Proceedings of the Second Annual Conference on Advances in Cognitive Systems ACS-2013 (73-90)

Anomaly-Driven Belief Revision by Abductive Metareasoning

Joshua Eckroth ECKROTH@CSE.OHIO-STATE.EDU
John R. Josephson JJ@CSE.OHIO-STATE.EDU
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210 USA

Abstract

Cognition may reasonably be distinguished into world-estimation and planning tasks. Our focus in this work is the world estimation task. One aspect of being smart in this task is noticing one's mistakes and correcting them. A smart world estimator may be realized by dividing its operation into a base-level reasoning system and a metareasoning system. The base-level system is responsible for processing inputs from the world and recording its conclusions in a doxastic state, which maintains the system's beliefs or world estimate. The metareasoning system monitors the base-level system so that it can detect symptoms of errors in the doxastic state and attempt belief revisions. In this work, we describe and evaluate such a system. The base-level system is an abductive reasoner responsible for finding explanations for inputs that take the form of reports of putative observations. When no plausible, consistent explanation is forthcoming for some reports, we say these reports are anomalous. The presence of anomalies is a symptom of errors in the doxastic state since, in the usual case, all reports should be explainable. However, sometimes the anomalous reports are not reports of observations but rather false or noisy reports. An abductive metareasoning facility attempts to determine whether anomalies are caused by noise or errors, and if the latter, applies the corresponding revisions. We evaluate this two-level system in a pair of object tracking tasks, one simulated and one based on aerial surveillance. Both tasks are challenging due to limited sensor capabilities and a very high level of noise. The proposed two-level system exhibits significantly improved world estimation accuracy and noise detection, as compared to a system that lacks abductive metareasoning.

1. Introduction

Cognitive systems that include a metareasoning or self-reflective 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 & Raja, 2011). Such architectures typically consist of a base-level reasoner, which reasons about object-level concepts, and a meta-level reasoner that monitors and manipulates the base-level reasoner. In this work, we describe and evaluate a cognitive system that is divided into a *abductive* base-level reasoner and an *abductive* metareasoner. The base-level abductive reasoner receives reports of putative observations and attempts to infer their true explanations, i.e., what caused the phenomenon that was observed and reported. The abductive reasoning is domain-general reasoning, concerning itself with generic evidence and hypotheses rather than object-level phenomena. In this work, domain-

© 2013 Cognitive Systems Foundation. All rights reserved.

specific concepts are encapsulated in a separate module that appears as a black box to the base-level reasoner (and metareasoner). Sometimes, no plausible, consistent explanations for some reports are generated by the domain-specific module, leaving those reports unexplainable. We call such reports anomalous and describe how they may be symptoms of errors in the system's world estimate.

The presence of anomalies activates the metareasoning system, which treats anomalies as evidence of possible errors and attempts to determine their causes. Thus, the metareasoning system is itself abductive, and uses the same machinery (algorithms and code) as the base-level reasoner. We enumerate the possible causes of anomalies that are specific to abductive base-level reasoners. Besides errors in the doxastic state, an anomaly might be caused by noise, i.e. a false report that is anomalous partly because it is false. Thus, an anomaly does not always indicate an error in the doxastic state. In our system, abductive metareasoning concludes that an anomaly is the result of noise as a fallback explanation when no other explanation of the anomaly is found that is consistent or sufficiently plausible. Thus, noise is not explicitly detected by domain-specific properties, but rather implicitly detected by way of domain-general considerations. We demonstrate that this two-level abductive system is effective at detecting and correcting errors, and detecting noise, even when faced with limited sensor capabilities and a very high level of noise.

Although other cognitive systems include an anomaly-driven metareasoning component to guide belief or theory revision and learning (Schmill et al., 2011; Bridewell, 2004; Cox & Ram, 1999), the system discussed here is, to our knowledge, the first to apply abductive metareasoning to abductive reasoning where both the base-level and metareasoning components use the same machinery. Some benefits of this combined system are its generality, simplicity, and effectiveness at its task.

The remainder of this paper is organized as follows. We present an abductive reasoning system in the next section. This is followed in Section 3 by an examination of the metareasoning system, which outlines the possible causes of anomalies in the base-level abductive reasoning system, and their corresponding belief revisions. Section 4 describes the experimental domains and is followed by Section 5 with the experimental methodology and Section 6 with experimental results. The final two sections discuss related work and offer concluding remarks.

2. Abductive Reasoning

By *abduction*, and *abductive inference*, we mean reasoning that follows a pattern approximately as follows (Josephson & 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.

2.1 Doxastic States

In order to keep track of evidence, possible explanations (equivalently, *hypotheses*), and the plausibility and status of each hypothesis, we construct a *doxastic state* as characterized in Definition 2.1. Note that, for simplicity's sake, we treat reports as hypotheses that explain nothing but are themselves considered unexplained if they are accepted. Such hypotheses are initially accepted, but may be rejected (ignored) during metareasoning if they are subsequently deemed to be noise.

Definition 2.1. A *doxastic state* is a tuple $D = (H, X, P, S, V, I)$, where $H = \{h_1, \ldots, h_n\}$ is a (finite) set of hypotheses and X is a relation over $H \times H$, where $(h_j, h_i) \in X$ means h_j could explain h_i. The relation X is constrained so that the resulting explanation graph is acyclic. Next, $P : H \to [0, 1]$ is a plausibility function, $S : H \to \{Accepted, Rejected, Undetermined\}$ gives the belief status of a hypothesis, and $V \subseteq H$ is a set of evidence hypotheses that, when accepted, are considered to require an explanation. I is an irreflexive, symmetric relationship over H where $(h_j, h_i) \in I$ means h_j is incompatible with h_i and vice versa. The sets I and X are constrained so that $X \cap I = \emptyset$. Additionally, $(\forall (h_j, h_i) \in I)(S(h_j) = Accepted \to S(h_i) = Rejected)$. We say that if $(h_j, h_i) \in X$ and $S(h_j) = S(h_i) = Accepted$, then h_j explains h_i and h_i is explained by h_j. This use of *explained* and *explained by* does not require or imply that either h_j or h_i is unique in its respective role.

