Improving Causal Inference by Increasing Model Expressiveness

David D. Jensen
College of Information and Computer Sciences
University of Massachusetts Amherst
140 Governors Drive
Amherst, Massachusetts 01003-9264
jensen@cs.umass.edu

Abstract

The ability to learn and reason with causal knowledge is a key aspect of intelligent behavior. In contrast to mere statistical association, knowledge of causation enables reasoning about the effects of actions. Causal reasoning is vital for autonomous agents and for a range of applications in science, medicine, business, and government. However, current methods for causal inference are hobbled because they use relatively inexpressive models. Surprisingly, current causal models eschew nearly every major representational innovation common in a range of other fields both inside and outside of computer science, including representation of objects, relationships, time, space, and hierarchy. Even more surprisingly, a range of recent research provides strong evidence that more expressive representations make possible causal inferences that are otherwise impossible and remove key biases that would otherwise afflict more naive inferences. New research on causal inference should target increases in expressiveness to improve accuracy and effectiveness.

Introduction

Learning and reasoning about the effects of actions—often referred to as *causal inference*—is a central activity of intelligent agents. Causal inference is used in AI to enable learning and reasoning of intelligent agents in applications such as robotics, planning, and game playing. Causal inference is a fundamental tool of essentially all sciences, particularly those that focus on devising effective interventions, such as medicine, public health, sociology, and economics. It is also a direct subject of study in psychology, cognitive science, statistics, and philosophy. Indeed, it is difficult to think of a field of human endeavor in which causal inference is not central.

Given its importance, it is surprising how starkly the goals of causal inference diverge from the majority of work in machine learning. Typical work in machine learning constructs models of statistical associations. Such models allow accurate inferences about unobserved variables (e.g., a class label) given that new data instances are drawn from the same distribution as the instances in the training data. In contrast, causal inference aims to learn a model that allows accurate inferences about the effects of explicit *changes* to the process that generated the training data. These changes, often termed *interventions*, alter the data generating process in ways that can be accurately represented as edits to a valid causal model. That is, valid causal models allow explicit representation of interventions and enable accurate inferences about the effects of those interventions.

Despite the prevailing non-causal focus of machine learning, a broad array of methods has been developed both inside and outside of AI that facilitate causal inference from both experimental and observational data. One class of methods—often labeled the potential outcomes framework (POF)—has largely been developed within statistics and the social sciences and typically focuses on estimating the average effect of a possible cause on a possible effect (Rubin 2005; Imbens and Rubin 2015). Another class of methods—sometimes referred to as structural causal models (SCMs)—has largely been developed within AI and typically focuses on reasoning about the causal dependencies among a set of variables and supports a diverse set of inferences (Pearl 2009; Spirtes, Glymour, and Scheines 2000). A rich body of work has developed around these and other paradigms, including technical contributions from fields such as philosophy of science, psychology, epidemiology, sociology, and economics.

Over the past several decades, researchers in causal inference have developed a range of innovations, including methods for deriving observable statistical consequences from classes of causal models (e.g., d-separation), a variety of designs and design elements (e.g., instrumental variable designs (Angrist, Imbens, and Rubin 1996), regression discontinuity designs (Hahn, Todd, and Van der Klaauw 2001)), practical estimation methods (e.g., G-computation, propensity score matching (Rosenbaum and Rubin 1983), doubly robust methods (Bang and Robins 2005)), algorithms for learning model structure (e.g., PC, FCI (Spirtes, Glymour, and Scheines 2000), GES (Chickering 2002), MMHC (Tsamardinos, Brown, and Aliferis 2006), and many others), and methods for inferring identifiability and adjustment sets based on partially specified models (e.g., (Pearl 2009)).

Despite these innovations, accurate causal inference remains extremely challenging. Most causal inferences require overcoming inherent challenges such as non-random treatment assignment, latent confounding, and endogenous se-

lection bias. In practice, conflicting analyses and errors are common because different analyses rely on assumptions that are difficult or impossible to empirically test. Innovations in causal inference that are particularly valuable are those that allow practitioners to overcome or check these assumptions.

