Received: 4 May 2017 | Revised: 20 July 2017 | Accepted: 20 September 2017

DOI: 10.1111/phc3.12470

WILEY

ARTICLE

Causal discovery algorithms: A practical guide

Daniel Malinsky[1] | **David Danks**[2]

[1]Department of Philosophy, Carnegie Mellon University
[2]Departments of Philosophy and Psychology, Carnegie Mellon University

Correspondence
David Danks, Departments of Philosophy and Psychology, Carnegie Mellon University, Pittsburgh, USA.
Email: ddanks@cmu.edu

Abstract

Many investigations into the world, including philosophical ones, aim to discover causal knowledge, and many experimental methods have been developed to assist in causal discovery. More recently, algorithms have emerged that can also learn causal structure from purely or mostly observational data, as well as experimental data. These methods have started to be applied in various philosophical contexts, such as debates about our concepts of free will and determinism. This paper provides a "user's guide" to these methods, though not in the sense of specifying exact button presses in a software package. Instead, we explain the larger "pipeline" within which these methods are used and discuss key steps in moving from initial research idea to validated causal structure.

1 | INTRODUCTION

Causal knowledge is critical for answering many important questions, particularly when we are interested in methods and policies for effective control. However, it has long been recognized that causal knowledge can be quite difficult to obtain. A standard maxim in statistics is "Correlation is not causation," and it is thereby often implied that causal information can never be learned from observational or non-experimental data. This implication is incorrect: Although correlation (or more precisely, probabilistic association) is not identical with causation, there are important and useful relationships between these concepts. Under appropriate conditions, we can even make inferences from correlation to causation. Algorithms that search for causal structure information—typically represented using causal graphical models—are spreading widely in statistics, machine learning, and the social and natural sciences; there are now numerous success stories in which causal structure was inferred from non-experimental data, and then subsequently confirmed by direct experimentation. These methods do not generally pin down all the possible causal details, but they often provide important insight into the underlying causal relations.

Increasingly, we philosophers have begun to collect and analyze data, whether from psychological experiments, sociological surveys, pedagogical studies, or other settings. And in many cases, we are particularly interested in causal questions, such as "does belief in determinism *cause* reduced attributions of moral responsibility?" This article aims to provide a practical "user's guide" to causal search algorithms. We will not extensively discuss the theory of causal search (see Eberhardt, 2009 for an introduction), nor precise details of the algorithms (see, e.g., Spirtes, Glymour, & Scheines, 2000 or Pearl, 2009) nor philosophical objections to the very possibility of causal search (e.g., Cartwright, 2007).

Instead, we will provide pointers for the practical use of these methods and indicate some common pitfalls and mistakes. In so doing, we will touch on many issues that are common "practical knowledge" among users of these algorithms but are not widely discussed (or known) within relevant philosophical communities.

There are many different causal search algorithms that have been developed over the past 30 years, and further development is the subject of active research. Causal search algorithms should be understood as tools: Some algorithms are useful in one situation, and alternative methods are appropriate for different situations. Before using an algorithm, it is important to check whether its assumptions and preconditions are plausibly satisfied for the particular dataset(s) at hand. In this regard, causal search algorithms are no different than other, better-known statistical or data analysis methods, such as various regression techniques. We focus here on algorithms implemented in standard software, primarily the `TETRAD` program[1] and the `pcalg` package in `R`, but there are numerous other algorithms and software packages that may be of interest to particular readers.[2] We strongly encourage interested readers to try out these algorithms (from any of the referenced packages) with their own data, as many complexities of causal search emerge only in practice.

This article is written in a relatively domain-general way and so surely fails to address every problem or complication that can arise in causal search. Nonetheless, it will be helpful to have two running examples. First, consider data on lay people's explicit beliefs about determinism, the possibility of free will, the nature of moral responsibility, and related notions. These various beliefs are undoubtedly related in many people's minds, but we might plausibly wonder about the causal connections. Does belief in determinism *cause* a reduction in attributions of moral responsibility? Or does a desire to attribute moral responsibility *cause* a belief in free will? Or some other connection? In this case, we can presumably directly measure participants' beliefs through their reports (even if those might be error-prone), and so these causal questions concern variables or features that were directly measured.