We do not require that a *could explain* relation $(h_j, h_i) \in X$ be interpretable as material implication or logical entailment, as is sometimes the case in other treatments of abduction (Aliseda, 2006; Kakas, Kowalski, & Toni, 1992). Instead, we suppose that hypotheses represent possible causal relations (h_j is a possible cause of h_i). Furthermore, these causal relations are not necessarily predictive. It is not presumed that h_j is a sufficient condition for h_i.

In order to experimentally evaluate abductive reasoning and metareasoning across different problem domains, we separate the problem domain from the reasoning system. The system architecture is shown in Figure 1. A problem domain is defined as follows.

Definition 2.2. A *problem domain* M is an opaque structure such that the functions OBSERVE and GENERATEHYPOTHESES are defined as follows.

$$\text{OBSERVE}(M, F) = (H_{\text{reports}}, P, I). \tag{1}$$

The OBSERVE function generates reports H_{reports} of observed properties of the world. These reports come with plausibilities defined by P and incompatibility relations defined by I. The set F may be used to focus the observations on particular features of the world. As used by abductive metareasoning, detailed in Section 3, F is a set of unexplainable reports. When $F = \emptyset$, the OBSERVE

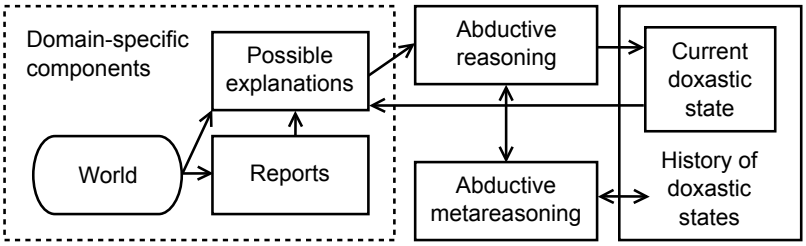


Figure 1. System architecture. Domain-specific components are separate from domain-general reasoning and metareasoning components. Reports, which might be noisy, are obtained from the world. The plausibility of each report is calculated according to domain knowledge and current beliefs. Each report requires explanation, as do any unexplained beliefs. Possible explanations are generated by a domain-specific function and then reviewed (accepted or rejected) by the abductive reasoning procedure. 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 ask the domain-specific components to generate new hypotheses, revise the belief state, and/or leave the anomalies unexplained. Reports that remain unexplained are deemed to be noise.

function can be thought of a general scan of the world without a particular focus. In an object tracking domain, this general scan might report highly plausible object detections that survive a high-threshold noise filter. In a medical diagnosis domain, this general scan might contain the set of answers to typical questions about the patient's symptoms and history. When $F \neq \emptyset$, it contains unexplainable reports that might require corroborating reports or more detailed reports in order to be explainable. For example, F might contain results from a medical test that, on their own, have no plausible explanation, but might be better explained if results are gathered from a followup test.

$$\text{GENERATEHYPOTHESES}(M, H_{\text{unexplained}}, H_{\text{accepted}}) = (H_{\text{hypotheses}}, X, P, I). \tag{2}$$

The GENERATEHYPOTHESES function produces new hypotheses that purport to explain reports in $H_{\text{unexplained}}$ such that these explanations are consistent with accepted hypotheses H_{accepted}. The new explanatory relations, plausibilities of the generated hypotheses, and incompatibility relations are given by X, P, and I, respectively.

Abstractly, we can think of M as containing the knowledge about a problem domain. As far as the abductive reasoning algorithms are concerned, M and its corresponding functions are black boxes. This separation enables us to experiment with different domains without modifying the reasoning engine.

Abductive reasoning, as used in our system, is characterized by the following definitions.

Definition 2.3. A *partial abduction* is an operation on doxastic states where $\text{PARTIALABDUCE}(D_1) = D_2$ means D_2 is produced from D_1 either by changing nothing ($D_1 = D_2$) or *accepting* one hypothesis and *rejecting* incompatible hypotheses, should any exist. A partial abduction meets the following constraints.

1. If no unexplained evidence have hypotheses, or no evidence is unexplained, then the doxastic state is unchanged.
2. If there exists a hypothesis that is undecided (not accepted or rejected) for some unexplained evidence and is sufficiently plausible to be accepted, then some hypothesis is accepted.
3. One or zero hypotheses are accepted.
4. When a hypothesis is accepted, all incompatible hypotheses are rejected.

Lemma 2.1. If $\text{PARTIALABDUCE}(D_1) = D_2$, then either $\text{UNEXPLAINED}(D_2) \subset \text{UNEXPLAINED}(D_1)$ or $D_1 = D_2$. In other words, the PARTIALABDUCE function, applied to a doxastic state, reduces the number of unexplained reports or leaves the doxastic state unchanged.

Proof sketch. Assume $\text{PARTIALABDUCE}(D_1) = D_2$ and suppose $D_1 \neq D_2$. Then some hypothesis h was accepted by the partial abduction operation, and the reports explained by h are no longer unexplained. □

Definition 2.4. A doxastic state D is *finalized* if $\text{PARTIALABDUCE}(D) = D$.

Algorithm 1 The FINALIZE and ABDUCE functions.

function FINALIZE(D)
  $D' \leftarrow \text{PARTIALABDUCE}(D)$
  while $D' \neq D$ **do**
    $D \leftarrow D', \quad D' \leftarrow \text{PARTIALABDUCE}(D)$
  end while
  return D'
end function
function ABDUCE($M, D_0, \eta, DoMetareasoning?$)
  $(H_{\text{reports}}, P, I) \leftarrow \text{OBSERVE}(M, \emptyset)$
  $D_1 \leftarrow \text{ADDREPORTSTODOXASTICSTATE}(D_0, H_{\text{reports}}, P, I)$
  $H_{\text{unexplained}} \leftarrow \text{UNEXPLAINED}(D_1), \quad H_{\text{accepted}} \leftarrow \text{ACCEPTED}(D_1)$
  $(H_{\text{hypotheses}}, X', P', I') \leftarrow \text{GENERATEHYPOTHESES}(M, H_{\text{unexplained}}, H_{\text{accepted}})$
  $D_2 \leftarrow \text{ADDHYPOTHESESTODOXASTICSTATE}(D_1, H_{\text{hypotheses}}, X', P', I', \eta)$
  $D_3 \leftarrow \text{FINALIZE}(D_2)$
  if $DoMetareasoning?$ **then**
    $D_4 \leftarrow \text{METAREASON}(D_3)$ ▷ Refer to Section 3
    return D_4
  else
    return D_3
  end if
end function

The abductive reasoning procedure, defined by the ABDUCE function (Algorithm 1), is responsible for obtaining evidence and hypotheses, and by way of the FINALIZE function, iteratively calling the PARTIALABDUCE function until the doxastic state is finalized. The algorithm is parameterized in part by η, the minimum plausibility threshold for a possible explanation to be initially considered. This parameter η is taken into account by the ADDHYPOTHESESTODOXASTICSTATE function, which rejects hypotheses (after adding them to the doxastic state) whose plausibilities are less than η.

