Commonsense Abductive Reasoning and Metareasoning Using Knowledge from Bayesian Networks

Joshua Eckroth and **John R. Josephson**
Department of Computer Science & Engineering
The Ohio State University
Columbus, Ohio, USA

Abstract

This work describes a method for finding causal explanations for world events that are reported to an agent. These reports may come from the agent's sensors or other agents. Some of these reports might be false, so the agent should exercise caution in believing what it is told. The agent's knowledge about the causal relations and plausibilities of events are represented in a Bayesian network. When the agent obtains a report, it builds an explanation by abductive reasoning, which is a kind of commonsense logic that enables reasoning from effects to causes. We show how abductive reasoning can use knowledge in a Bayesian network to infer plausible, decisive, consistent, relevant, and complete explanations. An abductive metareasoning process monitors and controls the abductive reasoning process in order to detect, explain, and respond to possible errors in reasoning and to detect and isolate false reports. The combined abductive reasoning and metareasoning process is capable of inferring very accurate explanations, often surpassing the accuracy of the "most probable explanation," due to its ability to exercise caution in forming beliefs and to accurately detect and ignore false reports, and performs well even when knowledge of plausibilities is less precise.

This work describes how an agent can use commonsense abductive reasoning to find explanations for reports that putatively describe events in its world. These explanations can play the role of qualitative beliefs that are processed by a knowledge-based system for, e.g., robot planning tasks. A Bayesian network is used to represent causal world knowledge. Inferring explanations from a Bayesian network is not as straightforward as finding the most probable explanation, i.e., the most probable complete set of beliefs about every aspect of the world, for a variety of reasons. First, explanations should include only the causal ancestors of the reported events (Pearl 1988, pp. 285–286). Second, the most probable explanation for some report might only be marginally more probable than a different explanation. In these cases, it might be helpful to gather more evidence to differentiate among the possible explanations before deciding. Third, the most probable explanation for some report might have very low probability. For example, a report might describe a very improbable but true event. But if reports might be false, i.e., noisy, then the agent might find it wise initially to discount very improbable explanations. These commonsense considerations are integrated into the abductive reasoning process.

We also enhance the abductive reasoning process by adding an abductive metareasoning component. Metareasoning, or thinking about thinking, is a way for an agent to reflect on its beliefs and how it came to those beliefs, in order to detect possible mistakes, and to respond appropriately. Systems that include a metareasoning component have enjoyed renewed interest in recent years, as evidenced by workshops such as the AAAI-2008 Metareasoning Workshop and the volume that followed, *Metareasoning: Thinking about thinking* (Cox and Raja 2011). In that volume, Perlis connects commonsense reasoning with metareasoning:

> Commonsense reasoning is the form of metareasoning that monitors an activity for mistakes and then deals with them, sparing the main activity the embarrassment of making a fool of (or destroying) itself (Perlis 2011).

In the system we describe, a metareasoning component monitors the base-level abductive reasoning process for failures to explain. An agent's goal is to make sense of reports, i.e., to infer, with some confidence, the causes of those reports, and the causes of those causes, and so on. It does so by consulting the Bayesian network and inferring states of variables that are ancestors to the reported variables. If the agent is unable to accomplish this task, then maybe the agent can identify the cause or causes of the impasse. Perhaps the reports are false, perhaps more evidence would shed more light on the situation, or perhaps the agent was too cautious and should reconsider less plausible explanations.

After describing the system, we evaluate its performance across a variety of randomly generated Bayesian networks. Our evaluation shows that commonsense abductive reasoning and metareasoning is very effective at finding accurate and complete explanations for reports, while also detecting noisy reports. It also out-performs the most probable explanation (MPE) in terms of accuracy, since the MPE is determined without the various commonsense considerations outlined above, such as plausibility, decisiveness, and the possibility that some reports obtained by the agent are actually just noise.

Abductive Reasoning with Bayesian Networks

By *abduction*, and *abductive inference*, we mean reasoning that follows a pattern approximately as follows (Josephson and Josephson 1994):

> *D* is a collection of data (findings, observations, givens).
>
> Hypothesis *H* can explain *D* (would, if true, explain *D*).
>
> No other hypothesis can explain *D* as well as *H* does.
>
> —
>
> Therefore, *H* is probably correct.

In a process of trying to explain some evidence, the object is to arrive at an explanation that can be confidently accepted. An explanation that can be confidently accepted is an explanation that can be justified as being *the best explanation* in consideration of various factors such as plausibility, consistency, and completeness, and in contrast with alternative explanations. Thus, an explanation-seeking process—an abductive reasoning process—aims to arrive at a conclusion that has strong abductive justification. We hope that readers recognize abductive reasoning is a distinct and familiar pattern, and has a kind of intuitively recognizable evidential force. It is reasonable to say it is part of commonsense logic. It can be recognized in a wide range of cognitive processes including diagnosis, scientific theory formation, language comprehension, and perception.