As a second kind of example, suppose that we are in a pedagogical context and are interested in what educational interventions might improve quintessentially philosophical skills, such as argument analysis. In contrast with explicit beliefs, we arguably cannot directly measure someone's skills, as any particular response could be correct (or incorrect) just due to luck. Rather, we must infer the student's proficiency on different skills from performance on multiple problems. In addition, it is unlikely that any single pedagogical intervention changes someone's skills in a measurable way; instead, skill acquisition is typically the result of multiple educational experiences. That is, the relevant (potential) cause and effect are both indirectly measured by multiple features and are not themselves directly (individually) measured. Moreover, there may be multiple, distinct "latent" skills (e.g., mathematical reasoning skills vs. verbal skills), and these aggregates might themselves be causally related. The common theme in this example is that the *measured* variables or features are not necessarily the relevant *causal* variables or features. These types of "indirectness" complicate causal search in various ways but are quite common in real-world studies.

We now turn to the particulars of how to do causal search. We begin by examining when these algorithms are appropriate and then explore each step of the data analysis process. We conclude by discussing how to interpret (and in some cases, *not* interpret!) the outputs of causal search algorithms.

2 | WHEN IS CAUSAL SEARCH THE RIGHT OPTION?

Causal search algorithms are principally used to investigate hypotheses concerning type-level causal relations, such as the causal relationship between *X* (attributions of moral responsibility) and *Y* (beliefs about free will) in the context of other relevant factors like *Z* (beliefs about determinism). Note that this task is importantly different from the task of determining causal relationships (attributions) among token-level events such as my personal beliefs, and also different from making accurate non-causal, statistical predictions.

Suppose that the researcher is interested in some set of variables *X*, *Y*, *Z*, where she may already know some causal facts, such as "*X* does not cause *Z*" (perhaps because *X* is later in time). When the researcher is ignorant of at least some causal relations, causal search algorithms can narrow down the set of causal models consistent with the data, using patterns of observed probabilistic independence and dependence. Essentially all present-day causal search algorithms assume that these causal models are expressible in the (very flexible) language of causal graphical models. A popular

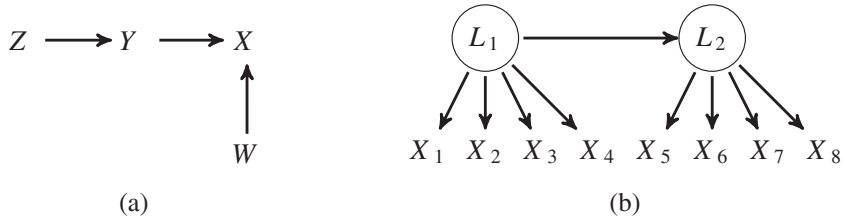

FIGURE 1 (a) A Directed Acyclic Graph model relating variables Z, Y, X, and W. (b) A Directed Acyclic Graph model with latent variables L_1 and L_2 and measured variables $X_1, \ldots, X_8$

kind of graphical model is the Directed Acyclic Graph (DAG) that represents the direct causal relations between the variables. For example, suppose the variables in Figure 1a correspond to measured beliefs about determinism (Z), beliefs about free will (Y), attributions of moral responsibility (X), and ethical context (W). In that case, the model in Figure 1a represents the following hypothesis: Ethical context and beliefs about free will both influence attributions of moral responsibility, but beliefs about determinism are only indirectly relevant, via beliefs about free will. These structures are connected with statistical observations by bridge principles—such as the Causal Markov Condition and Faithfulness described below—that can be exploited (with background knowledge) by causal search algorithms (Eberhardt, 2009).

Often, a researcher's interest in type-level causal relations stems from her interest in learning about the potential outcomes of various manipulations to the system under study (Woodward, 2003). These manipulations might be medical treatments, therapies, social programs, economic policies, or other interventions. In more philosophical contexts, these might be externally generated reasons or externally driven changes in one's beliefs or commitments. In those cases, the researcher should focus on variables and quantities that are at least in principle (independently) manipulable; for example, a clever experimentalist might compel a patient to take some medicine without otherwise directly impacting her.

Causal search can also lead to certain kinds of causal explanations, even when interventions are not practically possible or relevant. For example, one may wonder whether correlation patterns among math, logic, and writing test scores can be explained by a single latent factor like "general intelligence" or whether multiple factors are required to explain the data. For example, the DAG in Figure 1b could represent the hypothesis that verbal reasoning skill (L_1) directly causes argumentative abilities (L_2), where $X_1, \ldots, X_4$ are measurable responses on verbal tests, and $X_5, \ldots, X_8$ are various measures of argumentation ability. Competing explanations can be represented with different graphical models, and then the task of causal search is to select the best (or better) model on the basis of the data.