Theorem 2.1. The FINALIZE function is guaranteed to terminate.

Proof. From lemma 2.1, we have that either the partial abduction leaves a doxastic state unchanged, causing termination of the loop, or the partial abduction reduces the set of unexplained evidence. Since the set of hypotheses, which includes the evidence, is finite, it follows that the set of unexplained evidence is finite, and therefore the algorithm is guaranteed to terminate. □

2.2 The EFLI 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 (the set I in our definition of the doxastic state) cannot efficiently be solved. Specifically, they prove that it is NP-complete to determine whether a consistent, complete set of explanations exists for such an abduction problem. They also prove that it is NP-hard to find a most-plausible consistent and complete set of explanations.

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 hypothesis and rejects incompatible hypotheses, until either all evidence is explained or no other hypotheses for the unexplained evidence are available. Hypotheses are grouped into *contrast sets*, where each contrast set contains all the plausible hypotheses for some report. Essential hypotheses are accepted first. An essential hypothesis is the sole member of a contrast set, so it is the only plausible hypothesis for some report. Unless it is accepted, some evidence would remain unexplained. Then, hypotheses are ordered for acceptance by the degree to which the best hypothesis in a contrast set surpasses the second best hypothesis, in terms of plausibility. Although not shown here, the steps in the *EFLI* algorithm satisfy the requirements of a PARTIALABDUCE function.

3. Metareasoning

We call *anomalies* those reports and other evidence that remain unexplained in a finalized doxastic state. The system checks for their presence in a metareasoning function, METAREASON, that is activated from the ABDUCE function (Algorithm 1). In some cases, domain knowledge or background knowledge is insufficient, causing explanations not to be generated for some true reports. However, in this work we assume that domain knowledge is sufficient to generate true hypotheses for all true reports, assuming the current world estimate is accurate. Under this assumption, in domains where

all reports are guaranteed to be true (noise-free conditions), anomalies are necessarily the result of errors. However, in more realistic environments, not all unexplainable reports are true reports; some reports might be unexplainable partly because they are noise and do not warrant any explanation. Part of the challenge of metareasoning is to identify which reports are unexplainable due to errors in the doxastic state and which are due to false reports (and hence, not due to errors). 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*, as produced by a virtual problem domain M_{meta}. Such a metareasoning system is able to use the same abductive reasoning machinery employed by the base-level reasoner.

The following four sections detail the possible causes of anomalies, their corresponding belief revisions, and the criteria that tell us (as experimenters) whether the revision is correct. Note that an anomaly might have multiple possible causes. The METAREASON function handles the following tasks. For each possible cause, a meta-hypothesis is generated, which specifies the cause, the revision, the subset of anomalies it is said to explain, and an estimated plausibility score. The meta-hypotheses are added to a *meta-doxastic state*, and abductive reasoning commences, yielding a set of accepted meta-hypotheses. The belief revisions that are specified by the accepted meta-hypotheses are applied to the original doxastic state, which is then finalized. If any anomalies remain (or new anomalies appear), metareasoning is activated again on the new doxastic state. Care is taken not to generate meta-hypotheses that have already been evaluated. This ensures that the procedure halts, although we do not provide a proof here.

In the following sections, we use the following notation. Let D be a doxastic state, $A = \text{ANOMALIES}(D)$, and $H_A = \bigcup_{h \in A} \text{HYPOTHESES}(D, h)$, i.e., the set of possible explanations of anomalies, which if any exist, necessarily were rejected.

3.1 Implausible Hypotheses

Some reports may be anomalous due to the rejection of one or more implausible hypotheses. These rejected hypotheses are characterized by $H_P = \{h | h \in H_A \wedge P(h) < \eta\}$, where P is the plausibility function of D and η is the minimum plausibility threshold. Unrejecting one or more of these implausible hypotheses, thus possibly allowing their acceptance, might eliminate some anomalies.

For each $h \in H_P$, we hypothesize that the rejection of h is responsible for some reports having no explanation. The plausibility is estimated by $p = \left[P(h) + \sum_{r \in R} P(r)\right] / (\|R\| + 1)$, where $R \subseteq A$ contains the anomalies that are eliminated by applying the revision and finalizing the resulting doxastic state. This estimate is higher when the plausibility of h and what it explains are higher. We have found that this plausibility estimate works reasonably well in empirical studies. For evaluating experimental results, we stipulate that the revision is correct if $R \neq \emptyset$, a greater number of the eliminated anomalies in R are true evidence rather than false, and the hypothesis h that is unrejected is a true hypothesis.

3.2 Incompatible Hypotheses

Some reports may be anomalous due to some of the hypotheses being rejected upon the acceptance of other hypotheses. Let $H_I = \{h | h \in \text{ACCEPTED}(D) \wedge \text{INCOMPATIBLE}(D, h) \cap H_A \neq \emptyset\}$.

The set H_A contains all possible explanations of the anomalies. The set H_I contains accepted hypotheses that are incompatible with members of H_A. An accepted hypothesis $h \in H_I$ may have been responsible for rejecting some possible explanations of some anomalies; thus, if the status of h is changed to undecided and then rejected (to prevent it from being accepted again), some anomalies might be eliminated.

For each $h \in H_I$, we hypothesize that the acceptance of h is responsible for some reports having no explanation. The set $R \subseteq A$ contains the anomalies that actually are eliminated by rejecting h. The plausibility is estimated by $p = (1 - P(h)) * \left(\sum_{r \in R} P(r)/\|R\|\right)$. This estimate is higher when h is less plausible and the anomalous reports are more plausible. We count the revision as correct if $R \neq \emptyset$, a greater number of the eliminated anomalies in R are true than false, the hypothesis h that is to be rejected is false, and some hypothesis that h had precluded is in fact true.