Notation

This work investigates the use of Bayesian networks (Pearl 1988) to represent causal relations and probabilistic knowledge for abductive reasoning. In the Bayesian networks we consider, each variable has two discrete states. We denote variables with uppercase letters, e.g., X, and states with lowercase letters, e.g., x. Bold uppercase letters such as $\mathbf{X}$ indicate sets of variables and bold lowercase letters such as $\mathbf{x}$ indicate their corresponding states. The set $\bar{\mathbf{x}}$ denotes the complement of states $\mathbf{x}$. In the network, an edge from X' to X means that X is causally dependent on X'. We may also say that X' is a parent of X. The set of parents of a variable X is denoted $\Pi(X)$. Each variable has a conditional probability table which defines the probability of the variable holding each of its states given the states of its parents. Finally, we denote the agent's doxastic state, i.e. its belief state, as the set of variable states $\mathbf{b}$.

Certain states of certain variables may be incompatible with certain states of other variables. A set of variables $\mathbf{x}$ is incompatible with another set $\mathbf{y}$, for our purposes, if and only if some $x \in \mathbf{x}$ is incompatible with some $y \in \mathbf{y}$. When the agent's beliefs do not include both states of an incompatible pair, we say its beliefs are consistent (we are only considering pairwise inconsistency). Incompatibility can be represented in the Bayesian network with constraint variables (Pearl 1988, pp. 225–226). Each pair of variables that have incompatible states, say variables X and Y with incompatible states x and y, are parents of a unique constraint variable C_{xy}, which is fixated to observed state c_{xy}. Constraint variables have conditional probability tables that ensure if either of the incompatible states x and y is observed or assumed, then the other state has zero probability. The conditional probability table is as follows.

	$\bar{x}, \bar{y}$	$\bar{x}, y$	$x, \bar{y}$	x, y
c_{xy}	1.0	1.0	1.0	0.0
$\bar{c}_{xy}$	0.0	0.0	0.0	1.0

It should be noted that the constraint variables produce side effects, in the sense that their presence alters the distributions of variables that are ancestors to those in the incompatible pair. Crowley, Boerlage, and Poole (2007) argue that ancestor variables should not be affected because C_{XY} is not evidence for the ancestors of X and Y. They provide an algorithm for building "antifactors" and "antinetworks," essentially more variables in the network, that counteract such effects of the constraint variables. However, we do not agree with their reservations. It seems reasonable that since only one of x or y is true, or neither is true, each of their causes are less likely than if x and y were not incompatible states.

Definition 1. Suppose X is believed to have state x. Then a *possible explanation* of x is a partial instantiation $\mathbf{y} \neq \emptyset$ of parents $\mathbf{Y}$ of X such that $\mathbf{b} \cup \mathbf{y}$ is consistent.

Definition 2. A possible explanation has an associated *plausibility*. As we use the term here, plausibility is a graded confidence that does not necessarily have all the formal properties of a probability. Suppose that $\mathbf{y}$ is a possible explanation of x. We investigate two ways to calculate the plausibility of $\mathbf{y}$, denoted $Pl(\mathbf{y})$. The first method is simply the posterior (where \ denotes set difference),

$$Pl(\mathbf{y}) = P(\mathbf{y}|\mathbf{b} \setminus \mathbf{y}).$$

The second method, which we call the *estimated prior*, is defined as,

$$Pl(\mathbf{y}) = \max P(x|\mathbf{v}),$$

where $\mathbf{v}$ is a set of states of X's parents $\mathbf{V} = \Pi(X)$ that agree with $\mathbf{y}$ (so $\mathbf{y} \subseteq \mathbf{v}$). In other words, estimated prior is the maximum prior probability of x with parent state assignments (partially) specified by $\mathbf{y}$. We investigate the use of estimated plausibilities because they avoid the potentially large computational cost of computing posteriors (Cooper 1990). Calculating the estimated prior is simply a matter of finding the row in the conditional probability table for X, among those rows that agree with states $\mathbf{y}$, that gives the maximum conditional probability for x.

In order to establish the *ground truth* of the probabilistic model, which is to be inferred by the agent, a direct sample of the whole network is taken, respecting conditional probabilities and incompatibilities. This process involves fixating variable states, starting from the top variables and proceeding down the graph in the direction of parent variables to child variables and randomly selecting a variable state based on the conditional probability table for each variable, while respecting incompatibilities. Some variable states in the ground truth set are communicated to the agent in the form of reports. The agent's goal is to explain the reports.

Furthermore, we say that any belief requires explanation if it has parents in the network. Possible explanations range across the various combinations of states of its parents. The agent generates possible explanations for reports and beliefs, calculates their plausibility, and then executes the abduction algorithm in order to decide which possible explanations should turn into beliefs.

Abduction Algorithm

One might imagine that a practical goal of an abductive reasoner is to find the most plausible, consistent, and complete composite explanation of the evidence. However, Bylander et al. (1991) show that abduction problems that involve an incompatibility relation among pairs of hypotheses cannot be practically solved. Specifically, they prove that it is NP-complete to determine whether a consistent, complete explaining set exists for such an abduction problem. They also prove that it is NP-hard to find a most-plausible consistent, complete, composite explanation.

