Anomaly-Driven Belief Revision by Abductive Metareasoning

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Joshua Ryan Eckroth, M.S.

Graduate Program in Computer Science and Engineering

The Ohio State University

2014

Dissertation Committee:

John R. Josephson, Advisor

B. Chandrasekaran

Neil Tennant

Abstract

Abduction, or *inference to the best explanation*, is, plausibly, part of commonsense reasoning, and a means by which a cognitive system may arrive at estimates of its world from observational and other evidence. We take this "world estimate" to be the cognitive system's *beliefs*. Since such reasoning is fallible, and world estimates will sometimes contain errors, an abductive reasoning system might improve its performance if it has a way to engage in belief revision when new evidence, or further reasoning, indicates the existence of a problem.

In this study, we develop, implement, and experimentally validate a *metareasoning* system that monitors and attempts to correct beliefs established by the base-level abductive reasoning system. We first identify that the presence of an *anomaly*, which we define as an observation or other evidence that cannot plausibly and consistently be explained, as a signal that the cognitive system's world estimate might be incorrect or, alternatively, that the unexplainable datum is noise. The metareasoning system responds to the presence of anomalies by asking exactly that question: which anomalies are due to mistakes in the world estimate, and warrant specific belief revisions, and which anomalies are due to noise, and should not instigate belief revisions? Various considerations regarding the nature of the anomalies and the system's reasoning history are brought to bear to answer this question.

Fundamentally, we see the metareasoning question ("what explains these anomalies: mistaken beliefs, or noise?") as structurally similar to the cognitive system's original question, "what explains these observations?" Thus, the metareasoning system is an abductive reasoning system, just like the base-level system. The anomalies constitute *meta-evidence* which may be explained by *meta-hypotheses*. These meta-hypotheses describe the various

kinds of causes of anomalies and specify particular belief revisions in order to resolve the anomalies. The same abductive reasoning algorithms employed by the base-level reasoner are activated to find the best explanation for the anomalies. An anomaly is judged to be the result of noise when no meta-hypothesis is judged to be a good enough explanation. In this manner, the cognitive system may engage in corrective belief revision and noise identification via abductive metareasoning.

We experimentally validate both the abductive reasoning and combined abductive reasoning and metareasoning systems with a software implementation. The software architecture maintains a formal separation of the reasoning procedures and the problem domain, i.e., what specific class of problems is being solved with the reasoning system. We explore three intentionally-simplified problem domains: simulated object tracking, aerial tracking, and inference to the best explanation with arbitrary Bayesian networks. These domains are intentionally simplified so that we can clearly identify how performance in these tasks is affected by various parameterizations of the reasoning and metareasoning systems. Our experiments show that (1) abductive reasoning is an effective way of reasoning in these problem domains, and (2) abductive metareasoning brings a significant boost in accuracy and noise identification. These experimental results, plus the system's architectural simplicity, together give strong evidence that abductive metareasoning is an appropriate and effective strategy for a cognitive system to revise its beliefs and arrive at more accurate estimates of its world.

Dedication

This thesis is dedicated to my parents and grandparents,

who gave me the freedom to pursue my passions.

Acknowledgments

In August, 2008, two weeks before my start in graduate school, I visited John Josephson's office and told him I wanted to work on "good old-fashioned AI" and high-level cognition, and that I thought his work in abductive reasoning was very interesting. His response was, "Well, it looks like we're karmically stuck to work together." I have never regretted connecting with JJ nor have I been jealous of the attention and assistance given to other graduate students from their advisors. In fact, I feel incredibly lucky (karmically or otherwise) to have been able to work with someone who is so deeply sophisticated in the issues relating to this thesis. I am continually inspired by the direct and honest approach he takes to his work and student advising. I recall asking him once if he had any advice for how I should write a certain research article. He said, "Be as clear as possible." That phrase has stuck with me like a zen koan and now guides all of my writing, teaching, and thinking. I want to thank JJ for being as clear as possible.

I also want to thank B. Chandrasekaran (Chandra) and Neil Tennant for serving on my committee and providing guidance over the years. Both Chandra and Neil bring an incredible depth of knowledge ensuring that every conversation gives me a new perspective on my work. They both are experts at finding the deeper meanings of surface-level phenomena and dissecting and later synthesizing the various efforts of other researchers. I also have benefited from their close inspection of the present work though, of course, all muddied arguments, *ad hoc* designs, and weak analyses remain my own.

My undergraduate advisor, Guy-Alain Amoussou, also deserves acknowledgment for his forethoughtful insistence that I participate in undergraduate research and teaching. At the time, I did not attribute any particular long-term value on my participation in these

programs. However, I now believe that Guy-Alain had a grand plan for me and the other students he advised. I attribute my entrance and success in graduate school to the work we did together. And six years after I finished my undergraduate education, I found myself discussing those activities in job interviews. Guy-Alain's guidance so long ago helped me secure an assistant professor position, and I would not be surprised if that was his plan all along.

I would also like to thank Norsys Software Corp. (www.norsys.com) for their development and support of Netica, for which they kindly provided me a *gratis* license. The Netica library was used extensively in this work for efficient probabilistic reasoning with Bayesian networks.

Finally, I want to acknowledge the unbounded support from my wife, Tracy Dawson, and father and step-mom, John Eckroth and Kathryn Golden. My parents systematically removed every obstacle barring my success (except my own laziness, of course), while my wife was capable even of stamping out my laziness (well, sometimes). She has always seen the light at the end of the tunnel and, if I am able to return the favor, the destination beyond the tunnel will be very bright, sunny, and warm indeed.

Joshua Eckroth

Columbus, Ohio

First hint of Spring, 2014

Vita

2013 ..... M.S. Computer Science, The Ohio State University
2008 ..... B.S. Computer Science, Humboldt State University
2008 ..... B.A. Mathematics, Humboldt State University

Publications

J. Eckroth, J. R. Josephson. Anomaly-Driven Belief Revision by Abductive Metareasoning. *Advances in Cognitive Systems*, to appear.

J. Eckroth, J. R. Josephson. Commonsense Abductive Reasoning and Metareasoning Using Knowledge from Bayesian Networks. *AAAI-14 Spring Symposium, Knowledge Representation and Reasoning in Robotics*, to appear.

B. G. Buchanan, J. Eckroth, R. G. Smith. A Virtual Archive for the History of AI. *AI Magazine*, 86–98, Summer 2013.

J. Eckroth. Abductive Metareasoning for Truth-Seeking Agents. *AAAI-2012/SIGART Doctoral Consortium*, 2388–2389, 2012.

J. Eckroth, L. Dong, R. G. Smith, B. G. Buchanan. NewsFinder: Automating an AI News Service. *AI Magazine*, 43–54, 2012.

J. Eckroth, D. Reddy, J. Josephson, R. Chellappa, T. Miller. From background subtraction to threat detection in automated video surveillance. *ARL CTA Report*, 2009.

J. Josephson, J. Eckroth, T. Miller. Estimation of adversarial social networks by fusion of information from a wide range of sources. *12th International Conference on Information Fusion*, 2144–2152, 2009.

J. Eckroth, R. Aytche, G.-A. Amoussou. Toward a science of design for software-intensive systems. *DESRIST 2007 Proceedings*, 39–53, 2007.

J. Eckroth, G.-A. Amoussou. Improving software quality from the requirements specification. *Proceedings of the 2007 Symposium on Science of Design*, 38–39, 2007.

Awards

Graduate Associate Teaching Award, Graduate School, The Ohio State University, 2012.

Eleanor Quinlan Memorial Award, Computer Science & Engineering Department, The Ohio State University, 2012.

Student of the Year, Computing Sciences Department, Humboldt State University, 2008.

Fields of Study

Major Field: Computer Science; Specialization: Artificial Intelligence
Minor Field: Cognitive Science
Minor Field: Mathematical Logic

Table of Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1: Introduction

Cognition can be conceptualized as four distinct phases, as shown in Figure 1.1:

1. Observing the world via sensors or reports from other agents.

2. Making sense of these observations and other evidence in order to arrive at a world estimate, i.e., a set of beliefs about the world.

3. Planning to act in the world in order to achieve goals. Clearly, success in this planning phase requires having accurate beliefs about the current state of the world. Planning may also involve planning to observe in order to better make sense.

4. Acting out the plan. This leads back into observing the world in order to both ensure that the plan has produced the desired effects and to discover any changes in the state of the world.

In this research, we focus on the *making sense* or *world estimation* problem. We investigate cognitive systems that attempt to make sense of observations that putatively describe properties of the world. In particular, we investigate cognitive systems that do not manipulate sensors, communicate with other agents, or otherwise act in the world. These cognitive systems are *passive*, in the sense that they can only observe the world and not manipulate it. Even so, they are interesting for the following reasons:

- Passive world estimation may be the whole task. For example, the goal of accident investigation is to produce an estimate, once all the evidence is gathered, of the events

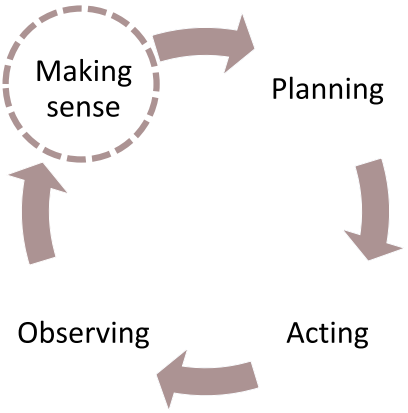


Figure 1.1. Four phases of cognition. This work focuses on the *making sense* phase.

that caused the accident. Passive world estimation for dynamic worlds includes tasks like tracking people or projectiles with fixed cameras or radar antennae.

- All cognitive systems, even those that plan and act in the world, must solve the world estimation problem as a subtask. All cognitive systems need to make sense of their inputs.

The world estimation systems described herein are explicitly *abductive*. Abduction, or equivalently, *abductive reasoning* or *abductive inference*, is reasoning to the best explanation. Given some surprising or unexplained data D, abductive reasoning seeks an explanation e of D. For e to be the best explanation, it should itself be plausible but also more plausible than alternative explanations of the data. We take *explanation* to be an undefined relation. What does and does not count as an explanation is beyond the scope of this report, although an explanation has been said to provide a *causal story* for how D occurred or *gives understanding* for why D is the way it is (Lombrozo, 2006).

When explanations take the form of causal explanations, abductive inference is capable of reasoning from effects to causes. The passive world estimation task can be construed

as reasoning from effects to causes. The data D are world properties as observed and reported by sensors or other agents. The potential explainers E describe events or properties of the world that could be causes of the reported world properties. For example, a person walking from one place to another could be an explanation for a set of camera reports of visually similar person-shaped blobs at different locations over time.

In this work, we investigate a family of abductive reasoning systems that perform world estimation. This investigation has two major components. First, we show that abductive reasoning is suited for the world estimation task and that it can be performed efficiently and with reasonable accuracy and completeness. Second, we show that this world estimate can be made more accurate and complete, in many cases, by engaging in *abductive metareasoning* when the base-level abductive reasoning process is unable to arrive at a consistent world estimate that explains all the data.

A *metareasoning* system is one that monitors and controls the base-level reasoning system. The purpose of the metareasoning system is to boost the accuracy and completeness of the beliefs formed by the base-level system. The question that one might ask is, why would the base-level system make mistakes, and what could possibly be done to identify and fix them? The base-level abductive reasoning system might make mistakes, or produce an incomplete world estimate that does not explain all the data, for a variety of reasons:

- The reasoning system does not have the right hypotheses or the plausibility estimates of the available hypotheses fail to differentiate hypotheses or are wildly inaccurate.
- Some of the data are inaccurate, i.e., noisy, and the reasoning system was unable to identify and isolate this noise.
- The reasoning system did not consider all possible ways to combine the various hypotheses in order to find the most plausible, most complete explanation. An ideal but

computationally intractable reasoning system would, presumably, be able to examine all possible explanations, and pick out the most plausible. A practical reasoning system must utilize heuristics to make the reasoning task tractable. These heuristics might be the cause of mistakes or incomplete world estimates, but are not the only possible source as evidenced by the two previous points.

Some scenarios may be detected and repaired some of the time by a metareasoning system. We build and analyze a metareasoning system that responds to the presence of reports that cannot be confidently and consistently explained. We call unexplainable reports *anomalies* and show that the presence of anomalies is a good indication that something is wrong with the cognitive system's world estimate. Furthermore, we treat the metareasoning task as an *abductive* one, in which the anomalies are *meta-evidence* that the world estimate might be inaccurate. The metareasoning system then constructs *meta-hypotheses* that can explain the anomalous reports by positing the reason why the report is anomalous. Normal abductive reasoning works on these meta-hypotheses to arrive at the most plausible, most complete consistent explanation of the anomalies. All accepted meta-hypotheses detail particular belief revisions, which are then applied. We show that these *anomaly-driven* revisions significantly increase the accuracy and completeness of the cognitive system's world estimates.

1.1 Central claims

Our central claims in this report are as follows:

- The base-level abductive reasoning system that we develop is an effective and efficient way to handle the world estimation task across various problem domains.

- The presence of anomalies is a good sign that the cognitive system's world estimate is incomplete and/or inaccurate.
- The abductive metareasoning facility that we develop significantly increases accuracy and completeness by finding explanations for anomalies and applying corresponding belief revisions.

1.2 Methodology

As described in our central claims, we aim to show that the combined abductive reasoning and abductive metareasoning system yields high *accuracy* for reasoning tasks across various problem domains. The metareasoning system is implemented as a separate module that can be selectively enabled. In order to show that metareasoning offers benefits over the base-level reasoning system, we first measure performance of the base-level system on a reasoning task with metareasoning disabled. Then we compare those results to performance on the exact same task with metareasoning enabled.

Each experiment has the reasoning system obtaining evidence about the world and reasoning about this evidence to form beliefs. Evidence is obtained piece-wise across discrete *time steps*. The system is expected to reason about current evidence based on beliefs arrived at in light of earlier evidence. Beliefs may be corrected later, when more evidence is available, and indeed this is partly the responsibility of the metareasoning system. We note that in an ideal world, decisions can always be delayed until all evidence is obtained. Such delayed decisions should result in maximally accurate beliefs. But agents-in-the-world often do not have access to complete evidence about a particular aspect of the world and must make decisions based on partial evidence. Reasoning with partial evidence is more realistic but also more prone to error. As will be shown, our metareasoning system is designed partly to correct mistaken beliefs based on insufficient, weak, or false evidence. These cor-

rections are possible only when more complete and accurate evidence is obtained at a later time step. We wish to investigate realistic reasoning tasks and to explore a variety of ways that metareasoning can repair bad beliefs. Thus, all of our experiments are designed so that the reasoning system arrives at beliefs about the world, established at each time step, and often based on partial evidence.

Some problem domains are simulated domains, meaning we do not use real data from actual sensors in the world, but instead generate a simulated world that is monitored by simulated sensors. This approach is not new, as simulation has been used to evaluate software systems for as long as software systems have been evaluated. Simulation provides the following benefits:

- Since we generate the world and sensor reports, and even generate noisy reports, we (as experimenters) have access to the *truth* of the case. This truth is not necessarily equivalent to the *ground truth* as designated by a human who has manually marked up a dataset (e.g., labeling words in a speech signal or tracks in a video). Rather, the truth of the simulation is an objective truth. So, we can objectively measure the accuracy of the cognitive system's world estimate.

- A wide range of cases can be explored just by tuning the parameters for world generation. We can test the cognitive system on easy cases, hard cases, pathological and rare cases, etc. Real-world datasets usually do not include pathological and rare cases (by definition) so experiments with real datasets often fail to show how the system behaves in particularly difficult scenarios. Yet, it is often these scenarios we explicitly want to test, since if a system is going to fail, it will likely fail in those kinds of cases.

The simulated worlds and corresponding reasoning tasks are designed specifically to

highlight interesting properties of abductive reasoning and metareasoning. Since these simulated worlds are original, we are unable to compare the performance of abductive reasoning and metareasoning with other approaches from other researchers. Our goal is to show that abductive metareasoning boosts performance over just abductive reasoning without metareasoning. But since we have defined both the problem and the solution, we effectively have a conflict of interest. How can we be sure that we did not create easy or specially-tuned cases just to show that our system performs well in those cases?

