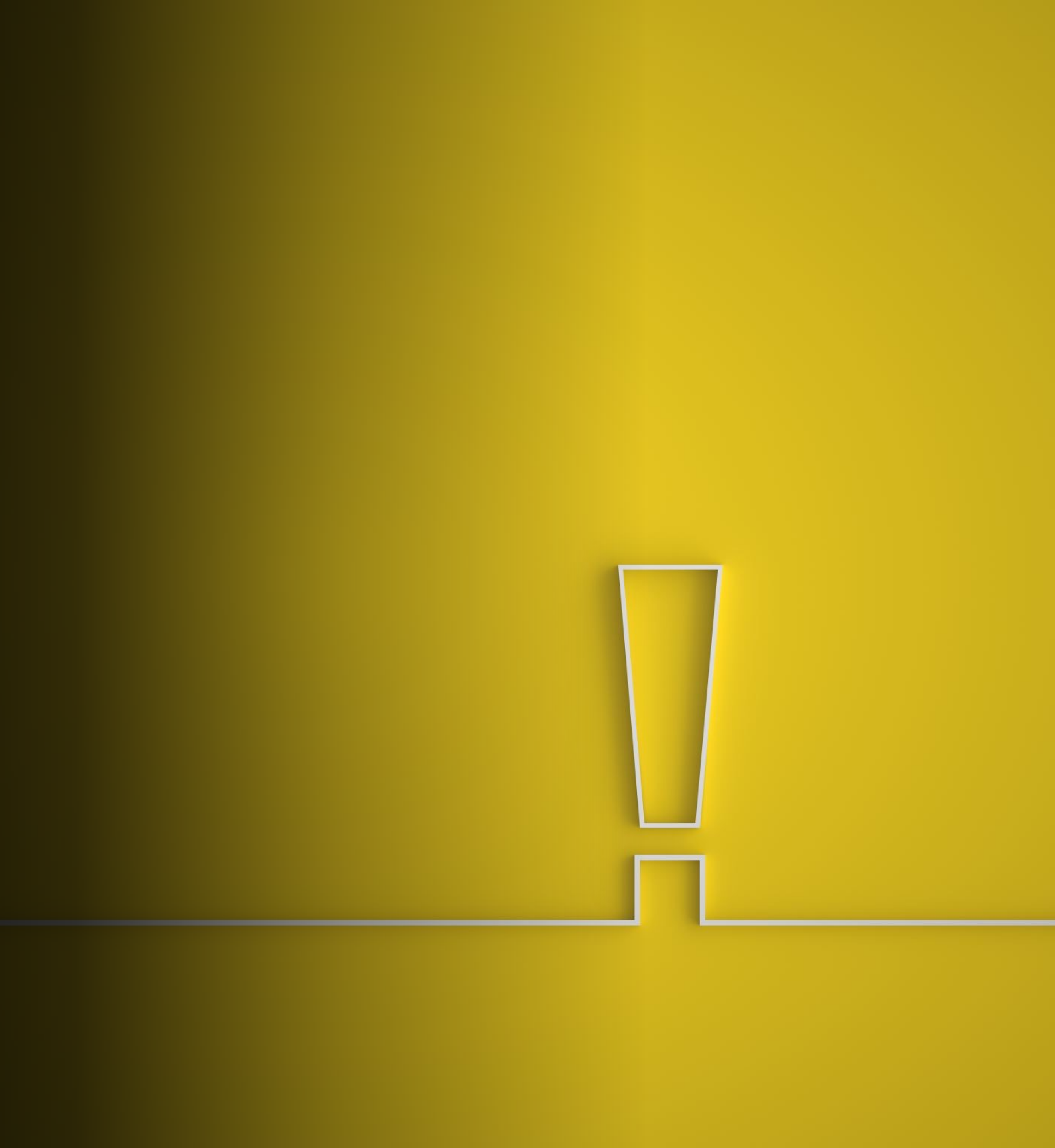


Unifying Emerging Trends

- HCXAI & XAI (incl. CoT) adopts a pragmatics of explanation grounded in folk theories of mind.
- Intentional explanations of intentional systems only go so far as to claim that “on occasion a purely physical system can be so complex, and yet so organized, that we find it convenient, explanatory, pragmatically necessary for prediction, to treat it *as if* it had beliefs and desires and was rational” (Dennett, 1987, pp.91-92, *emphasis added*).

Diagnosis

False appropriation is not the cause of a lack cognitive forcing functions, it is the lack of the *right* cognitive forcing functions relative to the kind of explanation furnished by an XAI/HCXAI model.



Diagnosis cont’d.

False appropriation is not just caused by misunderstanding the content of the explanation.

False appropriation can arise in virtue of implied content given the nature of the dialogue and or explanation context.