3.3 Insufficient Evidence

The first two possible causes of anomalies, implausible hypotheses and incompatible hypotheses, describe scenarios where some reports have known possible explanations but those hypotheses were made unavailable. Other anomalous reports, however, might have no possible explanations at the base-level, so the first two scenarios are inapplicable. An anomaly may have no possible explanations for one of two reasons: either prior accepted explanations limit possible explanations for the new reports to the point that none are offered, or there are insufficient reports to reasonably narrow the set of hypotheses. This latter case will be described first.

Any reasonable implementation of the function GENERATEHYPOTHESES will generate only those hypotheses that can meaningfully explain the reports. Given a dearth of evidence, one would expect a reasonable problem domain not to generate an infinite (or very large) set of overly-specific hypotheses. For example, in a medical diagnosis domain, the single report "headache" might be explainable by any one of (or combinations of) many diseases and ailments. But no bounded rational agent would methodically consider each of these hypotheses. Instead, more evidence would be gathered.

The *insufficient evidence* meta-hypothesis claims that some anomaly is unexplainable because no hypotheses were offered due to insufficient evidence. Thus, the corresponding revision requires first seeking more evidence, preferably reports that corroborate or add detail to the anomalous report in question, and then generating new hypotheses. The plausibility of such a meta-hypothesis is estimated by $p = P(r)$, where r is the anomaly that *might* be eliminated by gathering more evidence. We count the revision as correct if r is true.

Whether or not such a revision is effective (more reports are obtained and new hypotheses for some anomalies are generated and accepted) is not known until the meta-hypothesis is accepted during abductive metareasoning and the revision is applied. This conservative approach is taken because sometimes obtaining more evidence (e.g., performing medical tests) is costly and/or harmful and should be performed only if other meta-hypotheses are ruled out.

3.4 Order Dependency

The final possible cause of an anomaly is that no hypotheses were ever offered due to the cognitive system mistakenly believing that the report is unexplainable (i.e., impossible) given its current world estimate. We suppose that the problem domain's GENERATEHYPOTHESES function is defined to generate only those hypotheses that are consistent with the current doxastic state. Thus, if some of the accepted hypotheses in the doxastic state (which were accepted to explain earlier reports) are false, then the GENERATEHYPOTHESES function might fail to generate hypotheses for true reports.

The *order dependency* meta-hypothesis claims that one or more anomalies have no possible explanations because prior accepted hypotheses were in error, and that they should be reconsidered *in light of* recently-obtained reports. In other words, the anomalies are the result of the particular order the reports were obtained. The revision involves identifying a previous doxastic state to revert to (and thereby erasing recently generated and accepted hypotheses), then injecting recent reports, generating new hypotheses (given the less committed doxastic state and more reports), and finalizing the doxastic state. It is not clear, at the time of this writing, how far back the doxastic state must be reverted. In the experiments detailed here, the system reverts to the doxastic state immediately preceding the introduction of the reports that ultimately proved to be anomalous. The plausibility of such a meta-hypothesis is estimated as $p = \sum_{r \in R} P(r)/\|R\|$, where $R \subseteq A$ is the set of anomalies that *might* be eliminated by reconsidering previously-accepted hypotheses in light of subsequent reports. We say that the revision is correct if more of the eliminated anomalies in R are true than false. Like the insufficient evidence meta-hypothesis, an order dependency revision is not applied until the meta-hypothesis is accepted as an explanation of some anomalies. This is because generating new hypotheses in light of subsequent reports might be costly.

3.5 Noise Detection

Not all anomalies should trigger belief revisions. 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 hypotheses, for example, the correct action is to reject the anomalous reports so that they are no longer considered unexplained evidence. This is achieved by treating the noise hypothesis as a fallback meta-explanation when no other meta-hypothesis is sufficiently plausible. We find that a minimum plausibility threshold for abductive metareasoning, η_{meta}, is effective for filtering out implausible meta-hypotheses. Anomalies that remain unexplained after abductive metareasoning are considered to be noise, and in turn rejected. Because these reports are rejected, they no longer manifest as anomalies.

4. Object Tracking

In this work, we experiment with the described cognitive system in two object tracking domains, one simulated and one based on aerial video surveillance. Although we believe that many different kinds of cognitive systems tasked with different problem solving goals can benefit from an abductive metareasoning system, object tracking tasks expose the usefulness of abductive metareasoning in the following ways.

- The task is easily framed in explicitly abductive terms, in which object detections make up reports and object movements serve as the hypotheses.
- As will be shown, it is practical to establish a minimum plausibility for movement hypotheses, though anomalies due to implausible hypotheses might result.
- In the simulated tracking domain, movement hypotheses are incompatible if they describe the same object in two different locations at the same time. Thus, anomalies resulting from incompatible hypotheses are possible.
- It is often practical for surveillance systems to filter out sensor detections that do not meet a minimum threshold of plausibility in an attempt to filter out noise. In some cases, however, this filtering causes anomalies due to insufficient evidence.
- Future movement hypotheses depend on the system's current estimate of the situation, i.e. its beliefs, which are formed from the acceptance of prior movement hypotheses. Consequently, false beliefs might cause order dependency anomalies.

In the simulated tracking domain, the cognitive system obtains reports about moving objects in a 10x10 discrete grid. This grid constitutes the world, and is fully observable, with one caveat described below. The objects' movements are random walks. At each time step, each object makes a fixed number of random 1-step movements, which we call *grid steps* (diagonals not allowed). Virtual sensors report the final location of each object's walk in that time step. No two objects are allowed to occupy the same grid cell at the completion of their walk. Thus, there is no need to handle merges and splits.

For simplicity, each object bears a unique color. An object's color is a stand-in for any variety of more realistic object features that support its identification. However, the center 50% of the grid is watched only by sensors that do not detect color. All objects in that area are seen as gray, and are therefore indistinguishable. The outer 50% is watched by sensors that do report color. When objects move into this outer area, they can be identified. Noisy reports are simulated by introducing fake reports which describe non-existent objects and by distorting (randomly modifying) reports about actual objects. Each object report has a *Noise %* chance of distortion, which in our experiments ranged from 0% to 30%. Two examples of anomalies are shown in Figure 2.