In order for our argument to be convincing, we must validate each component as we go: the simulated worlds must cover a wide range of easy and hard cases, and not be designed specifically to work well with abductive reasoning. The base-level abductive reasoning system must also be shown to be perform well on the task. The base-level system should not be handicapped just to allow the meta-level system to bring improvements. Then, should abductive metareasoning actually provide some boost in accuracy, we will know that this boost is not simply the result of correcting easy and obvious mistakes but rather that it is due to some essential feature of abductive metareasoning: specifically, that it is capable of finding the causes of anomalies, and fixing them, in a way that is simply not possible with the base-level reasoning system (since the base-level system is not self-reflective).

Each of the three experimental chapters (Chapters 6–8) follows a common argument. The problem domain is introduced. Then, various *validation* experiments are conducted with the base-level abductive reasoning system to demonstrate that easy cases are easy and hard cases are hard, and for the right reasons. Finally, abductive metareasoning is added and shown to give a significant boost in accuracy. In Chapter 7, we look at an experimental domain that uses aerial surveillance imagery for a pedestrian tracking task. This domain utilizes real-world data rather than simulated data, yet we are still able to show that abductive metareasoning provides benefits. We note in Chapter 10 that more

domains and a greater variety of domains would be helpful to provide further evidence that abductive reasoning with abductive metareasoning is a good way for a cognitive system to engage in world estimation.

1.3 Overview of prior work

Recently, Don Perlis listed several properties a "commonsense reasoning system" should possess, including a metareasoning facility (Perlis, 2011). Two of these properties will now be discussed. Quoting Perlis,

> [The metareasoning facility] can be fairly simple, based on a core set of general kinds of things that can go wrong and general kinds of fixes for them. Such a [facility] could be general-purpose, not built for any specific system or domain. This is because humans are not specific in that sense. We manage to muddle through in a wide variety of unanticipated changes within, and even of, arenas of action.

The abductive metareasoner, detailed in Chapter 4, is general-purpose, apart from its assumption that the base-level reasoner is abductive. The metareasoner receives information about the reasoner's history and its difficulties in finding plausible explanations of reports. The metareasoner does not have access to the *content* of reports or their possible explanations. The same metareasoner, and same reasoner, both work well without modification in different world estimation problems. This will be demonstrated experimentally.

Again, quoting Perlis,

> If such a [facility] were to be built and put to use with a given AI system, that system would become far less brittle, and vastly better at dealing with anomalies, than any AI systems at present.

Perlis, with Schmill, Anderson, and others (Schmill et al., 2011), have worked towards building a system that meets his criteria. They suggest that there exists a level of abstraction at which the variety of reasoning failures is finite. They then propose a taxonomy of such possible failures. Each kind of failure is a deviation from an expectation. In order for the metareasoner to detect failures, it must be able to detect expectation violations. Furthermore, in order for the metareasoner to attempt repairs, it must know which parameters of the base-level reasoner may affect its performance and thus require adjustment.

Schmill et al.'s metareasoner detects expectation violations and abduces the most probable failure and the repairs that are least costly. A Bayesian network provides possible failures, their causes, and corresponding repairs. The structure of the network is fixed. It encodes the domain-general taxonomy that the authors hypothesize is sufficient to describe symptoms, causes, and repairs for failures in any domain. However, the parameters of the network must be determined ahead of time, either through training or by manual specification. One might suspect that these parameters are domain-specific. Schmill et al. hope to prove the generality of the taxonomy and the parameters of the associated Bayesian network but have not yet done so. The authors also do not provide a domain-general strategy for deciding when an attempted repair is successful, or when a different repair should be tried.

Our work is similar to Schmill et al.'s in some ways but does not suffer some of the limitations of their current work. Specifically,

- We are also proposing a novel metareasoning facility that responds to symptoms of failure, i.e., anomalies.
- In the systems we investigate, the base-level reasoner is abductive and domain-general. The metareasoner is designed to work with such an abductive base-level reasoner. The only parameters of the base-level reasoner that are adjustable by the

metareasoner are domain-general parameters (e.g., the minimum required plausibility of a possible explanation).

- The decision to proceed with an attempted repair or to try another is addressed in this work. Again, the decision is made using only domain-general information.

- Our claims of domain-generality are supported by experiments from different domains, in which the exact same abductive reasoning and metareasoning systems are used without modification.

A different perspective on metareasoning comes from the philosophical tradition known as *belief revision*, of which there is an extensive literature. We refer to this perspective as *strict* belief revision in order to differentiate it from probabilistic belief revision as in Bayesian belief networks. Strict belief revision involves categorically accepting as *beliefs* certain statements and categorically rejecting other statements. Often there is also a third set of statements that are neither believed nor disbelieved. This approach is contrasted with belief revision in a Bayesian belief network, in which each statement is partially believed according to some probability and a revision involves updating the probabilities in light of observations. The reasoning systems we investigate categorically accept, reject, or have no opinion on statements. Thus, our work is more similar to strict belief revision.

Although the literature in strict belief revision will be addressed more fully in Chapter 5, it deserves a brief mention at this point. Strict belief revision usually refers to the addition and retraction of beliefs in order to maintain consistency. In some approaches, consistency need not be maintained at all times, leading to systems involving paraconsistent logics. Often, some kind of *preference ordering* (such as preferring statements with fewer assumptions) allows one to determine the unique revision that maximizes preference and preserves consistency. Pragmatic concerns such as resource constraints or concerns about plausibility are typically left out of the theoretical investigations.

In the abductive cognitive systems that we investigate, we take accepted explanations to be *beliefs*. Thus, beliefs are not provided by an external source as inputs but rather are inferred by the system. Most of the belief revision literature is incompatible with this arrangement because the literature usually assumes that beliefs, rather than evidence which forms the basis of beliefs, are the inputs. However, Pagnucco (1996) has addressed this difference. In his characterization, inputs (evidence) are added as beliefs when they can be explained; the explanations are also added as beliefs simultaneously. He provides expansion, contraction, and revision operations that respect this distinction between evidence and explanations.

Pagnucco's work, like most work in strict belief revision, is theoretical and does not provide an experimental evaluation of the system's performance characteristics in a variety of tasks. Much of the belief revision literature deals with belief sets that are infinite and revision operations that are undecidable or at least intractable (Doyle, 1992). Our goal, in contrast, is to build reasoning systems that are tractable and demonstrably accurate.

There have been some efforts in tractable belief revision. One such effort is Tennant's work in efficient belief contraction (2012). He investigates how a "logical paragon" would change its beliefs, where a logical paragon is a rational agent that always does the right thing given the inputs and background knowledge available to it. The agent can make mistakes if the inputs are misleading (e.g., when there is noise in the form of false reports), but it would never fail to apply the correct logical operations, given enough time. Thus, it would always eventually come to the best possible beliefs given the information available. Tennant's concern is not with how an agent comes to its beliefs, but rather how it justifies its beliefs, and how its beliefs change whenever a previously-justified belief must be retracted.

Tennant represents agents' beliefs as *nodes* that are related by way of justifications. Nodes are "featureless and unstructured"; all that matters is whether a node is believed, disbelieved, or neither, and how nodes relate in terms of justifications. The justificatory

relations may take various forms but generally match our intuitions. If belief in p justifies belief in q, and p is believed, then the logical paragon would consider q to be justified and therefore believes it (assuming q is not already disbelieved or justifiably presumed to be false). Should the agent discover that q is in fact false, the agent must perform a contraction on its beliefs with respect to q by retracting at least q and possibly other beliefs. In the trivial case illustrated, this would require retraction of belief in p as well, in order to take away the justification for believing q.

Tennant proves that finding a contraction is intractable when it is required to be *minimally mutilating*, that is, when the contraction must retract the fewest beliefs possible while conforming to the desiderata of the contraction operation. Tennant develops an efficient contraction algorithm that very often produces the best possible results. This contraction algorithm addresses one concern relevant to the present work. An abductive reasoning process may accept some explanation e that at a later time proves to be problematic. For example, it may be that e is found to be incompatible with a future essential explanation e' of some report r'. Because e' is the only available explainer of r', but e' cannot be accepted due to incompatibility with the accepted explainer e, we find that r' is unexplainable, i.e., anomalous. Metareasoning is triggered and tasked with deciding whether r' is to be ignored as noise or whether the prior acceptance of e should be retracted in order to free up e' as an explainer. Tennant's algorithm may be able to efficiently find a best or near-best contraction with respect to e. The metareasoning system may choose to simulate such a contraction and evaluate the quality of the result, and compare this repair with others in order to determine the best course of action. Although Tennant's work may be useful as part of an abductive metareasoning process, we have not yet explored its effectiveness in that task.

Another effort in tractable belief revision is Johnson and Shapiro's work (2005) in dependency-directed reconsideration (DDR). Their "reconsideration" operation removes

order-dependency effects resulting from reasoning about a changing world. The reasoning system cannot be sure when it has received all the information it will ever receive, and beliefs resulting from reasoning about the information received up to some point may be contradicted by information received later. Furthermore, beliefs that were revoked may be candidates for restoration when more information is available that sheds more light on these beliefs. Their DDR technique efficiently determines which belief revisions should be reconsidered when new information becomes available. The outcome of the DDR process is a consistent belief set that produces an "optimal" knowledge state (where "optimal" is defined with respect to some preference order on beliefs).

Johnson and Shapiro's work considers beliefs, rather than evidence, to be the inputs to the system, similar to most work in belief revision. However, their concern with efficient computation and, in particular, their handling of order dependencies, agrees with the approach we take with abductive metareasoning. We note in the analysis of reasoning errors (in Chapter 2) that order dependency is a source of some errors. In these cases, prior accepted explanations need to be reconsidered. Abductive metareasoning, as we describe it, is responsible for noticing a possible reasoning error, hypothesizing that the error may be due to an order dependency (among other possible causes), and invoking a reconsideration of prior accepted explanations (among other attempted repairs).

In summary, prior work in metareasoning and strict belief revision suffers from one or more limitations that we address in this work:

- The representation of beliefs makes commitments about the world. For example, the use of propositional logic makes the representation of probabilistic, modal, and "default" beliefs very difficult or impossible. However, many researchers choose to employ propositional logic when studying belief revision in order to simplify concepts such as *incompatibility* and preference orderings. Pagnucco also uses propo-

sitional logic even though his proposed reasoning processes are abductive. In that case, explanations imply what they explain. Again, this simplifies analysis but may be inapplicable to some problem domains (for example, in medical diagnosis where presence of a disease does not guarantee each of its symptoms will manifest). Tennant's work is an important exception, although he relies on a *justification* relation, whose compatibility with *explanatory* relations has not been fully explored.

In this work, "A explains B" and "A is incompatible with B" are undefined relations, and the components A and B can take any form. Each explanation is associated with a *plausibility estimate*, which may be interpreted as binary possibility, probability, or other interpretations. We believe these three concepts are widely applicable and make the fewest possible commitments about the representations of beliefs, evidence, and explanations.

- Much work in strict belief revision does not address computational concerns, although Tennant's and Johnson and Shapiro's efforts were noted as exceptions. Other researchers have also shown experimental results regarding computational tractability (Dixon, 1994; Zhuang et al., 2007). The methods reported here are shown to be tractable.

- Consistency requirements and preference orders do not necessarily yield the most accurate beliefs, and we are not aware of any attempts to evaluate the impact that different theoretical approaches have on accuracy. On the other hand, Schmill et al.'s work in metareasoning (which has no significant overlap with strict belief revision) does mention plans to experimentally evaluate improvements in accuracy resulting from metareasoning.

  The work reported here is both formal and experimental, and attempts to be domain-general. We measure accuracy by comparing the cognitive system's beliefs with the

truth of the case. With this experimental setup, we show that abductive metareasoning significantly increases accuracy and completeness across a variety of domains.

The remainder of this report is organized as follows. Chapter 2 formalizes abductive reasoning while Chapter 3 explores prior work in abduction. Chapter 4 details abductive metareasoning, and Chapter 5 examines prior work in metareasoning, including work in strict belief revision. The following three chapters (Chapters 6–8) document experiments in three problem domains, respectively: simulated object tracking, aerial tracking, and abduction with Bayesian networks. Finally, Chapter 9 provides concluding remarks and Chapter 10 outlines plans for future work.

Chapter 2: Abductive reasoning

By *abduction*, *abductive reasoning*, and the like, we mean reasoning that follows a pattern approximately as follows:

> *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.

The strength of the conclusion *H*, the force of the *probably* in the conclusion statement, reasonably depends on the following considerations, borrowed from Josephson and Josephson (1994):

- how decisively the leading hypothesis surpasses the alternatives,
- how good this hypothesis is by itself, independently of considering the alternatives,
- how thorough was the search for alternative explanations,
- confidence in the accuracy of the data (although *noise* can be considered an alternative explanation of the data).

Besides confidence in its correctness, willingness to accept a conclusion of such abductive reasoning also reasonably depends on practical considerations, including:

- the expected costs of being wrong and benefits of being right,

- the expected costs of waiting before deciding (urgency), and the expected benefits of waiting, especially the benefits of obtaining further evidence before deciding,
- the expected costs of not deciding to believe.

Factors that might reasonably contribute to the evaluation of hypotheses, either in isolation, or in contrast with rivals, include: explanatory power, plausibility (precedent, consistency with background knowledge, consistency with data), parsimony, internal consistency, specificity, and productive promise.

However abduction is specifically described, or formalized, we hope readers recognize it as a distinctive and familiar pattern, and as having a kind of intuitively recognizable inferential (evidential) force. It seems an appropriate way to describe the evidence-combining characteristics of a variety of cognitive and perceptual processes, such as diagnosis, scientific theory formation, language comprehension, and inferring intentions from behavior. Thus, abductive inferences appear to be ubiquitous in cognition, although many of them may be implicit. Moreover, such inferences seem to be extremely common at or near the surface of typical arguments offered in science, diagnosis, forensic investigation, and ordinary life. In fact, one can readily observe that people commonly justify their conclusions by direct or barely disguised appeal to this pattern, which shows that speaker and hearer share a common understanding of it, cross-culturally. Thus, abduction seems to be part of commonsense logic.

It will contribute to clarity to distinguish abduction as a pattern of argumentation or justification, from abduction as a reasoning process. In a process of trying to explain some experience, or pattern of experiences, 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, 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*.

Characteristic subgoals, subfunctions, and subprocesses of abductive reasoning, include: distinguishing data needing explanation, generating explanatory hypotheses, evaluating hypotheses, comparing hypotheses, and deciding whether to accept a hypothesis as being sufficiently justified. The term *abduction* has sometimes been used for the hypothesis-generation part alone.[1] However, our primary interest in this work is the processes of evaluating, comparing, and deciding to accept hypotheses. The processes of deciding what data need explanation and generating hypotheses are not under investigation.

2.1 Abduction in static and dynamic worlds

In this work, we investigate abductive reasoning in both static and dynamic worlds. In static worlds, the world properties to be estimated or inferred do not change. Even so, the reasoning system might acquire evidence about the static world over time and might be required to commit to intermediate estimates of the world. These estimates might require revision as more evidence is acquired. In dynamic worlds, the world properties to be estimated might change over time, and the system's world estimate might need revision simply because the world changed. Thus, regardless of whether or not the world to be estimated is static or dynamic, it is useful (and sometimes necessary) to have a reasoning process that is able to revise prior estimates, i.e., revise beliefs.

We can represent the estimation problem for static and dynamic worlds as shown in Figures 2.1 and 2.2, respectively. In the static diagram, we have one world state, while in the dynamic diagram, we have a progression of world states. In either case, the reasoning system does not have direct access to the world states; rather, it relies on observations of

[1] C.S. Peirce: "Abduction is the process of forming an explanatory hypothesis. It is the only logical operation which introduces any new idea" (Peirce, 1958, par. 171).

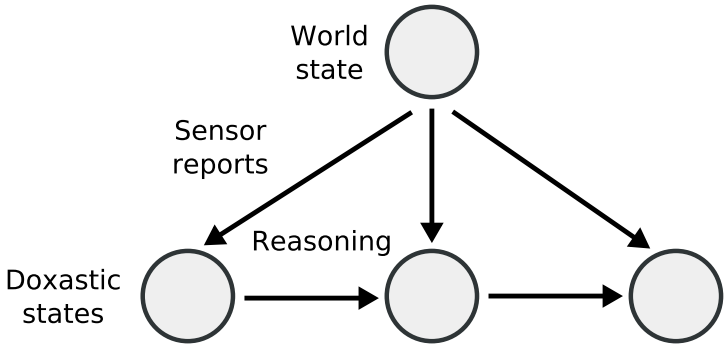


Figure 2.1. Reasoning state diagram for static worlds.