Thus, we take an efficient greedy approach to the abduction problem, similar to that implemented in Josephson & Josephson's PEIRCE-IGTT system (1994). Their system realizes an algorithm called *EFLI*: "essentials first, leveraging incompatibility," which iteratively accepts one explanation and rejects incompatible explanations, until either all evidence is explained or no other candidate explainers for the unexplained evidence are available. Explainers are grouped into contrast sets, where each contrast set contains all the plausible explanations for some report. Essential explanations are accepted first. An essential explanation is the sole member of a contrast set, so it is the only plausible explanation for some report. Without accepting it, some evidence would remain unexplained. Then, explanations are ordered for acceptance by the degree to which the best explanation in a contrast set surpasses the second best explanation, in terms of plausibility.

We say that the *best explanation* for evidence x is the most plausible, most decisive partial assignment of parent variable states of x. Of course, the explanation must be consistent with prior beliefs. We find it advantageous to require that the best explanation meet a minimum plausibility η and a minimum decisiveness $\delta \geq \Delta$ Threshold. The EFLI abduction algorithm automates the process of finding the best explanation according to these desiderata.

We summarize the EFLI algorithm after introducing the following definitions. Further details of the algorithm may be found in previous work (Eckroth and Josephson 2013).

Definition 3. A *contrast set* $C = \{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$ is an ordered set of possible explanations for some state x, ordered by plausibility, most plausible first. The *decisiveness* of the contrast set is defined as,

$$\delta = \begin{cases} [Pl(\mathbf{y}_1) - Pl(\mathbf{y}_2)] \,/\, \sum_i Pl(\mathbf{y}_i) & \text{if } \|C\| > 1, \\ 1.0 & \text{otherwise.} \end{cases}$$

The *most decisive contrast set* is the contrast set with the greatest δ. Finally, the Δ Threshold parameter specifies a lower bound on δ; any contrast set with decisiveness less than Δ Threshold will not be considered.

Definition 4. *EFLI abduction algorithm.*

```
1. Generate possible explanations for unexplained reports and beliefs.
2. Generate contrast sets.
3. For each possible explanation $\mathbf{y}$ that does not satisfy the minimum plausibility requirement, i.e., $Pl(\mathbf{y}) < \eta$, remove it from any contrast sets in which it is a member.
4. If there are no non-empty contrast sets, then there is nothing to accept, so halt.
5. Otherwise,
   (a) Find most decisive non-empty contrast set $C$; let $\delta$ be its decisiveness.
   (b) If $\delta < \Delta$ Threshold, then halt.
   (c) Otherwise,
       i. Find the most plausible possible explanation, $\mathbf{y}$ in $C$.
       ii. Add the states specified by $\mathbf{y}$ to the belief state.
       iii. For each state $z$ incompatible with any of the states specified by $\mathbf{y}$, add $\bar{z}$ to the belief state.
       iv. Go to step 1.
```

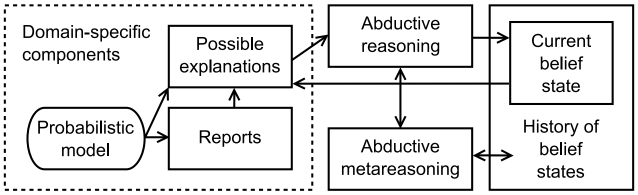

Figure 1: System architecture. Domain-specific components are separate from domain-general reasoning and metareasoning components. Reports are obtained from the world, which might be noisy. The plausibility of each report is calculated with the probabilistic model. Each report requires explanation, as do any unexplained beliefs. Possible explanations are generated and reviewed by the abductive reasoning procedure, typically resulting in an expanded belief state. Newly-acquired beliefs might themselves require explanation, and the process starts again. If any reports or beliefs remain unexplained, which we call anomalies, abductive metareasoning is activated in order to determine the causes of the anomalies. Abductive metareasoning might find new explainers, revise the belief state, and/or leave the anomalies unexplained. Reports that remain unexplained are deemed to be noise.

Abductive Metareasoning

Reports and beliefs that require explanation and remain unexplained after abductive reasoning are considered anomalous. The presence of anomalies activates an abductive metareasoning procedure that attempts to both determine the causes of the anomalies and resolve them. Possible resolutions include temporarily lowering the minimum plausibility η requirement to allow consideration of certain possible explanations, gathering more evidence, and retracting previously accepted explanations. Alternatively, anomalies may be left unresolved; anomalous reports that remain unresolved are deemed to be noise.

If an agent could be sure that the reports it receives about the world are truthful, then every anomaly would be resolvable and should be resolved, even if their resolutions required retracting previously accepted explanations or reconsidering implausible explanations. However, in more realistic environments, it is not necessarily the case that all reports are true reports, so the agent should be somewhat cautious about revising its beliefs or accepting implausible explainers. Some reports might be unexplainable partly because they are noise and do not warrant any explanation. Thus, part of the challenge of metareasoning is to identify which reports are unexplainable due to false beliefs and which are due to false reports (and hence, not due to false beliefs). The metareasoning task can be construed as an abductive one by treating the anomalies as a kind of meta-evidence that require explanation by meta-hypotheses. Such a metareasoning process is able to utilize the same abductive reasoning machinery as the base-level reasoner.