Reports provided by the OBSERVE function take the form "an object was detected at location x, y at time t with color c (or gray)." Reports are assigned random plausibility scores such that noisy reports typically score low and true reports score high. A threshold is established so that an initial call to the OBSERVE function returns only highly plausible reports, but additional calls (focusing on certain anomalies) return all nearby reports regardless of their plausibility. Hypotheses generated by the GENERATEHYPOTHESES function take the form "The object with color c moved from x, y at time t to x', y' at time t'." Two movement hypotheses are incompatible if they claim that two objects moved into the same location or out of the same location (the domain does not handle merges and splits), or that the same object (identified by color) is in two different locations at the same time. Movement hypotheses are scored based on the distance of the movement. The system is trained on

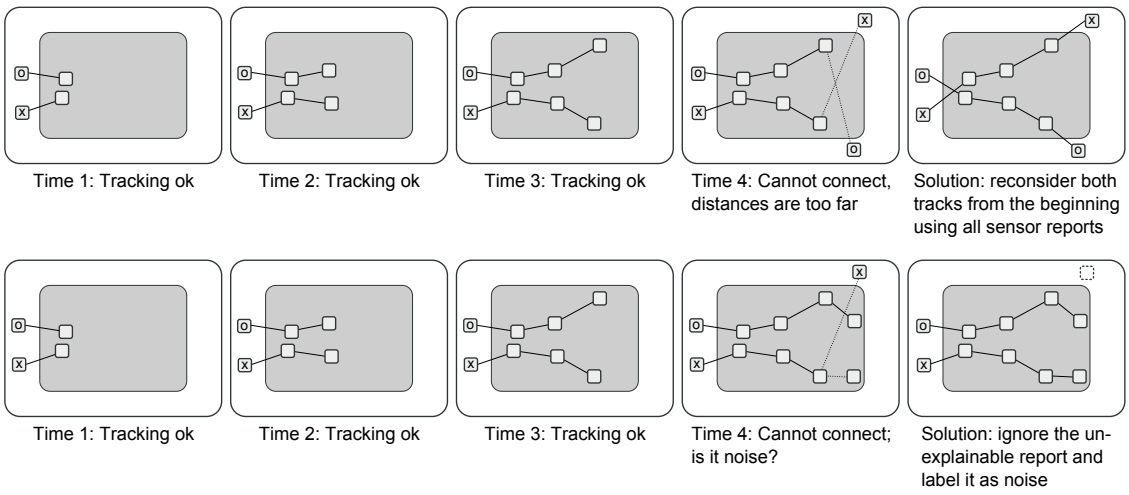


Figure 2. Anomalies due to order dependency (top) and noise (bottom) in the simulated object tracking domain. Reports for objects in the middle gray area do not specify the object's color (here represented by 'x' and 'o'), thus any object within a short range could possibly account for the detection.

examples of real object movements and builds a model of the probability of movements of various distances.

In addition to the simulated object tracking domain, we experimented with object tracking from aerial imagery using the KIT AIS Data Set[1] (Schmidt & Hinz, 2011). In each frame, hundreds of very small person-like objects are detected by a *Gentle AdaBoost* classifier (Friedman, Hastie, & Tibshirani, 2000). We take the detections as reports and perform object tracking in the same manner as the simulated tracking domain. Each detection is assigned a plausibility by the classifier, and the OBSERVE function initially provides only highly plausible detections. If queried for more observations nearby to some anomalous reports, the threshold limitation is eliminated for reports in that region. Note that no pair of hypotheses in this domain are incompatible with each other.

5. Experimental Methodology

We performed 30 random simulated tracking scenarios for each minimum plausibility η value and a single experiment for each η value with the aerial tracking dataset. In simulated object tracking, each simulation has 10 time steps and six different objects moving about. At each time step, each object took a random walk of six grid steps. In either domain, when metareasoning is enabled, some reports might also be rejected as noise, which we call *noise claims*. The levels of noise in both domains are very high. In the simulated domain, about 25% of reports are false. In the aerial domain, about 73% of reports are false.

1. `http://www.ipf.kit.edu/downloads_People_Tracking.php`

We define the following metrics to gauge the accuracy of the cognitive system's explanations and noise detection:

$$\begin{aligned}
\text{Precision} &= \frac{\|\text{Actual movements} \cap \text{Accepted movement hypotheses}\|}{\|\text{Accepted movement hypotheses}\|} \\
\text{Recall} &= \frac{\|\text{Actual movements} \cap \text{Accepted movement hypotheses}\|}{\|\text{Actual movements}\|} \\
\text{Noise precision} &= \frac{\|\text{Actual noisy reports} \cap \text{Noise claims}\|}{\|\text{Noise claims}\|} \\
\text{Noise recall} &= \frac{\|\text{Actual noisy reports} \cap \text{Noise claims}\|}{\|\text{Actual noisy reports}\|}
\end{aligned}$$

Three metareasoning strategies are compared: *abd-estimate*, as described in Section 3; *ignore*, which simply rejects all anomalies, effectively labeling them as noise; and *oracle*, which is the same abductive metareasoning strategy as *abd-estimate* but where the plausibilities of meta-hypotheses are set so that true meta-hypotheses have score 1.0, false have score 0.0, and $\eta_{\text{meta}} = 0.01$ to ensure that only true meta-hypotheses are accepted. Oracle metareasoning performance is the maximum performance that abductive metareasoning can be expected to achieve; however, it does not guarantee perfect performance because some errors and noise never manifest as anomalies.

The following hypotheses guide our experimental investigations:

Hypothesis I: Abductive reasoning with abductive metareasoning gives good performance in terms of Precision, Recall, and noise detection metrics in both domains. Furthermore, best performance is achieved when the minimum plausibility $\eta > 0$, thus demonstrating the utility of this parameter.

Hypothesis II: Metareasoning gives better performance on these metrics compared to no metareasoning. We also expect that oracle metareasoning consistently performs best.

Hypothesis III: Each of the four types of meta-hypotheses are accepted, at least in some cases, indicating that each kind of meta-hypothesis has a useful role.