Traditionally, causal discovery in the social sciences has proceeded through hypothetico-deductive inference or hypothesis testing. The researcher posits a handful of, perhaps just one, possible models (whether for explanatory features, or prima facie plausibility) and derives the models' implications for observational data. Then the researcher typically performs a hypothesis test: Are the data consistent with one or another model's implications? Models may be straightforwardly rejected using this procedure, but if they are not rejected, we still should not categorically accept them. There may be alternative, unconsidered models that are consistent with the data but differ in their explanatory features or counterfactual implications. In contrast, causal search algorithms effectively consider the entire class of models consistent with background assumptions (not just a handful), test them against the data, and thereby select a (hopefully small) set of models consistent with both the observed data and background assumptions. The advantages of this broad search are nicely shown using philosophical data in Rose, Livengood, Sytsma, and Machery (2012).

3 | PREPARING THE DATA

The first step in any causal search is to ensure that the data are appropriately prepared for analysis. Many issues in this step are not specific to causal inference: For example, if variables have missing values, then one must either address

them in the dataset (e.g., by interpolating values or by excluding some datapoints) or choose an algorithm that is robust to that possibility. We leave aside those challenges and focus on data preparation issues that arise specifically for causal search algorithms.

First, causal search algorithms assume that the variables are "semantically independent"—not logically or mathematically interdefinable, and in theory, independently manipulable (Spirtes & Scheines, 2004)—and so the researcher may need to remove some "redundant" variables. For example, suppose that one measures *HDL cholesterol*, *LDL cholesterol*, and *Total cholesterol*, where *Total* is the sum of *HDL* and *LDL*. These three variables are mathematically related and not independently manipulable, and so we should include only *HDL* and *LDL* in our dataset but not *Total* (Spirtes & Scheines, 2004; Zhang & Spirtes, 2008). There is no universal rule for determining which variables to remove, as this step depends partly on background domain knowledge (see also Woodward, 2016). One general guideline is to ensure that there is no collinearity in the dataset (i.e., there is no subset of variables such that one is perfectly or very highly predictable from the others).[3] Detecting collinearity is tricky, but one common symptom of perfect collinearity in a data set is that the covariance matrix will not be invertible (or almost non-invertible), leading most statistical programs to output some type of "matrix inversion" error.

Second, variables are either continuous in value (e.g., *height* can be any number greater than 0) or categorical in value (i.e., having only a finite number of possible values, often only two). Almost all current causal search algorithms assume that all variables are of the same type, continuous or categorical. In practice, however, our data often include a mix of types of variables, and so we face a choice. One option is to use one of the recently developed causal search algorithms for such mixed datasets (e.g., Andrews, Ramsey, & Cooper, in press; Sedgewick, Shi, Donovan, & Benos, 2016), though their reliability on real-world data has not been fully tested. We encourage interested readers to explore available causal search software to learn about the newest algorithms.

The second, and presently more common, response to mixed datasets is to transform some variables so we can use standard algorithms. If most of the variables are continuous, we can treat the categorical variables as if they were continuous, as long as the values of each categorical variable can be placed on a "scale."[4] For example, variable values of "low," "medium," and "high" can usually be treated as continuous; variable values such as "cat," "dog," and "horse" (e.g., values for the variable *species*) cannot. If most variables are instead categorical, then we may choose to discretize continuous variables. However, discretization should be done very carefully, if at all; different discretizations can yield quite different independence judgements and thus different inferred causal structures. Discretization can also make nonlinear causal dependencies difficult to detect. Ideally, the discretization should divide the continuous variable values into causally-appropriate "bins" that preserve relevant causal relationships, or else the causal search algorithm can give very misleading results. Many causal search algorithms discussed below work for either continuous or categorical datasets, though readers should be aware that some methods are restricted to one type or the other. For example, algorithms of the "semi-parametric assumptions" variety can only be used with continuous variables. Decisions about how to treat variables should, of course, be driven by the factors discussed here, rather than mere availability (or not) of a causal search algorithm.

Third, we might have multiple proxy measurements for some unobservable variable of interest. For example, we cannot directly measure *argumentative skill*, only performance on various related tasks (e.g., "identify the premises of the argument"). If the unobserved factor is the relevant potential cause for other variables of interest, then we may want to combine the measured variables into a single estimate of that factor, so that we can include it in our causal search. Psychological scales are, for example, constructed for exactly this reason. But if one uses such scales or estimates, then it is critical to ensure that the proxies are actually *accurate* estimates of a *single* unobserved causal factor, else our causal search will, in general, be unreliable.[5]

Fourth, the researcher must know whether her datapoints represent measurements of different individuals or of the same individual or system over time. Time series data (e.g., a month-by-month series of measurements of an economy) provide additional constraints for causal inference since we have timing information. However, they also present a number of distinctive challenges and so can require quite different causal search algorithms (e.g., Entner & Hoyer, 2010; Hyttinen, Plis, Jarvisalo, Eberhardt, & Danks, 2017; Moneta, Chlaß, Entner, & Hoyer, 2011). We focus here exclusively

on the static case of measurements of different individuals, as most data collected or analyzed by philosophers are of this form.