An anomaly might have multiple possible causes. Each is detailed in its own section below. For each possible cause, a meta-hypothesis is generated, which specifies the cause, the resolution, the anomaly or anomalies it is said to explain, and an estimated plausibility score. Two meta-hypotheses are incompatible if they are capable of explaining (and thus resolving) one or more anomalies in common. Abductive reasoning is activated on the set of meta-hypotheses, which compete to explain the anomalies. The result is a set of accepted meta-hypotheses, and the resolutions that are specified by the accepted meta-hypotheses are applied. If any anomalies remain (or new anomalies appear), metareasoning is activated again. Care is taken not to generate meta-hypotheses that have already been considered, ensuring that the metareasoning cycle halts.

Implausible Explainers

A report or belief x might be anomalous because at least one possible explanation in the contrast set for x does not meet the minimum plausibility requirement η. These implausible possible explanations of x are characterized by $E = \{\mathbf{y} | \mathbf{y} \text{ is a possible explanation of } x \wedge Pl(\mathbf{y}) < \eta\}$. The corresponding resolution is to perform abductive reasoning again but allow members of E to explain x, i.e., ignore the minimum plausibility requirement η just in order to find an explanation for x.

For each anomaly x that might be anomalous due to implausible explainers (i.e., $E \neq \emptyset$), a meta-hypothesis is established. The meta-hypothesis is a possible explanation of x only if the resolution, when applied in a simulated environment, actually produces an explanation for x. The plausibility of the meta-hypothesis is estimated by $[Pl(x) + Pl(\mathbf{y})]/2$, where $\mathbf{y}$ is the most plausible member of E. This plausibility estimate has been empirically shown to work reasonably well.

Incompatible Explainers

Another possible cause of an anomaly x is that each possible explanation $\mathbf{y}$ of x is already disbelieved (i.e., $\bar{\mathbf{y}} \subseteq \mathbf{b}$ for every such $\mathbf{y}$), because some other explanation $\mathbf{z}$, which is incompatible with $\mathbf{y}$, was accepted to explain evidence x'. These incompatible explanations of x are characterized by the set of pairs $E = \{(\mathbf{y}, \mathbf{z})\}$ where $\mathbf{y}$ is a possible explanation of x that is incompatible with some accepted explanation $\mathbf{z}$. The corresponding resolution is to perform abductive reasoning again after retracting the acceptance of $\mathbf{z}$ and removing it from consideration.

For each anomaly x that might be anomalous due to incompatible explainers, a meta-hypothesis is established that is said to explain x only if the resolution yields an explanation for x. Its plausibility is estimated by $1 - \delta$ where δ is the decisiveness of the contrast set in which $\mathbf{y}$ (the incompatible accepted explainer) was most plausible.

Insufficient Evidence

The last possible cause of an anomaly x is that the contrast set for x is not sufficiently decisive, i.e., $\delta < \Delta$ Threshold, indicating there is not enough evidence supporting one possible explanation over others. Perhaps if more evidence is obtained, some possible explanation with become sufficiently more decisive than the alternatives, or new possible explanations will be found.

For each anomaly x whose contrast set was not sufficiently decisive, a meta-hypothesis is established with estimated plausibility $(1 - \delta) * Pl(x)$. The corresponding resolution involves gathering more evidence about variables that might shed more insight on the causes of x. Relevant variables are the parents and children of x. However, these other variables are not necessarily observable; even if some reports about these variables are obtained, these reports might be noisy. Furthermore, in practice, observing states of variables might be costly. It might involve reorienting sensors or asking humans for feedback. Thus, caution must be exercised when evaluating whether an anomaly is due to insufficient evidence. The other two possible causes and resolutions of anomalies examined and manipulated the agent's doxastic state (a relatively cheap operation) and did not interact with the world. An insufficient evidence meta-hypothesis is different in this way, and its resolution is only attempted if it is the estimated to be the most plausible explanation for an anomaly.

Noise Detection

Not all anomalies should trigger revision of the doxastic state or gathering more evidence. Some reports might be unexplainable because they are false, i.e., noisy reports. Though some of these noisy reports might be explainable by accepting implausible but false explainers, for example, the correct action is to ignore the anomalous reports and simply leave them unexplained. When no meta-hypothesis is a possible explanation of an anomaly, the anomaly is deemed to be noise, and left unexplained. In this way, noise is a "fall-back" explanation of an anomaly, employed only when no other explanation is forthcoming.

Experimental Methodology

A Bayesian network representing probabilistic knowledge is generated randomly for each experiment. Random generation includes random network configurations and variable states, and random conditional probability tables. Averaged across 360 randomly generated networks, the networks have 33.7 variables (ignoring constraint variables), with standard deviation (s.d.) of 5.2, and 1.3 parent variables (s.d. = 0.1). The network has an average depth of 6.3 (s.d. = 1.3), and contains 3.0 (s.d. = 2.2) incompatible variable state pairs.

Some experiments include different variations of the quality of plausibility estimates. Plausibilities are calculated either as posteriors or estimated priors, which were defined earlier. Plausibility values are represented either as decimal values (with precision to two decimal places) or on a 5-point scale, i.e., 0 = very implausible, 0.25 = implausible, 0.5 = uncertain plausibility, 0.75 = plausible, 1.0 = very plausible. When the 5-point scale is in use, the original plausibility value is translated to the nearest value on the 5-point scale. When plausibilities are represented as decimal values, we say that they have granularity = 100; when the 5-point scale is used, we say they have granularity = 5.

