Causal Cognition in Humans and Machines: Intelligence and Intelligent Systems

Tim Lawson

tim.lawson@bristol.ac.uk

University of Oxford
11–12 January 2024

Contents

Thursday 11 January 2024

Limited Human Causal Knowledge is About Interventions

Steven Sloman

Brown University

How can we elicit causal models from humans? In the first study, they ask people to assign functions to the parts of an object, e.g., a light source. Then, they ask people to provide causal rules that link the parts of an object. From this, they show a causal graph, in response to which the user can change their inputs. The effectiveness of the resultant causal models is measured in terms of the ability of a robot to construct the object from the model. The specific metrics used are the 'hit rate', i.e., the proportion of causal connections elicited that match the ground truth, and the 'false alarm rate', i.e., the proportion that don't appear in the ground truth. Finally, the performance of the robot is measured in terms of the number of steps saved in the assembly. The results indicate that people often fail to generate true causal relations but rarely generate false ones. There are problems with this methodology, e.g., it requires a long tutorial and takes a long time to complete, with the outcome that participants often failed to complete it. Moreover, the resultant causal models tend to be quite 'flat', i.e., unsophisticated. In general, people didn't find the interface user-friendly.

So, how else can we elicit causal models from people? A second method takes advantage of the intuitiveness of *intervention*. In other words, what's the effect of a change like breaking the lightbulb? We ask people whether removing parts affects the functionality of other parts, and of the whole object (removal queries). If no, then there's a causal relationship between the two parts – it's a counterfactual. However, this only really makes sense for small systems with relatively few parts. How can we infer causal models from these relationships? We start by determining a graph or adjacency matrix to represent all of them. From this, we can infer and underlying causal model by removing direct connections that are explained by indirect connections. That is, if A affects B and B affects C, then we don't need the direct link between A and C. This method improves the hit rates and slightly improves the false alarm rates, so it seems like the interventional approach to eliciting causal structure is more effective. We also gave the participants the functions in the second case. Working on combining the two in that way. The suggestion is that people do have access to causal knowledge, but it's interesting to think about the nature of its mental representation. The lack of success with the explicit graphical method suggests that people can't necessarily describe their causal knowledge, i.e., don't have access to it. This approach is explicitly linked with counterfactuality.

'The knowledge illusion', or the illusion of explanatory depth. It seems like we should know the causal structure of simple objects! But people overestimate their understanding. Extending this

observation to the domain of political policy, we find that a causal explanation reduces people's sense of understanding and their hubris, but simple reasoning doesn't. See Rozenblit and Keil (2002) and Fernbach et al. (2013). With regard to scientific consensus, the more objective knowledge people have, the more they tend to agree with it. But we find the opposite correlation between *subjective* knowledge and people's opposition to scientific consensus. Notably, we didn't find this effect for climate change, but we did for genetically-modified foods, gene therapy, childhood vaccination, and COVID-19.

We can link these observations to decision-making strategies. We assume that we should take a *consequentialist* or utilitarian approach – in other words, the best decision is the one with the best consequences. This approach relates to causal or probabilistic beliefs. On this view, the role of causal models is to indicate the likely consequences of a decision. An alternative to the consequentialist approach is one based on *sacred values*. These values are absolute, i.e., trade-offs based on the likely consequences are considered taboo. This approach relates to deontology, although there isn't necessarily an explicit theory from which the moral precepts are derived. This approach has a long history in Western thought, among others. It may be that sacred values arise from emotional responses rather than theoretical derivations. On this view, values refer to *actions* rather than *outcomes*. Value commitments can 'fill in' causal gaps – good leaders can take advantage of these values, e.g., to foster an ideological community based on a sense of outrage.

The proposed model is to frame issues in one of two ways (sacred values versus consequentialism), which affects people's subjective understanding, which leads to either compromise or outrage. We find that subjective understanding increases aversion to compromise and tendency to outrage when sacred values are violated. See Fernbach et al. (2018). We present participants with a mock Reddit page with discussion on a topic, where opposing views are presented either in terms of sacred values or consequences. People have a higher sense of understanding when arguments are framed in terms of sacred values. Sacred-values frames also decrease people's willingness to compromise (perceived tractability), i.e., how likely it is that the parties could resolve their disagreement. In another study, we ask people to choose a reason for a policy using words from either of two sets that imply either a sacred-values or consequentialist frame. We then ask people to indicate their willingness to donate to a charity when it's framed that way. Generally, we find that the results follow the direction of the prediction – people are more likely to be willing to donate, and to donate more, when they generate a reason for the policy based on sacred values. A meta-analysis of related studies finds a significant effect. We also look at companies that brand themselves in terms of sacred values versus those that don't (specifically, Patagonia and H&M). Again, we ask people to write a reason for an argument based on different sets of terms, and find that changing the frame changes what people care about.

In summary, people are bad at constructing causal models of everyday objects. People can, however, answer questions about the effects of interventions. Decision-making by sacred values is relatively common, and sacred-values frames increase subjective understanding, reduce perceived tractability, and increase willingness to act.

Questions Are we talking about the same kind of causal reasoning in the two projects? The inferred models make a big assumption in terms of causal theory. It's clear that people create too many causal relations in the original graphs, and it's less clear how we could infer a simpler (more correct) causal model without eliminating direct connections that are explained by direct connections. When do

people revert to a greater level of detail? It's also assumed that people are generally the same.

Learning Causes and Using Them

Samantha Kleinberg

Stevens Institute of Technology

Photo of a neurological intensive care unit (ICU). Most ICUs don't store data, or don't store it electronically. It's difficult to argue that it's valuable to capture that data without access to it. There's something of a trend in AI to not concern ourselves with *why* something works. However, papers commonly use a lot of causal language, despite their ostensible lack of interest in causality (Cofield et al. 2010; Han et al. 2022). Generally, papers use causal language in the abstract and discussion but not elsewhere! Perhaps causality is a desirable property that it's valuable to pretend you have.

Alarm fatigue: the more alerts you get, the less likely you are to pay attention to them. In the context of healthcare, patients and doctors really want to know why something is happening and what they can do about it – action requires causality. How can we extract causal information from data? What about when we can't intervene or control what data is recorded?

We looked at the causal structure of brain physiology following a brain injury (Claassen et al. 2016). The model is very complicated and includes a lot of dependencies on time, but it doesn't capture anything that wasn't explicitly measured. Doctors wanted a simulation system to test the effects of interventions. We thought we needed causal models to help experts to make decisions and take actions, but when we produced a causal model, it wasn't useful to them. As scientists, we're pleased when we come up with a comprehensive diagram, but then people use the complicated diagrams to demonstrate why a policy is unworkable. Are these kinds of models actually useful for decision-making?

There isn't much evaluation in the ML and AI community regarding causal representations. There are positive results in psychology on how children learn causal models. Can causal information aid decision-making in familiar scenarios? We presented people with the same textual information and, in some cases, a causal model – it's like an 'open-book exam', i.e., we assumed that people already had the model somehow. But people did worse with the text and/or diagram (Zheng et al. 2020). 'Open-book exam', assume people already have the causal model. People did worse with the text and/or diagram! What if we change what people think they know? (Kleinberg and Marsh 2020). Dual-process theory (system-1 and system-2). Once it becomes familiar, the causal model is unhelpful. It's not that people don't have a causal model internally, but that they already have something that's difficult to explain and not necessarily reconcilable with an external model.

How do you focus people's attention? We tried tailoring diagrams to the question, i.e., to only show the causal paths that are relevant (Kleinberg and Marsh 2021). Simpler diagrams improved people's performance. Causal chains: intervene on roots versus direct causes. E.g., showing that 'seeing food ads' leads to 'not eating healthy' makes people more likely to choose 'fast-forward through ads' as a response to how to lose weight. We found the same effect by highlighting the relevant parts of the model rather than removing the rest of it. We tried letting people choose what information they would find useful – more information leads people to overthink.

What causes do algorithms find? Stronger causes, earlier causes, modifiable factors. Are some causes more valuable than others? We don't typically optimize for these, which is important given that

we can only give people so much information if they're going to use it effectively. A complete model is not always a useful one. We have done some work to investigate how the strength of causal relations influences people's decisions, but it's non-trivial to indicate the strength of relationships visually.

Questions Is there a difference between causal and predictive models? How does causality relate to the aims of explainable AI and interpretability?

Steven Sloman: Does providing causal knowledge help people reason about things? Do nodes and links specifically help people? We found that using nodes and links helps people to agree on an overall model. Physicians tend to be experts in a narrow domain, so that framework is useful for integrating and representing their combined knowledge. Showing the nodes and links doesn't help people reason all that much, even though they're good at reasoning locally. But maybe it does help at a collective level. Samantha Kleinberg: We don't think the result is due to the diagrammatic form only, we find the same with text. Information conflicts with people's pre-existing knowledge and/or it isn't easy for people to combine it with their own. We tried eliciting mental models from people, but it's difficult to get things that are more complex or consistent with people's decision-making.

Causality Beyond Shannon

Stefan Leijnen

University of Applied Sciences Utrecht

Shannon entropy underpins most, if not all, AI. It 'brackets out' the causality of the processes that underlie the creation of the probability distribution. This was explicitly mentioned in "A mathematical theory of communication": "Semantic aspects of communication are irrelevant to the engineering problem. " (1948). Selecting from among the possible messages. This idea treats information as a substance rather than a relation: quantifiable but without semantics ('aboutness'). 'Aboutness' is the presupposed relation between the source and destination that makes the signal *useful*. The 'bracketing out' of this relation yields the 'semantic debt' of modern-day AI, which leads to problems with understanding, alignment, and safety. Where does this presupposed relation come from?