Potential remedy → Dual-Process Theory + Cognitive Forcing Functions (but, not just any CFFs, the *right* CFFs).

Overreliance Vs. Anthropomorphism

“I grant that a misconstrued understanding of an AIs agentic capacities is much more serious than an over reliance on AI. In its most serious of manifestations, this misconstrued understanding of the agentic capacities of artificial intelligence is the belief that AIs are conscious persons that outperform humans and will continue to outperform humans in basic and complex cognitive tasks at increasingly rapid rates (it is a science fiction nightmare). Of course, this perception is intertwined with over reliance.”

Inspiration → Dennett (2023) “The Problem With Counterfeit People”

Remedy

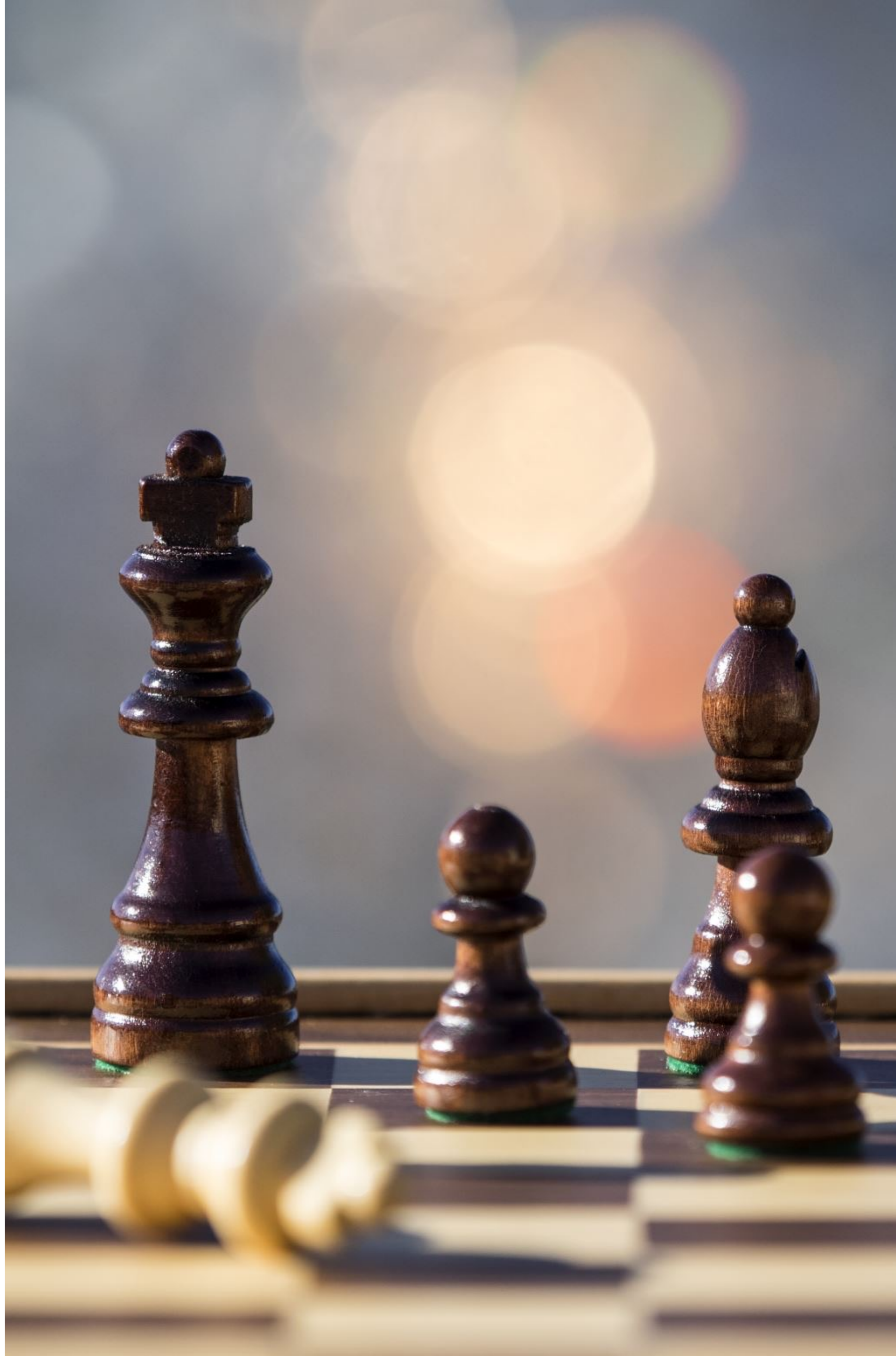
Using CFFs to limit anthropomorphic biases.

