Argumentation for Interactive Causal Discovery

Fabrizio Russo
Imperial College London
fabrizio@imperial.ac.uk

Abstract

Causal reasoning reflects how humans perceive events in the world and establish relationships among them, identifying some as causes and others as effects. Causal discovery is about agreeing on these relationships and drawing them as a causal graph. Argumentation is the way humans reason systematically about an idea: the medium we use to exchange opinions, to get to know and trust each other and possibly agree on controversial matters. Developing AI which can argue with humans about causality would allow us to understand and validate the analysis of the AI and would allow the AI to bring evidence for or against humans' prior knowledge. This is the goal of this project: to develop a novel scientific paradigm of interactive causal discovery and train AI to recognise causes and effects by debating, with humans, the results of different statistical methods.

1 Motivation

Statistical models have been used for decades to aid decision-making. The advent of Machine Learning (ML) has often improved traditional regression and classification models when judged by accuracy metrics. However, statistical methods were always assessed by a statistician who would judge if the results made sense, e.g. the direction of the weights in a regression. ML models are instead often black-boxes that do not allow the modeller to understand the relationships that the model is leveraging to predict the target outcome. This poses a problem particularly when ML is deployed for high-stakes decisions, like granting credit or parole to individuals. It has thus been advocated to simply stop using black-box models for high-stakes decisions [Rudin, 2019]. We would add that, for high-stakes decisions affecting people lives, we should only consider stable relationships [Pearl, 2009]: those that describe the problem's causes and effects and can be acted upon with the expectation of changing an outcome.

Causal models, following Pearl's paradigm [Pearl, 2009], consist of two main components: a causal graph and a set of structural equations (SEM). The terms in the latter respect the relationships specified in the former. Causal Discovery (CD, see [Glymour *et al.*, 2019] for an overview) aims at uncovering the causal graph, the structure of the relationships among the variables underpinning a phenomenon and its Data Generating Process. Agreeing on a causal graph is key in order to perform causal inference: the graph represents the causal assumptions needed to build a SEM. The SEM will in turn describe the links in the graph through equations that represent the effects of actions, like granting parole, on future behaviour, e.g. committing further crimes. When dealing with causal models, disagreements can arise from multiple sources: incomplete or multiple datasets, different focuses towards the problem (and therefore variables considered) or different judgments. Previous work aimed at resolving conflicts within either causal graphs or models (see [Alrajeh *et al.*, 2020] for work towards the latter setting, and references on the former), but we employ computational argumentation to resolve these conflicts, by debating inconsistencies in both data and experts' views.

Argumentation (see [Atkinson *et al.*, 2017] for an overview) provides a very flexible framework that allows for both transparency and soundness of modelling in domains where there is conflicting information. Properties of argumentation have been studied extensively [Baroni *et al.*, 2018] and its suitability to support eXplainable AI (XAI) has been advocated by many (see [Cyras *et al.*, 2021] for a recent survey). Strong theoretical foundations, together with the flexibility of argumentation frameworks (AFs) to represent any information as arguments and dialectical relations, make it an ideal candidate to support decision making. This is particularly true, in our view, when dealing with inconsistencies that arise from data but can and should be complemented by causal knowledge.

Our work aims at creating algorithms that involve both humans and machines in a debate about causality.

2 Contributions

We made the following three contributions at the intersection of argumentation, causal discovery and XAI:

1. We formulated a method to extract argumentative explanations from causal models [Rago *et al.*, 2021].
2. We demonstrated how the explanations from [Rago *et al.*, 2021] can be used to get insights into ML models [Rago *et al.*, 2023].
3. We devised a method to *inject* a causal graph into a feed-forward neural network [Russo and Toni, 2022].

In [Rago *et al.*, 2021] we reinterpret properties of AFs and invert them to extract *explanation moulds*: templates that satisfy these desirable properties for explanation. We demonstrate our methodology using the property of bi-variate reinforcement in bipolar AFs and show how the resulting *reinforcement explanations* (RXs) can be used to give insights into some variable of interest, given a causal model describing its behaviour as a function of its parents.

We extend this work by evaluating RXs empirically when the causal model represents either a Bayesian or Neural Network. In [Rago *et al.*, 2023] we show the advantages of imposing specific properties onto explanations: RXs manage to expose the interactions between type of model and training data in the selection of the greatest contributors to the model's output. Hence, we use argumentation to draw insights from ML models and causal representations thereof, but the latter are abstractions of the former's workings. What if, instead, the ML model already followed a causal representation?