Aristotle on causality (Phys II and Metaph V). Understanding reality requires causal knowledge, yielded by causal investigation. The study of existence looks for why something is created and destroyed and why it is. Aristotle distinguishes between intrinsic and coincidental causes: material, formal, efficient, and final causes. The material cause is compositional. Causality is about understanding the regularities and habits that underpin 'goodness', i.e., quality or suitability. Interestingly, Aristotle considers these four causes to be irreducible but ordered, i.e., the final cause offers the best explanation. We tend to observe that final causes are formal, and tend to conflate them. That is, we ignore the processes that underpin the relation that's explained by final causality, or explain them in terms of formal and efficient causes. We focus on the substance rather than the purposeful relation.

There is good reason for science to mistrust teleological explanations, due to fear of the theological connotations. With respect to AI, *telos* as objective, coincidence as randomness. Substrate-independent information processing where the underlying thermodynamics and synchronicity are abstracted away leads to technology that is not grounded in reality beyond the efficient but thin causal link with transistors. We try to do away with the causal link to model the others. So we need to study the

mechanisms of primary causes, e.g., what the final causal relation is about: what are the physical properties that underpin the regularities/synchronicities of the model.

The CAUSE research programme, established 2021, aims to 'explore different socioeconomic gardens for AI to blossom'. It is named for: Creativity, Autonomy, Understanding, Sentience, Ethics. Projects: mechanisms of final causality; convivial AIs; causal reasoning through constrained learning.

Operationalization of Steerable Kernels: Causal Reasoning Through Constrained Learning

Coert van Gemeren

University of Applied Sciences Utrecht

The steerability of equivariant kernels is a powerful concept to extend the notion of translation equivariance to Euclidean and non-Euclidean geometries (Weiler and Cesa 2019; Cesa et al. 2022). Geometric deep learning: with regular convolution, rotation invariance only comes from data augmentation. A model is *invariant* if a group action on the domain results in the same activation in the co-domain. A model is *equivariant* if the result of the operation is the same when actions are applied on the domain and co-domain. G-steerable kernels constrain the representation space to outcomes computable by the steering operation (G-action). An affine roto-translation steerable kernel constrains the kernel representation in Euclidean space to the precise rotations allowed by the affine rotation matrix. Invariance is a special case of equivariance when the action on the co-domain is the identity.

Constraints help align a CNN to a geometrical goal. Discretely computable outcomes for a specific degree of freedom of the kernels. Weiler and Cesa (2019) show that a G-steerable kernel guarantees invariance of neuronal activation for equivalent samples. What about other domains, e.g., color, age, and gender? Fairness constraints. But it's difficult to create a kernel to morph data to represent different ages. We can consider causality as a stack of steerable operations. We can stack translation-equivariant layers in a CNN, so perhaps we could model causal reasoning through operationalization of features in a stack of kernels. The 'Causal reasoning' space could be constrained to find valid chains of causes and effects. Instead of optimization, we could find models tailored to utility due to ethical considerations. Is there a relationship to analogical relationships in embedding spaces? Objective functions are a simple proxy for the utility of a model.

Bridging Between Quantum and Living Systems Towards Human Cognition and AI

Tomas Veloz [1] , Olga Sobetska [2]

1 *Universidad Tecnologica Metropolitana, Chile*

2 *Centre Leo Apostel for Interdisciplinary Studies, Brije Universiteit Brussel*

Is causality an invariant of experience? This talk is about emergence, Self, embodiment, and goals. Boole introduced the notion of possible experience. There are constraints on the states of

affairs that data can point to, and likewise some things that cannot happen (cf. Bell inequalities). The conditions of possible experience and, relatedly, causality, have been formalized in various ways. See, e.g., Abramsky's formalization of Bell inequalities in databases (2013). These formalisms imply a classical probabilistic representation, i.e., hidden variables, which are always satisfied if experiments are performed on a single sample. This concept is fundamental to causal reasoning, e.g., expert systems, fuzzy set and possibility theories, Bayesian inference, and database consistency. The conditions of possible experience are a *meta-invariant* for causality. However, there is experimental evidence that challenges the validity of these conditions in quantum physics and cognition.

Consider the example of the conjunction fallacy, which is well-established in various circumstances (Kahneman and Tversky 1972). Veloz and Sobetska (2023) give a literature review of instances of the conjunction fallacy and visualized the differing probabilities of A, B, and $A \cap B$ that were used. Quantum theory can model *interference* between A and B, i.e., that the combination of the two is somehow its own entity. Interference is maximized when you know least about the two elements individually. They also find that these instances exhibit a consistent deviation from rational (classical) expectations, with differing forms of the de Morgan laws.

What about an 'emergentist' approach to machine learning, that centres agency, autonomy, and goal-directedness? The dominant paradigm in ML is that optimization is a defining feature of the notion of 'goal', but the process of configuring goals is not considered. Goals are intrinsic to the self but intimately related to environmental interactions. These interactions explain the emergent complexity of goals. Autopoeisis (self-creation). The idea of affordances: an entity has a particular structure, which can *afford* reality in a particular way, and has certain goals. A faculty of this nature implies dangers and related objectives.

Complex adaptive systems, and the example of chemical organization theory. In order for something to be, it has to be stable (self-maintaining). The network needs to be closed, i.e., the things that it produces must be part of the same network. Stationary states of such a system are organizations, and its structural dynamics are movements between organizations, triggered by perturbations. The set of organizations is a partially-ordered set. How does such a set evolve? Random walk over structures and the possibility of stable structures developing (Veloz et al. 2023). Resilience to perturbations and the basis of a more 'constructivist' theory of cognition.

Nietzsche's three transformations of the human spirit, for the AI era. The evolution of self-producing complex adaptive systems provides a framework to explain the emergence of purposeful intelligence. Can mathematics from quantum theory help to explain this emergence?

Amortized Inference of Bayesian Causal Models for Quantitative AI Safety Guarantees

Yoshua Bengio

Université de Montréal

The content of this talk represents the intersection of a more recent interest in AI safety, and a long-standing interest in bringing high-level human cognitive abilities to machines. These abilities are missing in the state-of-the-art. The aim is to create AI systems with *epistemic humility*, i.e., knowing

what they don't know, and quantitative safety guarantees. How can we avoid catastrophic outcomes from AI? We need to solve the *alignment* and control problem, which is technical and political (due to the costs of research and development). There's also the *coordination* challenge: making sure safe and ethical protocols are followed, and preparing for a situation in which a 'rogue' AI emerges anyway. This is a socio-technical challenge – even if we knew how to make AI safe, how would we get people to do it? Today is mostly about the technical aspects.

Statistically, the true and estimated objective or reward functions may appear similar on a random test set. But maximizing the estimated reward function is likely to reveal the greatest mismatches between the true and estimated rewards! This happens empirically, and the difference can be very large. In a sense, we want the system to 'know' that what it's optimizing 'isn't the real thing' and that there's some uncertainty about the true objective. We need CIRL-like solutions – human intentions are a latent variable. Reinforcement learning leads to 'reward hacking' with lots of compute. Cohen et al. (2022) – optimizing a policy to maximize expected reward necessarily leads to reward hacking as compute increases. In the extreme case, advanced artificial agents could intervene in the provision of reward. In other words, if we tried to interfere with how the system was rewarded, then it could try to stop us. This is a problem with maximum-likelihood models. They may be confidently wrong (hallucinate). It doesn't happen too often, but it's a concern if the potential outcomes are severe. Optimizing for maximum-likelihood discovers 'fantasy treasures', and doesn't provide any quantitative safety guarantees.

For some time, I have been a proponent of end-to-end AI training. But there's no explicit world model or ability to plan, which humans have. An idea is to separate an inference engine, which can answer questions in natural language, and the model itself. The optimal capacity of the model is far less than the optimal capacity of the inference engine. For instance, the causal rules that define the game of Go are very simple, but the game requires a very large machine-learning model to make causal inferences. Typically, end-to-end deep learning confounds both of these parts: it overfits the world model and under-fits the inference engine. Moreover, it doesn't trivially offer a way to incorporate *inductive biases*. I think of causality as an inductive bias. We can constrain how an AI system understand the world by encoding some mathematical properties of causality and interventions that restrict the distributions that can occur in the world. In terms of 'system-1' and 'system-2' thinking, deep learning is doing well at the first, but not at the second. It's not great at deliberation, planning, reasoning, etc.

A causal model is joint over variables and interventions – it's a family of distributions over variables, which are indexed by interventions. They all share parameters. The idea is that generalization over possible interventions may equate to out-of-distribution generalization (to distributions that correspond to unseen interventions). Causal dependencies have special constraints: the causes are marginally independent, and interventions are special causes that change the mechanism to ignore or modulate the other causes of an outcome. The actions of agents lead to interventions. Latent dynamics in the world can lead to cyclic dependencies. Markov equivalence class (MEC) of causal models. Even with an infinite amount of data, still can't determine the right causal graph. Finitude of interventional and observational data. Bayesian posterior over causal models. Need to understand that there are multiple explanations for the data we're seeing. For humans, we call that *epistemic humility*.

Bayesian posteriors for safe uncertain decisions. Maximum likelihood means choose whichever theory. If you know that the theories coexist, and you don't know which is correct, then you can

avoid catastrophic consequences for actions. Reject action if probability of harm given an action, data, context, is greater than some threshold. How can we efficiently estimate these Bayesian posterior probabilities conservatively? We can't afford to underestimate it, so we need to make sure we make errors in the right direction. The gold-standard for Bayesian calculations uses Markov chain Monte Carlo simulation. We think of the distribution as over theories about the data – if that distribution has many modes, which will be the case, we have the 'mixing problem'. We can explore trajectories by making small changes to hypotheses (sample). How do you get between modes? It's exponentially difficult. This happens in practice in high-dimensional spaces.