Finally, although not strictly an issue of data preparation, the researcher should determine her background knowledge about the potential causal relations, as that information can be used by most causal search algorithms. Post hoc application of background knowledge can lead to inefficiencies and inconsistencies, so one should incorporate it from the outset. Time order and experimental design are two common, and important, sources of background knowledge. Different software implementations allow different types of background knowledge, but all can incorporate knowledge of the definite presence or absence of connections, as well as "ordering" information (e.g., *X* occurs before *Y*, so *Y* cannot cause *X*).

4 | VARIETIES OF SEARCH ALGORITHMS

The next step in causal search is to use the right algorithm. All methods assume that the data-generating process (the "true" causal graph, whatever it is) satisfies the much-discussed Causal Markov Condition (CMC): Every variable *X* in **V** (the set of variables in the causal graph) is independent of its non-effects conditional on its direct causes. Philosophically, the CMC is a formalization and generalization of Reichenbach's Principle of the Common Cause (Reichenbach, 1956) and is widely (if not always explicitly) assumed in causal modeling (Spirtes, 1995). Many search algorithms also assume Faithfulness: The only independencies among the variables in **V** are those entailed by the CMC. Faithfulness can be violated when parameters on causal pathways are exactly "tuned" to cancel out; typically, such parameterizations are excluded from consideration, though there are methods that weaken this restriction. We will not review arguments for and against the CMC and Faithfulness here; see Eberhardt (2009) for an overview of the literature.

The CMC and Faithfulness together imply a tight connection between (causal) graphical structure and conditional independencies in the data (Geiger, Verma, & Pearl, 1990), and this connection can be exploited in search. Specifically, each causal graph implies a specific, determinate set of (conditional) independencies, though different graphs might imply the same set.[6] Pathways in the graph correspond to probabilistic dependence, and graphical non-adjacencies imply (conditional) independence. There are then three principal approaches to causal search: constraint-based algorithms; Bayesian or score-based algorithms; and procedures that exploit semi-parametric assumptions.[7]

Constraint-based algorithms focus directly on the connection between graphs and implied independence facts: Specifically, they aim to discover the set of causal graphs that imply exactly the (conditional) independencies found in the data by performing a sequence of hypothesis tests. For example, if possible causal models are all DAGs over measured variables **V**, then one could (in principle) derive the implied conditional independencies for each graph in the space and then compare the implied pattern against the data using a statistical test. In practice, there are too many DAGs to exhaustively enumerate and test them, so algorithms like `PC`[8] use a clever schedule of tests to effectively explore the whole space of DAG models in an efficient way (Spirtes, Glymour, & Scheines, 2000). Alternately, the restriction to DAGs over **V** may be unwarranted; for example, we may know or suspect that there are unmeasured variables that are common causes of two or more measured variables (i.e., there is "causal insufficiency" or "latent confounding"). In that case, we need to use a different constraint-based algorithm (e.g., Fast Causal Inference or `FCI`) to search through a larger space of alternatives that includes unobserved common causes in the DAGs.

An important feature of constraint-based algorithms is that, in principle, they need not make any assumptions about the parametric form or functions for the causal connections. In practice, though, these algorithms often rely on statistical tests of conditional independence that make more restrictive assumptions; for example, popular tests assume Gaussian or multinomial distributions for the data. However, new statistical tests of conditional independence (particularly nonparametric tests) are continually being developed, which yield constraint-based algorithms for nonlinear and non-Gaussian data (e.g., Zhang et al., 2011; Ramsey, 2014). The choice of conditional independence tests for constraint-based algorithms should reflect information about the data distribution gleaned from background knowledge and exploratory data analysis.