6. Results

Results for the simulated and aerial tracking domains are shown in Figure 3 and Table 3, respectively. For the simulated tracking domain, each of the three kinds of metareasoning that we compared exhibited the best trade off between Precision and Recall at $\eta = 0.10$, where this trade off is calculated as the harmonic mean of the two metrics (often called the F1 measure). Table 1 shows the metrics at $\eta = 0.10$. Best performance in the aerial domain was found at $\eta = 0.80$ (Table 3).

Performance in simulated tracking is low in noisy conditions, but this is to be expected due to the inherent difficulty of the task. The object movements are random and in half of the grid the objects are indistinguishable. Furthermore, some reports are noisy. Table 1 shows performance in both noise-free and noisy conditions. Noise-free conditions yield considerably higher performance. In order to better understand how noise impacts simulated tracking performance, we experimented

Table 1. **Simulated tracking.** Performance in noise-free (N % = 0) and noisy conditions (N % = 30), $\eta = 0.10, \eta_{\text{meta}} = 0.40$. Under noise-free conditions, Noise Precision and Noise Recall are necessarily 0. Asterisks in *abd-estimate* scores indicate the difference with *ignore* scores is statistically significant (* $p < 0.05$, ** $p < 0.01$).

N %	Metareasoning	Precision	Recall	Noise Precision	Noise Recall
0	abd-estimate	0.875 ± 0.021	0.777 ± 0.022 *		
0	ignore	0.862 ± 0.023	0.715 ± 0.032		
0	oracle	0.916 ± 0.013	0.843 ± 0.016		
30	abd-estimate	0.664 ± 0.017 **	0.526 ± 0.014 **	0.588 ± 0.012 **	0.709 ± 0.015
30	ignore	0.634 ± 0.018	0.481 ± 0.016	0.537 ± 0.011	0.715 ± 0.013
30	oracle	0.767 ± 0.015	0.623 ± 0.013	0.679 ± 0.011	0.780 ± 0.012

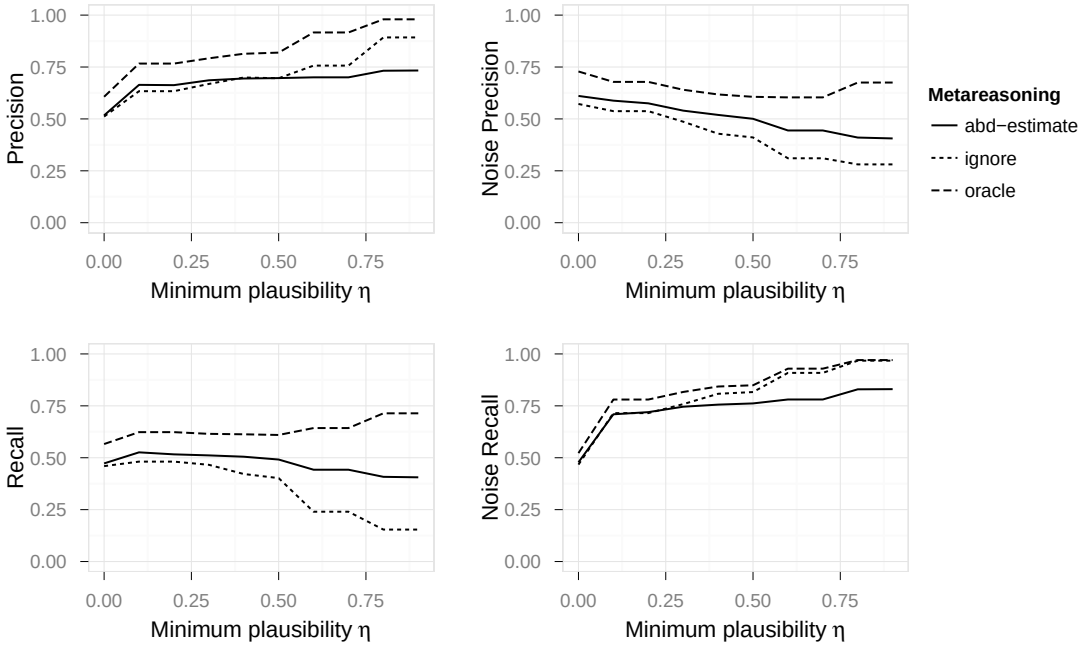
Table 2. **Simulated tracking.** Impact, as a ratio r, of increased noise level % on Precision, Recall, and Noise Precision metrics for different minimum plausibility η values. The Noise Recall metric was not significantly impacted by increased noise level. The ratio and R^2 values are derived from linear regression models. A ratio r means that for every 1% increase in the noise level, the metric increased by r.

η	Precision (R^2)	Recall (R^2)	Noise Precision (R^2)
0.00	-0.009 (0.71)	-0.007 (0.62)	0.012 (0.59)
0.10	-0.007 (0.60)	-0.007 (0.68)	0.011 (0.65)
0.20	-0.006 (0.52)	-0.007 (0.57)	0.012 (0.72)
0.30	-0.006 (0.45)	-0.006 (0.52)	0.011 (0.72)

Table 3. **Aerial tracking.** Performance in noisy conditions, $\eta = 0.80, \eta_{\text{meta}} = 0.60$. Results are from a single experiment on the whole dataset. The original dataset was not modified to add or remove varying degrees of noise.

Metareasoning	Precision	Recall	Noise Precision	Noise Recall
abd-estimate	0.660	0.825	0.972	0.853
ignore	0.636	0.700	0.911	0.853
oracle	0.680	0.850	0.972	0.853

Figure 3. **Simulated tracking.** Average performance across 30 random cases in the simulated object tracking domain for various minimum plausibility η values and $\eta_{\text{meta}} = 0.40$.



with different noise levels. Table 2 summarizes the results. We see that increased noise causes decreased performance in Precision, Recall, and an increase in Noise Precision. Noise Recall was not significantly impacted.

In summary, Hypothesis I is confirmed by the evidence. Furthermore, Tables 1 and 3, for simulated tracking and aerial experiments, respectively, show that *abd-estimate* metareasoning is better than *ignore* metareasoning, at the right η values, and worse than *oracle* metareasoning. For the simulated tracking domain, in which we executed multiple experiments with different random variations, we see that in noisy conditions, *abd-estimate* performs significantly better in terms of Precision, Recall, and Noise Precision. Thus, Hypothesis II is confirmed.