We want increasing compute to increase safety – MCMC works in principle but doesn't scale. Neural-network amortized inference approaches the Bayesian posterior over causal models as the network size and training time increase. We need neural networks that can represent rich Bayesian posteriors. We have a forthcoming paper about variational approaches (forthcoming). The correct calculations, inferences, etc. will be intractable, but we can approximate them with a neural network. This exploits mathematical results that tell us that the larger the network and the longer the training, the closer we approximate the true distributions.

Why rich amortized predictors? Compositional so exponential number of models, and it's difficult to mix modes with MCMC. Traditional variational inference produces a single mode-seeking approach. Machine-learning predictors can generalize from small fraction of visited modes by exploiting generalizable structure. Generative Flow Networks (GFlowNets) implicitly marginalize over models – amortized predictors are fast at run-time, i.e., we pay up front for training, but inferences are cheap. Intersection of RL and variational methods: the policy they learn stochastically samples proportional to the reward function, i.e. the more rewarded a theory is, the more it's sampled. By Bayes's rule, we learn the Bayesian posterior. See Deleu et al. (2022) and Deleu et al. (2023). How can we enforce interpretability? E.g., by autoencoding-like losses. We want to produce explanations in something like natural language, and to be able to use human theories. The theories that you care about most are the ones that could produce harmful results. We can use losses to provide confidence intervals.

Questions

- Probabilistic programming methods aren't new – why now? Originally thought it was hopeless because intractable, but GFlowNets changed that. Have a neural network generate the program from the set of programs that explain the data.
- What types of representations? This is a neural network that generates hypotheses and answers to questions. Current LLMs do that, but they're not 'trained right'. We don't care too much about the specific representations.
- Formal methods, logic arising from natural language, theorem-proving. Mathematical propositions, etc. are 'shortcuts': they save us a lot of compute. There are models at different levels – it's intractable, for example, to model macro-scale behaviour based on quantum theory, due to the computation required. We can think of pure maths as an abstract thing that searches for compact ways of expressing relationships.
- A limitation of probability theory is that it struggles to represent ambiguity. It's hard to come up with a single definition of 'war', for example, but could have a rich distribution over possible

definitions – thus the Bayesian approach. Not just probabilistic but Bayesian about everything that the AI observes.

- What about other forms of cognition, i.e., not just verbal and mathematical? We're good at generating images but not 'object-centric' vision.
- What assumptions do GFlowNets make about the underlying data? We want to avoid assumptions. A lot of previous work assumed that the posterior distribution would be uni-modal, but we want to represent arbitrary distributions. E.g., the space of programs is infinite, but you generally want shorter ones. Don't want to limit to a particular parametric form.

Causal AI for Actionable Decision-Making

Francesca Raimondi

causaLens

This talk covers causality for decision-making, discovery, model discovery, industry applications (especially healthcare), fairness, etc. There has been an uptick in the popularity of causal AI, e.g., the Nobel prize in economics in 2021. Causal graphs, confounders and spurious correlations. Confounders affect both outcome and treatment. Spurious correlations aren't true drivers of the outcome and can make estimates of the treatment effect worse. We really care about the *causal* question of, e.g., taking the COVID vaccine. This doesn't have a mathematical formulation in classical statistics. Instead, use 'do-calculus' or interventional statistics. Traditionally, decision-making with AI was just collecting data, building a black-box model, and coming up with post-hoc explanations. This doesn't address the above but also leads to overfitting, lack of trust, bias, etc. Instead, collect data, create a causal graph, build a structural causal model. This approach is less prone to spurious correlations, existence of confounders is clearer, explainable *ex ante*, etc.

How can we infer a causal graph, i.e., a directed acyclic graph (DAG)? E.g., controlled interventions, observational data, domain knowledge, or all three. Causal discovery. Reichenbach's common cause principle. It's hard to fully automate causal discovery. A could cause B, vice versa, C could cause both, or the correlation could be spurious. Full-graph causal discovery: entirely from observational data. Avoid expensive interventions, i.e., randomized controlled trials (RCTs). How can we infer causal DAGs in practice? Can introduce some human domain knowledge. E.g., constraints and inductive biases we want to enforce, classes of relationships between variables, the type of noise. Causal sufficiency, faithfulness, positivity, tiers. This can be an interactive process – human-guided causal discovery, in between fully automated and human-drawn. The main causal discovery methods are constraint-based (conditional independence tests); score-based (evaluate e.g. with Bayesian information criterion); and continuous optimization (functional learning). Kleinegesse et al. (2022) – reduce computational expense and graph sizes by pruning graph according to known and forbidden edges (causal relationships)

How to estimate the treatment effect? Instrumental variables; propensity score matching; back/front-door adjustment criteria; structured causal models (SCMs), etc. are unified under the name 'do-calculus'. Pragmatic approach: speed up algorithms and use domain knowledge, which helps adoption, interpretability, etc. What is a structured causal model? See, e.g., Peters et al. (2017). Exogenous,

hidden/unobserved variables that we suppose are causing observed variables. More informative than a predictive model. Causal modelling and moving beyond predictions: CausalNet, doubly robust ML. causaLens provide the only implementation of SCMs with doubly-robust training. The advantages with respect to traditional machine learning are that it's intrinsically explainable and easier to debias, whereas techniques like LIME and SHAP are approximate and post hoc. Post-hoc explanations aren't trustworthy or human friendly, and they're generated too late in the process.

Algorithmic recourse: what to do to produce a specific outcome? causaLens provide the only enterprise implementation of the method from this paper. E.g., in healthcare, analysis of controversial TOPCAT trial (Raimondi et al. 2022b). This trial was controversial because the results varied widely by region. Complex trials are difficult with traditional statistical analysis, particularly for multi-site clinical trials. We demonstrate regional differences in causal mechanisms.

Causality is important with respect to compliance with stricter regulations. Traditional AI struggles with fairness, e.g., it can find proxies to protected attributes, which means that removing protected variables doesn't ensure fairness. We can divide variables into groups based on whether they're valid explanatory variables. Fairness metrics, disparate outcomes versus treatment (equality and equity). One way to certify fairness is counterfactually: i.e., if X had been Y, would the decision have been the same? See Raimondi et al. (2022a). LLMs – productivity tool i.e. suggests things to data-scientist operators. Can use causal graph as context for the LLM to help prevent hallucinations etc. What about variables that aren't labelled or meaningful?

Consciousness in Causal Cognition

Johannes Kleiner

LMU Munich Center for Mathematical Philosophy (MCMP)

LMU Graduate School of Systemic Neurosciences (GSN)

Institute for Psychology, The University of Bamberg

Association for Mathematical Consciousness Science (AMCS)

Kleiner presents the claim that "consciousness is relevant to, or part of, human cognition". On this view, a proper understanding of cognition requires understanding of consciousness, and proper modelling of cognition requires modelling of consciousness. Historically, the notion of consciousness is strongly linked with the idea of 'what it is like' to be something (Nagel). Wittgenstein: ostensive definition (pointing).

The 'old-school' view of consciousness and its relation to cognition is that there are incoming signals; some kind of 'forward' cognitive processing; and conscious experience is somehow produced 'along the top'. The 'new-school' view is that cognitive processing is 'bottom-up' as well as 'top-down'. Active inference, Bayesian brain, predictive processing, free energy principle, etc. See Allen and Friston (2018) and Tull et al. (2023). How do theories of consciousness impact this picture?

- Global neuronal workspace theory: selects among multimodal parallel free energy subprocessors whose result is to distribute to all other processors.

- Higher order theories: monitoring, distinguishing, selecting among parallel free energy subprocessors (again) to make available to downstream processing.
- PP-consciousness (predictive-processing) proposals: consciousness is the generative model containing a model of itself and how itself interacts with the world. Self-evidencing, internal states and actions on the environment maximize the evidence for their own existence.
- Integrated information theory: consciousness is the causal structure that does the cognitive processing.

Questions Top-down versus bottom-up: we're only representing one part of the process in string diagrams. We think that systems show that systems can't be conscious. But we're moving towards computation that doesn't have an instruction set that determines the form of its computations.

Friday 12 January 2024

How Counterfactual Explanations Affect People's Understanding of an AI System's Decisions

Ruth Byrne

Trinity College Dublin

'If only' thoughts: we mentally simulate how things could have turned out differently, usually after bad outcomes. Counterfactual possibilities help us to explain the past, and to work out some causal relations. Partly, that helps us to prepare for the future, learn from mistakes, form intentions, and make decisions. 'Zooming in' on potential causes for outcomes. E.g., Kahneman and Tversky (1982), Roese and Epstude (2017), and Byrne (2016). I'm going to focus on the explanatory aspects. We understand a lot about how people reason about counterfactuals but comparatively less about how they explain them. There's been an explosion of interest in counterfactuals in AI, e.g., there are more than 100 computational methods to generate them. It's surprising as a cognitive scientist that there's a lack of empirical evidence with human users. Of 120 XAI counterfactual papers, less than 20% included test human users, and many of these were informal. E.g., Wachter et al. (2017/2018), Karimi et al. (2020), and Keane et al. (2021). 'How people reason with counterfactual explanations for decisions'. E.g., Warren et al. (2023). Specifically, with respect to decisions made for people by AI systems. Three strands of research (see subheadings).

Are counterfactual explanations better than causal ones?

Causal explanations are the dominant form in XAI whereas counterfactuals are newer. Would you expect them to be better? There's an assumption of a close relationship between counterfactual and causal explanations. Some recent studies have shown differences between them. The contents of counterfactual thoughts don't necessarily match causal ones. Exceptionality bias: return to normal rather than abnormal behaviours. Drunk driver caused a crash rather than 'if only they had driven by a different route'. Argued that causal thoughts focus on strong causes (necessary and sufficient), whereas counterfactuals focus on enabling. Difference in mental representations: dual versus single possibilities. Recover the pre-supposed or known facts. Causally, people are initially just thinking about the causal link, rather than what did happen and what's supposed could have happened. E.g., Kahneman and Tversky (1982) and Mandel and Lehman (1996), Byrne and Deighan (2023), in prep.