In contrast, **score-based algorithms** compare models on the basis of some measure of model fit. The most common score is the Bayesian Information Criterion (BIC) score, which approximates the posterior probability of the model

given the data (assuming a uniform prior probability distribution over DAG-space). For computational reasons, we cannot exhaustively enumerate and score every model in the space of DAGs, so we typically use a "greedy" algorithm like Greedy Equivalence Search (GES) or the much faster, optimized Fast GES (FGES; Ramsey, Glymour, Sanchez-Romero, & Glymour, 2017) implemented in TETRAD. These greedy methods search in a local, heuristic way but converge on the globally optimal scoring model in the limit of infinite data (Chickering, 2002). Score-based algorithms are less developed for the case of possible unobserved common causes, though there have been recent advances on that front (e.g., Ogarrio, Spirtes, & Ramsey, 2016). In simulation studies, score-based causal search algorithms (e.g. GES, also Greedy FCI or GFCI) are generally more accurate than constraint-based algorithms making the same assumptions (e.g., PC, FCI) for datasets with small sample sizes, though much depends on the particular simulation. Note that score estimation is typically tied to parametric assumptions—algorithms like GES use a BIC score as calculated for Gaussian or multinomial data—though relaxing such restrictions is also a frontier of current research.

Finally, **causal search algorithms with semi-parametric assumptions** use these additional assumptions to learn causal relationships more efficiently or in more detail. For example, when the data are generated by linear mechanisms but with non-Gaussian noise, Linear Non-Gaussian Model (LiNGaM) algorithms can recover the DAG structure using a signal analysis technique called independent components analysis (Shimizu, 2014 provides an overview). Some techniques in this category can even account for "feedback" loops of different types (Lacerda, Spirtes, Ramsey, & Hoyer, 2008) or latent common cause variables (Hoyer, Shimizu, Kerminen, & Palviainen, 2008). Compared with constraint-based and score-based methods, these algorithms do not rely on the controversial Faithfulness assumption. They also can often learn substantially more about the underlying causal structure and sometimes even determine a unique model. There are drawbacks, though, as the semi-parametric assumptions can themselves be controversial or difficult to test. In addition, these algorithms typically require much larger sample sizes (e.g., thousands or tens of thousands of samples) to be accurate and so may not be usable in cases of interest for philosophers.

One final set of algorithms deserves special note. When the goal is to discover latent variables and their relationships, as in the pedagogical example, the researcher may consider algorithms like BPC, FOFC, and FTFC.[9] These methods are explicitly *causal* analogues to traditional factor analysis: They "cluster" *variables* (not individuals) into groups that are proxy measures of one or more latent variables. For example, they can discover that $X_1, \ldots, X_4$ in Figure 1b are measures of an underlying latent, such as verbal reasoning ability. The algorithms employ constraint-based logic: Latent variable models imply particular patterns of observed correlation or conditional independence (e.g., "tetrad constraints" and generalizations), and the algorithms find models with implications consistent with observed data. Importantly, these algorithms do not require knowledge of the number of "variable clusters" (in contrast with factor analysis), though they do require some assumptions about the measurement model (Kummerfeld & Ramsey, 2016; Kummerfeld, Ramsey, Yang, Spirtes, & Scheines, 2014; Silva, Scheines, Glymour, & Spirtes, 2006). A researcher may use one of these algorithms to first "discover" latent constructs, and then perhaps subsequently learn the causal connections among the latent constructs themselves.

4.1 | Tuning parameters and other statistical decisions

Most causal search algorithms involve parameters of one sort or another that "tune" the search. Software implementations almost always include defaults for these "tuning knobs," but it is important not to blindly employ the defaults. We focus here on the choice of independence tests and hypothesis test thresholds.

When using constraint-based or semi-parametric causal search algorithms, one should choose conditional independence tests that match knowledge of the empirical distribution of the data. For continuous variables, informal techniques like Q-Q plots and formal tests of multivariate Gaussianity should be used to inform choice among options like the Fisher Z-test and non-parametric alternatives.[10] With categorical variables, the options are more limited: Chi-square tests and tests based on logistic regression are popular options; non-parametric alternatives are possible but less developed. Where feasible, one should try multiple options to determine if results are robust to the choice of independence test.

Similarly, proper choice of algorithm parameter settings requires a mixture of background knowledge and exploration of multiple settings to determine robustness. Constraint-based methods require specification of a decision threshold α for the conditional independence tests. If the researcher were performing only one test, then α would correspond to the probability of Type I error (false positive). However, a constraint-based search algorithm performs many tests, and multiple-testing considerations complicate matters. In some scientific domains, α may be set by convention (e.g., at 0.05, 0.01, or some other threshold), but fixed-α conventions are not straightforwardly justified nor usually advisable in the model search context. Recent work has explored modified constraint-based algorithms in which the tuning parameter controls the global False Discovery Rate (FDR; Strobl, Spirtes, & Visweswaran, 2016; see also Drton & Perlman, 2007).[11]