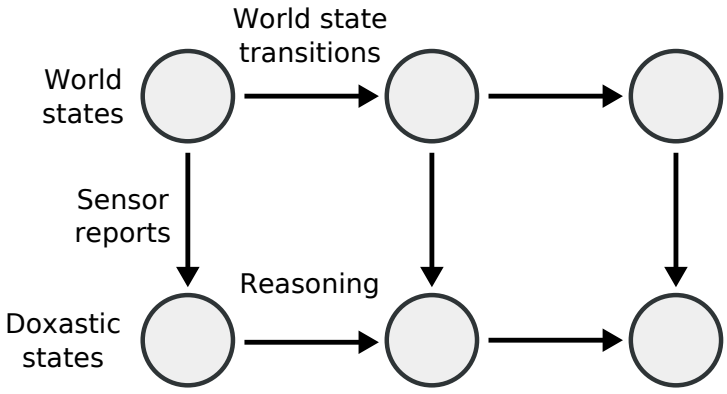


Figure 2.2. Reasoning state diagram for dynamic worlds.

the states, as provided by sensor reports. For simplicity's sake, world states change at discrete time steps and sensors report a group of observations at each time step. The system maintains a sequence of *doxastic* or *belief states* (defined in formal terms below), which are indicated by the bottom nodes in the diagrams. The most recent doxastic state represents the system's most recent world estimate. Doxastic states are the result of applying abductive reasoning based on the sensor reports and prior doxastic state. For the sake of efficiency, the prior world estimate is kept and expanded to incorporate more recent sensor reports.

2.2 Notation

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 2.2.1.[2] Note that, for simplicity's sake, we treat reports as hypotheses that explain nothing but are themselves initially 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 2.2.1. A *doxastic state* is a tuple $D = (H, X, Pl, S, V, I)$, 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, $S : H \to \{Accepted, Rejected, Undetermined\}$ gives the belief status of a hypothesis, $Pl : H \to [0,1]$ is a plausibility function, and $V \subseteq H$ is a set of *evidence* hypotheses that, when accepted, are considered to require an explanation. Note that V may contain hypotheses that could explain other hypotheses but nevertheless require explanation themselves. I is an irreflexive, symmetric relationship over H where $(h_j, h_i) \in I$ means h_j

[2]This formalism was first introduced by Eckroth and Josephson (2014).

is incompatible with h_i. The sets I and X are constrained so that $X \cap I = \emptyset$. Additionally, $(\forall (h_j, h_i) \in I)(S(h_j) = Accepted \rightarrow S(h_i) = Rejected)$. 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. Note that $(h_j, h_i) \in X \wedge S(h_j) = Accepted$ does not imply $S(h_i) = Accepted$ because h_j may have been accepted to explain some other accepted hypothesis $h_k \neq h_i$. 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 (e.g., Aliseda (2006); Kakas et al. (1992)). Logical entailment does not capture what it means to *explain* (Mayer and Pirri, 1996). 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.

Plausibilities do not necessarily span the full range of $[0,1]$. They may be restricted to a subset of cardinality $n+1$ where the values come from $\{i/n | i \in \{0,1,\ldots,n\}\}$.

Definition 2.2.2. The **plausibility precision** is the value n that redefines the plausibility function to be $Pl : H \rightarrow \{i/n | i \in \{0,1,\ldots,n\}\}$.

It will be shown experimentally that small values of n typically produce just as good abductions as a very large plausibility precision.

We can represent explanatory relations and hypothesis belief status of a doxastic state with an explanation graph. An explanation graph is a structure in which vertices are taken to be hypotheses. Edges with arrowheads are *could explain* relationships (the tail of the arrow could explain the head of the arrow). Edges that appear as dashed lines are *incompatibility* relationships. No two vertices can have both an explanatory and an incompatibility relationship. Every vertex has one of three states at any time: *Accepted*, *Rejected*, or *Undetermined*. The plausibility of a hypothesis is sometimes shown as a number in

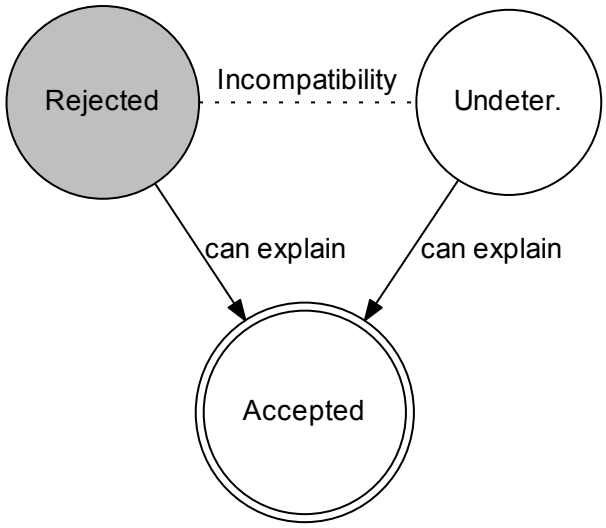


Figure 2.3. Illustration of explanation graph diagrams. “Undeter.” means “Undetermined.”

the hypothesis vertex. An explanation graph is not necessarily connected. Different sets of observations may have different explainers without any explanatory or incompatibility relations between them. Figure 2.3 shows an illustration to guide interpretation of explanation graph diagrams.

Note that top-level vertices, that is, vertices with no incoming edges, do not need to be explained if they are accepted. An explanation graph is constructed so that vertices with no incoming edges have no relevant explainers. In every abductive task, there is a limit to what is considered a relevant explainer. A medical doctor, for example, wants to diagnose the patient’s symptoms in terms of a catalogue of diseases. Perhaps she even seeks an explanation of how the patient acquired the disease. E.g., “excessive exposure to the sun caused skin cancer which caused the growth on the patient’s arm.” But the cause of the patient’s excessive sun exposure may not be relevant.

The following definitions of *explaining set*, *consistent explaining set*, and *complete explaining set* guide further analysis into the computational properties of abduction.

Definition 2.2.3. An **explaining set** $A \subseteq H$ of a doxastic state is the set of hypotheses that are *Accepted*. A **consistent explaining set** does not include both hypotheses in any incompatible pair. Note that $A = \emptyset$ is a consistent explaining set.

Definition 2.2.4. A **complete explaining set** is a consistent explaining set that has the following property: $(\forall h_i \in A)((\exists h_j)((h_j, h_i) \in X) \rightarrow (\exists h_k)((h_k, h_i) \in X \wedge h_k \in A))$. Informally, every accepted hypothesis that has any possible explainers has at least one accepted explainer.

A complete explaining set may not exist. The explanation graph in Figure 2.4 is a minimal example. The proof of its minimality is trivial. Recall that no hypothesis can be incompatible with itself, and that no two hypotheses can have both an explanatory and incompatibility relationship.

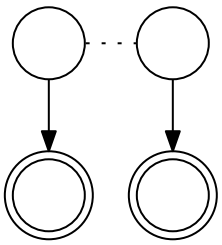


Figure 2.4. An explanation graph in which no complete explaining set exists.

Definition 2.2.5. A **parsimonious explaining set** is a complete explaining set that has the property that no proper subset of the explaining set is also complete.

Definition 2.2.6. A **minimal-cardinality explaining set** is a complete explaining set that has the property that no other complete explaining set (for the same explanation graph) has fewer vertices.

Definition 2.2.7. A **best complete explaining set** A is a complete explaining set that has the property that $\sum_{v \in A} Pl(v)$ is maximal among all complete parsimonious explaining sets.

Both parsimonious and best complete explaining sets for some graph are not necessarily unique. Furthermore, not every parsimonious explaining set is minimal-cardinality. The explanation graph in Figure 2.5 is a minimal example. Likewise, not every parsimonious explaining set is best complete. The explanation graph in Figure 2.6 is a minimal example. This last example also shows that parsimonious explaining sets are not necessarily unique.

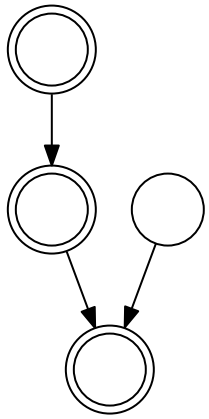


Figure 2.5. Counterexample to the claim that every parsimonious explaining set is minimum-cardinality.

2.3 Equivalence with satisfiability

The kinds of abduction problems that can be expressed with explanation graphs include all statements that can be expressed with propositional logic. That is, given some propositional statement, we can build an explanation graph so that the explanation graph has a complete explaining set if and only if the propositional statement is satisfiable. Of course,

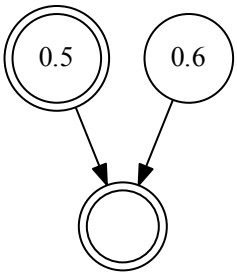


Figure 2.6. Counterexample to the claim that every parsimonious explaining set is best complete.

propositional logic does not allow the expression of *plausibility* or any other kind of graded confidence. So explanation graphs are capable of expressing more than propositional logic.

The details of how to translate a propositional statement into an equivalent explanation graph are provided by Bylander et al. (1991). They show how to translate a 3SAT problem into an equivalent abduction problem. Their abduction problems are specified slightly differently than our explanation graphs (i.e., non-graphically) but the formalisms are easily seen to be equivalent.

Figure 2.7 shows the basic building blocks of propositional statements as partial explanation graphs. These partial explanation graphs should be linked so that outgoing arrows connect to incoming arrows. Some vertices must have incompatibility relationships with others and some must be "accepted" by default. An example of a complex propositional statement expressed as an explanation graph is shown in Figure 2.8. Should this explanation graph have a complete explaining set, then the propositional statement is satisfiable. Furthermore, the acceptance status of the literal vertices (e.g., p, and $\sim p$ "not p") gives the truth-status of the literals so that the propositional statement is satisfied.

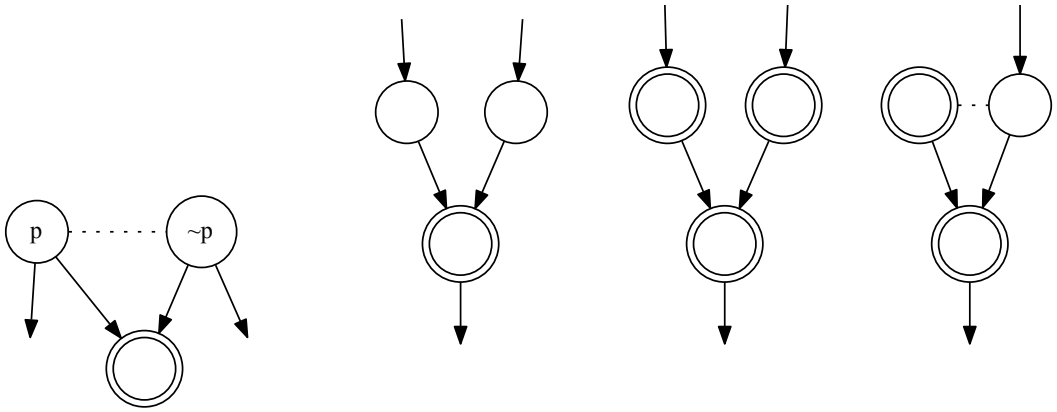


Figure 2.7. Basic building blocks for constructing explanation graphs equivalent to propositional statements. Outgoing arrows must connect to incoming arrows. For each distinct literal, such as p, both p and $\sim p$ (“not p”) must be included in the explanation graph in the manner shown in the first diagram.

2.4 Complexity of abduction

Bylander et al. (1991) classify certain set-covering abduction problems as *independent incompatibility* problems; these abduction problems include an incompatibility relation among pairs of hypotheses just like our use of I. The preceding formalization of abduction satisfies the independent incompatibility problem classification. Bylander et al. showed that it is NP-complete to determine whether a complete explaining set exists for an independent incompatibility problem. Additionally, they show that it is NP-hard to find a best complete explaining set.

Our formalization of abduction is also classified as (a subset of) cost-based abduction (CBA) problems (Charniak and Shimony, 1990, 1994). The goal is to find least-cost proofs (LCP), or equivalently, best complete explaining sets. Not surprisingly, finding least-cost proofs is NP-hard (Charniak and Shimony, 1994). Our formalization is equivalent to CBA

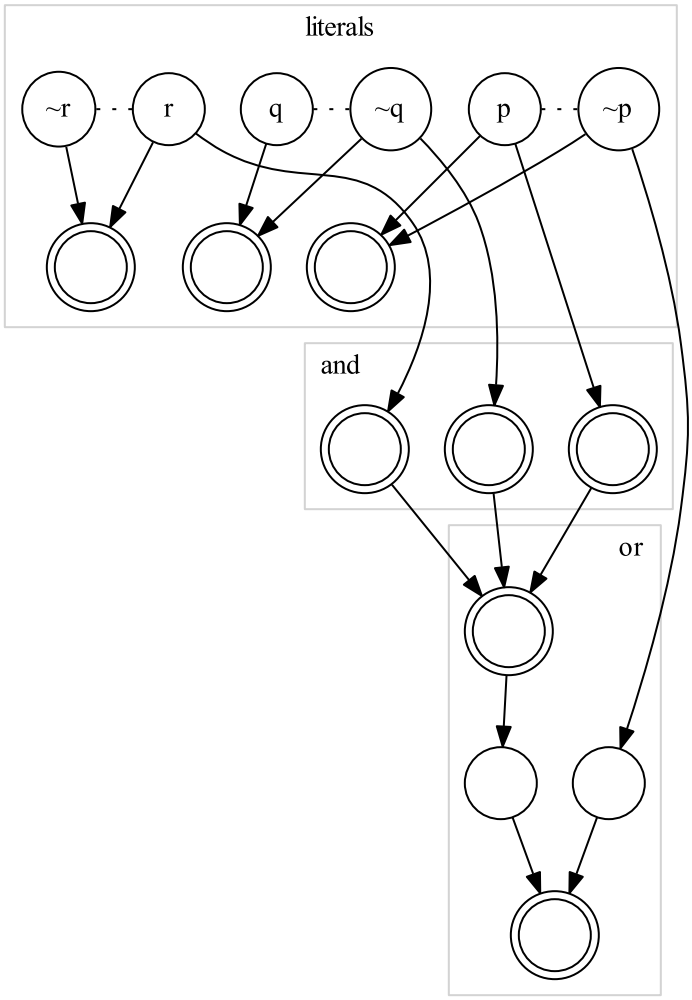


Figure 2.8. Explanation graph equivalent to the proposition $\neg p \vee (p \wedge \neg q \wedge r)$.

methods that are restricted so that every rule has only one antecedent. Abdelbar (2004) has shown that approximating LCPs within a fixed ratio c of the cost of an optimal solution is NP-hard for any $c > 0$. This result applies to general CBAs and the restricted CBAs where each rule has one antecedent. Finally, Eiter and Gottlob have shown (1995), for general propositional abduction, that determining whether a complete explaining set exists is in the complexity class Σ_2^P.

Thus, the state of affairs is clear. In the words of Eiter and Gottlob, "abduction is harder than deduction" (op. cit.). Furthermore, their results "clearly show that the major variants of logic-based abduction are very hard—in most cases even harder than classical propositional reasoning. Hence, there is no hope for complete and efficient algorithms that solve these problems" (op. cit.).

Josephson & Josephson have come to a similar conclusion. They propose to change the task definition. They write,

> [A]bductive inference appears to be ubiquitous in cognition. Moreover, humans can often interpret images, understand sentences, form causal theories of everyday events, and so on, apparently making complex abductive inferences in fractions of a second. [...] Clearly there is a basic tension among the intractability of the abduction task, the ubiquity of abductive processes, and the rapidity with which humans seem to make abductive inferences. [...] The new characterization of the abductive-assembly task is "explaining as much as possible," or, somewhat more precisely, "maximizing explanatory coverage consistent with maintaining a high standard of confidence." [...] The tractability of the task under the new description is demonstrated by giving an efficient strategy for accomplishing it (Josephson and Josephson, 1994, Ch. 9).