What do people think about immediately, given causal or counterfactual explanations? Short stories that people hear with some quadrants in front of them, with images or text. Wear headsets

that determine where people's eyes move among these quadrants. Where people look gives them a clue about what they're thinking. Interested in how different mental representations are vindicated by this. Once people hear 'roses', their eyes move to the part of the screen with 'roses' on it. Picture is very different for counterfactuals: become interest in either roses or no roses. Alternating between those two quadrants. So we want to argue that they're thinking about those two possibilities from the outset, i.e., counterfactuals are represented in different ways. Richer mental representations, so require more cognitive resources (McEleney and Byrne 2006).

See Celar and Byrne (2023), Warren et al. (2023), and Dai et al. (2022). E.g., user interface that shows either causal or counterfactual explanations for an outcome given input variables. Perturb action decisions until they cross the boundary and then generates a counterfactual. People judge counterfactual explanations as more helpful. Rate satisfaction with explanation and overall trust in application (1–5 or Hoffman scale) We also measured whether counterfactuals actually help people by asking people to predict given new input variables. Found that counterfactuals were no more helpful than causal ones. People feel like they're being given a better explanation, but it's not necessarily helping them more. This raises questions about fairness and trust. Conclusion: 'People prefer counterfactual explanations, but accuracy improves with either'.

Do people prefer simple counterfactual explanations?

Explanatory virtues: e.g., simplicity and breadth. Simplicity is favoured in XAI, e.g., sparsity. E.g., Lombrozo (2007). Most satisfying explanation was overwhelmingly the simple one: i.e., a single cause rather than two independent ones. See also Liefgreen and Lagnado (2023), Stephan (2023), and Johnson et al. (2019) There are some domains, e.g., legal, where people prefer a more complex or thorough explanation. What about counterfactuals, with the same scenario? Reformulate the same explanations as both causal and counterfactuals. Do people prefer the simple one in counterfactual as well as causal? Why is simplicity preferred? E.g., as a proxy for probability – if the causes are rare, then one is more likely than two – you wouldn't predict a difference between causal and counterfactual. Alternatively, simplicity from representational elegance: easy to consider the common cause as opposed to multiple causes. Less simplicity preference for counterfactuals, because they require that you're thinking about what might have happened already. Dai, Keane and Byrne (in prep) reproduce Lombrozo (2007), but find a smaller effect for counterfactual explanations than causal explanations.

People prefer simple explanations when they diagnose causes from known effects. AI systems make diagnoses, e.g., medicine, machine faults, cybersecurity. But they also make predictions, e.g., credit risk, turnover, weather. Do people also prefer simple explanations when they predict effects from known causes? I.e. common effect rather than multiple effects. Warren, Keane, and Byrne (in prep) again find a simplicity bias for diagnostic inference and a less strong one for prediction. How do people reason about continuous and categorical variables, i.e., not just binary. Analogous scenario with diagnostic and predictive tasks. Find the same bias towards simple explanations, with a significant difference between causal and counterfactual explanations.

Do counterfactuals for AI decisions lead people to switch from risky to safe choices?

We chose example where people are well-known to predict safety versus risk, taken from Kahneman and Tversky (1982), who found that people choose the 'safe bet'. Contrast with people die rather

than people are saved – when losses are in prospect, people take the chance. People are risk-averse for gains, but risk-seeking for losses. What if an AI system recommends the atypical option, i.e., the one that people don't tend to choose? Only apply to participants who make the typical choice. People are blamed more if they make the atypical choice and things go wrong.

Dai, Keane, Shalloo, Ruelle, and Byrne in prep. What percentage of people make the switch? AI is more likely to persuade people to make a sure decision than a risky one. The counterfactual explanation increases the likelihood of persuading in both cases. Different when people tell you to do things, i.e., a medical expert rather than an AI system. The implication is that an AI recommendation needs to be explained for people to follow it.

Questions Same content in a different linguistic expression, what about linguistic/cultural background? How generalizable are these findings? Native English-speaking and resident in a predominantly English-speaking country. People tend to both causally and counterfactually explain things but generate more causal explanations, because they're easier. Does language have a big impact on counterfactual thinking? Have found similar things with other European languages. Previously, there was a view in the West that there was no way to express counterfactuals in a Chinese language. That's not true, but the linguistic markers for counterfactuals aren't used very often. Also found there didn't seem to be a difference in the inferences people made, but the expressions were more cumbersome. Interested in individualistic cultures (e.g. American) but again there doesn't seem to be a difference. But there are cultural differences. Particularly when judging others' decisions, guided by norms in society.

Is it negation or extra information that causes the difference between the explanations? Can use implicit negations, find similar results. In the 1970s, people tended to remember the opposite of the counterfactual. Anchor counterfactual imagination in reality. Does negation require symbols to be represented, or is it represented in terms of the alternative? Analogy to blood-alcohol example that's unfamiliar to people. People find explanations less helpful in an unfamiliar domain. Performance improves a bit but not much.

Conversational implicature, Bryce. Someone's telling you something, there must be a reason for that. Conjunction fallacy, people think that you're saying 'bank teller only'. Conditional reasoning literature - invited inferences. Two prominent viewpoints: possibilities and mental models versus figuring out prior beliefs with probability calculation. Think neither depends heavily on implicatures. How much can we explain with just reasoning instead of also linguistic overlay?

Framing problem with explanations. Degree to which people accept an explanation depends on framing. Implications for XAI program: if everything's explainable then it's safe. Are you trying to explain the global system, the domain it somehow encodes, or just a local explanation for a decision? Are you just trying to justify a decision? When people have knowledge, no effect for correct answer, but helps with atypical one. XAI: difficulty of prediction being made in people's understanding. Benefit of explanation disappears often when you have a material set that's easier, i.e., people can already get the answer and the explanation's redundant. How do explanations get turned off and on? People are intolerant of errors, i.e., expect decision to be correct. Trust gets undermined quite quickly.

Philosophical: counterfactual versus conserved-quantity view. Causation transfers something from cause to effect. Psychological literature suggests that in physical systems, people think more about the latter. Does that distinction explain some of the differences? How are mechanistic explanations repre-

sented? Mechanism, intentionality, other explanations. What kinds are people especially persuaded by in social domains. Have mostly looked at AI helping to make certain decisions. Especially used to assume other people's perspectives.

Educational Neuroscience: Teaching the Brain to Understand Causality

Michael Thomas

Birkbeck, University of London

Educational neuroscience; the brain's 'implementation constraints'; and an example of translation(al) research: scientific concepts and inhibitory control. What is educational neuroscience? It's part of the new science of learning, which is an interdisciplinary field. Psychology and education have a long history of interaction. But the twin developments in neuroscience (*in vivo* imaging) and machine learning (computational POV) were new. How do neuroscience and education interact? Mostly within a cognitive neuroscience framework: improving psychological theories of learning, which influence education in turn. Possibility of a direct route: e.g., brain health, the brain as a metabolic device with nutritional needs, stress response, etc. It's important to be modest as a neuroscientist: classrooms are a complicated context. We can't contribute to that much of it. How much time is dedicated to engaged learning? Constraints on learning in education – Bronfenbenner's 'Ecological Systems Theory' and Michie et al. 'Behaviour Change Wheel'. Hierarchy of constraints, e.g., individual attributes up to national policies. 50% of the variation in learning outcomes is explained by genetics, school is about 10% and home/working environment is around twice that, i.e., there's only so much educational neuroscience can do.

Why is it important to know how the brain works? There are some unintuitive things about how the mind works: memory, framing, plasticity. There are other ways to run cognitive systems and control bodies (machines). Our cognitive system works that way it does due to the brain (biology). Biology works the way it does because of evolution. Evolution only has locally optimal solutions. Thinking with neurons might not be the best way: metabolic costs, etc. The brain isn't like a digital computer (per 1980s cognitivism). The brain has 'content-specialised' devices. A computer has a general code (abstraction) and domain-general processes. The brain's knowledge is built into its structure – you can't move information back and forth between domain-general mechanisms. Modulation and reconfiguration of content-specific circuits. Hierarchies, hubs, maps, and networks.

How does the brain work? The sensory system is a hierarchy of sensory modalities. Coarse features at the bottom, higher-order invariances above. Bidirectional connectivity. E.g., sensory neurons do something like a deep convolutional neural network. Then we have an association cortex to do mappings between them. We're social, so a lot of what the brain is built for is social: expressions, motion, space, intentions, narrative, etc. And then we have a motor hierarchy. That similarly goes from immediate actions to sequences of actions, contingent actions, and plans. Might want to use perceptual information later. A bunch of what the brain does is control: task schema selection and maintenance. Valence map of emotional values to different associations. Motor smoothing: 80% of the brain's neurons! Computationally, moving a chess piece is more complicated than deciding where to move it in the first place.

Learning: one thing in the classroom, many things in the brain. Multiple mechanisms, networks, roles of different factors. Concepts also get messy from the POV of the brain. Educational domains are combinations of sensorimotor operations, concepts, and procedures. E.g., mathematics combines symbols, words, bodily actions, amounts, facts, calculations. The tricky thing in neuroscience is translation, i.e., how do you translate insights into brain implementation constraints into useful heuristics for teachers? The brain's design priorities are sensorimotor, emotion, social, and cognitive – in that order. To optimize learning, align the first three priorities with the fourth. 'Cognitive load theory' based on 1980s cognitive psychology. Static memory, got to get knowledge through the bottleneck of working memory into long-term memory. Neuroscience you want to add other factors like executive functions, changes in knowledge structures, agency, emotions, social context, adolescence, etc., which all differ between individuals.