In this brief paper, I argue that the accuracy and effectiveness of causal inference can be dramatically improved by addressing long-standing deficiencies in the representation of causal models. Specifically, existing causal models omit a range of representational innovations that are common both inside and outside of computer science, including representation of objects, relationships, time, space, and hierarchy. While traditional ML has long found ways of reducing more complex representations to simple variable-based representations (i.e., "feature vectors"), causal models make higher demands on the representational fidelity of learned models. These higher demands are due to the central task of causal models—valid reasoning under intervention. Specifically, effective causal models accurately represent the modular structure present in the real-world systems that they model. That is, interventions should be representable as changes to only a small number of model components and those changes should not affect the causal mechanisms in the remainder of the model.

In addition, a set of recent research results indicate that using more expressive representations can directly improve causal inference. First, more expressive representations can improve structure learning—the ability to accurately infer the causal structure among candidate model components. Second, more expressive representations can be used to reduce biases that afflict analyses that use simpler models. These are described in more detail below.

Example

To help make clear the potential benefits of more expressive representations for causal models, consider the everyday task of reasoning about the interactions among physical objects. People employ this form of reasoning in an array of everyday tasks, such as stacking dishes, climbing stairs, setting down coffee cups, and reorganizing closets. This form of reasoning also appears in some video games. For example, consider the popular game of Angry Birds. In each level of the game, players use a slingshot to fire projectiles (birds) at structures consisting of stacked blocks of various shapes and sizes (see figure 1 for an example level). The goal in each level is to collapse the structures so that they crush or otherwise destroy target objects (green pigs). Players receive points for each target and block that is destroyed and for each unused projectile.

A video game may seem a peculiar choice for an example, but we reference Angry Birds for several reasons. First, even though it is not a real physical system, it exemplifies an idealized form of reasoning about physical causation. Second, it is a system in which complex causal reasoning and inference is clearly possible. Human players can reason quite successfully about how the structures will respond to different projectile impacts, despite substantial complexity of the simulated structures and only approximate knowledge of the key parameters of those structures (e.g., the simulated mass

Figure 1: A sample level of Angry Birds. The slingshot at left is used to shoot projectiles (birds) at the structures at right with the goal of destroying the targets (green pigs).

of objects). Human players also rapidly improve their causal knowledge when new birds, blocks, or other entities are introduced into game levels. Third, it is an environment that enables repeatable experiments, and thus it is possible to derive ground-truth causal effects. Such environments are difficult to find in practice, which is one reason that empirical evaluation of causal inference methods has proven so difficult (Dorie et al. 2019; Gentzel, Garant, and Jensen 2019). Finally, Angry Birds has spawned a long-running competition at AAAI and IJCAI in which autonomous agents compete in their ability to play the game (Renz et al. 2016).

Despite the ability of humans to play Angry Birds effectively, the game differs quite substantially from the systems that can be successfully modeled by nearly all current methods for causal inference. Accurately modeling the causal structure of even a single level of Angry Birds using either the potential outcome framework or structural causal models would be difficult or impossible, and learning the structure and parameters of that model from limited data would be even more challenging. Why this gap?

Modularity: A Representational Imperative

One answer can be found in one of the most foundational and yet elusive concepts in causal inference: *Modularity*. Modularity posits that a causal system consists of components whose underlying mechanisms are invariant to the mechanisms of other components (Pearl 2009). Thus, intervening in the mechanism of one component will not affect the *manner* in which other components respond, even though that intervention may indirectly change the state of those other components. Modularity is also referred to as *autonomy* (Aldrich 1989), *causal invariance* (Peters, Janzing, and Schölkopf 2017), and *independent causal mechanisms* (Schölkopf 2019). For a given set of interventions, valid causal models have a structure that correctly reflects the modularity of the causal system they are intended to represent.