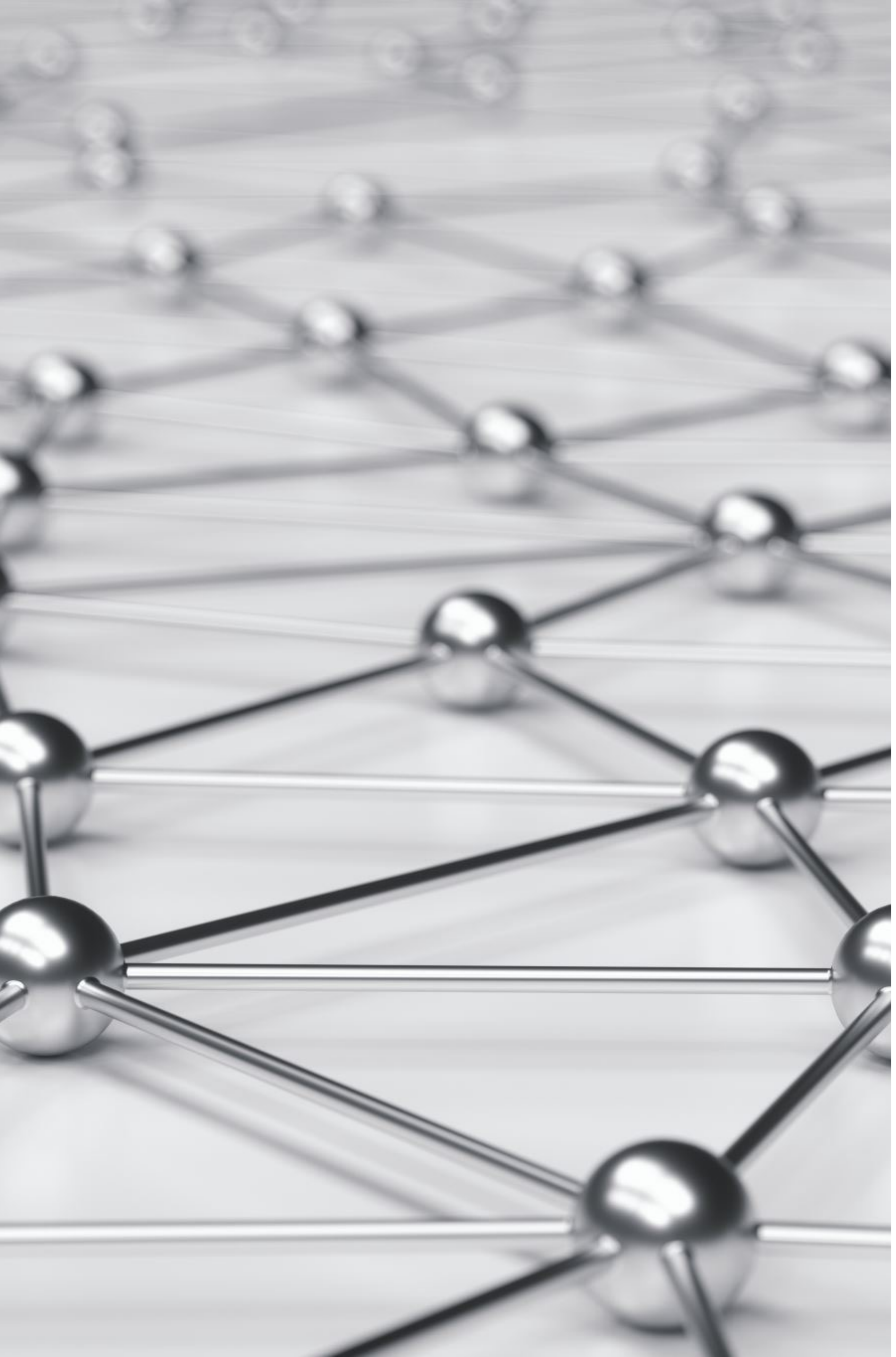
How? This remains an open inquiry.

- Stop idealizing human-centric techniques;
- Do not trade faitfullness for decreased cognitive load of HCXAI technique;
- “Correct” does not entail “causally complete;”
- Re-think the value of intentional explanations;

Inferences in HCXAI techniques

- “Self-explanation” via human-computer-interaction (HCI)
- Cognitively “force” system two cognition via contrastive and counterfactual reasoning.
- Cognitively “force” *practical inferences* to correct for false beliefs that stem from the conversational nature of the explanation.





What’s next?

- Continued work on values in user experience (UX) design and development of novel XAI & HCXAI techniques
- **Project 2:** Use HCI as a learning opportunity for understanding human and machine decision structures
- **Project 3:** Questioning insights and consequences of sharing technical vocabularies between disciplines that study human and artificial minds, i.e., “conceptual borrowing.”

Contact Information

Carson Johnston, cjohn445@uwo.ca

PhD Candidate, University of Western Ontario (Western University)

Department of Philosophy & Rotman Institute of Philosophy