'Why don't kids like school?' The brain continuously computes whether a task is worth the mental effort given the expectations of success and reward. Attention is metabolically expensive. Learning scientific concepts. Foundation of intuitive concepts since infancy. New conceptual repertoire, abstractions, vocabulary, notation. Transitions and transformations: temporal dimension. Cognitive control/executive function; importance of language and group-based learning; dissociations between prediction and explanation. Sometimes facts are inconsistent or unintuitive. Learning about maths and science sometimes involves inhibiting prior beliefs or direct perceptual information.

Experts get better at inhibiting pre-potent responses rather than replacing prior concepts with new ones. Similar bits of the brain are activated in misconception-type problems. The wider role of inhibitory control in reasoning. Syllogistic reasoning with erroneous premises or counterfactual conclusions involves cognitive control. E.g., *modus ponens*. Use inhibition to focus on language properties of syllogisms rather than semantics (Houdé and Borst 2015). The UnLocke project: large-scale neuroscience-based intervention study.[1] Key idea is to train children to use existing inhibitory skills better in the context of math and science. Content-specific circuits: not general skills but in the context of maths and science, so need to do in that context. Relation to dual-process theory. Detail effect as it progresses across middle childhood. Stop and Think paper. Broader evidence-based approach to education.

Educational neuroscience uses understanding of implementation constraints to facilitate classroom practices and improve learning outcomes. Classroom learning involves many brain mechanisms, educational knowledge is hybrid. Causal reasoning relates to sensorimotor biases, early developing of intuitive knowledge, the importance of language to control attention and context, cognitive control, etc. Maybe AI can do better than humans? But probably also less well-adapted to human contexts.

Questions

- 'Suspending judgment is the hallmark of morality'. Is it also ethical education?
- What about language? Complexities of language, expect some kind of neural correlates. Can't really find language in the brain as a structure, as opposed to other animals. Evolutionary POV.
- What do we want the educational system to do? Are you worried about the gaps between people rather than the mean? Seems easier to shift the whole distribution. Don't seem to know why

[1] See `https://unlocke.org/`

Finland is good. Do those techniques for example generalize well to other populations.

- Domain-general versus -specific. Psychologists that talk about innate concepts need to explain how it's encoded in DNA etc. Something's innate if it's more robust to variation in environmental conditions. Pre-natal processes. Cortex is more generally plastic and has a lot of highly-conserved processes of activity-driven self-organization.
- Meaning-making and causal thinking emerges from social interaction. Sensorimotor bias to learn better if couched in that. Layer over abstract things. Basis for more abstract processes.

To What Extent is General Intelligence Relevant to Causal Reasoning? A Developmental Study

Selma Dündar-Coecke

Quantum Brain Art

Quantinuum

King's College London

Determined the best model with EQS. Psychometric route to explore intelligence. Carroll's 1993 three-stratum model. Correlations suggest the existence of a common higher-order factor of 'general intelligence', that subsumes broad ability factors. Fluid/non-verbal intelligence and crystallized/verbal intelligence measures. Analogous tasks to physics, biology, and chemistry. Bespoke assessment. Explained variance by g-factor (general intelligence): subsumed the variances of both verbal measures but not fluid or non-verbal intelligence. Cognitive abilities are important for causal thinking but couldn't explain the whole picture. Developmental differences in the g-factor.

Causal Cognition via String Diagrams

Sean Tull

Quantinuum

Formal aspects of causal models and reasoning. Particularly how we can model them using string diagrams (category theory). *Causal Models in String Diagrams* (Lorenz and Tull 2023); *Active Inference in String Diagrams* (Tull et al. 2023). Builds on a lot of developments in applied category theory.

Introduction to string diagrams. Monoidal category, e.g., $\mathbf{Mat}_{\mathbb{R}_+}$. Depiction of a probabilistic process $M : X \to Y$. A channel when normalized for each input. Turn probability theory into pictures. Sequence (summing over) versus parallel (independent). Copy and discard (marginalize over) variables.

A causal model would usually be represented by, e.g., a Bayesian network (DAG). As diagrams. DAG with chosen observed vertices is equivalent to a network diagram with no inputs. A causal model in C is an interpretation of such a diagram as channels in C. Causal model is a diagram in a certain category. Generalize to other categories. Why this description? In the previous account, you have a

DAG and probability theory and switch between them. With the diagrams, you can do a lot with just diagrams, i.e., the same formalism. Natural way to study causal models.

Interventions – modifications of the mechanisms of the model. E.g., a do-intervention setting some variable, replacing mechanism with fixed value. Equivalent to probability distribution over variables before/after intervention. Or, e.g., policy that changes the process.

Counterfactuals. Need more than a Bayesian network: a functional causal model. Deterministic channels (functions) and hidden exogenous variables. Background conditions that lead to randomness of model. Another diagram that describes the distribution, the actual world is conditioned on what happens, plus an imaginary world with the alternative. Normalization over the rest to get an actual distribution. 'All else the same': exogenous variables fixed. Generalize to definition of counterfactuals in terms of string diagrams.

E.g., active inference – agent uses a causal generative model. Observations and future observations, states and future states, action policies, and habits. Framework in AI and cognitive science. Update process to future actions, based on observation and future preferences. Updating prior over habits? Pearl-style updating (or Bayesian). Can't do this exactly, so want to approximate via free energy.

Questions Causal and other compositional models may help interpretability in AI. More general concept. Is the epistemic status of the two parallel worlds represented in the model?

Challenges and Prospects for the Projective Consciousness Model

Grégoire Sergeant-Perthuis

LCQB, Sorbonne Université

Phenomenological aspects of consciousness, computational phenomenology. Consciousness involves a subjective perspective, characterized by viewpoint-structured organization, a sense of unity (holistic), embodiment, and an internal representation of the world in perspective from a specific standpoint. K Williford's MoC4 presentation.

Goals: implement a subjective perspective for synthetically adaptive agents; in a way compatible with axioms of consciousness (Alexander's axioms). Rudrauf, Williford, Bennequin, Friston. Mathematical model of embodied consciousness 2017; projective consciousness and phenomenal selfhood 2018; moon illusion explained by the projective consciousness model 2020.

A step towards robotics. World models, recalling the Bayesian brain hypothesis: adaptive systems have evolved to preserve their integrity, which necessitates the prediction of environmental behaviour. By forming hypotheses about the world and updating with new observations, to assist in decision-making on how to act. Friston active inference FKH06 DPS+20; Markov decision process and partially-observed. Imbuing perspective-taking in social agents. Plan in the future with respect to a prioris in environment.

Problem: associating actions and movement with perspective-taking (projective transformation). Subproblem: how to relate configuration and projective perspectives on the environment. Impose a set of axioms in the agent's frame of reference. Centre the subject after the transformation. Preserve axes of Euclidean frame associated with the agent. No points appear to be truly at infinity, only when subject directly represents a half space. Objects near to the agent have the same size in Euclidean frame.

Axioms impose a set of projective transformations. Rudrauf et al 2022a, 2022b.

Model maladaptive behaviours. Imagination and empathy: take others' perspective; simulate it. Big claims but in a restricted setting.

Group-structured world models: incorporate actions and 'perspectives' of agents within the agent's internal space. Reconstruct sensory input. A way to put your own actions in the geometry of the space. Plan the way you're moving inside of your space and how it will affect the world. Think of it as a group acting on the space.

Rudrauf et al 2023.

What changes with respect to classical perspective? POMDP don't have complete knowledge of environment via observations. Specialization of POMDP: some actions relating to the agents, over which the agent has control, affect its state space.

Euclidean vs projective case changes the behaviour of the agent to find an object. Get closer to get more evidence about where object is. Don't add any drive other than curiosity (mutual information).

POMDP is a particular form of causal model. Group-structured world models are a form of inductive bias. Equivariance to symmetries of objects that allow better prediction and action planning.

Challenges: difficult to have exact filtering and planning with theoretical guarantees. Approximate method can be simplified. Optimizing over possible group actions is also challenging, example learning by maximal likelihood. How to learn the possibility of taking different perspectives on the data?

Bayesian network: factorizing of probability distributions over variables (factor graph, Markov random field). Each variable is assumed to have a group acting on it, but the action is undetermined. Fixed DAG but not forcing probability distributions on it. Assumption about how group acts on model. Composing action of group on factorized probability distribution.

Deep learning – learn group-invariant transformations. Not trying to be invariant *to* something.

Quantum Causality in Pictures: Everyone Can Do It!

Bob Coecke

Quantinuum, University of Oxford

Introduction

Quantinuum's Oxford office has 40–50 people working on 'compositional intelligence' and quantum-inspired and -enabled systems. The principle of *compositionality* is typically attributed to Frege, but it goes back to Boole. Fregean compositionality is 'bottom-up': the meaning of the whole is determined by the meanings of its parts. We use it in a more general sense, which we haven't yet precisely defined. Compositionality encompasses different variants of causality. See, for example, Kissinger et al. (2017), Cho and Jacobs (2019), Schmid et al. (2020), and Lorenz et al. (2023). Causality isn't solely the story of *interactive relationships*, but they're important. In this talk, we'll discuss examples: humans reasoning about quantum processes, and quantum machines performing human-like cognition.