For example, consider the case of objects in Angry Birds. Most human observers would correctly assume that inter-

vening to directly change an attribute of a given block (e.g., its mass, position, or velocity) would not affect the *causal mechanisms* of another block (e.g., how that block responds to a given force). While the intervention may indirectly affect the position or velocity of another block (because of some causal chain of interactions), it will not affect the fundamental *manner* in which that other block responds to external forces.

Conversely, imagine a causal model that implicitly assumed a specific gravitational constant in its definition of the mechanisms by which blocks interacted. If an intervention altered the gravitation field, then the causal mechanisms of every block in the model would need to change for the model to accurately infer the effect of this intervention. This is one reason why physicists use *mass* rather than *weight* as a fundamental attribute for describing objects. Mass is "modular" with respect to gravitation.

Causal models that exhibit modularity have several advantages:

- *Intervention* — Interventions can be represented as minimal changes to a model rather than requiring wholesale redefinition. For example, an intervention on the mass of a block may have implications for the position or velocity of other blocks, but does not require redefining the mechanism of those interactions.
- *Composition* — Components of the model can be recombined in new ways without needing to redefine the entire model. For example, blocks can be assembled into a completely new physical structure and yet they still constitute a valid model.
- *Novelty* — Entirely new components can be introduced into a model without requiring the redefinition of existing components. For example, an iron block can be introduced into the game without requiring redefinition of wood and stone blocks.

Modularity has long been identified as a key property of SCMs. Specifically, SCMs consist of a set of conditional probability distributions (CPDs) that fully define the model. It is a formal property of an SCM that each CPD represents an independent causal mechanism—intervening on one CPD does not alter another CPD. In an accurate SCM, the CPDs are defined such that they match the modularity of the causal system that they model. However, the extent to which this correspondence holds is entirely a matter of the efficacy of the learning algorithm, the extent to which an appropriate set of variables has been defined, and the extent to which the SCM formalism can accurately represent the modularity of the system being modeled. This last issue is our focus here: Can existing modeling formalisms accurately represent the modularity necessary to reason about the effects of interventions in a wide variety of causal systems?

Expressiveness of Current Causal Models

Unfortunately, both the potential outcomes framework and structural causal models use an extremely limited model representation. By far the most common model representation in these two frameworks is *propositional*, in which data instances correspond to flat feature vectors. Vectors represents the attributes of a given type of entity, often referred to as a *unit of analysis* (e.g., person, organization, purchase). Typically, each instance of a feature vector is assumed to be marginally independent of every other instance. It is worth nothing that data analysts employing these propositional representations sometimes use units of analysis that consist of multiple entities (e.g., married couples or a student and their school) or features that exist over time and space (e.g., temporally separated treatments and outcomes), but this complexity is almost never explicitly represented within the model representation itself. For example, structural causal models are sometimes represented with nesting plates (representing hierarchical structures of objects), but such plates are merely a convenient device for producing a "ground graph" that is analyzed as if the original object structure did not exist.

Despite this, researchers who work primarily with structural causal models have long argued that a key capability of this model class is its ability to explicitly represent complex patterns of causal dependence and allow automated reasoning about that dependence. Indeed, knowledge representation and reasoning are some of the most important intellectual contributions that AI can bring to causal inference. Yet many of the key representational innovations of researchers inside and outside of AI are not yet used in causal modeling.

These representational innovations include:

- *Objects* — Multiple types of entities with associated attributes (e.g., birds, blocks, pigs, and platforms);
- *Relations* — Discrete relationships that relate two or more objects to each other (e.g., loaded-in-slingshot, supported-by), perhaps with associated attributes;
- *Time* — Changing sequences of attribute values and durations of existence of objects and relations (e.g., the sequence of events produced by a single shot);
- *Space* — Embeddings of objects within spatial fields to indicate smoothly varying attributes such as relative position, gravitation, explosive force, etc.; and
- *Hierarchy and Composition* — Combinations of all of the above that can be represented as objects, relations, etc. unto themselves (e.g., multi-object platforms).