In practice, researchers commonly take one of two approaches when the algorithm parameters (e.g., α) cannot be chosen a priori: (a) report the causal information that is stable over a range of parameter settings; or (b) use a tuning parameter selection procedure (proposed by Maathuis, Kalisch, & Bühlmann, 2009) that systematically searches over multiple parameter values and returns the result with best overall model fit (BIC score).[12] Additionally, a different kind of robustness check can inspire confidence in (parts of) the output model: find the parts of the model that are invariant under subsampling of the data. That is, choose a parameter setting that is relatively "lenient" (e.g., large α), run the causal search algorithm on multiple (random) subsets of the data, and retain the causal connections that are consistently discovered (Stekhoven et al., 2012). The general idea underlying all of these robustness or stability checks is that the most reliable causal connections are those whose discovery is not tied to the specific setting of a parameter or undermined by small variations in the data.

5 | INTERPRETING THE OUTPUT

The final step in the causal search process is correct interpretation of the algorithm outputs, which can involve several complications. First, most causal search algorithms can output multiple causal models, though encoded in one compact graphical representation, and we need to "translate" the output. For example, the `PC` algorithm might return a Pattern[13] which (unlike a DAG) can contain undirected edges, for example, $Z - Y$ in Figure 2a. An undirected edge $Z - Y$ in a Pattern indicates that there is at least one DAG model that is consistent with the data and has $Z \rightarrow Y$ and another DAG consistent with the data with $Z \leftarrow Y$. With three variables, the pattern $X - Y - Z$ encodes three distinct causal possibilities: $X \rightarrow Y \rightarrow Z$; $X \leftarrow Y \rightarrow Z$; and $X \leftarrow Y \leftarrow Z$.[14] We might wish that causal search algorithms always found the unique underlying causal model, but that is often not possible given observational data alone, as there can be multiple causal models that are equally consistent with the data. In such cases, the algorithms correctly output multiple possibilities, and so we need to attend to what is shared or differs across them. For example, the three possibilities above agree that X and Z are not directly causally connected but disagree about the $X - Y$ and $Y - Z$ causal directions.

Some algorithms that allow for latent confounders (e.g., `FCI` and `GFCI`) produce a representation called a Partial Ancestral Graph (PAG); see Figure 2b for an example. PAGs are more complicated to interpret than DAGs or Patterns. They can contain double-headed arrows that represent latent confounding: $X \leftrightarrow Y$ indicates that neither causes the other, but there is some unobserved variable (or variables) that causes both X and Y.[15] In a PAG, there may also be edges

FIGURE 2 (a) A Pattern representing two different Directed Acyclic Graph, one with $Z \rightarrow Y$ and one with $Z \leftarrow Y$. (b) A Partial Ancestral Graph model, where circle marks indicate uncertainty about the possibility of latent confounding

with circle marks on them to represent ambiguity. Z o–o Y means that either $Z \leftrightarrow Y$, $Z \rightarrow Y$, or $Z \leftarrow Y$ (or the first combined with either of the latter two). W o→ X means that either $W \leftrightarrow X$ or $W \rightarrow X$ (or both).

Importantly, these causal graphs encode only qualitative conclusions about direct causation—that is, whether a manipulation of C will have *any* (probabilistic) impact on E when the other variables in $\mathbf{V}$ are held fixed at *some* values. In many settings, we also want to know about the strengths or contexts of causal influence—when does C matter for E, and how much?—perhaps to support effective control. For example, the causal graph $C_1 \rightarrow E \leftarrow C_2$ is consistent with C_1 being a strong cause of E while C_2 is a weak one, or the opposite, or C_1 being a strong cause only when C_2 takes on specific values (i.e., context-specific causation).[16] In practice, we thus often use our data to estimate causal model parameters that provide quantitative strength information (e.g., linear coefficients), though this can be complicated (and sometimes underdetermined) when there are potential unobserved factors (Maathuis, Kalisch, & Bühlmann, 2009; Malinsky & Spirtes, 2016).

Most causal search algorithms can provide some information about *why* certain causal models or connections were included or excluded. For example, standard implementations of constraint-based algorithms can list the (conditional) independencies used to find the output. If certain key independencies hold only weakly, then we might place less confidence in the algorithm outputs (though causal search algorithms cannot generally provide true confidence intervals; see Zhang, 2008). Although this information can be quite helpful in assessing the robustness of the output, such assessments typically require relatively detailed knowledge of the algorithms.