Josephson & Josephson developed the PEIRCE-IGTT abductive reasoning system in

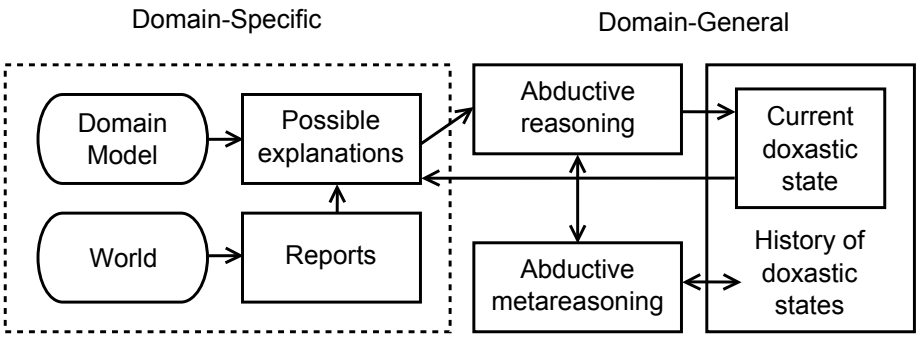


Figure 2.9. System architecture.

order to realize an efficient strategy for abductive reasoning. (Section 3.10 examines the PEIRCE-IGTT system in more detail.) The present work distills the PEIRCE-IGTT system into a collection of algorithms, defined below. Before we look at these algorithms in Sections 2.7 and 2.8, we first examine the architecture of our abductive reasoning system.

2.5 System architecture

Figure 2.9 shows the 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 may 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 (see Chapter 4). Abductive metareasoning might ask the domain-specific components to generate new hypotheses, revise the doxastic state, and/or leave the anomalies unexplained. Reports that remain unexplained

are considered by the cognitive system to be noise.

The next section addresses the domain-specific components. When implementing a new domain, only those components must be modified; the abductive reasoning and metareasoning systems can be reused. Section 2.2 previously defined *doxastic states*. Section 2.7 examines the abductive reasoning algorithms, and Chapter 4 provides complete detail about the metareasoning system.

2.6 Problem domains

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

Definition 2.6.1. A *problem domain M* is an opaque structure that provides the functions OBSERVE and GENERATEHYPOTHESES, defined as follows.

$$\textsc{Observe}(M) = (H_{\text{reports}}, Pl, I).$$

The OBSERVE function generates reports H_{reports} of putatively observed properties of the world. These reports come with plausibilities defined by Pl and incompatibility relations defined by I.

$$\textsc{GenerateHypotheses}(M, H_{\text{unexplained}}, H_{\text{accepted}}) = (H_{\text{hypotheses}}, X, Pl, I).$$

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, Pl, 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 func-

tions are black boxes. This separation enables us to experiment with different domains without modifying the reasoning engine.

The next two sections detail the abductive reasoning algorithms.

2.7 General abduction algorithms

We define the following functions to query and manipulate a doxastic state $D = (H, X, Pl, S, V, I)$.

$$
\begin{aligned}
\textsc{Explains}((\cdot, X, \cdot, \cdot, \cdot, \cdot), h_1) &= \{h_2 | (h_1, h_2) \in X\}. \\
\textsc{Hypotheses}((\cdot, X, \cdot, \cdot, \cdot, \cdot), h_1) &= \{h_2 | (h_2, h_1) \in X\}. \\
\textsc{Incompatible}((\cdot, \cdot, \cdot, \cdot, \cdot, I), h_1) &= \{h_2 | (h_1, h_2) \in I\}. \\
\textsc{Evidence}((\cdot, \cdot, \cdot, S, V, \cdot)) &= \{h | h \in V \wedge S(h) = \mathit{Accepted}\}. \\
\textsc{Accepted}((\cdot, \cdot, \cdot, S, \cdot, \cdot)) &= \{h | S(h) = \mathit{Accepted}\}. \\
\textsc{Rejected}((\cdot, \cdot, \cdot, S, \cdot, \cdot)) &= \{h | S(h) = \mathit{Rejected}\}. \\
\textsc{Undetermined}((\cdot, \cdot, \cdot, S, \cdot, \cdot)) &= \{h | S(h) = \mathit{Undetermined}\}. \\
\textsc{CandidateExp}(D, h) &= \textsc{Hypotheses}(D, h) \cap \textsc{Undetermined}(D). \\
\textsc{Unexplained}(D) &= \{h | h \in \textsc{Evidence}(D) \wedge \\
&\qquad \textsc{Hypotheses}(D, h) \cap \textsc{Accepted}(D) = \emptyset\}. \\
\textsc{ContrastSets}(D) &= \{\Gamma | h \in \textsc{Unexplained}(D) \wedge \Gamma = \textsc{CandidateExp}(D, h) \wedge \Gamma \neq \emptyset\}. \\
\textsc{Accept}((H, \cdot, \cdot, S, \cdot, \cdot), A) &= (H, \cdot, \cdot, S', \cdot, \cdot), \text{ where} \\
&\qquad S' = \{(h, s) | h \in H \wedge s = \mathit{Accepted} \text{ if } h \in A, \\
&\qquad\qquad s = S(h) \text{ otherwise}\}. \\
\textsc{Reject}((H, \cdot, \cdot, S, \cdot, \cdot), R) &= (H, \cdot, \cdot, S', \cdot, \cdot), \text{ where} \\
&\qquad S' = \{(h, s) | h \in H \wedge s = \mathit{Rejected} \text{ if } h \in R, \\
&\qquad\qquad s = S(h) \text{ otherwise}\}. \\
\textsc{RelatedHyps}(D, h) &= \{h\} \cup \bigcup_{h' \in R} \textsc{RelatedHyps}(D, h') \cup R, \text{ where} \\
&\qquad R = \textsc{Incompatible}(D, h) \cup \textsc{Hypotheses}(D, h). \\
\textsc{Undecide}((H, \cdot, \cdot, S, V, \cdot), h) &= (H, \cdot, \cdot, S', V, \cdot), \text{ where} \\
&\qquad R = \textsc{RelatedHyps}((H, \cdot, \cdot, S, V, \cdot), h), \\
&\qquad S' = \{(h', S(h')) | h' \notin R\} \cup \{(h', \mathit{Undetermined}) | h' \in R\}.
\end{aligned}
$$

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

Definition 2.7.1. A *partial abduction* is an operation that transforms one doxastic state into another. The statement $\textsc{PartialAbduce}(D_1) = D_2$ means that D_2 was 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 pieces of 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 such hypothesis is accepted.

3. One or no hypothesis is accepted.

4. When a hypothesis is accepted, all incompatible hypotheses are rejected.

Algorithm 2.1 The generic partial abduction function.

```
function PartialAbduce(D_0)
    Φ ← ContrastSets(D_0)
    if Φ = ∅ then                                  ▷ No contrast sets, nothing to accept
        return D_0
    else
        Let Γ ∈ Φ.                                 ▷ Pick out a contrast set
        Let h ∈ Γ.                                 ▷ Pick out some hypothesis
        D_1 ← Accept(D_0, {h})                     ▷ Accept h
        D_2 ← Reject(D_1, Incompatible(D_0, h))    ▷ Reject incompatible hypotheses
        return D_2
    end if
end function
```

Theorem 2.7.1. The function PARTIALABDUCE is a partial abduction according to Definition 2.7.1.

Proof. Each constraint in the definition of a partial abduction will be addressed in turn. Let D_0 be the doxastic state under consideration.

1. *If no unexplained evidence have hypotheses, or no evidence is unexplained, then the doxastic state is unchanged.* Should no hypotheses be available, or nothing is unexplained, then CONTRASTSETS(D_0) will return the empty set, and PARTIALABDUCE will leave D_0 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.* If there do exist possible explainers, then let Φ be the set of contrast sets for D_0. The PARTIALABDUCE function picks out one hypothesis from one such contrast set; let h be that hypothesis. Then, h is accepted, and no other hypothesis is accepted.

3. *One or no hypothesis is accepted.* As shown previously, only one hypothesis is accepted, if any hypotheses are acceptable. Otherwise, no hypothesis is accepted.

4. *When a hypothesis is accepted, all incompatible hypotheses are rejected.* The PARTIALABDUCE function rejects all hypotheses incompatible with h whenever some hypothesis h is accepted.

□

Lemma 2.7.1. If PARTIALABDUCE$(D_1) = D_2$, then either $D_1 = D_2$ or, instead, $D_1 \neq D_2$ and $\|\text{CONTRASTSETS}(D_2)\| < \|\text{CONTRASTSETS}(D_1)\|$. In other words, the PARTIAL-

ABDUCE function, applied to a doxastic state, either leaves the doxastic state unchanged or reduces the number of contrast sets.

Proof. Assume $D_1 \neq D_2$. It suffices to show that $\text{UNEXPLAINED}(D_2) \subset \text{UNEXPLAINED}(D_1)$. Since $D_1 \neq D_2$, some hypothesis h was accepted to explain some evidence e. Then $e \in \text{UNEXPLAINED}(D_1)$ but $e \notin \text{UNEXPLAINED}(D_2)$. Now suppose $e' \in \text{UNEXPLAINED}(D_2)$. It must be the case (by definition) that $\text{HYPOTHESES}(D_2, e') \cap \text{ACCEPTED}(D_2) = \emptyset$. The PARTIALABDUCE function does not add or remove hypotheses from the doxastic state, so $\text{HYPOTHESES}(D_1, e') = \text{HYPOTHESES}(D_2, e')$. We also know that $\text{ACCEPTED}(D_1) \subset \text{ACCEPTED}(D_2)$; actually, D_2 has exactly one more accepted hypothesis than D_1, namely, h. Therefore, $\text{HYPOTHESES}(D_1, e') \cap \text{ACCEPTED}(D_1) = \emptyset$, so $e' \in \text{UNEXPLAINED}(D_1)$. □

Definition 2.7.2. A doxastic state D is *finalized* if $\text{PARTIALABDUCE}(D) = D$.

The abductive reasoning procedure, defined by the ABDUCE function (Algorithm 2.2), 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 2.7.2. The FINALIZE function is guaranteed to terminate.

Proof. From Lemma 2.7.1, we have that either the partial abduction leaves a doxastic state unchanged, causing termination of the loop, or the partial abduction reduces the number of contrast sets. Since by Definition 2.2.1, the set of hypotheses, and hence the number of contrast sets, are finite, the algorithm is guaranteed to terminate. □

Algorithm 2.2 Various general abduction functions.

function ADDREPORTSTODOXASTICSTATE($(H_0, X_0, Pl_0, S_0, V_0, I_0), H, Pl, I$)
  $H_1 \leftarrow H_0 \cup H$
  $Pl_1 \leftarrow Pl_0 \cup Pl$
  $V_1 \leftarrow V_0 \cup H$
  $I_1 \leftarrow I_0 \cup I$
  $S_1 \leftarrow S_0 \cup \{(h, Accepted) | h \in H\}$ ▷ Accept all reports
  return $(H_1, X_0, Pl_1, S_1, V_1, I_1)$
end function
function ADDHYPOTHESESTODOXASTICSTATE($(H_0, X_0, Pl_0, S_0, V_0, I_0), H, X, Pl, I, \eta$)
  $H_1 \leftarrow H_0 \cup H$
  $X_1 \leftarrow X_0 \cup X$
  $Pl_1 \leftarrow Pl_0 \cup Pl$
  $I_1 \leftarrow I_0 \cup I$
  $S_1 \leftarrow S_0 \cup \{(h, Rejected) | h \in H \wedge Pl_1(h) < \eta\}$ ▷ Reject implausible hypotheses
  return $(H_1, X_1, Pl_1, S_1, V_1, I_1)$
end function
function FINALIZE(D)
  $D' \leftarrow$ PARTIALABDUCE(D)
  while $D' \neq D$ **do**
    $D \leftarrow D'$
    $D' \leftarrow$ PARTIALABDUCE(D)
  end while
  return D'
end function
function ABDUCE($M, D_0, \eta, DoMetareasoning?$)
  $(H_{\text{reports}}, Pl, I) \leftarrow$ OBSERVE($M, \emptyset$)
  $D_1 \leftarrow$ ADDREPORTSTODOXASTICSTATE($D_0, H_{\text{reports}}, Pl, I$)
  $H_{\text{unexplained}} \leftarrow$ UNEXPLAINED(D_1)
  $H_{\text{accepted}} \leftarrow$ ACCEPTED(D_1)
  $(H_{\text{hypotheses}}, X', Pl', I') \leftarrow$ GENERATEHYPOTHESES($M, H_{\text{unexplained}}, H_{\text{accepted}}$)
  $D_2 \leftarrow$ ADDHYPOTHESESTODOXASTICSTATE($D_1, H_{\text{hypotheses}}, X', Pl', I', \eta$)
  $D_3 \leftarrow$ FINALIZE(D_2)
  if *DoMetareasoning*? **then**
    $D_4 \leftarrow$ METAREASON(D_3) ▷ Refer to Chapter 4
    return D_4
  else
    return D_3
  end if
end function

2.8 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, as we saw earlier (Section 2.4), 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, hill-climbing 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. We call this its *decisiveness*.

Depending on the specifics of a task domain, one may wish to establish a minimum plausibility η (seen previously in Algorithm 2.2) and/or a minimum decisiveness threshold δ. The EFLI algorithm adds support for δ, which was missing in the generic ABDUCE function. Note that a large η or δ threshold might cause some reports or other beliefs to remain unexplained because either their possible explainers are too implausible or no al-

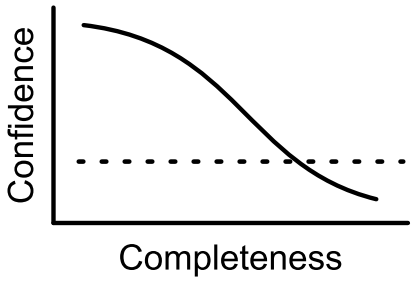


Figure 2.10. Completeness–confidence trade-off. This trade-off is dictated by the parameters η and δ. Greater values for these thresholds yield more confidence but may sacrifice completeness. For some abductive reasoning task, there may exist an ideal trade-off, represented by the dotted line.

ternative is sufficiently decisive. However, one would expect greater accuracy, thus greater confidence in the belief state. Figure 2.10 illustrates this trade-off.

The EFLI algorithm (Algorithm 2.3 extends the PARTIALABDUCE function (Algorithm 2.1) by involving the following additional parameters:

- δ, the *minimum decisiveness* parameter that establishes how decisive the best hypothesis in a contrast set must be in order to be accepted.
- $\geq_{\text{hyp}}$, a preference relation on vertices in a contrast set.
- $\geq_{\text{contrast}}$, a preference relation on contrast sets.

These relations are defined as,

$$\{h_1,\ldots,h_m\} \geq_{\text{contrast}} \{h'_1,\ldots,h'_n\} \quad \text{iff} \quad Pl(h_\alpha) - Pl(h_\beta) \geq Pl(h'_\alpha) - Pl(h'_\beta),$$
$$h \geq_{\text{hyp}} h' \quad \text{iff} \quad Pl(h) \geq Pl(h'),$$

where h_α and h_β are the first- and second-most plausible hypotheses, respectively, in the contrast set $\{h_1,\ldots,h_m\}$, and h'_α, h'_β likewise for the contrast set $\{h'_1,\ldots,h'_n\}$.

We define *arbitrary abduction* as the algorithm where $\geq_{\text{contrast}}$ and $\geq_{\text{hyp}}$ are random binary relations (i.e., $\geq_{\text{contrast}}$ and $\geq_{\text{hyp}}$ have a 50% probability of being true for any two contrast sets or hypotheses) and setting $\delta = 0$.

Algorithm 2.3 The EFLI partial abduction function.

```
function PARTIALABDUCE_EFLI(D_0)
    Φ ← CONTRASTSETS(D_0)
    if Φ = ∅ then                                  ▷ No contrast sets, nothing to accept
        return D_0
    else
        Γ ← max_{≥contrast} Φ                      ▷ Find most decisive contrast set
        h ← max_{≥hyp} Γ                           ▷ Find best hypothesis
        if ||Γ|| = 1 then                          ▷ If h is an essential explainer...
            D_1 ← ACCEPT(D_0, {h})                 ▷ Accept h
        else
            h' ← max_{≥hyp} Γ \ {h}                ▷ Get second best hypothesis
            if Pl(h) − Pl(h') ≥ δ then             ▷ If h is sufficiently decisive...
                D_1 ← ACCEPT(D_0, {h})             ▷ Accept h
            end if
        end if
        D_2 ← REJECT(D_1, INCOMPATIBLE(D_0, h))    ▷ Reject incompatible hypotheses
        return D_2
    end if
end function
```

Theorem 2.8.1. The function PARTIALABDUCE$_{EFLI}$ is a partial abduction.