Finally, Hypothesis III is confirmed by the following evidence. Experiments with the simulated tracking domain show that each meta-hypothesis played a role. At $\eta = 0.10$, on average 1.68 ± 0.27 (standard error) *implausible hypotheses* meta-hypotheses were accepted during each experiment, 4.68 ± 0.29 *incompatible hypotheses*, 8.56 ± 0.72 *insufficient evidence*, and 3.42 ± 0.32 *order dependency* meta-hypotheses. In the aerial domain, we find that six *implausible hypotheses* and two *insufficient evidence* meta-hypotheses were accepted at $\eta = 0.80$. *Order dependency* and *incompatible hypotheses* meta-hypotheses were never considered to be possible causes of anomalies because there are no incompatible pairs among the hypotheses in this domain.

7. Related Work

Computational approaches to abductive reasoning can be divided roughly into three styles. Pagnucco (1996) distinguishes between two styles, *logic-based* and *set-covering*. We add to these *probabilistic* abduction, and give a short description of each. Our approach, characterized by the *EFLI* algorithm, most closely matches the set-covering style.

Logic-based abduction typically reifies the concept p *explains* q as $\Theta \cup \{p\} \vdash q$, where Θ is the background theory and $\Theta \cup \{p\}$ is consistent. The task is to find p given that q has been reported and requires explanation (i.e., $\Theta \nvdash q$). One kind of logic-based abduction is Abductive Logic Programming (ALP), introduced by Kakas et al. (1992) and directly influenced by earlier work on THEORIST (Poole, Goebel, & Aleliunas, 1986). Their approach limits the set of possible explanations in two ways: hypotheses can only come from a finite set of atomic sentences called "abducibles," and domain-specific integrity constraints must be satisfied. ALP has been implemented as an extension to Prolog (Fung & Kowalski, 1997) and subsequently integrated with constraint programming (Kakas, Michael, & Mourlas, 2000; Endriss et al., 2004). Another approach to logic-based abduction utilizes semantic tableaux as detailed by Aliseda (2006). Semantic tableaux allow testing if a formula q follows from a certain set of formulae Θ. Aliseda shows that when a tableau indicates that q does not follow from Θ, we can "read off" the tableau the information necessary to determine p such that $\Theta \cup \{p\}$ entails q. In Aliseda's terminology, this makes p an *abduction*. Further, if it's not the case that p itself entails q, but only entails q when combined with Θ, then p is *explanatory*.

Set-covering abduction operates on set of effects, a set of causes, and a relationship between the two sets that specifies which causes can possibly explain which effects. The task is to find a subset of the possible causes that is both internally consistent (some causes might be incompatible with each other), minimal in some sense, and explains all of the effects. Reggia et al. (1983) give an early example of this approach, although in that work, no possible causes were incompatible. It is clear that *EFLI* is a kind of set-covering abduction algorithm, though it is more directly influenced by the *hypothesis assembly* approach of Bylander et al. (1991). However, *EFLI* adds *pragmatic* concerns by preferring hypotheses that more significantly surpass their rivals in terms of plausibility, and by establishing a minimum plausibility threshold η.

Probabilistic abduction involves finding the most probable set of propositions (one from each variable of interest) given some evidence. This set of propositions is said to explain the evidence (Pearl, 1987). Pearl refers to this process as both belief revision and abductive inference (Pearl, 1994). Poole (1993) connects probabilistic abduction and logic-based abduction.

Belief revision has been a subject of study in the philosophy and artificial intelligence communities for at least thirty years. The classic "AGM" variety of belief revision (Alchourrón, Gärdenfors, & Makinson, 1985) operates on a *belief set*, which is the logical closure of a set of sentences. They define postulates that determine the characteristics of belief revision operators. The idea is to ensure that a belief revision only takes back and adds the fewest beliefs necessary to incorporate a new belief. However, Tennant (2006) shows that AGM belief revision postulates fail at that core task.

In any event, AGM belief revision is not directly applicable to our work since we do not use a propositional language with an entailment operator to represent beliefs. Rather, our doxastic states contain hypotheses that relate to each other only in terms of explanatory relations and incompatible relations. Additionally, AGM belief revision is *prioritized* belief revision, meaning that the new

input (in our case, an explanation of a report) is assumed to be true and must be accepted into the doxastic state. However, in our work, we want to judge whether or not it is appropriate to explain a report by examining how well it fits with the existing doxastic state and how plausible is the explanation on its own. In cases where no explanation is accepted, we call the unexplained report *noise*. Some work has explored *non-prioritized* belief revision (Hansson, 1999), in which the new input is not necessarily accepted. Eloranta et al. (2008) describe an interesting approach which modifies the input before accepting it. Our work is analogous to efforts in non-prioritized belief revision, but since our doxastic states do not represent belief sets, we do not find the AGM-style approaches to be directly useful for our purposes.

Abductive reasoning and (logical) belief revision have been shown to be closely related. Abduction is a way to do belief revision (Aliseda, 2000; Boutilier & Becher, 1995; Paglieri, 2003; Pagnucco, 1996). Furthermore, truth maintenance systems (Doyle, 1979) essentially combine abduction and belief revision (de Kleer & Reiter, 1987; Dixon & Foo, 1993).

Anomaly-driven theory revision has been explored previously. An early system by Karp (1989) responds to unexplained experimental outcomes, i.e., prediction failures, by designing modifications to the theory that, when applied, produce a theory that is able to predict the observed outcome. Bridewell (2004) describes a system with a similar goal. Bridewell's work maintains the assumption that reports from the world are noise-free. Karp's method, on the other hand, might simply fail to produce an acceptable theory revision. In this sense, it is similar to how abductive metareasoning decides whether or not a report is noisy.

Abductive metareasoning was investigated experimentally by Bharathan (2010). However, the metareasoning facility there does not utilize the same machinery as its base-level reasoner, and is somewhat *ad hoc* in its design. Additionally, the system only considers order dependency meta-hypotheses and does not attempt to detect noise.