Noise

On average, under noise-free conditions, 16.5 (s.d. = 1.7) variable states are reported, with abductive reasoning performed after each batch of between 1 and 5 reports. These reported states are randomly chosen and may come from anywhere in the network. They are taken from the sampled variable states that represent the ground truth. When the noise level is set to $n\%$, each report has an $n\%$ chance of being subject to perturbation or deletion, in order to simulate noise. We utilize four types of noise:

Distortion noise: A true variable state x is reported as $\bar{x}$.

Duplication noise: Along with a true report x, the state $\bar{x}$ is also reported.

Insertion noise: A random variable Y, with true state y, is reported to have state $\bar{y}$. The true state y might or might not be reported as well.

Deletion noise: An observation is simply deleted from the set of reported observations.

We say that the agent has detected a noisy report if it fails (refuses) to explain the report. Each of distortion, duplication, and insertion noise may be detected by the agent, but deletion noise cannot be detected in this manner since there is no corresponding report.

Metrics

We refer to the following metrics.

Accuracy: The percent of beliefs in the agent's final belief state that are true according to the direct sample of variable states that constitutes the ground truth.

MPE Accuracy: The percent of variable states in the most probable explanation (MPE) that are true. All reports (some of which might be noisy) are treated as observed states when calculating the MPE.

Coverage: The percent of beliefs, not including reports, that require explanation and are in fact explained.

AccCov Mean: $2 * \text{Accuracy} * \text{Coverage}/(\text{Accuracy} + \text{Coverage})$, i.e., the harmonic mean of Accuracy and Coverage.

Other metrics measure the accuracy of noise detection. Let $\mathbf{r} = \mathbf{r}_T \cup \mathbf{r}_F$ be the set of reports, with $\mathbf{r}_T$ the set of true reports and $\mathbf{r}_F$ the set of false reports (noise). Let $\mathbf{n} \subseteq \mathbf{r}$ be the set of reports that the agent either disbeliefs (r is reported but the agent believes $\bar{r}$) or leaves unexplained. This set $\mathbf{n}$ is the set of reports that the agent believes to be noise.

Noise True Positives:
$\text{TP} = \|\mathbf{n} \cap \mathbf{r}_F\|$, noise correctly identified

Noise False Positives:
$\text{FP} = \|\mathbf{n} \cap \mathbf{r}_T\|$, true reports deemed to be noise

Noise True Negatives:
$\text{TN} = \|\mathbf{b} \cap \mathbf{r}_T\|$, true reports correctly identified

Noise False Negatives:
$\text{FN} = \|\mathbf{b} \cap \mathbf{r}_F\|$, true reports deemed to be noise

Noise False Positive Rate: Noise $\text{FPR} = \text{FP}/(\text{FP} + \text{TN})$

Noise True Positive Rate: Noise $\text{TPR} = \text{TP}/(\text{TP} + \text{FN})$

Noise Mean: The harmonic mean of (1 - Noise FPR) and Noise TPR.

Note that noise detection accuracy is shown with receiver operating characteristic (ROC) curves as a way of visualizing the trade off between low Noise FPR and high Noise TPR.

Experimental Hypotheses

The following hypotheses guide our experimental investigations:

Hypothesis I: Abductive reasoning with abductive metareasoning gives good performance in terms of accuracy, coverage, and noise detection. Metareasoning gives better performance on these metrics compared to no metareasoning. Best performance is achieved when the minimum plausibility $\eta > 0$ and Δ Threshold > 0, thus demonstrating the utility of these parameters. Finally, each of the three types of meta-hypotheses were accepted, at least in some cases, proving that each kind of meta-hypothesis has a useful role.

Hypothesis II: Abductive reasoning with abductive metareasoning yields greater Accuracy than MPE Accuracy, particularly in noisy conditions and parameters $\eta > 0$ and Δ Threshold > 0. This is because the MPE method does not exercise caution when assigning variable states; abductive reasoning and metareasoning, on the other hand, only accept possible explanations if they are plausible and decisive. Additionally, MPE does not distinguish between true and noisy reports, but rather includes all reports in the most probable explanation.

Hypothesis III: Good performance is maintained under various conditions: more and less noise, reduced plausibility granularity, and less accurate plausibilities using estimated priors instead of posteriors. This outcome would demonstrate the robustness of the approach.

