Insights from the Causal Cognition in Humans and Machines Conference

11-12 January 2024

University of Oxford

Understanding causality is deemed crucial for making informed decisions and shouldering responsibilities in both artificial and real-world contexts. The applications span a wide range, from fostering reasoning and learning to identifying cause-effect relationships, enhancing problem-solving skills, augmenting predictive power, and refining the accuracy of causal inferences. As artificially enhanced systems increasingly integrate into our daily lives, such competencies are no longer exclusive to humans.

Following the first workshop in 2019, the second conference on Causal Cognition in Humans and Machines brought together experts in cognitive science, machine learning, AI, and related fields, students from diverse disciplines, and entrepreneurs for in-depth discussions on causal intelligence.

The aim was to explore;

- *Innovative methodologies for cognition and intelligent systems*, how to enhance analytical tools, novel frameworks, and experimental designs in the context of causality.
- *We embraced diverse perspectives through* seven keynote presentations, nine short research talks, nine posters, followed by a panel discussion,
- *Fostering global collaboration* among participants from various countries was *facilitated via the interdisciplinary nature of the topic***,** which provided valuable insights into the connections between intelligence and causal thinking across different fields, including cognitive science, computer science, and AI.

We found that despite the methodological diversity, most of the talks were quite accessible to the wide-ranging audience. The conference successfully served as an interdisciplinary platform, fostering fruitful discussions and shared insights among participants while bridging connection between the fields.

As we transition from human cognition to machines, key discussions emerged regarding;

- **The significance of counterfactual and causal explanations**: In an insightful talk, we explored the impact of counterfactual explanations on decision-making processes when individuals encounter an AI system's choices between certain and risky options. Further examination of subjective preferences and objective measures reveals distinctions in how people perceive and comprehend counterfactual versus causal explanations of AI decisions. The narrative also extended to empirical findings, highlighting differential preferences for simple causal explanations in predictive inferences compared to diagnostic inferences, providing valuable insights into the realm of causal cognition in humans and machines.

- **Cognitive bias and emotional influences**: In a discovery, it was observed that individuals frequently overvalue their proficiency, particularly in social settings, attributing this overestimation to a fusion of comprehension and personal values. The perceived understanding of individuals tends to escalate when topics are presented through the lens of sacred values rather than causal consequences. These findings underscore the influence of cognitive biases and emotional factors, resulting in a complex interplay between emotions, values, and decision-making processes, and offering insights into the multifaceted nature of intelligence in social contexts.

- **Insights from educational neuroscience on causality, brain mechanisms, and sensory hierarchies**: Teaching the brain to think causally and comprehend causality is a complex process through which develop largely based on experiences. It is evident that these intuitive concepts collide largely with scientific knowledge imparted in formal education, there arises a need to understand how the brain learns and processes causal information. A particularly enlightening aspect of the discussion focused on the work done within and adjacent work of the Centre for Educational Neuroscience. The approach underscored the importance of inhibiting intuitive concepts through content-specific brain circuits to facilitate access to counterintuitive ideas acquired in formal educational settings. The key takeaway emphasized the trainability of these circuits. On a more contemporary note, sensory hierarchies are recognized as fundamental to causal thinking from a neuroscience standpoint, while the brain mechanisms involved in learning are largely considered to be domain general. This insightful discussion showcased the symbiotic relationship between educational outcomes and the underlying neural

processes, advocating for an aligned understanding of how the mind learns and processes causal relationships.

- **The interplay between general intelligence mechanisms and causal thinking**: We have robust cognitive models, and extensive research in intelligence solidifies our understanding of human cognition. Recent studies, casting a spotlight on the nexus between general intelligence and causal thinking: illuminate a profound impact of the former on the latter.

- **Human-like AI or Beyond?** The quest for human-like AI beckons us to question the very essence of artificial intelligence. Are we on the path to creating machines that merely mimic human intelligence, or are we venturing into uncharted territories, surpassing the confines of human cognition? This dilemma sparks discussions that delve into the potential of AI, not only to replicate, but also to transcend human capacities, a topic that could be explored further in the next conference.

- **New methodologies I: Diagrammatic insights: unveiling compositionality and causality linking mathematical and linguistic representations:** A specialized team at Oxford is actively working on the realm of compositionality, offering Compositional Intelligence (CI) as an alternative to AI. Their focus lies in comprehending how CI systems can adeptly decompose and recompose elements to derive nuanced meanings. The significance of understanding how CI contributes to the holistic intelligence of machines, and what are the implications for achieving a deeper comprehension of causality is largely unexplored. The team posits that causality can be viewed as an instance of interactive relationships, effectively visualized through their diagrams. A compelling experiment involving human participants demonstrates the power of use of these diagrammatic representations in breaking down complex mathematical concepts. Taking the example of quantum physics, which was previously inaccessible to anyone without advanced training in physics and mathematics, now these representations make it more accessible and alleviate the mathematical barriers. Notably, the experiment, conducted with younger learners (high school students), indicates that this formalism aligns more closely with how humans naturally think, reason, and conceptualize complex scientific content.