Proof. The proof is trivial, since PARTIALABDUCE$_{EFLI}$ is a variant of PARTIALABDUCE that just specializes how hypotheses are preferred. □

2.9 Complexity of abduction with EFLI

Suppose we have r reports, and that each of the r contrast sets contains n hypotheses. The $\max_{\geq_{\text{contrast}}}$ operation must sort each contrast set, requiring time $O(rn\log n)$. Since each contrast set is already sorted, $\max_{\geq_{\text{hyp}}}$ requires $O(1)$ time. Therefore, a single iteration of PARTIALABDUCE$_{EFLI}$ requires $O(rn\log n)$ time. To find explainers for all the reports,

the partial abduce function is repeated $O(r)$ times. Thus, the computational complexity of abduction with EFLI comes to $O(nr^2 \log n)$.

2.10 Errors

Now that we have defined abductive reasoning, we might want to ask, "how can it fail to arrive at true beliefs?" We define an *error* as the acceptance of a false hypothesis or the failure to accept a true hypothesis. Of course, the cognitive system does not have independent access to truth, but we as experimenters may examine a doxastic state and identify its true and false beliefs according to the ground truth established for an experiment. We can classify errors according to the following definitions.

Plausibility

A false hypothesis was accepted but one of its lower-plausibility rivals (from the same contrast set) was true. Or, a true hypothesis was not accepted but it was a lower-plausibility rival when a false hypothesis was accepted. It is worth nothing that the EFLI algorithm is not at fault in this case. Rather, the domain gave inaccurate plausibility estimates.

MinPlausibility

A true hypothesis was rejected because its plausibility did not meet the minimum plausibility η. No errors of this sort are possible when $\eta = 0$.

MinDecisiveness

A true hypothesis was not accepted because it was not sufficiently decisive according to the minimum decisiveness δ threshold. No errors of this sort are possible when $\delta = 0$.

NoExplOffered

A false hypothesis was accepted because no true hypothesis for a true sensor report was ever offered.

Noise

A false report was accepted, or a false hypothesis was accepted in order to explain a false report. No errors of this sort are possible when noise is absent.

There exists one more case in which either a true hypothesis is not accepted or a false hypothesis is accepted. There are two possible scenarios that realize this case:

1. A true hypothesis h_T is rejected owing to incompatibility with a false but accepted hypothesis h_F. Or,

2. A false hypothesis h'_F is accepted to explain some d but the true explainer of d, which we call h_T, was rejected owing to incompatibility with some false but accepted hypothesis h_F.

Suppose, for example, that some false hypothesis h_F is an *essential* hypothesis for some true evidence d. By *essential*, we mean at the time h_F was accepted to explain d, no other hypothesis was available to explain d. Further suppose that the true hypothesis h_T for that report was previously rejected but not because of to not meeting the η threshold. If we look further back in the chain of acceptances and rejections, we inevitably find that the true hypothesis h_T was rejected for a particular reason. It may be that a false incompatible hypothesis was accepted, causing h_T to be rejected. Then, owing to the cascade of errors, h_F was accepted as a downstream result of a Plausibility error that initially affected h_T. Thus, in some cases, we must look at prior acceptances and rejections in order to determine the original cause for an error.

2.11 Detectable and undetectable errors

The kinds of errors defined above can be measured from the experimenter's viewpoint, but not from the system's viewpoint since each kind is defined in terms of the *truth* of certain hypotheses. The system does not have access to truth, rather it is trying to estimate it. Nevertheless, a reasonable question to ask is, under what circumstances can the system *detect* and *correct* its errors?

Though errors themselves cannot be directly detected, perhaps errors can be discovered via some kind of proxy indicator. Pomeranz and Reddy (2010) describe how undetectable faults in logic circuits can influence the behavior of detectable faults. They demonstrate that more detectable faults can be identified if the range of possible undetectable faults are considered in tandem with the detectable faults. In this work, we take an analogous approach. While errors cannot be directly detected, we might be able to find *symptoms of errors*. In this work, we take unexplainable reports and beliefs to be *anomalous*, and show that the presence of such anomalies indicates that the system might have committed an error (accepted a false belief or failed to accept a true belief). The presence of anomalies is considered evidence for errors, but does not imply errors. The question of whether the anomalies point to errors is handled in Chapter 4, where we outline abductive metareasoning.

2.12 Conclusions

Abductive reasoning is, arguably, a kind of commonsense reasoning that is capable of inferences from evidence to causes. The "making sense" phase of cognition is rich with abductions since the goal of sense-making is to reason from evidence, in the form of observations, claims made by other agents, etc., to beliefs about the world. In our formalization, abductive reasoning is a process that involves these steps:

1. Gathering evidence with the OBSERVE function.

2. Gathering possible explainers of the evidence with the GENERATEHYPOTHESES function.

3. Iteratively accepting and rejecting hypotheses to arrive at a plausible, complete or as-complete-as-possible composite explanation.

Our formalization is domain-general. The abductive reasoning process has no access to the *contents* of the evidence and hypotheses; that is to say, the algorithms only operate on the two relationships *explanatory* and *incompatible* between and among hypotheses and evidence. Because the problems of finding a complete explaining set and the best complete explaining set (in terms of plausibility) are intractable, we built a greedy, hill-climbing (non-backtracking) abductive reasoning algorithm.

The result of abductive reasoning is some explaining set. This explaining set contains the hypotheses that the cognitive system *believes*. We have identified the reasons these beliefs might be in error, including inaccurate plausibility estimates and noisy reports. Many errors are not detectable, but in some cases the existence of *unexplainable* reports or other beliefs indicates the possibility that some beliefs are false. Unexplainable reports, or *anomalies*, are easily detected, and their presence activates an *abductive metareasoning* system that attempts to correct false beliefs, and identify noise, by finding explanations and hence repairs for the anomalies. Abductive metareasoning will be discussed in Chapter 4 after a brief look at prior work in abductive reasoning.

Chapter 3: Prior work in abductive reasoning

A wide variety of abductive reasoning procedures have been designed for artificial intelligence applications. Abduction has been used to support diagnosis, planning, and probabilistic inference. In this chapter, we summarize several approaches related to diagnosis, since diagnosis is essentially the kind of "making sense" that we focus on in this work. In diagnostic abduction, the input to the reasoning process is a set of evidence and hypotheses, or a knowledge base from which hypotheses may be generated, and the output is a set of categorical beliefs (accepted hypotheses). The varieties of diagnostic abduction differ in how they represent evidence, hypotheses, and beliefs, how composite explanations are assembled, and what makes one explanation better than another. The last approach that we examine, the PEIRCE-IGTT framework, is most similar to our abductive reasoning system detailed previously (Chapter 2).

3.1 A framework for comparing approaches

This review analyzes various computational abductive reasoning systems across four dimensions:

Explanation: What consitutes an explanation? In some accounts, an explanation must logically entail what it explains. In other accounts, an explanation must increase probability of the explananda. Sometimes, the relationship is more nuanced, and sometimes less.

Trigger: When should abductive reasoning be invoked? One might expect, for example, that if the "making sense" phase of cognition can be achieved in a particular situa-

tion with deductive reasoning, then there is no reason to pursue abductive reasoning. What I call a trigger has previously been called a "cognitive irritant" (Garcez et al., 2007).

Hypothesis generation: How are the possible explainers generated?

Criteria for the "best" explanation: How is one explanation preferred over another? There are many possible criteria for what makes one explanation better than another: it is simpler, it posits fewer new causal mechanisms, it has higher plausibility or probability, etc. In the work we review, the criteria for the best explanation are always decidable but not necessarily efficiently computable.

3.2 Abductive logic programming

Abductive logic programming (ALP), introduced by Kakas et al. (1992) and influenced by earlier work on THEORIST (Poole et al., 1986) and inductive generalization (Plotkin, 1970), treats abduction as reverse deduction. ALP has been implemented as an extension to Prolog (Fung and Kowalski, 1997) and subsequently integrated with constraint programming (Endriss et al., 2004; Kakas et al., 2000). The ALP extension is activated when the underlying Prolog system fails to find a proof for some query. The ALP process then determines whether asserting one or more of a reserved set of *abducibles* (atomic sentences) would allow the proof to succeed. An *explanation* is the smallest set of abducibles sufficient to prove the query. Integrity constraints may be specified to limit possible explanations.

- *Explanation*: A conjunction of abducibles A explains ϕ if ϕ is not already provable, ϕ would be provable if each of $\alpha \in A$ were true, $\neg\alpha$ is not provable for each $\alpha \in A$, and integrity constraints are satisfied. In symbols, A is an explanation if,

$$(\Theta \nvDash \phi) \wedge (\Theta \cup A \models \phi) \wedge (\forall \alpha \in A : \Theta \nvDash \neg\alpha) \wedge (\Theta \cup A \models I),$$

where Θ is the background theory and I are integrity constraints.

- *Trigger*: Explanations are sought when a query cannot be proved.
- *Hypothesis generation*: Abducibles are specified upfront, as are integrity constraints.
- *Criteria for "best"*: Fewest number of abducibles (i.e., a minimal explanation).

We will see later that abduction-as-reverse-deduction does not capture all the properties of commonsense abductive reasoning. Nevertheless, ALP gives a clean, programmatic means of representing and solving abduction problems.

3.3 Semantic tableaux

Semantic tableaux are another computational model for abduction-as-reverse-deduction. Though semantic tableaux have been around for several decades (Beth, 1961; Hintikka, 1955), Aliseda has recently utilized them as a semantically and computationally convenient way for generating abductive explanations (Aliseda, 2006, Ch. 4). Semantic tableaux allow testing whether a formula follows from a set of other formulae. A tableau $\mathscr{T}$ for a set of formulae Θ, written $\mathscr{T}(\Theta)$, is constructed by listing the formulas "vertically" in a single-branch "tree," then building branches at leaf nodes such that for every occurrence in higher parts of the tree of a two-part disjunction $a \vee b$ (where $\mu \rightarrow \psi$ is written $\neg\mu \vee \psi$), two branches are added at each leaf node, representing the case when a is true and the case when b is true. For each conjunct $a \wedge b$, both a and b are added "vertically" to the leaves of the tree but without introducing new branches. If a branch (traced back to the root) contains both a and $\neg a$ (for any formula a), then that branch closes. A closed branch indicates a contradiction may be reached from its formulae. If all branches in $\mathscr{T}(\Theta)$ are closed, then there is no assignment of truth-values to the atoms of the language that satisfies Θ. If any branch remains open, then the initial formulae in Θ are jointly satisfiable. Thus, in

order to test if a formula μ follows logically from premises Θ, a tableau is constructed as $\mathscr{T}(\Theta \cup \{\neg\mu\})$. If that tableau is closed (all branches are closed), then μ follows from Θ (because $\Theta \cup \{\neg\mu\}$ is not satisfiable); otherwise, μ is not a valid consequence from Θ, because $\Theta \cup \{\neg\mu\}$ is satisfiable.

Aliseda shows that when a tableau indicates that μ does not follow from Θ, that is, when $\Theta \cup \{\neg\mu\}$ has one or more open branches, we can read off the tableau the exact truth-assignments of atomic formulae that show μ not to be a consequence of Θ. Then we only need to find which formulae to take out of Θ, to get Θ', say, so that $\mathscr{T}(\Theta' \cup \{\neg\mu\})$ is completely closed; that is, we just need to close the open branches (while keeping $\neg\mu$ around). If we can do so, then we have effectively come up with formulae Ψ such that $\Theta \cup \Psi$ entails μ. This makes Ψ an *abduction*. Further, if it's not the case that Ψ itself entails μ, but only entails μ when combined with Θ, then Ψ is an *explanation* in Aliseda's terminology.

Semantic tableaux do not, in general, yield polynomial-time abductions. In fact, the computational complexity of semantic tableaux is sometimes even worse than truth tables (D'Agostino, 1992). However, as Aliseda has argued, semantic tableaux yield convenient algorithms for producing explanations.

- *Explanation*: Unlike ALP, an explanation α found with a semantic tableaux need not be an atomic sentence. However, like ALP, an explanation must entail the explanandum.

$$(\Theta, \alpha \models \phi) \wedge (\alpha \not\models \phi) \wedge (\Theta \not\models \neg\alpha) \wedge (\Theta \not\models \phi)$$

- *Trigger*: Explanations are sought when a query cannot be proved.

- *Hypothesis generation*: Aliseda states: "First compute abductions according to the plain version [see below] and then eliminate all those that do not comply with the various additional requirements [i.e., consistent and explanatory]."

Plain version: $\mathscr{T}(\Theta \cup \{\neg\phi, \alpha\})$ is closed, i.e., $(\Theta, \alpha \models \phi)$.

Consistent: Plain abduction in addition to $\mathscr{T}(\Theta \cup \{\alpha\})$ is open, i.e., $(\Theta \not\models \neg\alpha)$.

Explanatory: Plain abduction in addition to:

1. $\mathscr{T}(\Theta \cup \{\neg\phi\})$ is open, i.e., $(\Theta \not\models \phi)$.
2. $\mathscr{T}(\{\neg\phi, \alpha\})$ is open, i.e., $(\alpha \not\models \phi)$.

- *Criteria for "best"*: Aliseda does not specify criteria for the best explanation. Instead, she considers any formulae that meet the criteria above ("Hypothesis generation") to be warranted explanations.

3.4 A knowledge-level account

Abductive logic programming and semantic tableaux take a *logical* approach to abduction. They treat abduction essentially as reverse deduction, and exploit features of the symbol-level representations of knowledge to do their work. Levesque (1989) attempts to abstract away from symbol-level details such as how evidence and explanations are represented. He establishes a formalism based on *beliefs* that are expressed in a standard propositional language $\mathscr{L}$. Sentences of the language of beliefs, $\mathscr{L}^*$, have the form $\mathbf{B}\alpha$ where $\alpha \in \mathscr{L}$. This language $\mathscr{L}^*$ allows us to describe what is and is not believed. Beliefs may differ by type, so we write $\mathbf{B}_\lambda \alpha$ to indicate a belief of type λ. An epistemic state e determines which atomic sentences from $\mathscr{L}^*$ are believed; we write $e \vDash \mathbf{B}_\lambda \alpha$ to indicate that $\mathbf{B}_\lambda \alpha$ is true in the epistemic state e. Levesque does not commit to how propositions are defined. Rather, he uses the notation $\|\alpha\|$ to denote the proposition that α expresses.

Levesque argues that if abduction is simply "reverse deduction," there is a problem of uniqueness and relevance.[1] Consider a medical domain where sentences in $\mathscr{L}$ stand

[1]This point is also made by Mayer and Pirri (1996).

for properties of the patient. Suppose that we know that $male$ and $hepatitis \rightarrow jaundice$. If we observe *jaundice* in the patient, we want to reason abductively to $hepatitis$. Were abduction simply reverse deduction, abduction would not necessarily generate a unique explanation or even finitely-many explanations. For example, $(\neg\neg hepatitis \wedge migraines) \vee (hepatitis \wedge \neg migraines)$ also accounts for *jaundice* in that it is consistent with what is known and, were it to be true, then *jaundice* would also be true. One can imagine other arbitrary propositions that also seem to *account for jaundice* but do not seem to be explanatory.

The preferred explanation is, naturally, $hepatitis$. Levesque shows that this preference cannot be realized from only logical considerations. Rather, some non-logical criteria such as *simplicity* are required to select the best explainers. Formally, explanation is defined by Levesque as follows, where e is an epistemic state and λ is a type of belief:

Definition 3.4.1. α $\text{expl}_\lambda \beta$ wrt e iff $e \models [\mathbf{B}_\lambda(\alpha \supset \beta) \wedge \neg \mathbf{B}_\lambda \neg \alpha]$.

An explanation α is *simpler* than an explanation α' whenever α effectively "contains fewer propositional letters" than does α'. Because Levesque states that abduction should produce the simplest explanations, he defines the abduction operation *Explain* as follows, where α $\text{min-expl}_\lambda \beta$ is true whenever α is a minimal (simplest) explanation of β.