Hypothesis IV: When plausibility estimates are missing or, equivalently, $Pl(\mathbf{y}) = 1$ for every possible explanation $\mathbf{y}$, performance is significantly and strongly degraded. This

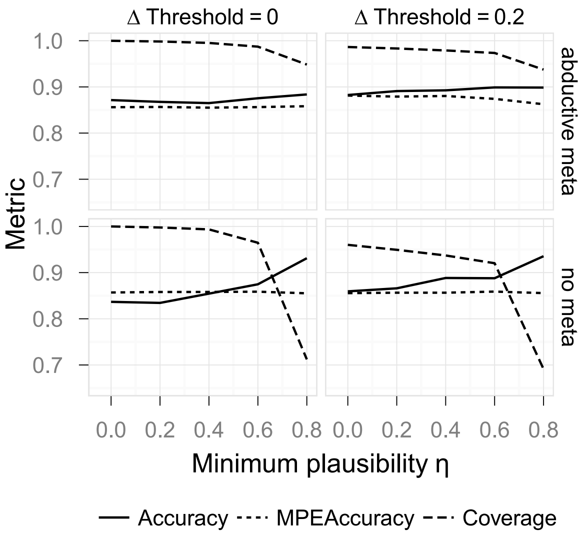

Figure 2: Accuracy, MPE Accuracy, and Coverage metrics for a range of minimum plausibility η settings and two Δ Threshold settings. Plausibilities were calculated as posteriors, and the noise level was 20%.

outcome would show that the probabilistic model is contributing information about plausibility and not just explanatory relations as determined solely by the network structure.

Results

Hypothesis I

Figures 2 and 3 support the following conclusions:

- Abductive reasoning with abductive metareasoning performs well. Maximum performance for both AccCov Mean and Noise Mean is found at Δ Threshold $= 0.2$ and minimum plausibility $\eta = 0.6$. In these cases, average AccCov Mean is 0.93 (s.d. = 0.05), with Accuracy 0.90 (s.d. = 0.08) and Coverage 0.97 (s.d. = 0.04). Average Noise TPR for the same parameters is 0.80 (s.d. = 0.26), Noise FPR is 0.19 (s.d. = 0.11), and Noise Mean 0.78 (s.d. = 0.20).
- When abductive metareasoning is active, Δ Threshold $>$ 0 gives better Accuracy, at the cost of Coverage, and minimum plausibility $\eta > 0$ gives better noise detection.
- Metareasoning gives better accuracy, coverage, and noise detection than no metareasoning. The differences are minor but significant. With Δ Threshold $= 0.2$ and $\eta = 0.6$, abductive metareasoning yields an average of 0.01 greater Accuracy, 0.05 greater Coverage, 0.03 greater AccCov Mean, 0.03 greater Noise TPR, 0.09 less Noise FPR, and 0.09 greater Noise Mean ($p < 0.01$ for each metric except Noise TPR, which has $p < 0.05$).

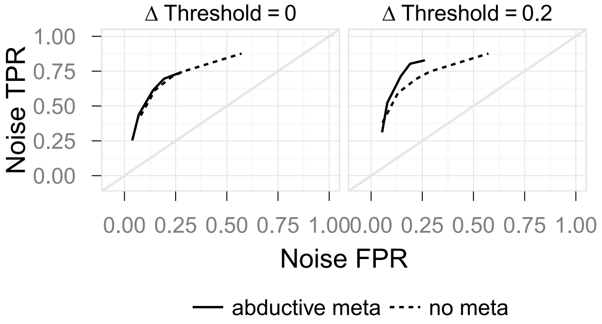

Figure 3: ROC curves for noise detection. Plausibilities were calculated as posteriors, and the noise level was 20%. The curves are drawn according to increasing minimum plausibility η setting, from 0 to 0.8. A line is shown where FPR=TPR, representing the expected accuracy of a random noise detection process.

η	Accuracy (R^2)	MPE Acc. (R^2)	Noise FPR (R^2)
0	-0.0041 (0.57)	-0.0040 (0.63)	0.0029 (0.50)
0.2	-0.0030 (0.41)	-0.0038 (0.61)	0.0021 (0.35)
0.4	-0.0020 (0.20)	-0.0037 (0.59)	0.0016 (0.11)
0.6	-0.0013 (0.09)	-0.0035 (0.58)	0.0014 (0.08)
0.8	-0.0010 (0.05)	-0.0033 (0.53)	0.0015 (0.06)

Table 1: Impact of increased noise level % on Accuracy, MPE Accuracy, and Noise FPR metrics for different minimum plausibility η values. Only the Accuracy, MPE Accuracy, and Noise FPR metrics were significantly impacted by increased noise level. Δ Threshold $= 0.2$ in all cases. The ratio and R^2 values are derived from linear regression models. A ratio r means that for every 1% increase in the noise level, the metric increased by r.

- Each meta-hypothesis was accepted in order to explain anomalies. With Δ Threshold $= 0.2$ and $\eta = 0.6$, for example, the incompatible explainers meta-hypothesis was accepted an average of 2.19 times (in a single experimental run), with s.d. = 0.17; implausible explainers was accepted 5.04 times (s.d. = 0.42); and insufficient evidence was accepted 9.48 times (s.d. = 0.49).

These outcomes essentially validate the abductive reasoning and metareasoning system design and operation.

Hypothesis II

As seen in Figures 2 and 4, abductive reasoning generally outperforms the most probable explanation (MPE) for the same Bayesian network in terms of Accuracy. This is true in all represented parameters except for the combination of low values of η, no Δ Threshold, and either no noise or no metareasoning. Across all cases represented in Figure 4, abductive reasoning with metareasoning yields, on average, Accuracy greater than MPE Accuracy by 0.03, and the difference is significant ($p < 0.001$).