- **New methodologies II: Generative Flow Networks**: These networks not only relate to variational models and inference but also offer promising avenues for non-parametric Bayesian modeling, generative active learning, and unsupervised or self-supervised learning of abstract representations. GFlowNets are implementing System 2 inductive biases, crucial for incorporating causality and rationalizing out-of-distribution generalization. They empower neural networks to model distributions over complex structures like graphs, enabling sampling and estimation of complex probabilistic

quantities that would otherwise seem intractable. While initially challenging, GFlowNets seems to represent a powerful paradigm, with ongoing research and publications unveiling their potential across various applications, from explaining causal factors to dealing with out-of-distribution scenarios and bridging the gap between state-of-the-art AI and human intelligence.

- **Linguistic Dominance in AI Development:** The discourse surrounding the evolution of artificial intelligence (AI) prompts profound inquiries, pushing the boundaries of what it means for AI to be truly human-like or perhaps even surpass human capabilities. The overarching themes include the pivotal role of the predominant focus on Large Language Models (LLMs), where the degree of causality within the language domain remains ambiguous. The prevailing methods in training and enforcing current machines and intelligent systems often revolve around linguistic and algorithmic approaches. However, this approach may overlook the significance of developmental experiences, spatial-temporal cognition, and the conveyance of a substantial body of knowledge through nonverbal routes. Are we limiting machines by compelling them to learn and develop predominantly through linguistic tools? Does the current emphasis on linguistic methods neglect other vital dimensions of cognition? These are some questions that could be explored further in the next conference.

- **The Linguistic and Mathematical Conundrum:** A considerable amount of time and research has been invested in combining mathematical and linguistic tools to enhance the efficiency of AI. While AI excels in delivering swift solutions for certain operations in mathematics and language, this has predominantly focused on speed and accuracy. The current landscape raises concerns about whether we are devoting enough attention to fostering complex and diverse capacities in AI that mirror the capabilities of the human brain.

- **Optimizing Causal Inference in AI and Cognitive Intelligence:** As we contemplate the future trajectory of AI and Causal Intelligence, an intriguing proposition emerges: Can AI and Causal Intelligence collaboratively focus on optimizing causal inference and benchmarking models? The majority of existing models often rely on probabilistic or Maximum Likelihood approaches. The idea is to transcend traditional methodologies, seeking innovations that propel AI beyond traditional LLMs, which often lack explicit causal reasoning. Such an idea aims to address challenges related to bias, interpretability, and robustness in AI applications.

These discussions serve as a call to action, encouraging the scientific community to rethink existing paradigms, explore new dimensions, and collectively shape the future of AI and cognitive systems.

Future Prospects

Encouraged by the success of the first two conferences, we are excited about the possibilities that lie ahead. We are in the early stages of planning another special journal issue that will compile the wealth of research and discussions that emerged during the 2nd conference (Please refer to the first conference special issue here).

As we continue our journey, we are also considering a more focused conference format for the next edition. This involves defining specific questions in advance and welcoming keynote speakers who would like to address them.

Focusing on at least two themes, one aspect of the upcoming conference envisions a compelling exploration of augmenting human cognition and learning, including brain-machine interface (BMI) systems. This anticipates a future where these interfaces play a pivotal role in enhancing human capabilities. One aspect could be exploring the symbiotic relationship between humans and machines, contemplating whether human brains, when interconnected with machines through interface systems, can amplify cognitive capabilities. This forward-looking perspective seeks not only to uncover the potential cognitive enhancements for individuals but also to explore the reciprocal influence, questioning how the remarkable capacities of human brains might transcend into machines.

Another theme we intend to explore is the realm of non-linguistic AI. it is not the case -as is widely assumed- that verbal communication dominates dialog or understanding. Evidence suggests that nonverbal elements account for up to 70% of human communication. This elucidates why modern technologies lack the capacity to deliver the variety of messages that humans use concurrently. This theme will focus on the burgeoning field of AI that goes beyond LLMs, encompassing areas such as visual perception, sensory processing, spatial-temporal reasoning, and so on.

Communication and collaboration

We invite you to explore our websites for detailed information and insights:

The link of the first conference: https://www.causalcognitioninhumansandmachines.com
The special issue compiled the work: https://www.frontiersin.org/research-topics/9874/causal-cognition-in-humans-and-machines

The link of the second conference: https://amcs-community.org/events/causal-cognition-humans-machines/

If you have any inquiries or require additional information, please do not hesitate to reach out via **cchmconference@gmail.com**

We greatly appreciate your time and consideration and look forward to the possibility of collaborating with you to share the exciting developments in the field.