Finally, researchers may want to examine other causal models in a post hoc manner. Causal search algorithms output the "best" causal models according to their criterion (e.g., fit-to-data for score-based algorithms). However, other causal models might be almost as good (on that criterion) and so should perhaps be explored. For computational reasons, no algorithm can consider every possible causal model, but we can examine a subset—perhaps those "close" to the output or suggested by background knowledge—to evaluate their quality. This process can provide further insight into which causal connections are plausible or implausible.

6 | CONCLUSION

Causal search algorithms can provide significant insight into the causal relationships underlying our data and so are beginning to be used for problems of philosophical interest. There are a number of distinctive issues that arise in the selection, application, and interpretation of these algorithms; they are more complex than running a simple descriptive statistical test. At the same time, they do not require advanced training in causal modeling or algorithm design, particularly given the numerous software implementations that are now available. Rather, they require care to ensure that we make appropriate choices of the variables to include in the dataset, the search algorithm and parameter settings to use, and the interpretation of the output.

ENDNOTES

[1] http://www.phil.cmu.edu/tetrad/, as well as versions in `R`, `python`, or made available through the Center for Causal Discovery (http://www.ccd.pitt.edu).

[2] See, for example, the Bayes Net Toolbox (https://github.com/bayesnet/bnt) for `Matlab`. More generally, novel algorithms can be found in journals such as the *Journal of Machine Learning Research* (JMLR) and proceedings of conferences including Uncertainty in Artificial Intelligence, Knowledge Discovery and Data Mining, and Neural Information Processing Systems. In most cases, authors provide code for their implementations.

[3] Or one may use an algorithm that is reliable even given deterministic relationships; see Glymour (2007).

[4] This option often arises with *Condition* variables, where we must determine whether the experimental conditions can be ordered. Alternatively, we can code the conditions using binary (two-valued) variables, as those can be regarded as either continuous or categorical.

[5] For this reason, there are causal search algorithms that can discover these "proxy" relationships, rather than requiring the researcher to know them in advance. We discuss these algorithms in Section 4.

[6] That is, the *graphs* ↦ *independencies* map is many-to-one.

[7] For reasons of space, we do not discuss methods that combine approaches or otherwise fall outside these categories (e.g., Janzing et al., 2012).

[8] Named for its inventors, Peter (Spirtes) and Clark (Glymour).
[9] Build Pure Clusters; Find One-Factor Clusters; and Find Two-Factor Clusters
[10] In high-dimensional settings, computational considerations may also be relevant since some non-parametric tests can be quite slow.
[11] Score-based approaches do not require an α setting but often allow users to tune a parameter that places an extra penalty on complex models (Ramsey et al., 2010), or such a complexity penalty may be prudent from a computational point-of-view in high-dimensional problems.
[12] This approach resembles the general method of cross-validation used in (supervised) machine learning.
[13] Sometimes also called a Completed Partial DAG (CPDAG).
[14] Notice that there is a fourth possibility that is excluded: $X \rightarrow Y \leftarrow Z$. This DAG model is not consistent with the data because it implies a different set of conditional independence facts (i.e., it is not *Markov equivalent* to the other three models). For an explanation of Markov equivalence, see Spirtes, Glymour, and Scheines (2000).
[15] PAGs can also contain undirected edges that indicate selection bias. We omit discussion of those edges here, as they are rarely discovered in practice.
[16] We ignore complications about how to measure "causal strength."

ACKNOWLEDGEMENTS

Thanks to Edouard Machery and an anonymous reviewer for helpful comments on an earlier version of this paper.

ORCID

David Danks http://orcid.org/0000-0003-4541-5966

WORKS CITED

Andrews, B., Ramsey, J., & Cooper, G. (in press). Scoring Bayesian networks of mixed variables. *Journal of Data Science and Analytics*.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge: Cambridge University Press.

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

Drton, M., & Perlman, M. D. (2007). Multiple testing and error control in Gaussian graphical model selection. *Statistical Science*, *22*, 430–449.

Eberhardt, F. (2009). Introduction to the epistemology of causation. *Philosophy Compass*, *4*, 913–925.

Entner, D., & Hoyer, P. (2010). On causal discovery from time series data using FCI. In *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models*, pp. 121–8. Helsinki, Finland.

Geiger, D., Verma, T., & Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks*, *20*, 507–534.

Glymour, C. (2007). Learning the structure of deterministic systems. In Gopnik, A., & Schulz, L. eds. Causal Learning: Psychology, Philosophy, and Computation, 231–240. Oxford: Oxford University Press.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., & Palviainen, M. (2008). Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, *49*, 362–378.