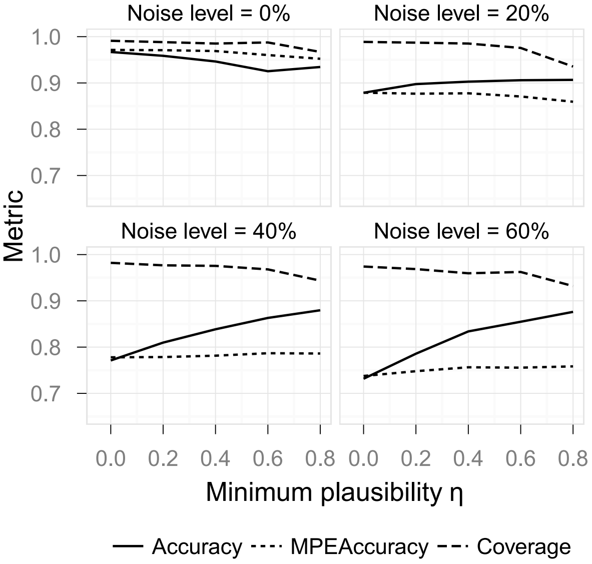

Figure 4: Accuracy, MPE Accuracy, and Coverage metrics for abductive reasoning with abductive metareasoning and a range of noise levels. Plausibilities were calculated as posteriors. In every case, Δ Threshold $= 0.2$.

Hypothesis III

Figure 5 supports the claims of Hypothesis III that abductive reasoning with abductive metareasoning maintains good performance even when plausibilities are estimated according to the estimated priors calculation, and not the more accurate Bayesian posterior calculation. Furthermore, performance is maintained in experiments with lower plausibility granularity, in which plausibilities are represented as one of five discrete values (very implausible, implausible, uncertain plausibility, plausible, very plausible; or, equivalently, [0.0, 0.25, 0.5, 0.75, 1]). This result agrees with previous findings (Pradhan et al. 1996).

Table 1 shows that increased noise degrades performance, but the impact is reduced with larger minimum plausibility η values. Nevertheless, the impact on performance is small. For example, with $\eta = 0.6$, a 10% increase in noise only decreases Accuracy by 0.013 (average value at 20% noise is 0.899) and increases Noise FPR by 0.014 (average value at 20% noise is 0.19). Coverage and Noise TPR are not significantly affected by the noise level. Furthermore, MPE Accuracy suffers more than the Accuracy achieved abductive reasoning with metareasoning. Note that the MPE Accuracy slightly responds to changes in minimum plausibility η. This is because the MPE is calculated according to all acquired reports, and when η is high, more anomalies are manifested, and metareasoning sometimes acquires more reports by applying the insufficient evidence resolution.

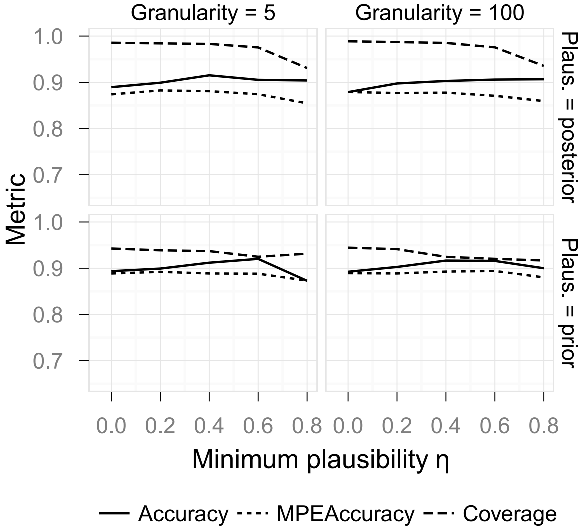

Figure 5: Accuracy and Coverage metrics for abductive reasoning with abductive metareasoning, Δ Threshold $= 0.2$. The plausibility calculation and granularity varied. The noise level was 20%.

Hypothesis IV

To evaluate whether mostly-accurate plausibility estimates are beneficial for abductive reasoning with abductive metareasoning, we compare performance with normal posterior plausibilities to performance with no plausibilities, or equivalently, every plausibility is set to 1. This change effectively eliminates any role for the minimum plausibility η parameter, as well as the Δ Threshold parameter since $\delta = 0$ in all contrast sets except those that have only one possible explanation (and hence $\delta = 1$). Performance was significantly worse across 240 cases, compared against abductive reasoning with metareasoning and $\eta = 0.6$, Δ Threshold $= 0.2$. Each experiment included a 20% chance of noise. Accuracy decreased, on average, by 0.14, Coverage increased by 0.03, Noise TPR decreased by 0.50, and Noise FPR decreased by 0.15 (for each metric, $p < 0.001$).

Related Work

Our use of the term *explanation* differs from other uses in the context of Bayesian networks. For example, the term sometimes refers to the most probable explanation (MPE), which is a complete assignment of variable states, or the maximum *a posteriori* (MAP) assignment of variable states for a subset of all variables. However, MPE suffers from the “overspecification problem” (Shimony 1993), as does MAP if too many variables are chosen to be members of the explanation set. The overspecification problem arises when variables that are irrelevant, e.g., downstream effects of the observations, are included in the explanation.