Quantum mechanics

The typical formalism of quantum mechanics today is due to von Neumann (1932). As soon as 1935, von Neumann stated that he didn't believe in Hilbert space (Rédei 1996). He explored other formalisms, which he considered as failures, despite making various advances in mathematics in the process. In 1935, Schrödinger stated that interaction (interference) was the characteristic trait of quantum mechanics. Whereas people usually start talking about quantum mechanics in terms of *measurement*, Schrödinger took a different view. Coecke's "Kindergarten Quantum Mechanics" (2006) interpreted Schrödinger's view in terms of Penrose-like diagrams, which Penrose invented to perform tensor calculations. This approach is described in *Picturing Quantum Processes* (2017), which is taught at Oxford. *Quantum in Pictures* (2023) contains the same material, but without any mathematics. There are accompanying videos on YouTube. The following slides are a rapid overview of several chapters of the book.

Wires and boxes

The combination of wires and boxes suffices to describe symmetric monoidal categories. This is an incarnation of Schrödinger's idea of composing things, i.e., 'plugging' systems into each other. 'Cup' and 'cap' boxes are special types of boxes, which we can introduce by 'bending' existing wires in the diagram. A cup is a *state*, i.e., something we can do something with. With two wires coming out of it, it's a two-state system, which is called an entangled, Bell, or EPR state in quantum mechanics. The combination of a cup and cap box is equivalent to a plain wire, which doesn't do anything. We can also move boxes around wires, which explains quantum teleportation. Creating two particles and moving them apart 'creates the wire'.

section 2: spiders are all you need. What can you do with a wire? If you connect two wires, you get another wire. A spider has n wires in and out. Again, you can combine or split them. Like connecting multi-plugs. Spiders of different colours: wires between them cancel out. Only need two colours. Quantum circuits/programs – that you stick in a quantum computer to do computations. The vertical wires are like qubits. Spiders connected together are like gates. Simplified circuit based on collapsing spiders and wires.

Measurement. Quantum is especially about the connection between classical and quantum worlds. What people always talk about: measurement about a way to go from quantum to classical world. Quantum as two wires and classical as one wire – measurement. The other way around is encoding. If you combine the two operations, nothing happens. But if the colours of the spiders are different (position and momentum), you chop out the legs. No legs between them means they're independent. E.g. repeated measurements.

Translated the formalism of QM into string diagrams. Neural about the ontology or philosophy of it. Well, the ontology is relational and process-based. Away from Democritus and towards Heraclitus.

Not just illustrational but computational tools. Is this language closer to the way we think? Lots is first-order topological?

DisCoCat

When presenting this work at a Quantum Interaction symposium, Joachim Lambek immediately pointed out that diagrams of this kind correspond to *grammar*. This formed the basis of the categorical

compositional distributional semantics framework, abbreviated as DisCoCat (Clark et al. 2008; Coecke et al. 2010). Mehrnoosh Sadrzadeh knew about the mathematics of grammar, Stephen Clark knew about distributional semantics, and Bob Coecke had developed the diagrammatic formalism for quantum foundations. The framework presents a way to combine grammar and meaning representations. This common 'quantum language' or structure is about interactive relationships – in quantum theory, the 'magic' comes from systems interacting; in language, it comes from meanings interacting. These phenomena aren't fully captured by a traditional view of causality. For example, in quantum theory, relativistic causality doesn't produce a probabilistic causal model. Likewise, it doesn't make sense in language, which doesn't represent traditional causal relations. As above, the idea of compositionality goes beyond the Fregean idea – information flows down, as well as up. In a sense, this is comparable to the transition from Wittgenstein's early to later philosophy – it moves from a reductionist, positivist approach to a context-based one.

Quantum NLP

The computational demands of quantum compositional structure increases exponentially, which makes it difficult to scale up language models. "Quantum Algorithms for Compositional Natural Language Processing" (2016) introduced a *quantum* algorithm for natural language processing, which demonstrates a quantum advantage for the example task of computing word similarity. This is performed with the `lambeq` library (Kartsaklis et al. 2021), which you can try out online via IBM quantum computers. Miranda et al. (2021) apply the same framework to music, instead of grammar. But we need to go beyond sentences to represent text more generally. *The Mathematics of Text Structure* (Coecke 2020) introduces *text circuits* as a new theory of language, 'distilling text into circuits'. We find that differences in word order, phrasing, and language disappear when we represent text in terms of circuits. We might therefore hypothesize that circuits are closer to mental representations and processes than text and speech, which are linear.

Questions

Some key advantages of traditional, predictive language models is that they learn both syntactic and semantic information, which makes them flexible in terms of being able to model phenomena like semantic change. In this formalism, you need to specify the grammatical structure of phrases, whereas phrases are not always grammatical (and grammar may evolve over time). As it stands, the formalism represents a fragment of the reality of language use. Also, they find that context-dependence 'saturates' at a certain point, i.e., you don't need the entirety of the context to capture context-dependent effects, only a subset of it. This potentially opens the way to more scalable models, by considering subsets of the text for training on classical machines, and performing inference on the whole, compositional structure on quantum machines, where the advantage is necessary.

How Does the Human Update? Via Pearl or via Jeffrey?

Bart Jacobs

Radboud University Nijmegen

About Pearl and Jeffrey

Pearl (1989) presented challenges in probabilistic logic, who was one of the founders of the field. There is a cultural difference between logic and probability theory. Embarrassingly, there is no probabilistic logic for symbolic reasoning. Update rules should form a part of such a logic. Update rules are difficult in traditional, monotonic logic, where adding more information cannot make a true statement become false. We need to switch to a likelihood-based view, i.e., one that is fuzzy instead of 'sharp'.

The outdated view of learning is that it is a process of adding information. Alternatively, predictive coding theory (Fristol et al.) supposes that the mind continually makes predictions and compares them to the actual outcomes. Then, updates occur in response to mismatches between the two (prediction errors). The underlying idea is that the brain is a Bayesian prediction and correction engine. But Jacobs's intuition is that it would be more appropriate to call the brain a Jeffreyan engine.

There are two update rules (Pearl/Bayes) and Jeffrey, which aren't well-distinguished in the literature. They both have clear formulations using channels. Unclear when to use each of them. Mathematically non-trivial but it's intriguing to wonder whether the mind uses one or the other within predictive coding theory. Cognitive science may provide an answer to this.

Papers: Bart Jacobs 2019, MFPS 2021, MFPS 2023.

E.g., a medical test has a certain sensitivity and specificity. Computing the predictive positive test probability gives an unintuitive result. What if the test is performed in unfavourable circumstances? What's the disease likelihood i.e. the posterior? Pearl and Jeffrey's rules give significantly different results! This makes people uncomfortable.

This relates to multiple test results as well as uncertain test results. Pearl's approach doesn't scale to many tests (without a conjugate prior situation). A possible interpretation for the difference: Pearl is about tests for an individual, whereas Jeffrey is about tests for a population, with different individuals.

Zooming out

These update rules as optimisations.

Pearl's rule uses evidence (predicate) to update a prior to a posterior, such that the validity (expected value) of the evidence increases. Formally: the validity of the evidence in the prediction based on the posterior is higher than in the predication based on the prior.

Jeffrey's rule uses an observed distribution/state to update form prior to posterior, such that the mismatch with the observation decreases. Formally: the KL-divergence between the observation and the prediction based on the posterior is lower than on the prior.

Jeffrey's rule reduces prediction errors, as in predictive coding.

Comparisons:

- Pearl's rule increases what's right (effect); you learn nothing from uniformity (no differences); successive updates commute.
- Jeffrey's rule decreases what's wrong; you learn nothing from what you already know (predict); successive updates don't commute.

Big question: does the human mind use one or the other? Bet is on Jeffrey but don't have any evidence for it. Maybe it's a combination or varies under different circumstances. But what are the circumstances? We're sensitive to the order in which we receive information (priming).

Underlying mathematics

Mathematics: distributions (finite and discrete). A distribution or state over a set Z is a formal finite convex sum:

$$\sum_{z \in Z} p(z)|z\rangle \text{where} \sum_{z \in Z} p(z) = 1 \text{and} p(z) \geq 0.$$

Distributions can also be described as functions to 0 1 with finite support and sum 1. Distribution monad on sets. Kleisli map is also called a channel, and written as $p \colon X \to Y$. Channels condition probabilities in a graphical calculus. Symmetric monoidal categories. Kleisli extension. Distribution on X and a distribution on Y. Also called bind, state extension.

Use sets $D = \{d, d^\top\}$ for disease or not and $T = \{p, n\}$ for positive or negative test outcomes. The prevalence state or distribution is a prior that's a combo of the two. Testing is done via the channel test. Captures sensitivity and specificity. The predicted test distribution.

Divergence between states. For $\omega, \rho \in \mathcal{D}(X)$, the KL-divergence is

$$\mathrm{KL}(\omega \mid \rho) = \sum_{x \in X} \omega(x) \log \frac{\omega(x)}{\rho(x)}.$$

It's a standard way to compare states. Basic divergence properties, not symmetric, etc.

We also need predicates and transformations. A predicate on a set Z is a function $\rho \colon Z \to \{0, 1\}$. Each subset forms a sharp predicate via the indicator function. Also have point predicates determined by an element. Given a channel and a predicate, you can define a predicate transformation. State transformation goes forward (direction of arrow) and predicate transformation goes backwards.

Validity and conditioning. Expected value: sum over all the probabilities and values of the predicate. If the expected value is non-zero, define the conditional distribution. Normalized product of w and p. Incorporating evidence into distribution.

Results about validity. Validity and transformations: for a channel $c : X \to Y$, state σ on X, and a predicate q on Y…Validity increase…By incorporating evidence, the result becomes more true. The notation is clumsy in the literature, and they don't have a way of expressing it. Hence, the new notation.

The 'dagger' of a channel: Bayesian inversion. Assume a channel and a state. For an element on the codomain of the channel: form the point predicate, its transformation as a predicate on X, and the updated state. Define this to be an inverted channel or dagger. Depends in the probabilistic case on the prior. Dagger functor.