Definition 3.4.2. $\text{Explain}_\lambda [\![e, \beta]\!] = \|\{\alpha | \alpha \text{ min-expl}_\lambda \beta \text{ wrt } e\}\|$.

Thus, Levesque has given a *knowledge level* account of abduction by abstracting away from *symbol level* details such as how propositions are written, what specific types of *belief* are involved, how an explanation can be said to account for what it explains, and so on. He also has shown that abduction is not merely reverse deduction. Paul (1993) has shown how this knowledge-level account of abductive reasoning can be modeled with an assumption-based truth-maintenance system (de Kleer, 1986).

However, his definition of *simpler* and requirement that the best explanation(s) be the simplest is too limiting. What makes an explanation best might depend on various factors, some normative and some pragmatic, such as those listed earlier and borrowed from Josephson and Josephson (1994). Moreover, Levesque's account does not include quantified beliefs.

3.5 Binary-choice Bayesian abduction (MEDAS)

An entirely different kind of abduction is Bayesian abduction. An early example of a probabilistic abductive reasoning system is the Medical Emergency Decision Assistance System (MEDAS) developed by Ben-Bassat et al. (1980). It was designed to "provide the clinician with decision aids from the time the patient is first seen in the emergency department until the immediate risk of life has been minimalized." Their model includes binary features and disorders, where a feature is any patient data such as age, sex, complaints and symptoms, results of medical tests, etc. Features that are not originally binary are transformed into a finite number of discrete features by specifying different ranges (e.g., heart rate $< 50/\text{min}$, between $50-100/\text{min}$, and $> 100/\text{min}$). A cost is assigned to each feature that specifies the financial cost or risk required to test for the feature.

Disorders are related to features by "characterizing patterns." A characterizing pattern is composed of a set of features which are relevant for diagnosing the disorder, as well as conditional probabilities of the form $P_{ij} = P(X_j|D_i)$ and $\bar{P}_{ij} = P(X_j|\neg D_i)$, where X_j is a feature among the set relevant to the disorder and D_i is the disorder. Additionally, each disorder has a prior probability of occurrence, $P(D_i)$.

The authors assume conditional independence among the features, so the probability of a disorder D_i given features $x_1, \ldots, x_k$ is as follows:

$$P(D_i|x_1,\ldots,x_k) = \frac{P(D_i)f_i(x_1)\cdots f_i(x_k)}{P(D_i)f_i(x_1)\cdots f_i(x_k) + (1-P(D_i))\bar{f}_i(x_1)\cdots \bar{f}_i(x_k)},$$

where,

$$f_i(x_j) = \begin{cases} P_{ij} & \text{if } x_j = 1, \\ 1 - P_{ij} & \text{if } x_j = 0. \end{cases}$$

$$\bar{f}_i(x_j) = \begin{cases} \bar{P}_{ij} & \text{if } x_j = 1, \\ 1 - \bar{P}_{ij} & \text{if } x_j = 0. \end{cases}$$

The authors provide an approximate method to efficiently handle cases where conditional independence does not hold among the features for a given disorder. Its details are not important for our purposes.

The MEDAS systems has two processes: a bottom-up process that determines which disorders are present, and a top-down process that determines which unobserved features should be tested. The bottom-up process compares $P(D_i)$ and $P(\neg D_i)$; if the former is greater than the latter, the disorder is *abduced* to be present. The top-down process evaluates the cost of each test and how well the test results may disambiguate which disorders are present.

The results of the abductive, bottom-up process are presented to the physician, who makes the decision about accepting that the abduced disorders are present or performing suggested tests, updating the feature set, and executing the bottom-up process again.

- *Explanation*: Disorders explain a set of features whenever the set of features is part of the disorder's characterizing pattern. Thus, this relationship is established up front by domain experts.

- *Trigger*: Physicians initiate the bottom-up abductive process.

- *Hypothesis generation*: Hypotheses take the form of disorders which are established up front.

- *Criteria for "best"*: No "best" disorder is automatically chosen. Rather, the results of abduction are presented to the physician for further consideration. Each possible explanation is scored according to the posterior probability of the disorder, given the features.

Other varieties of Bayesian abduction are addressed in Section 8.8.

3.6 Explanatory coherence

Yet another novel approach to abduction is a *coherence*-based approach, developed by Thagard (1989). Here, *explanation* and *explain* are taken as primitives. The goal is to arrive at a coherent explanation of some evidence.

> We should accept propositions that are coherent with our other beliefs, reject propositions that are incoherent with our other beliefs, and be neutral toward propositions that are neither coherent nor incoherent. Acceptability has finer gradations that just acceptance, rejection, and neutrality, however: The greater the coherence of a proposition with other propositions, the greater its acceptability (Thagard, 1989).

He defines coherence as follows. "Propositions P and Q cohere if there is some explanatory relation between them." More specifically, at least one of the following must be true:

1. P is part of the explanation of Q.
2. Q is part of the explanation of P.
3. P and Q are together part of the explanation of some R.

4. *P* and *Q* are analogous in the explanations they respectively give of some *R* and *S*.

Propositions may also have an incoherence relation, such as, but not limited to, propositions that contradict each other. The goal is to find greatest global coherence of a system of propositions, defined as "a function of the pairwise local coherence of those propositions."

Global coherence is found by encoding coherence and incoherence relations in a neural network structure with excitatory and inhibitory links, respectively. The initial weights of these links are parameters that must be established *a priori*. Each node represents a proposition. The evidence is activated, which propagates positive or negative influences to connected nodes. When the network is sufficiently "settled," we can read off the *acceptability* of each proposition by examining its final activation strength. Positive activations indicate acceptable propositions, negative activations indicate the opposite. In this way, global coherence is achieved.

Earlier work by Reggia (1985) takes a similar approach as Thagard's, but is concerned more with modeling associative memory with value-passing systems (such as neural networks). However, the following quote illustrates the conceptual similarity.

> In memory models implemented as value-passing systems, each processing node typically represents a "concept" or "hypothesis," and the level of activation associated with a node represents the relevance of or confidence in the concept/hypothesis represented by that node (Reggia, 1985).

We can summarize Thagard's system using our framework for comparing different approaches to abductive reasoning.

- *Explanation*: In Thagard's system, explanation is an undefined binary relation. Perhaps more important is *coherence*. Propositions *P* and *Q cohere* if at least one of the

four criteria above are met.

- *Trigger*: Abduction is activated by the user.
- *Hypothesis generation*: Hypotheses (propositions) are established ahead of time.
- *Criteria for "best"*: The best explanation is those propositions with high activations as found in a settled globally-coherent neural network.

Other researchers have also studied abduction with neural networks (Abdelbar et al., 2003; Garcez et al., 2007; Goel and Ramanujam, 1996).

3.7 INTERNIST

We have now examined logic-based, Bayesian, and coherence-driven abduction. More simplistic accounts may be found in the very early attempts at doing abduction in machines. The INTERNIST-I expert system (Miller et al., 1985; Pople et al., 1975), previously called DIALOG (DIAgnostic LOGic), was one of the earliest abductive reasoning systems. Its purpose was to assist internal medicine clinicians by inferring diseases from symptoms. The knowledge base includes "disease entities" and an associated list of "manifestations" known to be associated with each disease entity. A hierarchy of disease categories organizes the disease entities. In some cases, if a specific disease entity cannot be determined, a more general node in the disease hierarchy (e.g., liver disease) might be inferable. Note that no two disease entities may be incompatible.

- *Explanation*: Disease entities are said to explain manifestations. These relationships are stored in the knowledge base.
- *Trigger*: As each manifestation is entered into the system, nodes in the disease hierarchy are "evoked" (entered into consideration).

- *Hypothesis generation*: All possible manifestations and disease entities they evoke are stored in the knowledge base. A hypothesis might be ruled out if further tests do not produce outcomes predicted by the hypothesis.

- *Criteria for "best"*: Each evoked disease hypothesis is scored in terms of various criteria including an "evoking strength," how many observations it can explain, how many expected symptoms for this disease were observed to be present, etc.

Further development on the INTERNIST-I system led to INTERNIST-II (Pople, 1977). While INTERNIST-I sequentially built an explanation, it was not able to find and separate subproblems. Real-world experience showed that physicians are able to identify "obvious" subproblems involving a subset of the evidence that had uncomplicated solutions. INTERNIST-II uses heuristics to find subproblems and generate and score hypotheses for those subproblems independently. Thus, a disease entity does not necessarily have to be highly plausible for all the manifestations, but rather only a relevant subset. Multiple disease entities may make up the final composite explanation.

3.8 Set covering

The INTERNIST model can be described as a *set covering* model. This model was distilled and extensively developed by Reggia et al. (1985a,b) as Generalized Set Covering (GSC). This approach defines the abductive task as one of finding a minimal set of *disorders* that explain or "cover" the reported *manifestations*. An abductive problem, or *diagnostic problem*, is defined by the set of disorders, manifestations, their relations (which disorders explain which manifestations), and a set of reported manifestations. Note that there is no way to represent incompatibility among disorders. An *explanation* is a set of disorders such that all manifestations are explained and no subset of the explanation is also an explanation (the

explanation is therefore "parsimonious," i.e., has minimum cardinality, or "irredundant," that is, no proper subset of the explanation also covers all the manifestations).

- *Explanation*: Disorders are said to explain or "cover" manifestations. These binary relationships are stored in the knowledge base.
- *Trigger*: As each manifestation is entered into the system, nodes in the disease hierarchy are "generated" (entered into consideration).
- *Hypothesis generation*: All possible manifestations and disorders are stored in the knowledge base. A hypothesis (disorder) might be ruled out if further tests do not produce outcomes predicted by the hypothesis.
- *Criteria for "best"*: The best explanation is parsimonious or irredundant.

Finin and Morris (1989) also explore set-covering abduction in comparison with INTERNIST, MIDAS, and other approaches. They note that computing the best explanation in the GSC framework is exponential in the worst case, since finding the best might require evaluating all covering sets of disorders. The criteria for "best" might also be problematic. They note that the GSC suffers from the "rare disease problem," wherein the minimal covering explanation often posits a rare disease that explains all the symptoms. But, due to its rarity, it is often not the true explanation. They write, "all of these approaches go as far as their underlying theory will take them, and then use heuristics to carry on with differential diagnosis." The GSC is capable of specifying what constitutes a covering explanation but relies on heuristics to narrow down to the best (parsimonious) explanation. Heuristics might also be needed to control the search for a covering explanation. Finin and Morris state that it is important to have a solid underlying theory (such as set covering) but the tricky part is finding the right heuristics to complete the abductive task.

3.9 PEIRCE

Another variety of the set covering model of abduction, though significantly extended beyond GSC, is the PEIRCE system, named after C. S. Peirce who coined the term *abduction*. The PEIRCE system (Josephson and Josephson, 1994, Ch. 4) decompiles various tasks of abductive reasoning into a goal/subgoal architecture and allows definition of various *methods* that achieve particular subgoals. The top-level goal of finding the best explanation is decompiled into three subgoals (quoted from Josephson and Josephson):

1. generation of a set of plausible hypotheses
2. construction of a compound explanation for all the findings
3. criticism and improvement of the compound explanation

Each of these goals may be solved by some method. A method for solving the goal of "generation of a set of plausible hypotheses" might further break down the task into subgoals of,

1. query knowledge base for hypotheses that can explain one or more findings
2. determine relationships, such as incompatibility, among the hypotheses
3. estimate the plausibility of each hypothesis
4. filter out very implausible hypotheses

Each of these subgoals presumably has a method to solve it, which might "bottom out" into a particular action such as querying a database. Since several methods may be applicable for solving a particular subgoal, PEIRCE uses a "sponsor-selector" mechanism

for controlling method activation. Methods are grouped under "sponsors," which examine the active subgoals and provide the set of applicable methods. Then "selectors" go through this set and choose the most appropriate method(s) according to ratings provided by the sponsors. This design allows knowledge engineers to focus on adding, removing, and modifying sponsors, selectors, and methods independently without worrying about the details of the larger system.

Our normal breakdown of abductive reasoning into explanation, trigger, hypothesis generation, and criteria for "best," do not apply to the PEIRCE system because PEIRCE allows each of these factors to be multiply specified by various *methods*. Which method is chosen at a particular time to solve a particular subgoal depends on the sponsors and selectors. Thus, the PEIRCE system is much more general than the earlier systems we have examined.

It is worth nothing that PEIRCE also includes a built-in control mechanism known as "delay hard decisions." When some sponsors give different methods equal or near-equal ratings, i.e., different hypotheses seem equally plausible, then the system has encountered a "hard decision." Intuitively, arbitrarily choosing among equally good alternatives is not a winning strategy in the long run. Rather, if the decision can be delayed and "serendipitously" worked out by some other means (using Josephson and Josephson's terminology), then the hard decision can be avoided altogether. Such serendipity might come by finding that some other, very plausible hypothesis is accepted to explain some other evidence but happens to be incompatible with all but one of the hypotheses in the hard decision. The alternatives in the hard decision are thereby rejected, rendering the decision trivial since only one hypothesis remains. The hope is that by delaying hard decisions, more complete explanatory coverage and a more plausible final composite explanation may be found without resorting to a backtracking, try-all-combinations algorithm.

3.10 PEIRCE-IGTT

We pointed out in Section 2.4 that finding the best complete explaining set for abduction problems that involve pair-wise incompatibility relations is NP-hard. Yet, as Josephson and Josephson (1994, Ch. 9) point out, abductive reasoning, as described early in Chapter 2, is an intuitive and recognizable cognitive process of explaining evidence. By abductive reasoning, cognitive agents are able to reason from evidence to causes and acquire evidence-based knowledge about the world. If this process is truly NP-hard, then it seems impossible that cognitive agents are able to do it. Perhaps the goal is too strict. Josephson and Josephson reformulate the problem as one of "maximizing explanatory coverage consistent with maintaining a high standard of confidence" (op. cit.). They then extend the PEIRCE system as a next-generation system known as PEIRCE-IGTT (Integrated Generic Task Toolset) that realizes a computationally efficient process for solving this reformulated problem.

The PEIRCE-IGTT system incorporates the EFLI strategy described earlier (Section 2.8). It retains much of the generality of PEIRCE by supporting independent specification of various methods for achieving various subgoals. For each kind of subgoal, a default method is defined, though the default may be overridden in certain circumstances. The default is usually the most efficient way of solving the subgoal. For example, the EFLI algorithm is the default method for assembling a composite explanation.

- *Explanation*: An explanation is a consistent composite of hypotheses that plausibly explain the evidence.

- *Trigger*: After evidence is entered into the system, hypotheses are generated that are able to explain some of the evidence. In a "layered" problem, in which some hypotheses also require explanation should they be accepted, more hypotheses are

generated when necessary. The presence of unexplained evidence triggers the abduction machinery to find a best explanation.

- *Hypothesis generation*: The PEIRCE-IGTT system does not specify exactly how hypotheses are generated; rather, like the PEIRCE system, hypothesis generation is a subgoal that is achieved by specialized methods appropriate for the particular problem domain. The IGTT component might provide assistance by supporting database queries and hierarchical search among a knowledge base of cause–effect interactions.

- *Criteria for "best"*: The best composite explanation is defined as the one that is both maximally covering but also highly plausible. These criteria are operational; they guide the search process rather than evaluate the result. That is, the "best" composite explanation is *built* rather than *picked out* from a set of all consistent composites. Finding the set of all consistent composites is NP-hard (Section 2.4), so such a set is never found in practice.

In PEIRCE-IGTT, the criteria for "best" is defined operationally rather than structurally as in ALP, semantic tableaux, and similar approaches. It is difficult if not impossible to examine a set of consistent composite explanations and pick out the one that PEIRCE-IGTT would build. However, unlike the other systems we have examined, PEIRCE-IGTT can claim a cognitive plausibility, primarily due to its tractability, but also its "argumentative force" as described in the beginning of Chapter 2.