Yuan and Lu (2007) develop an approach they call the *Most Relevant Explanation* (MRE) and refine it in subsequent work (Yuan, Lim, and Littman 2011). The MRE "aims to automatically identify the most relevant target variables by searching for a partial assignment of the target variables that maximizes a chosen relevance measure." One such relevance measure is likelihood, though the authors show that the generalized Bayes factor (Fitelson 2007) provides more discriminative power. Our approach is similar in that we also evaluate partial assignments of parent (target) variables and evaluate each assignment according to a kind of relevance measure. In our system, when a possible explanation is sufficiently plausible and decisive, it is considered relevant, and if it is compatible with existing beliefs, it is accepted. Yuan and Lu's MRE differs from our system, however, in that the target variables for the MRE must be established ahead of time, while the abductive reasoning system automatically seeks explanations for unexplained reports and beliefs.

The variety of definitions for what constitutes an explanation in a Bayesian network illustrates the difficulty in using nothing more than the probability calculus to infer commonsense causal explanations. It is apparent that the most common use of a Bayesian network is not to arrive at commonsense causal explanations, as we have done, but rather to determine the most probable state assignments for a set of variables (the maximum *a posteriori*, or MAP, variable assignments). Abductive reasoning, as described in this work, explicitly seeks causal explanations for reports. The causal explanations found by abductive reasoning do not necessarily match the MAP variable assignments; i.e., they might not be the most probable assignments and might commit to different states for the same variables. On the other hand, abductive reasoning attempts to maximize explanatory coverage, while preferring more plausible explainers, by operating with a notion of *unexplained evidence* that is lacking in Bayesian MAP inference. The variables that make up a MAP assignment must be chosen ahead of time, but the probability calculus does not offer any insight. Other factors, such as the causal relations represented in a Bayesian network, should inform the process of determining which variables contribute to an explanation. Abductive reasoning exploits both the causal relations and plausibility of variable states in order to arrive at commonsense explanations of evidence, while a purely Bayesian MAP process does not.

Finally, the combined abductive reasoning and abductive metareasoning system used in our experiments was previously detailed and analyzed (Eckroth and Josephson 2013). In that work, rather than a Bayesian network the system was tasked with finding explanations for sensor reports of moving objects. Both simulated and real-world aerial tracking experiments were conducted, and similar performance gains were found with abductive metareasoning. The reasoning and metareasoning systems are domain-general, and experiments with object tracking and Bayesian networks use exactly the same inference algorithms and code-base.

Conclusion

Each of our experimental hypotheses is strongly supported by this work. They tell us that commonsense abductive reasoning and metareasoning is a very effective strategy for using Bayesian networks to infer the true causal explanations of reports. Due to the general usefulness of Bayesian networks, we expect that our system will prove beneficial in a variety of tasks. Further work aims to demonstrate its benefits in real-world applications.

Acknowledgments

We thank Norsys Software Corp. (www.norsys.com) for their development and support of Netica, which was used extensively in this work for efficient probabilistic reasoning with Bayesian networks.

References

Bylander, T.; Allemang, D.; Tanner, M.; and Josephson, J. 1991. The computational complexity of abduction. *Artificial Intelligence* 49(1-3):25–60.

Cooper, G. F. 1990. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial intelligence* 42(2):393–405.

Cox, M., and Raja, A. 2011. Metareasoning: An introduction. In Cox, M. T., and Raja, A., eds., *Metareasoning: Thinking about thinking*. MIT Press. chapter 1, 3–14.

Crowley, M.; Boerlage, B.; and Poole, D. 2007. Adding local constraints to Bayesian networks. *Advances in Artificial Intelligence* 344–355.

Eckroth, J., and Josephson, J. R. 2013. Anomaly-driven belief revision by abductive metareasoning. In *Proceedings of the Second Annual Conference on Advances in Cognitive Systems*, 73–90.

Fitelson, B. 2007. Likelihoodism, Bayesianism, and relational confirmation. *Synthese* 156(3):473–489.

Josephson, J. R., and Josephson, S. G. 1994. *Abductive inference: Computation, philosophy, technology*. Cambridge University Press.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Perlis, D. 2011. There's no me in "meta"—or is there? In Cox, M. T., and Raja, A., eds., *Metareasoning: Thinking about thinking*. MIT Press. chapter 2, 15–26.

Pradhan, M.; Henrion, M.; Provan, G.; Del Favero, B.; and Huang, K. 1996. The sensitivity of belief networks to imprecise probabilities: An experimental investigation. *Artificial intelligence* 85(1):363–397.

Shimony, S. E. 1993. The role of relevance in explanation I: Irrelevance as statistical independence. *International Journal of Approximate Reasoning* 8(4):281–324.

Yuan, C., and Lu, T. C. 2007. Finding explanations in Bayesian networks. In *The 18th International Workshop on Principles of Diagnosis*, 414–419.

Yuan, C.; Lim, H.; and Littman, M. L. 2011. Most relevant explanation: Computational complexity and approximation methods. *Annals of Mathematics and Artificial Intelligence* 1–25.