Pearl and Jeffrey, formulated via channels (JAIR 2019). Set up a channel with a prior state on the domain. Evidence on Y that we want to update to use sigma. Pearl's update rule: evidence is a predicate on Y, update state. Jeffrey's update rule: evidence is a state. Turn around the channel and do state transformation that way.

Main optimization rules: Pearl increases validity. Jeffrey decreases divergence. The proof of Pearl is easy but for Jeffrey it's remarkably hard. Jeffrey's KL decrease is missing in the predictive-coding literature, although it forms the basis of error reduction. Is this the result that underpins it?

Conclusions

Concluding remarks: updating is magic, pillar of AI revolution, requires a proper logic. Different update rules give wildly different outcomes but aren't well distinguished in the literature. Clear formulations in terms of channels.

EM and LDA decrease divergence via Jeffrey's rule. Extensions to continuous or quantum settings are next steps. Connecting to cognition-theory community!

Jeffrey's rule underpinning things, but people don't necessarily know that.

Passive Inference is Compositional, Active Inference is Emergent

Jules Hedges

University of Strathclyde

Joint work with Toby St.Clere Smithe. Hedges's second affiliation is the CyberCat institute non-profit for research into applied category theory.

Markov kernels

A Markov kernel $\phi : X \to Y$ is a conditional probability distribution $\phi(x)$ on Y given $x \in X$: $\mathbb{P}_\phi\ [y \mid x]$.

You can compose these things together. Markov kernels compose by integrating out the middle variable. We call this composite Markov kernel 'phi then psi'. These are hard to compute exactly, because integration in high-dimensional spaces is hard. But they're easy for Monte-Carlo simulation.

A special case is that, given a distribution $\pi : 1 \to X$, the pushforward distribution $\pi\phi : 1 \to Y$ is the distribution on Y that you get by sampling from π and then applying ϕ.

Bayesian inversion

Suppose we observe a specific output y of ϕ but we did not see the input x. Need to have some prior belief, i.e., the distribution that the inputs are being drawn from. Bayes's theorem tells us how to update prior belief to get the posterior belief. Also hard to compute and also hard for MC simulation.

For a fixed prior π, this defines a Markov kernel $\phi^\dagger_\pi : Y \to X$, called the Bayesian inverse:

$$\mathbb{P}_{\phi^\dagger_\pi}\ [x \mid y] = \mathbb{P}_\pi\ [x \mid \phi(x) = y] \tag{1}$$

Hedges considers Jeffrey's rule the proper application of Bayes's theorem. The two are commonly conflated in the literature (see previous talk). This depends on π in a non-linear way. Jeffrey's rule is efficient because it's linear. This is why fixing π was useful.

The Bayesian chain rule

Can we relate the Bayesian inverse of a composite kernel to the Bayesian inverses of the single kernels?

Theorem (Smithe):

$$(\pi; \phi)^\dagger_\pi = \psi^\dagger_{\pi;\phi}; \phi^\dagger_\pi \tag{2}$$

Bayes's theorem is for 'sharp' evidence, whereas Jeffrey's and Pearl's are for 'fuzzy' evidence. This gives a posterior distribution as an output, so I'm doing a Jeffrey's update on ϕ. I consider this good evidence that Jeffrey's rule is 'right'. Maybe there's a similar single formula for Pearl's rule. There's no evidence; this is just a fact about Bayes's theorem.

This tells us that we could compute posteriors in a compositional way: it's easy if we've already solved the sub-problems. I.e., if we can already invert the sub-kernels, we can efficiently invert the whole thing. Smithe wrote a general proof that works for a subset of Markov categories. This seems to be folkloric, i.e., lots of people know it, but it doesn't appear in the literature.

Compare the reverse-mode chain rule for the transpose Jacobian of smooth functions that underlies backpropagation.

Lenses and perception

It's useful to package up a kernel $\phi : X \to Y$ and an indexed family of kernels into a single entity (process). We call this a Bayesian lens from X to Y: $(\phi, \phi') : X \to Y$. Lenses can be composed by a chain rule. The general idea is that lenses point down the perceptual hierarchy, from higher- to lower-level descriptions. ϕ is the prediction kernel: it converts a higher-level belief into a lower-level prediction. ϕ'_π is the inference (Bayesian inverse) kernel: it converts a lower-level observation into a posterior higher-level belief. Lens composition tells you how to do prediction and inference through many layers of abstraction.

Approximate inference

The previous discussion concerned *exact* inference. It's common to replace ϕ'_π with a parameterized functional form $\phi'(\theta)$ for some parameters. The goal is to learn p so that $\phi'_\pi(\theta) \to \phi^\dagger_\pi$. The general way to do this is to find a loss function that's minimized by the exact Bayesian inverse. There are several ways to do this, e.g., variational free-energy minimization.

The general shape of the algorithm is to take an input x, assumed to be distributed according to some prior; input an observation y', assumed to be distributed according to some conditional. The sample and output x' from $\mathbb{P}_{\phi^\dagger_\pi}[x \mid y']$. Then update the prior to $\mathbb{P}_{\phi^\dagger_\pi}[x' \mid y']$. Finally, given (x, y', x'), update p by your favourite formula.

Deep inference

Does the chain rule still work for approximate inference? The theorem above is for exact inference. Conjecture: if we asynchronously learn p and q then:

$$\psi^\dagger_{\pi;\phi}(q); \phi^\dagger_\pi(p) \to (\phi; \psi)^\dagger_\pi \tag{3}$$

Variational free energy fails to be compositional in the 'nicest' way.

$$\mathrm{FE}(\psi^\dagger_{\pi;\phi}(q); \phi^\dagger_\pi(p)) = \mathrm{FE}(\phi^\dagger_\pi(p)) + \mathrm{FE}(\psi^\dagger_{\pi;\phi}(q)) + \text{cross term} \tag{4}$$

It's trivially true if we do backward induction: first hold p and learn q to convergence, then learn p to convergence. Want to do these asynchronously so it can be in parallel. By analogy with deep learning,

I propose to call this 'deep inference'.

Active inference is emergent. Everything so far is holding the prediction kernel constant (stationary). You could call this *passive* inference. We think that if you change that, this stops being true. Non-stationary prediction kernels break compositionality.

Conjecture: if $\phi(p) \to \phi$, then in general we do not have… In particular, we believe that this fails for active inference. To put a positive spin on this: active inference exhibits emergent behaviour. It behaves differently to the sum of its parts. A system doing active-inference compositionally can fail to converge to true beliefs, even in a stationary environment. Each piece of it is not in a stationary environment.

The hope is that you can do this extremely efficiently. Online learning is hard. Banking on doing this in real time.

Hot take: this is an interesting model of how the cortex works. The cortex is compositional according to this formalisation, through long hierarchies.

Choosing Your Evidence: Freeing but Fraught

Deanna Kuhn

Columbia University

I haven't ventured into AI modelling, but have instead studied what we call 'causal cognition' in the wild, e.g., children, adolescents, and adults. I began by using the paradigm that's typical of much causal-inference research, i.e., here's some kind of data display, what can you conclude? More recently, I began to wonder if this is really the best paradigm for causal cognition, especially in the wild – is that what we really do? I would argue maybe not. Considering much of what we read these days, and can't escape, we think 'how can *X* say that'! Or, if you have your own favourite claim, you think about what evidence supports it. I would describe that as a more typical form of causal cognition in everyday circumstances. Therefore, I have moved towards this paradigm in my design. We've used this kind of paradigm with children, adolescents, and adults. We drew on a favourite topic: obesity. I've been impressed of the importance of causal cognition in a multivariate format, and to get away from the 'school science', univariate kind of approach. If there were ever a topic where multiple causes contribute to an outcome, obesity is one.

Tables of three variables (exercise, diet, and parent). The subjects weren't shown the tables in this format. What's your theory? What evidence would you need to show that you're right? What we actually presented to the subjects was a set of data-cards, separately, which were movable, so the subjects could make comparisons, etc. 'You want to demonstrate that (e.g.) exercise matters – what cards do you need to show'? How many cells in the four-cell table did people need to make reference to (four cards). How many students felt they needed lots of cards – only 1 out of 35, despite them being students with a good socioeconomic status and education.

Showed that the ability to reason proportionally wasn't the problem. People didn't have trouble with that. Young children just use the biggest quantity rather than the biggest fraction. But the 35 in the adolescent sample didn't have that problem. Make the case (without getting into it) – the issue is that it's a cognitive laziness kind of phenomenon, rather than a competence deficit, as with younger

children. Think you can extend that to the stance of cognition in natural settings. What's the minimum that would make my case? Make the extension that's what we see in more consequential reasoning.

Moved on to allow participants to do something that they couldn't do in the previous study, which is to look at interaction effects. That's where the real answer lies.

Two take-home messages: what happens when you give people a richer database that allows people to reason about multiple causes? Elicited explanations and asked people with the same design what evidence they'd need. What they did was common-sensical, i.e., look at the output variable. Controlled comparison is perhaps not the answer for science education. Left explanation out of the picture – variable at the bottom. Proportion of belief revision increased with pre- and post-data explanation, less so with post-data explanation, almost none with no explanation. Explanation is a double-edged sword: some is helpful, too much has adverse effects. Story can help people to attend to the data more.

Bibliography

Abramsky, Samson (2013). "Relational Databases and Bell's Theorem". In: *In Search of Elegance in the Theory and Practice of Computation: Essays Dedicated to Peter Buneman*. Ed. by Val Tannen et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 13–35.

Allen, Micah and Karl J. Friston (2018). "From cognitivism to autopoiesis: towards a computational framework for the embodied mind". In: *Synthese* 195.6, pp. 2459–2482.

Byrne, Ruth M.J. (2016). "Counterfactual Thought". In: *Annual Review of Psychology* 67.1, pp. 135–157.