The Meta-Cognitive Loop (MCL) from Schmill et al. (2011) shares similarities with the present work. The MCL is a component that attaches to a host reasoning system and is informed by the host system about possible actions and expectations regarding the results of those actions. Then, *in situ*, the MCL component monitors the host system's actions and detects expectation violations, which are called anomalies. Causes of the anomalies, and appropriate responses, are determined by consulting domain-general ontologies, represented as Bayesian networks. The present work differs from the MCL component in that, in abductive metareasoning, the variety of possible causes of anomalies is significantly smaller than those represented in MCL's ontologies. The likelihood of each kind of anomaly must be learned in MCL, while in the present work, the plausibility of meta-hypotheses are estimated according to domain-general features. Additionally, abductive metareasoning detects noise by way of a generic fallback meta-explanation rather than domain-specific noise detectors.

8. Conclusion

This paper has described and evaluated a two-level abductive reasoning and abductive metareasoning cognitive system. It has also shown at least one way to determine the right belief revisions in the face of anomalies, by detailing possible causes of anomalies and estimates of their plausibilities.

Abductive metareasoning has proved to be very effective at correcting errors in the doxastic state and detecting (and ignoring) noisy reports. This effectiveness has been demonstrated in simulated and aerial object tracking tasks, both of which were challenging due to limited sensor capabilities (such as inability to detect color in the simulated domain and very low resolution in the aerial domain) and very high levels of noise.

Although not discussed here, we have evaluated abductive metareasoning in a different world estimation domain and noted similar benefits. In that domain, the cognitive system attempts to explain reports by consulting a Bayesian network world model to find possible explanations and their plausibilities. Although the problem domain is quite different from the object tracking domains presented here, the abductive reasoning and metareasoning components do not require any modifications to handle the task. The only domain-specific parameters that tune the abductive reasoning and metareasoning algorithms are η and η_{meta}. In future work, we aim to explore a wider variety of problem domains to determine the limits of the domain-generality of this architecture and implementation.

Acknowledgements

We would like to thank the anonymous reviewers of this work for their helpful feedback, as well as Bruce G. Buchanan and Reid G. Smith for their detailed comments.

References

Alchourrón, C., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of symbolic logic*, *50*, 510–530.

Aliseda, A. (2000). Abduction as epistemic change: A peircean model in artificial intelligence. *Abduction and Induction: Essays on their Relation and Integration*, 45–58.

Aliseda, A. (2006). *Abductive reasoning: Logical investigations into discovery and explanation*. Kluwer Academic Pub.

Bharathan, V. (2010). Belief revision in dynamic abducers through meta-abduction. Master's thesis, The Ohio State University.

Boutilier, C., & Becher, V. (1995). Abduction as belief revision. *Artificial intelligence*, *77*, 43–94.

Bridewell, W. (2004). *Science as an anomaly-driven enterprise: A computational approach to generating acceptable theory revisions in the face of anomalous data*. Doctoral dissertation, University of Pittsburgh.

Bylander, T., Allemang, D., Tanner, M., & Josephson, J. (1991). The computational complexity of abduction. *Artificial Intelligence*, *49*, 25–60.

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

Cox, M., & Ram, A. (1999). Introspective multistrategy learning: On the construction of learning strategies. *Artificial Intelligence*, *112*, 1–55.

de Kleer, J., & Reiter, R. (1987). Foundations for assumption-based truth maintenance systems: Preliminary report. *Proc. American Assoc. for Artificial Intelligence Nat. Conf* (pp. 183–188).

Dixon, S., & Foo, N. (1993). Connections between the atms and agm belief revision. *International Joint Conference on Artificial Intelligence* (pp. 534–534).

Doyle, J. (1979). A truth maintenance system. *Artificial intelligence*, *12*, 231–272.

Eloranta, S., Hakli, R., Niinivaara, O., & Nykänen, M. (2008). Accommodative belief revision. *Logics in Artificial Intelligence*, 180–191.

Endriss, U., Mancarella, P., Sadri, F., Terreni, G., & Toni, F. (2004). The CIFF proof procedure for abductive logic programming with constraints. *Logics in Artificial Intelligence*, 31–43.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *The annals of statistics*, *28*, 337–407.

Fung, T., & Kowalski, R. (1997). The IFF proof procedure for abductive logic programming. *The Journal of logic programming*, *33*, 151–165.

Hansson, S. (1999). A survey of non-prioritized belief revision. *Erkenntnis*, *50*, 413–427.

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

Kakas, A., Kowalski, R., & Toni, F. (1992). Abductive logic programming. *Journal of logic and computation*, *2*, 719–796.

Kakas, A. C., Michael, A., & Mourlas, C. (2000). ACLP: Abductive constraint logic programming. *The Journal of Logic Programming*, *44*, 129–177.

Karp, P. D. (1989). *Hypothesis formation and qualitative reasoning in molecular biology*. Doctoral dissertation, Stanford University.

Paglieri, F. (2003). Belief revision: cognitive constraints for modeling more realistic agents.

Pagnucco, M. (1996). *The role of abductive reasoning within the process of belief revision*. Doctoral dissertation, University of Sydney.

Pearl, J. (1987). Distributed revision of composite beliefs. *Artificial Intelligence*, *33*, 173–215.

Pearl, J. (1994). Belief networks revisited. *Artificial intelligence in perspective*, 49–56.

Poole, D. (1993). Probabilistic Horn abduction and Bayesian networks. *Artificial Intelligence*, *64*, 81–129.

Poole, D., Goebel, R., & Aleliunas, R. (1986). *Theorist: A logical reasoning system for defaults and diagnosis*. University of Waterloo, Canada.

Reggia, J., Nau, D., & Wang, P. (1983). Diagnostic expert systems based on a set covering model. *International Journal of Man-Machine Studies*, *19*, 437–460.

Schmidt, F., & Hinz, S. (2011). A scheme for the detection and tracking of people tuned for aerial image sequences. In U. Stilla, F. Rottensteiner, H. Mayer, B. Jutzi, & M. Butenuth (Eds.), *Photogrammetric image analysis*, 257–270. Springer.

Schmill, M. D., Anderson, M. L., Fults, S., Josyula, D., Oates, T., Perlis, D., Shahri, H., Wilson, S., & Wright, D. (2011). The metacognitive loop and reasoning about anomalies. In M. T. Cox & A. Raja (Eds.), *Metareasoning: Thinking about thinking*, chapter 12, 183–200. The MIT Press.

Tennant, N. (2006). On the degeneracy of the full AGM-theory of theory-revision. *Journal of Symbolic Logic*, *71*, 661–676.