3.11 Discussion

This review of prior work in abductive reasoning has addressed logic-based, Bayesian, coherence-based, and set covering abduction. A more thorough account of the various approaches to abduction is given by Schurz (2008), and Section 8.8 of this work looks at

Bayesian abduction in more detail. We saw that logic-based abduction over-emphasizes the "reverse deduction" understanding of abductive reasoning and, critically, treats explanation as entailment. We feel that on both accounts, logic-based abduction misses the commonsense, everyday aspect of abduction. Coherence-based abduction is interesting in its novelty but due to the very limited work in that area, we have chosen not to focus our attention there. Rather, we have focused on set covering abduction and, in particular, the PEIRCE-IGTT system. That system takes a distinctly pragmatic and computationally efficient approach by changing the goal from "find a most-plausible complete and consistent explanation" to "find a highly-plausible and complete or nearly-complete consistent explanation," and by utilizing the EFLI algorithm to accomplish this goal. Surely a different abduction methodology, such as ALP or GCS, could be adapted to solve most if not all of the problem domains that make up our experiments. The metareasoning strategy, which we detail in the next chapter, would also have to change in order to reason *about* whatever particular base-level reasoning system is used. We do not know for certain how performance would be impacted by such a change. However, we have good reason for adapting the PEIRCE-IGTT approach, and this approach is utilized again in the self-similar abductive metareasoning system.

Chapter 4: Abductive metareasoning

The unanswered question from Chapter 2, in which we defined an abductive reasoning process, was how can we detect and correct errors? The abductive reasoning process is not perfect, partly because it is an efficient non-backtracking algorithm that does not necessarily find the most complete, most plausible explaining set. But the abductive reasoning process might also introduce errors because the problem domain might not provide sufficient or accurately scored hypotheses. Nevertheless, are there ways to improve accuracy? We propose *abductive metareasoning* as one such way.

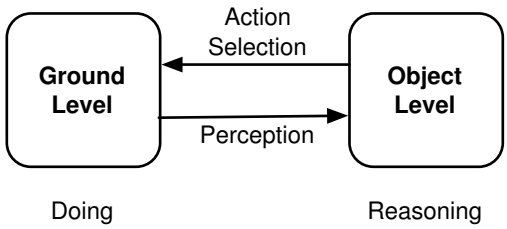


Figure 4.1. Action-perception cycle, after Cox and Raja (2011a).

Cox and Raja (2011a) depict action, perception, and reasoning as shown in Figure 4.1, and extend it to include metareasoning as shown in Figure 4.2. The latter diagram illustrates that *metareasoning* is another reasoning process that controls and monitors the base-level reasoning process. This metareasoning process might monitor for inconsistencies, failures to complete a reasoning task, or other signals of trouble. It might control the base-level reasoner by updating a knowledge base, restarting the reasoning process with different

parameters, etc.

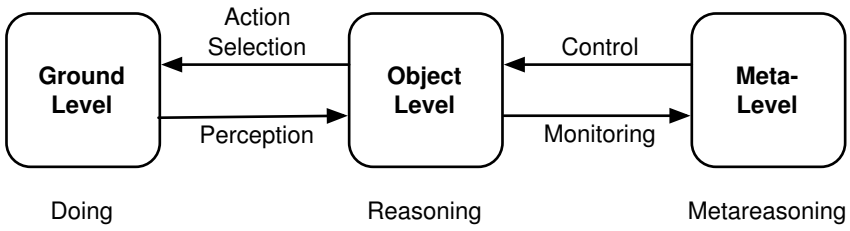


Figure 4.2. Action-perception cycle with metareasoning, after Cox and Raja (2011a).

In this work, the base-level reasoning system is an abductive reasoning system. Furthermore, the metareasoning system that we have designed is essentially a heuristic process tasked with identifying signals of trouble specific to abductive reasoning. We have isolated one such signal as the focus of this work. We take the presence of *anomalies*, i.e., reports or beliefs that require explanation but cannot be explained, as a signal of a problem. Our working hypothesis is that when anomalies are present, there is a good chance that the system has introduced an error. However, anomalies do not *imply* errors. Noisy reports might introduce anomalies, but noisy reports should not be explained and should remain anomalous. Thus, part of the challenge of a metareasoning process is to determine if an anomaly indicates errors, and where those errors exist in the doxastic state, or if the anomaly is due to noise and not an indication of errors.

The metareasoning system monitors the doxastic state for the presence of anomalies, and if they are found, reasons about the doxastic state and its anomalies. In some cases, the metareasoning system will modify the doxastic state and return to the base-level reasoning process. We modify the reasoning state diagram (Figure 2.2) to include the metareasoning component, as seen in Figure 4.3.

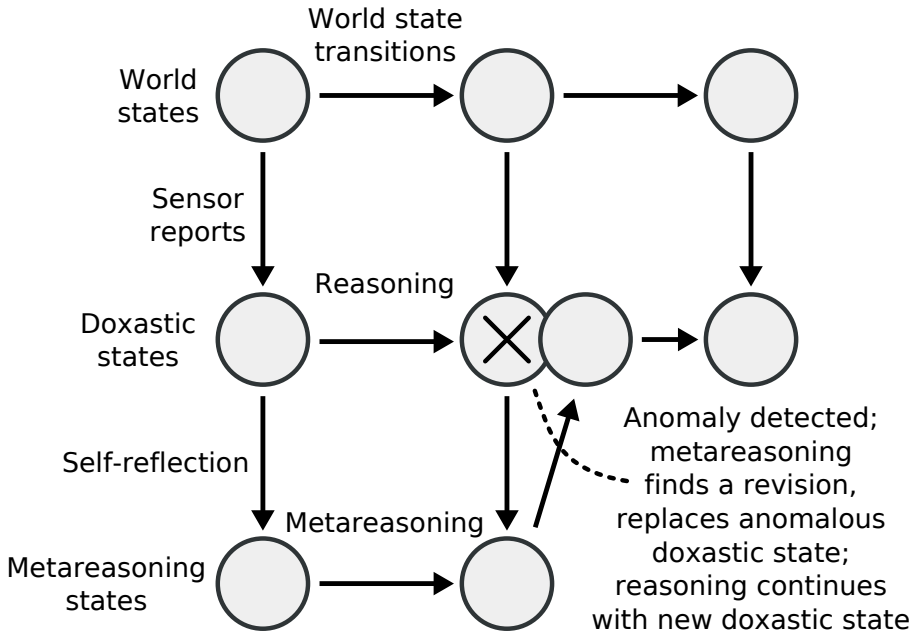


Figure 4.3. Metareasoning state diagram.

4.1 An abductive approach to metareasoning

We call *anomalies* those reports and other evidence that remain unexplained in a finalized doxastic state. The system checks for their presence in the METAREASON function (Algorithm 4.6), that is activated from the ABDUCE function (Algorithm 2.2). 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 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.

More specifically, an anomaly a may arise because,

1. no hypothesis was offered for a, or
2. all offered hypotheses were rejected due to,
    a) incompatibility with previously-accepted hypotheses, or
    b) not meeting the minimum plausibility requirement.

Note that if there do exist hypotheses that can explain a (but were necessarily rejected, rendering a anomalous), each of these hypotheses was rejected for only one reason, but not necessarily the same reason. Furthermore, if some rejected hypothesis is *undecided* (unrejected) and prevented from being rejected again for the same reason, then it might still be rejected for another reason. For example, a low plausibility hypothesis might originally be rejected for not meeting the minimum plausibility requirement, but if that condition is reversed (the minimum plausibility requirement is ignored for this specific hypothesis; see Section 4.2), it might next be rejected due to incompatibility. This suggests the need for *recursive* metareasoning in which multiple diagnoses and repairs are attempted, when necessary.

Note that an anomaly might have multiple possible causes. The abductive metareasoning system 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 (Definition 2.7.2; see also the FINALIZE function Algorithm 2.2). 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.

The following sections detail the possible causes of anomalies and their corresponding belief revisions. Let D be a doxastic state, $A = \text{ANOMALIES}(D)$, i.e., the set of anomalies in D, and $H_A = \bigcup_{h \in A} \text{HYPOTHESES}(D, h)$, i.e., the set of possible explanations of anomalies (if any exist). That is, H_A are possible explanations of the anomalies, but not possible explanations of their *anomalous status* (which is the role of meta-hypotheses). Note that each $h \in H_A$ is necessarily rejected in D.

4.2 Implausible hypotheses

Some reports may be anomalous due to the rejection of one or more implausible hypotheses. Figure 4.4 shows an example scenario as an explanation graph. These rejected hypotheses are characterized by,

$$H_P = \{h | h \in H_A \wedge Pl(h) < \eta\},$$

where Pl is the plausibility function of D and η is the minimum plausibility threshold. The IMPLHYPCANDIDATES function finds this set H_P (Algorithm 4.1). Unrejecting one or more of these implausible hypotheses, thus possibly allowing their acceptance, might eliminate some anomalies. This revision is given by REVISEIMPLHYP (Algorithm 4.2).

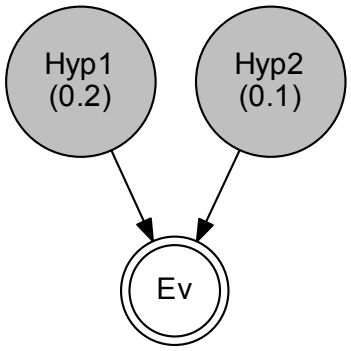


Figure 4.4. Example of an anomaly caused by implausible hypotheses.

For each $h \in H_P$, the metareasoning system hypothesizes that the rejection of h is responsible for some reports having no explanation.

Algorithm 4.1 Function that finds candidates for MetaImplHyp meta-hypothesis.

```
function IMPLHYPCANDIDATES(D)
    A ← ANOMALIES(D)
    H_A ← ⋃_{h∈A} EXPLAINS(D, h)        ▷ Known explainers of the anomalies
    H_P ← {h | h ∈ H_A ∧ Pl(h) < η}
    return H_P
end function
```

Algorithm 4.2 Revision function for the MetaImplHyp meta-hypothesis.

```
function REVISEIMPLHYP(D_0, h)
    D_1 ← UNDECIDE(D_0, {h})
    return D_1
end function
```

4.3 Incompatible hypotheses

Some reports may be anomalous due to some of the hypotheses being rejected upon the acceptance of other hypotheses. Figure 4.5 shows an example scenario as an explanation

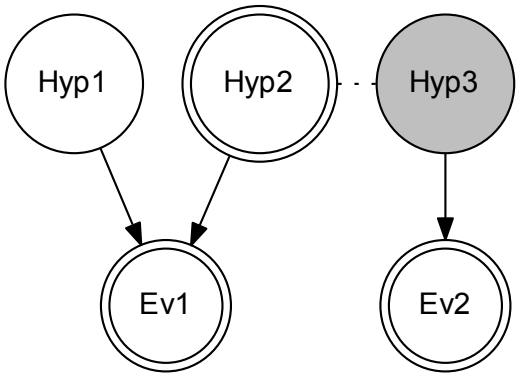


Figure 4.5. Example of an anomaly caused by incompatible hypotheses.

graph. 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. The INCOMPATHYPCANDIDATES function (Algorithm 4.3) finds this set H_I. For each $h \in H_I$, we hypothesize that the acceptance of h is responsible for some reports having no explanation. This revision is given by REVISEINCOMPATHYP (Algorithm 4.4).

4.4 Order dependency

The final possible cause of an anomaly is that no hypotheses were ever offered. We suppose that the GENERATEHYPOTHESES function is defined to generate only those hypotheses that are consistent with the current doxastic state. Thus, if some of the accepted hypotheses

Algorithm 4.3 Function that finds candidates for MetaIncompatHyp meta-hypothesis.

function INCOMPATHYPCANDIDATES(D)
  $A \leftarrow$ ANOMALIES(D)
  $H_A \leftarrow \bigcup_{h \in A}$ EXPLAINS(D, h) ▷ Known explainers of the anomalies
  $H_I \leftarrow \{h | \text{INCOMPATIBLE}(D, h) \cap H_A \neq \emptyset\}$ ▷ Hypotheses incompatible with H_A
  $H_C \leftarrow H_I \cap$ ACCEPTED(D) ▷ Incompatible hypotheses that were accepted
  return H_C
end function

Algorithm 4.4 Revision function for the MetaIncompatHyp meta-hypothesis.

function REVISEINCOMPATHYP(D_0, h)
  $D_1 \leftarrow$ UNDECIDE($D_0, \{h\}$)
  $D_2 \leftarrow$ REJECT($D_1, \{h\}$)
  return D_2
end function

in the doxastic state (which were accepted to explain earlier reports) are false and incompatible with true explainers of the anomaly, then the GENERATEHYPOTHESES function will not generate hypotheses for these reports. Figure 4.6 shows an example scenario as an explanation graph with a "decision boundary" (i.e., the period beliefs were formed by way of abductive reasoning between calls to GENERATEHYPOTHESES). The ORDERDEPCANDIDATES function (Algorithm 4.5) finds the subset A_O of anomalies that have no known explainers.

The *order dependency* meta-hypothesis suggests 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 in which 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 later, the system reverts to the doxastic state immediately preceding

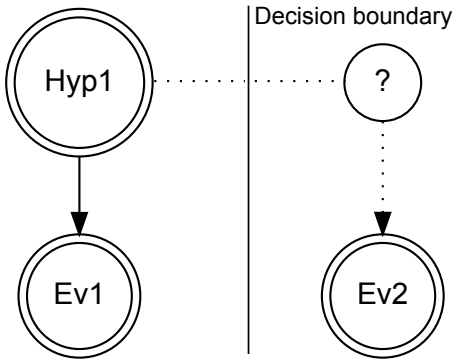


Figure 4.6. Example of an anomaly caused by an order dependency.

the introduction of the reports that ultimately proved to be anomalous.

Algorithm 4.5 Function that finds candidates for MetaOrderDep meta-hypothesis.

```
function ORDERDEPCANDIDATES(D)
    A ← ANOMALIES(D)
    A_O ← {a | a ∈ A ∧ EXPLAINS(D, a) = ∅}      ▷ Anomalies with no known explainers
    return A_O
end function
```

4.5 Plausibility estimate for meta-hypotheses

In each case, the plausibility of a meta-hypothesis h is estimated to be the average plausibility of the anomalies it explains, which we refer to as the set $A_h \subseteq A$. Hence,

$$Pl(h) = \sum_{a \in A_h} Pl(a) / \|A_h\|.$$

The intuitive support for this plausibility estimate is as follows.

- As detailed above, each of the three kinds of meta-hypotheses are only offered as possible explanations of anomalies when they are deemed applicable according to criteria that are specific to each kind of meta-hypothesis. For example, an *Implausible*

Hypotheses meta-hypothesis is only offered when an anomaly has possible explanations but they were rejected due to implausibility. Thus, it is not unreasonable to define a singular plausibility estimate function for all of the kinds of meta-hypotheses.

- It is more likely that an anomaly is true, and warrants an explanation, if it is plausible. This is because the plausibility estimate (which is specific to a domain) for reports and hypotheses generally scores true reports and hypotheses higher than false ones. If this is not true, then (1) the EFLI algorithm would likely produce poor results, and (2) a different plausibility estimate would be needed for meta-hypotheses. It is unclear, however, how such a plausibility estimate could be devised.

Experiments specific to each domain will show that abductive metareasoning still works reasonably well even if all anomalous reports have plausibility equal to 1.0. In this case, all meta-hypotheses that explain anomalous reports (as opposed to anomalous beliefs) also have a plausibility equal to 1.0, and thus cannot compete according to their plausibility, but only according to whether they can explain.

We have also experimented with a variety of more sophisticated plausibility functions for meta-hypotheses. Our experience is that these more sophisticated plausibility functions do not offer a significant advantage or disadvantage over the common plausibility estimate we described above.

4.6 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 at the end of an experiment (regardless of whether metareasoning was activated or not) are judged to be noise. Our domain-specific experiments (Chapters 6, 7, and 8) measure the *precision* and *recall* of such noise judgments.

4.7 Metareasoning algorithm

Algorithm 4.6 details the recursive abductive metareasoning procedure. This function is called at the end of the ABDUCE function (Algorithm 2.2); the METAREASON function likewise calls the ABDUCE function (but sets a flag to prevent this second call to ABDUCE from calling METAREASON yet again; the possibility of meta-metareasoning is addressed in Section 10.6). Note that most of the work of this algorithm is constructing a meta-doxastic state and calling the abductive reasoning procedure. The simplicity of the algorithm illustrates the power of a self-similar reasoning/metareasoning system.