Celar, Lenart and Ruth M. J. Byrne (2023). "How people reason with counterfactual and causal explanations for Artificial Intelligence decisions in familiar and unfamiliar domains". In: *Memory & Cognition* 51.7, pp. 1481–1496.

Cesa, Gabriele, Leon Lang, and Maurice Weiler (2022). "A Program to Build E(n)-Equivariant Steerable CNNs". In:

Cho, Kenta and Bart Jacobs (2019). "Disintegration and Bayesian Inversion via String Diagrams". In: *Mathematical Structures in Computer Science* 29.7, pp. 938–971.

Claassen, Jan et al. (2016). "Causal Structure of Brain Physiology after Brain Injury from Subarachnoid Hemorrhage". In: *PLOS ONE* 11.4, e0149878.

Clark, Stephen, Bob Coecke, and Mehrnoosh Sadrzadeh (2008). "A Compositional Distributional Model of Meaning". In: *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*. Ed. by Peter D. Bruza et al. Oxford, UK: College Publications, pp. 133–140.

Coecke, Bob (2006). "Kindergarten Quantum Mechanics: Lecture Notes". In: *AIP Conference Proceedings*. Vol. 810. Vaxjo (Sweden): AIP, pp. 81–98.

— (2020). *The Mathematics of Text Structure*. URL: http://arxiv.org/abs/1904.03478.

Coecke, Bob and Stefano Gogioso (2023). *Quantum in Pictures: A New Way to Understand the Quantum World*. Cambridge Quantum.

Coecke, Bob and Aleks Kissinger (2017). *Picturing Quantum Processes: A First Course in Quantum Theory and Diagrammatic Reasoning*. Cambridge: Cambridge University Press.

Coecke, Bob, Mehrnoosh Sadrzadeh, and Stephen Clark (2010). "Mathematical Foundations for a Compositional Distributional Model of Meaning". In: *Linguistic Analysis* 36.1–4, pp. 345–384.

Cofield, Stacey S., Rachel V. Corona, and David B. Allison (2010). "Use of Causal Language in Observational Studies of Obesity and Nutrition". In: *Obesity Facts* 3.6, pp. 353–356.

Cohen, Michael, Marcus Hutter, and Michael Osborne (2022). "Advanced Artificial Agents Intervene in the Provision of Reward". In: *AI Magazine* 43.3, pp. 282–293.

Dai, Jessica et al. (2022). "Fairness via Explanation Quality: Evaluating Disparities in the Quality of Post hoc Explanations". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '22. New York, NY, USA: Association for Computing Machinery, pp. 203–214.

Deleu, Tristan et al. (2022). "Bayesian structure learning with generative flow networks". In: *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*. PMLR, pp. 518–528. URL: https://proceedings.mlr.press/v180/deleu22a.html.

Deleu, Tristan et al. (2023). *Joint Bayesian Inference of Graphical Structure and Parameters with a Single Generative Flow Network*.

Fernbach, Philip M., L. Min, and S. A. Sloman (2018). *Values-based and consequence-based policy attitudes*. Tech. rep. Working paper.

Fernbach, Philip M. et al. (2013). "Political Extremism Is Supported by an Illusion of Understanding". In: *Psychological Science* 24.6, pp. 939–946.

Han, Mi Ah et al. (2022). "Causal language use in systematic reviews of observational studies is often inconsistent with intent: a systematic survey". In: *Journal of Clinical Epidemiology* 148, pp. 65–73.

Houdé, Olivier and Grégoire Borst (2015). "Evidence for an inhibitory-control theory of the reasoning brain". In: *Frontiers in Human Neuroscience* 9. URL: https://www.frontiersin.org/articles/10.3389/fnhum.2015.00148.

Johnson, Samuel G. B., J. J. Valenti, and Frank C. Keil (2019). "Simplicity and complexity preferences in causal explanation: An opponent heuristic account". In: *Cognitive Psychology* 113, p. 101222.

Kahneman, Daniel and Amos Tversky (1972). "Subjective probability: A judgment of representativeness". In: *Cognitive Psychology* 3.3, pp. 430–454.

— (1982). "The Psychology of Preferences". In: *Scientific American* 246.1, pp. 160–173. URL: https://www.jstor.org/stable/24966506.

Karimi, Amir-Hossein et al. (2020). "Model-Agnostic Counterfactual Explanations for Consequential Decisions". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 895–905. URL: https://proceedings.mlr.press/v108/karimi20a.html.

Kartsaklis, Dimitri et al. (2021). *lambeq: An Efficient High-Level Python Library for Quantum NLP*. URL: http://arxiv.org/abs/2110.04236.

Keane, Mark T. et al. (2021). *If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques*.

Kissinger, Aleks, Matty Hoban, and Bob Coecke (2017). *Equivalence of relativistic causal structure and process terminality*.

Kleinberg, Samantha and Jessecae K. Marsh (2020). "Tell me something I don't know: How perceived knowledge influences the use of information during decision making". In: *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*. URL: https://par.nsf.gov/biblio/10178793-tell-me-something-don-know-how-perceived-knowledge-influences-use-information-during-decision-making.

— (2021). "It's complicated: Improving decisions on causally complex topics". In: *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. Ed. by T. Fitch. et al. URL: https://par.nsf.gov/biblio/10279840-complicated-improving-decisions-causally-complex-topics.

Kleinegesse, Steven, Andrew R. Lawrence, and Hana Chockler (2022). *Domain Knowledge in A*-Based Causal Discovery*.

Liefgreen, Alice and David A. Lagnado (2023). "Drawing conclusions: Representing and evaluating competing explanations". In: *Cognition* 234, p. 105382.

Lombrozo, Tania (2007). "Simplicity and probability in causal explanation". In: *Cognitive Psychology* 55.3, pp. 232–257.

Lorenz, Robin and Sean Tull (2023). *Causal Models in String Diagrams*.

Lorenz, Robin et al. (2023). "QNLP in Practice: Running Compositional Models of Meaning on a Quantum Computer". In: *Journal of Artificial Intelligence Research* 76, pp. 1305–1342.

Mandel, David R. and Darrin R. Lehman (1996). "Counterfactual thinking and ascriptions of cause and preventability". In: *Journal of Personality and Social Psychology* 71.3, pp. 450–463.

McEleney, Alice and Ruth M. J. Byrne (2006). "Spontaneous counterfactual thoughts and causal explanations". In: *Thinking & Reasoning* 12.2, pp. 235–255.

Miranda, Eduardo Reck et al. (2021). *A Quantum Natural Language Processing Approach to Musical Intelligence*. URL: http://arxiv.org/abs/2111.06741.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press. URL: https://library.oapen.org/handle/20.500.12657/26040.

Raimondi, Francesca E. D., Andrew R. Lawrence, and Hana Chockler (2022a). *Equality of Effort via Algorithmic Recourse*.

Raimondi, Francesca E. D. et al. (2022b). *Causal Analysis of the TOPCAT Trial: Spironolactone for Preserved Cardiac Function Heart Failure*.

Rédei, Miklos (1996). "Why John von Neumann did not Like the Hilbert Space formalism of quantum mechanics (and what he liked instead)". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 27.4, pp. 493–510.

Roese, Neal J. and Kai Epstude (2017). "Chapter One - The Functional Theory of Counterfactual Thinking: New Evidence, New Challenges, New Insights". In: *Advances in Experimental Social Psychology*. Ed. by James M. Olson. Vol. 56. Academic Press, pp. 1–79.

Rozenblit, Leonid and Frank Keil (2002). "The misunderstood limits of folk science: an illusion of explanatory depth". In: *Cognitive Science* 26.5, pp. 521–562.

Schmid, David et al. (2020). *A structure theorem for generalized-noncontextual ontological models*.

Shannon, C. E. (1948). "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3, pp. 379–423.

Stephan, Simon (2023). "Revisiting the narrow latent scope bias in explanatory reasoning". In: *Cognition* 241, p. 105630.

Tull, Sean, Johannes Kleiner, and Toby St Clere Smithe (2023). *Active Inference in String Diagrams: A Categorical Account of Predictive Processing and Free Energy*. URL: http://arxiv.org/abs/2308.00861.

Veloz, Tomas, Simon Hegele, and Pedro Maldonado (2023). "A Markovian framework to study the evolution of complexity and resilience in chemical organizations". In: *ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference*. MIT Press.

Veloz, Tomas and Olha Sobetska (2023). "Analysing the Conjunction Fallacy as a Fact". In: *Trends and Challenges in Cognitive Modeling: An Interdisciplinary Approach Towards Thinking, Memory, and Decision-Making Simulations*. Ed. by Tomas Veloz et al. STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health. Cham: Springer International Publishing, pp. 101–111.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell (2017/2018). "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR". In: *Harvard Journal of Law & Technology (Harvard JOLT)* 31.2, pp. 841–888. URL: https://heinonline.org/HOL/P?h=hein.journals/hjlt31&i=859.

Warren, Greta, Ruth M. J. Byrne, and Mark T. Keane (2023). "Categorical and Continuous Features in Counterfactual Explanations of AI Systems". In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*. IUI '23. New York, NY, USA: Association for Computing Machinery, pp. 171–187.

Weiler, Maurice and Gabriele Cesa (2019). "General E(2)-Equivariant Steerable CNNs". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/hash/45d6637b718d0f24a237069fe41b0db4-Abstract.html.

Zeng, William and Bob Coecke (2016). "Quantum Algorithms for Compositional Natural Language Processing". In: *Electronic Proceedings in Theoretical Computer Science* 221, pp. 67–75.

Zheng, Min et al. (2020). "How causal information affects decisions". In: *Cognitive Research: Principles and Implications* 5.1, p. 6.