To this end, we propose two algorithms in [Russo and Toni, 2022]. The first *injection* algorithm allows to inject a causal graph into a feed-forward neural network so that the latter is guaranteed to use only the relationships in the former. Given the scarcity of causal graphs for real world applications, we propose a second algorithm for *human-AI collaboration* ont the causal discovery task with NN. The algorithms use CASTLE's architecture and CD methodology [Kyono *et al.*, 2020] to extract an initial graph. Once this is computed, a *contesting* algorithm is proposed, whereby subject matter experts (SMEs) discuss the results and request changes to the graph based on their prior knowledge. The refined causal graph is then fed back into the NN using the injection algorithm.

3 Ongoing Work

In [Russo and Toni, 2022] we proposed leveraging SMEs knowledge to validate the results of the CD algorithm underpinning [Kyono *et al.*, 2020]. Since human input is necessary when considering causality, our next steps aim at improving on existing CD benchmarks in terms of both accuracy and accessibility using interactive argumentation and visualisation.

Argumentative PC algorithm. Argumentation has been proposed to address sampling issues [Bromberg and Margaritis, 2009] in the independence tests underpinning the PC algorithm [Spirtes *et al.*, 2000]. Effectively, using AFs to make statistical methods more robust to incomplete data. We are extending [Bromberg and Margaritis, 2009] in three fundamental ways: we use Assumption Based Argumentation with preferences [Bao *et al.*, 2017] to incorporate rules directly in the AF; we use Bayesian posteriors and p-value corrections, rather than heuristics, for preferences; and we extend the debate to all three phases of PC (instead of only the first). With this work we use argumentation to resolve conflicts arising not only from the statistical tests, but from their interaction with graphical rules of causality within the PC algorithm.

Argumentation Semantics for Causal Discovery. Further work will involve the assignment of *dialectical strengths* to the arguments mapped from the CD algorithm. These are calculated using gradual semantics, e.g. [Jedwabny *et al.*, 2020], which allows for a more granular assessment of arguments. With this project we aim at developing a semantics that works in a probabilistic setting, allowing the assignment of actual probabilities to discovered edges, to guide the interactions with humans as well as the design of experiments.

Argumentative Causal Discovery. Finally, we will extend our AF to include several CD algorithms and augment it with SMEs' knowledge. This would all into the hybrid methods of CD [Glymour *et al.*, 2019], but we would mediate the results from the different methods with argumentation theory, while seamlessly allowing for humans to take part in the debate.

References

[Alrajeh *et al.*, 2020] Dalal Alrajeh, Hana Chockler, and Joseph Y Halpern. Combining experts' causal judgments. *Artificial Intelligence*, 288:103355, 2020.

[Atkinson *et al.*, 2017] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata. Towards artificial argumentation. *AI Mag.*, 38(3):25–36, 2017.

[Bao *et al.*, 2017] Ziyi Bao, Kristijonas Cyras, and Francesca Toni. Abaplus: Attack reversal in abstract and structured argumentation with preferences. In *Proc. PRIMA*, 2017.

[Baroni *et al.*, 2018] Pietro Baroni, Antonio Rago, and Francesca Toni. How many properties do we need for gradual argumentation? In *Proc. AAAI*, 2018.

[Bromberg and Margaritis, 2009] Facundo Bromberg and Dimitris Margaritis. Improving the reliability of causal discovery from small data sets using argumentation. *J. Mach. Learn. Res.*, 10:301–340, 2009.

[Cyras *et al.*, 2021] Kristijonas Cyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative XAI: A survey. In *Proc. IJCAI*, 2021.

[Glymour *et al.*, 2019] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, page 524, 2019.

[Jedwabny *et al.*, 2020] Martin Jedwabny, Madalina Croitoru, and Pierre Bisquert. Gradual semantics for logic-based bipolar graphs using t-(co) norms. In *Proc. ECAI*, 2020.

[Kyono *et al.*, 2020] Trent Kyono, Yao Zhang, and Mihaela van der Schaar. CASTLE: regularization via auxiliary causal graph discovery. In *Proc. NeurIPS*, 2020.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Rago *et al.*, 2021] Antonio Rago, Fabrizio Russo, Emanuele Albini, Pietro Baroni, and Francesca Toni. Forging argumentative explanations from causal models. In *Proc. AIxIA*, 2021.

[Rago *et al.*, 2023] Antonio Rago, Fabrizio Russo, Emanuele Albini, Pietro Baroni, and Francesca Toni. Explaining classifiers' output with causal models and argumentation. *IfCoLog Journal of Logics and their Applications*, 2023.

[Rudin, 2019] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019.

[Russo and Toni, 2022] Fabrizio Russo and Francesca Toni. Causal discovery and injection for feed-forward neural networks. *CoRR*, abs/2205.09787, 2022.

[Spirtes *et al.*, 2000] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search, Second Edition*. Adaptive computation and machine learning. MIT Press, 2000.