In eschewing these innovations, causal inference lags behind a surprising array of other fields that argue implicitly and explicitly for the utility of these representational innovations. For example, researchers in psychology and cognitive science argue for the importance of the object concept, temporal reasoning, and spatial reasoning as key stages in infant development. Similarly, philosophers have long argued for the fundamental importance of objects, relationships, and time in human reasoning, and philosophers of science have more recently explored the importance of reasoning about causal mechanisms in terms of how entities interact over time and space (Craver and Darden 2013; Illari and Williamson 2012; Machamer, Darden, and Craver 2000). A large number of practical methods for causal inference in social science reference objects, relationships, and time, including multi-level models (Gelman 2006; Goldstein

2010), within-subject designs (Greenwald 1976), interrupted time-series designs (McDowall et al. 1980), and difference-in-difference designs. Researchers in sub-fields of artificial intelligence, of course, have long argued for the utility of these representational innovations, including researchers in knowledge representation and reasoning, statistical relational learning (Getoor and Taskar 2007), object-oriented MDPs in reinforcement learning (Diuk, Cohen, and Littman 2008), and first-order and higher-order logics. Finally, other areas of computer science have long used these innovations, including work in type theory and object-oriented methods in programming languages and relational, temporal, and spatial models in databases.

Advantages of Expressive Representations

In addition to these general arguments for pursuing more expressive representations for causal models, there is a growing set of research results that demonstrate how increasing the expressiveness of causal models can *enable structure learning* and *decrease bias*. Structure learning concerns the specification of which components of a model directly cause other components. In the context of SCMs, methods for structure learning determine which variables directly cause which other variables, thus defining the form of the conditional probability distributions of the model.

Perhaps the best known structure learning results are those regarding small numbers of variables within the SCM paradigm. Given data about only two random variables X and Y, the direction of the causal dependence between them cannot be determined from observational data under standard assumptions.[1] However, given a third variable Z, much more information about the likely causal structure can be learned from observational data (Spirtes, Glymour, and Scheines 2000). As another example, practitioners of causal inference have long exploited temporal dependence to constrain possible causal dependencies by assuming that an event that occurs before a second event can only be a cause, but not an effect, of the second event.

More recently, researchers have discovered new ways in which structure learning can be enabled by using more expressive model classes than those typically used in POF and SCMs. Explicit representation of multiple types of objects and their relationships can allow the direction of causal dependence to be inferred whereas it would not be inferrable with only a flat, non-relational representation (Arbour, Marazopoulou, and Jensen 2016; Maier 2014). Hierarchical structure among entities (e.g., a company that employs multiple employees) can also be used to enable structure learning despite the existence of latent confounders (Jensen, Burroni, and Rattigan 2019; Witty et al. 2020). In both of these cases, variables alone are insufficient to make these causal dependencies discoverable.

In other work, researchers have discovered ways in which highly expressive representations can reduce the bias of estimates regarding specific causal dependencies. Said differently, *without* these more expressive representations, the estimated causal effects would be irretrievably biased. In one analysis (Maier, Marazopoulou, and Jensen 2014), researchers showed that large numbers of causal dependencies would be judged to be present (when in fact they were absent) unless the object-relational structure of a given data set was considered and the reasoning strategies (d-separation judgments) were adapted to this more expressive representation. In very recent work (Lee and Ogburn 2020), researchers showed that when object-relational structures existed, but were not represented and the effects accounted for, various standard estimators of causal effect could be strongly biased.

[1] Some relatively recent work has shown that, under certain very specific circumstances, the conditional distributions of the alternative models provide evidence about the likely direction of dependence (Peters, Janzing, and Schölkopf 2017).

Conclusions

Researchers have long sought to improve the accuracy and utility of methods for causal inference. However, nearly all of the research directions explored over the past several decades have assumed that conventional representations, based on variables alone, are sufficient to enable improvements. Recent results of several different studies now imply that more expressive representations of causal models can enable important new methods to improve identifiability and reduce bias. More work on this crucial direction is needed.