Hyttinen, A., Plis, S., Jarvisalo, M., Eberhardt, F., & Danks, D. (2017). A constraint optimization approach to causal discovery from subsampled time series data. *International Journal of Approximate Reasoning*, *90*, 208–225.

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., & Schölkopf, B. (2012). Information-geometric approach to inferring causal directions. *Artificial Intelligence*, *182*, 1–31.

Kummerfeld, E., Ramsey, J., Yang, R., Spirtes, P., & Scheines, R. (2014). Causal clustering for 2-factor measurement models. In Calders, T., Esposito, F., Hüllermeier, E., & Meo, R. eds. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 34–49. Berlin: Springer

Kummerfeld, E., & Ramsey, J. D. (2016). Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA.

Lacerda, G., Spirtes, P., Ramsey, J., & Hoyer, P. O. (2008). Discovering cyclic causal models by independent components analysis. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pp. 366–374. Arlington, VA: AUAI Press.

Maathuis, M. H., Kalisch, M., & Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, *37*, 3133–3164.

Malinsky, D., & Spirtes, P. (2016). Estimating causal effects with ancestral graph Markov models. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, pp. 299–309.

Moneta, A., Chlaß, N., Entner, D., & Hoyer, P. (2011). Causal search in structural vector autoregressive models. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, *12*, 95–114.

Ogarrio, J. M., Spirtes, P., & Ramsey, J. (2016). A hybrid causal search algorithm for latent variable models. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, pp. 368–379. Lugano, Switzerland.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge: Cambridge University Press.

Ramsey, J. D. (2014). A scalable conditional independence test for nonlinear, non-Gaussian data. arXiv preprint arXiv:1401.5031.

Ramsey, J., Glymour, M., Sanchez-Romero, R., & Glymour, C. (2017). A million variables and more: The fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, *3*, 121–129.

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A., & Glymour, C. (2010). Six problems for causal inference from fMRI. *Neuroimage*, *49*, 1545–1558.

Reichenbach, H. (1956). *The direction of time*. Berkeley: University of California Press.

Rose, D., Livengood, J., Sytsma, J., & Machery, E. (2012). Deep troubles for the deep self. *Philosophical Psychology*, *25*, 629–646.

Sedgewick, A. J., Shi, I., Donovan, R. M., & Benos, P. V. (2016). Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Bioinformatics*, *17*, 175.

Shimizu, S. (2014). LiNGAM: Non-Gaussian methods for estimating causal structures. *Behaviormetrika*, *41*, 65–98.

Silva, R., Scheines, R., Glymour, C., & Spirtes, P. (2006). Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, *7*, 191–246.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 491–498.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.

Spirtes, P., & Scheines, R. (2004). Causal inference of ambiguous manipulations. *Philosophy of Science*, *71*, 833–845.

Stekhoven, D. J., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M. H., & Bühlmann, P. (2012). Causal stability ranking. *Bioinformatics*, *28*(21), 2819–2823.

Strobl, E. V., Spirtes, P. L., & Visweswaran, S. (2016). Estimating and controlling the false discovery rate for the PC algorithm using edge-specific P-values. arXiv preprint arXiv:1607.03975.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.

Woodward, J. (2016). The problem of variable choice. *Synthese*, *193*, 1047–1072.

Zhang, J. (2008). Error probabilities for inference of causal directions. *Synthese*, *163*, 409–418.

Zhang, J., & Spirtes, P. (2008). Detection of unfaithfulness and robust causal inference. *Minds and Machines*, *18*, 239–271.

Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence*, pp. 804–813. Corvallis, OR: AUAI Press.

Daniel Malinsky is a PhD candidate in Logic, Computation, and Methodology at Carnegie Mellon University. He received his B.A. from Columbia University. Malinsky studies methods of causal inference, with a focus on data-driven techniques for tackling questions of policy, broadly construed.

David Danks is the L.L. Thurstone Professor of Philosophy and Psychology and the head of the department of Philosophy at Carnegie Mellon University. He received his PhD in Philosophy from University of California, San Diego, and an A.B. in Philosophy from Princeton University. His research largely falls at the intersection of philosophy, cognitive science, and machine learning, using ideas and frameworks from each to inform the others. His primary research in recent years has been in computational cognitive science, and the ethical and regulatory impacts of the introduction of autonomy into technological systems. He is the author of *Unifying the Mind: Cognitive Representations as Graphical Models* (MIT Press).

How to cite this article: Malinsky D, Danks D. Causal discovery algorithms: A practical guide. *Philosophy Compass*. 2018;13:e12470. https://doi.org/10.1111/phc3.12470