Acknowledgments

Thanks to the anonymous reviewers for their helpful feedback and suggestions. The author's work is supported by DARPA, the United States Air Force (USAF), and the Army Research Office (ARO) under the XAI (Contract No. HR001120C0031), CAML (Contract No. FA8750-17-C-0120), and SAIL-ON (Cooperative Agreement Number W911NF-20-2-0005) programs, respectively. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the DARPA, USAF, ARO, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright notation herein.

References

Aldrich, J. 1989. Autonomy. *Oxford Economic Papers* 41(1): 15–34.

Angrist, J. D.; Imbens, G. W.; and Rubin, D. B. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434): 444–455.

Arbour, D.; Marazopoulou, K.; and Jensen, D. 2016. Inferring causal direction from relational data. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 12–21. AUAI Press.

Bang, H.; and Robins, J. M. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4): 962–973.

Chickering, D. M. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research* 3(Nov): 507–554.

Craver, C. F.; and Darden, L. 2013. *In Search of Mechanisms: Discoveries Across the Life Sciences*. University of Chicago Press.

Diuk, C.; Cohen, A.; and Littman, M. L. 2008. An object-oriented representation for efficient reinforcement learning. In *Proceedings of the 25th International Conference on Machine Learning*, 240–247.

Dorie, V.; Hill, J.; Shalit, U.; Scott, M.; and Cervone, D. 2019. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science* 34: 43–68.

Gelman, A. 2006. Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics* 48(3): 432–435.

Gentzel, A.; Garant, D.; and Jensen, D. 2019. The case for evaluating causal models using interventional measures and empirical data. In *Advances in Neural Information Processing Systems 32*, 11722–11732.

Getoor, L.; and Taskar, B. 2007. *Introduction to Statistical Relational Learning*. MIT press.

Goldstein, H. 2010. *Multilevel Statistical Models*. Wiley, 4th edition.

Greenwald, A. G. 1976. Within-subjects designs: To use or not to use? *Psychological Bulletin* 83(2): 314.

Hahn, J.; Todd, P.; and Van der Klaauw, W. 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69(1): 201–209.

Illari, P. M.; and Williamson, J. 2012. What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science* 2(1): 119–135.

Imbens, G.; and Rubin, D. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Jensen, D.; Burroni, J.; and Rattigan, M. 2019. Object conditioning for causal inference. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press.

Lee, Y.; and Ogburn, E. L. 2020. Network dependence can lead to spurious associations and invalid inference. *Journal of the American Statistical Association* 1–15.

Machamer, P.; Darden, L.; and Craver, C. F. 2000. Thinking about mechanisms. *Philosophy of science* 67(1): 1–25.

Maier, M. 2014. *Causal Discovery for Relational Domains: Representation, Reasoning, and Learning*. Ph.D. thesis, University of Massachusetts Amherst.

Maier, M.; Marazopoulou, K.; and Jensen, D. 2014. Reasoning about independence in probabilistic models of relational data. *arXiv preprint arXiv:1302.4381* .

McDowall, D.; McCleary, R.; Meidinger, E. E.; and Hay Jr, R. A. 1980. *Interrupted Time Series Analysis.* Sage.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. New York, NY, USA: Cambridge University Press, 2nd edition.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of Causal Inference*. The MIT Press.

Renz, J.; Ge, X.; Verma, R.; and Zhang, P. 2016. Angry Birds as a challenge for artificial intelligence. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1): 41–55.

Rubin, D. B. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* 100(469): 322–331.

Schölkopf, B. 2019. Causality for machine learning. *arXiv preprint arXiv:1911.10500* .

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction and Search*. Cambridge, MA: MIT Press, 2nd edition.

Tsamardinos, I.; Brown, L. E.; and Aliferis, C. F. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning* 65(1): 31–78.

Witty, S.; Takatsu, K.; Jensen, D.; and Mansinghka, V. 2020. Causal inference using Gaussian processes with structured latent confounders. In *Proceedings of the International Conference on Machine Learning*, 1072–1082.