4.8 Completeness of the meta-hypotheses

The three kinds of meta-hypotheses detailed above, MetaImplHyp, MetaIncompatHyp, and MetaOrderDep, describe the only possible causes of anomalies in the abductive reasoning system. Recall that an *anomaly* is an unexplainable report or belief. Let a be some anomaly. Then according to the general abductive reasoning algorithms (Algorithm 2.2), a remains unexplained in the finalized doxastic state due to one of the following reasons:

1. There is no hypothesis in the doxastic state that could explain a. (MetaOrderDep looks for these cases.)

Algorithm 4.6 Abductive metareasoning algorithm.

function METAREASON((H, X, Pl, S, V, I))
  $D_0 \leftarrow (H, X, Pl, S, V, I)$
  $A \leftarrow$ ANOMALIES(D_0) ▷ Find any anomalies
  if $A = \emptyset$ **then** ▷ If there are no anomalies, we can stop here
    return D_0
  else
    $S_{\text{meta}} \leftarrow \{(h, Accepted) | h \in A\}$ ▷ Accept (as "reports") all anomalies
    $Pl_{\text{meta}} \leftarrow \{(h, Pl(h)) | h \in A\}$ ▷ Keep the anomalies' original plausibilities
    $D_{\text{meta}} \leftarrow (A, \emptyset, Pl_{\text{meta}}, S_{\text{meta}}, A, \emptyset)$ ▷ Create a dox. state with just the anomalies
    $D'_{\text{meta}} \leftarrow$ ABDUCE$(M_{\text{meta}}, D_{\text{meta}}, \eta_{\text{meta}}, \text{false})$ ▷ Perform abduction (Alg. 2.2)
    $H_{\text{accepted}} \leftarrow$ ACCEPTED(D'_{meta})
    if $H_{\text{accepted}} = \emptyset$ **then** ▷ No meta-hypotheses were accepted
      return D_0 ▷ Return the original doxastic state
    else
      $D_1 \leftarrow$ APPLYREVISIONS$(D_0, H_{\text{accepted}})$ ▷ Refer to the three functions below
      $D_2 \leftarrow$ FINALIZE(D_1)
      if ANOMALIES$(D_2) = \emptyset$ **then** ▷ Are all anomalies resolved?
        return D_2
      else
        $D_3 \leftarrow$ METAREASON(D_2) ▷ Find new hyps. for remaining anomalies
        return D_3
      end if
    end if
  end if
end function

2. There is a hypothesis in the doxastic state that could explain a, but it was rejected. (MetaImplHyp and MetaIncompatHyp look for these cases.)

There is no other possible reason why a remains unexplained in a finalized doxastic state. Furthermore, the only reason a hypothesis may be rejected is that either it is too implausible (which MetaImplHyp handles) or it is incompatible with an accepted hypothesis (which MetaIncompatHyp handles). Note that these contingencies are unrelated to EFLI. The EFLI algorithm might leave a report or belief unexplained because no hypothesis is sufficiently *decisive*, but such unexplained reports and beliefs are not considered *anoma-*

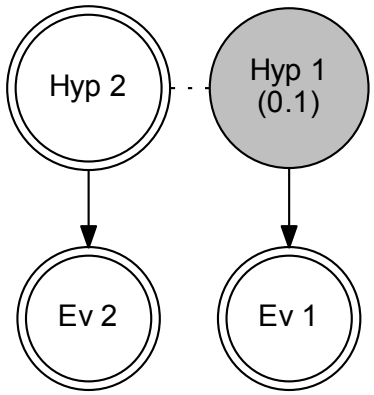


Figure 4.7. The report Ev 1 is an anomaly with multiple causes. Assume $\eta > 0.1$, so Hyp 1 was rejected originally due to low plausibility. No single meta-hypothesis is able to repair the anomaly.

lous. Thus, abductive metareasoning as defined is not specific to EFLI, but it is specific to the general abductive reasoning algorithms outlined earlier.

Therefore, MetaImplHyp, MetaIncompatHyp, and MetaOrderDep are the only meta-hypotheses needed in order to perform abductive metareasoning. However, an anomaly might have multiple causes in such a way that no single meta-hypothesis is able to repair it. An example is shown in Figure 4.7, where the report Ev 1 is anomalous due to both Hyp 1 being rejected for minimum plausibility and Hyp 1 being rejected due to incompatibility with Hyp 2. Of course, a hypothesis is never rejected twice during abductive reasoning, but neither MetaImplHyp nor MetaIncompatHyp can, on their own, repair this anomaly. Rather, both repairs must be applied (Hyp 2 must be rejected and Hyp 1 prevented from being rejected due to implausibility).

The abductive metareasoning procedure detailed in Algorithm 4.6 will work as follows on the case shown in Figure 4.7:

1. Two meta-hypotheses will be generated to explain the anomaly Ev 1: (1) a MetaIm-

plHyp meta-hypothesis that posits that Hyp 1 was rejected due to low plausibility, and (2) a MetaIncompatHyp meta-hypothesis that posits that Hyp 1 was rejected due to the acceptance of Hyp 2. Both of these meta-hypotheses will have the same plausibility (since they both explain the same anomaly).

2. If $\delta_{\text{meta}} > 0$, then neither of these two meta-hypotheses will be accepted to explain Ev 1 because neither stands out as decisive (they have the same plausibility). This seems reasonable since there is no *single* best way to repair the anomaly. However, if $\delta_{\text{meta}} = 0$, then the system will arbitrarily choose to accept one of these meta-hypotheses.

3. Suppose the MetaImplHyp meta-hypothesis is accepted. Then Hyp 1 will be undecided and abductive reasoning will be activated again. However, Hyp 1 will be immediately rejected again because it conflicts with the accepted hypothesis Hyp 2. So Ev 1 remains anomalous, and metareasoning will be activated again. Due to a special check in the code, the MetaImplHyp hypothesis will not be generated again, leaving only the MetaIncompatHyp. This brings us to the next case below.

4. Suppose the MetaIncompatHyp meta-hypothesis is accepted. Then Hyp 2 will be undecided and then rejected. But Hyp 1 will remain rejected because of low plausibility. Thus Ev 1 remains anomalous, and metareasoning will be activated again.

The anomaly ultimately may be explained if Hyp 2 is rejected first, freeing Hyp 1 to explain. Hyp 1 will still be rejected due to implausibility. Then on the second execution of metareasoning, Hyp 1 may be unrejected, and finally accepted. So abductive metareasoning may be able to resolve cases where anomalies have multiple causes. But what is missing is a distinct control strategy to ensure these kinds of cases are handled properly. We have noted this missing functionality and hope to address it in future work.

4.9 Performance expectations

The metareasoning system we have detailed has access to all the same information (reports, hypotheses) as the base-level reasoner. The metareasoning system is not able to request more information from the outside world or from other agents. However, the metareasoning system does differ from the base-level system in one important way: it has access to the *complete history* of the base-level reasoning system's decisions and alternative choices. The base-level reasoner was explicitly designed to be a greedy, non-backtracking process that efficiently arrived at a plausible, complete or nearly-complete explanation of the evidence. However, as we have seen, it might fail to explain all the evidence, i.e., it might leave some evidence anomalous. The metareasoning system responds to the anomalies and attempts to explain and thereby resolve them. It does so only by examining the existing reports and hypotheses, as well as the history of the base-level reasoning system's decisions.

It is important to note that, because the metareasoning system does not have access to external sources of information, we should not hold an expectation that the metareasoning system is able to correct all errors, correctly repair all anomalies, and/or identify all noise. The metareasoning system relies on the availability of the true hypotheses and the correctness of plausibility estimates. If the true hypotheses are not available or plausibility estimates have little relation to truth, then there is not much metareasoning can do to find explainers for the anomalies. Rather, the metareasoning system operates best when the problem domain provides mostly "reasonable" hypotheses and plausibility estimates. Abductive reasoning and abductive metareasoning both rely on a certain (but impossible to define) reasonableness of the problem domain since they codify intuitive "sense-making" arguments of the form illustrated at the beginning of Chapter 2. If, for example, false hypotheses, provided by the GENERATEHYPOTHESES function, are consistently scored

more "plausible" than true hypotheses, then the abductive argument which says that more "plausible" hypotheses are more likely to be true fails in this case.

For each of the three kinds of meta-hypotheses, we can identify the conditions when metareasoning will make the correct revision to the doxastic state. We are assuming for the moment that only a single kind of meta-hypothesis is considered; interactions among meta-hypotheses make the analysis significantly more difficult. In the following analysis, we assume $\delta = \delta_{\text{meta}} = 0$.

MetaImplHyp meta-hypotheses

In order for a MetaImplHyp meta-hypothesis to correctly revise a doxastic state, several conditions must be met.

- There is some true report or unexplained belief a.
- Some true hypothesis h, which can explain a, has low plausibility: $Pl(h) < \eta$. Thus, h is rejected. For simplicity, we will assume h is capable of explaining *only* a.
- Every other hypothesis that can explain a is rejected. Thus, a is an anomaly.
- Let h_{meta} be the MetaImplHyp meta-hypothesis that posits that h is true and should be accepted to explain a. Then for h_{meta} to be accepted, it must be the case that $Pl(h_{\text{meta}}) = Pl(a) \geq \eta_{\text{meta}}$. For simplicity, assume no other meta-hypothesis competes to explain the anomaly a.

If all of these conditions are met, then abductive metareasoning will correctly revise the doxastic state by accepting h even though the base-level reasoning system considered h to be too implausible. If any one of the conditions is not met, then abductive metareasoning will either not make the right revision, no meta-hypothesis will be accepted to explain a, or there will be no anomaly caused by a rejection due to the minimum plausibility threshold.

MetaIncompatHyp meta-hypotheses

In order for a MetaIncompatHyp meta-hypothesis to correctly revise a doxastic state, the following conditions must be met.

- There is some true report or unexplained belief a.
- h_T is the true explainer of a and was generated by GENERATEHYPOTHESES.
- Some false hypothesis h_F is incompatible with h_T but was accepted, and hence rejected h_T. Furthermore, h_F does not explain a.
- Let h_{meta} be the MetaIncompatHyp meta-hypothesis that posits that h_T is true, h_F is false, and h_F should be rejected (thus possibly allowing h_T to be accepted in order to explain a). It must also be the case that $Pl(h_{\text{meta}}) = Pl(a) > \eta_{\text{meta}}$. For simplicity, assume no other meta-hypothesis competes to explain the anomaly a.

Note that h_F must have been accepted in the same time step as h_T was hypothesized. While it is possible that h_F was known from a previous call to GENERATEHYPOTHESES before h_T was hypothesized, it could not have been accepted before h_T was hypothesized because GENERATEHYPOTHESES is required only to generate those hypotheses (including h_T) that are compatible with the current doxastic state (which therefore cannot include h_F as a belief).

MetaOrderDep meta-hypotheses

In order for a MetaOrderDep meta-hypothesis to correctly revise a doxastic state, it must be the case that some report or unexplained belief a has no known explainer. In other words, GENERATEHYPOTHESES yielded no hypotheses that can explain a. This may occur for one of two reasons: (1) there really is no possible explanation for a, regardless of

the current doxastic state, or (2) existing beliefs in the current doxastic state preclude all possible explanations of a. If the first case is true, then taking back recent decisions and generating new hypotheses will not make any difference (at least for explaining a). On the other hand, if the second case is true, then there is no other possible fix except by calling GENERATEHYPOTHESES with a less committed doxastic state.

MetaOrderDep meta-hypotheses are generated only to explain anomalies that have no known explainers. If a MetaOrderDep meta-hypothesis is accepted (which requires that $Pl(a) \geq \eta_{\text{meta}}$, as usual), then some prior (less committed) doxastic state will replace the current doxastic state, recent reports will be added to the older doxastic state, and new hypotheses will be generated. Note that all beliefs acquired and all revisions applied since the prior doxastic state will be lost. In this way, a MetaOrderDep meta-hypothesis might "undo" recent revisions from recent applications of metareasoning. After the prior doxastic state is brought up to date, if any anomalies remain, metareasoning will be activated again. It is possible that some of the same revisions that were undone will be applied again.

4.10 Self-similar metareasoning

The base-level abductive reasoning system and meta-level abductive metareasoning system both utilize the same reasoning process and, indeed, the same code in our experimental software. We feel that this is a good software architecture because common functionality is shared rather than duplicated in different system modules. This is shown in Figure 4.8. We also believe this architecture to be cognitively plausible. Since everyday commonsense reasoning seems to involve both abductive reasoning and abductive metareasoning, we find it plausible that a common *abductive reasoning process* is employed for both kinds of reasoning behavior.

It seems clear that *metacognition*, i.e., the act of reflecting on the entire cognitive cycle

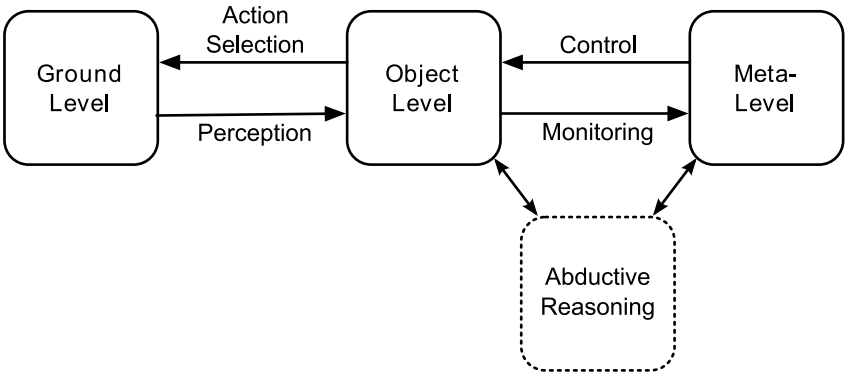


Figure 4.8. Action-perception cycle with abductive metareasoning, after Cox and Raja (2011a). Both the object level and meta-level components use abductive reasoning.

(Figure 1.1), involves at least *making sense* of one's beliefs and how they were acquired, the successes and failures of one's plans and actions, and so on. Metareasoning, as part of metacognition, therefore involves at least a kind of abductive reasoning, regardless of the nature of the base-level reasoning system (i.e., regardless of whether it is abductive or not). In this work, we have focused solely on making sense of the "making sense" phase of cognition rather than making sense of the entire cognition cycle. Our metareasoning system likewise responds to failures to make sense.

4.11 Conclusions

In this chapter, we introduced a *domain-separable* abductive metareasoning system that monitors and controls the base-level abductive reasoning system, and thereby is capable of belief revision and noise detection. Both the base-level and meta-level reasoning systems utilize the same abductive reasoning machinery (algorithms and code); in this way, the two levels are self-similar, which we have argued is a cognitively plausible architecture

for artificial cognitive systems. The metareasoning system monitors for the presence of *anomalies*, i.e., unexplainable reports and beliefs, in the base-level reasoner. Being abductive, the metareasoner treats these anomalies as *meta-evidence*. Possible explanations of the evidence are generated; these are called *meta-hypotheses*. Meta-hypotheses come in three types: MetaImplHyp, MetaIncompatHyp, and MetaOrderDep. Each type attempts to explain a specific kind of anomaly, and specifies a particular belief revision (in the cases of MetaImplHyp and MetaIncompatHyp) or specifies that a whole slew of recently acquired beliefs need to be retracted, at least temporarily, and new hypotheses need to be generated (in the case of MetaOrderDep). Normal abductive reasoning is applied to find the best meta-hypotheses to explain the meta-evidence, and if any such meta-hypotheses are accepted, their corresponding belief revisions or retractions are applied. Metareasoning is iterative, so should any anomalies remain, the process repeats, until no anomalies remain or there are no good meta-hypotheses. Any anomalies that remain at the end of this process are labeled "noise." Thus, noise detection is kind of a "fall back" explanation.

Abductive metareasoning brings added value to the base-level abductive reasoning system for a few reasons:

- Reasoning from existing beliefs, rather than starting with a blank doxastic state at every time step, is obviously more efficient. However, this strategy might be a source of mistakes. For example, a moving object might be mistracked through a low-visibility region, but the cognitive system might not be able to realize this error until much later. If the object is eventually reported at a different location than expected, this report might not be explainable, since existing beliefs indicate such a report is impossible (or noise). In other words, sometimes ambiguous reports may disambiguate at a later time when more evidence is available. The MetaOrderDep meta-hypothesis is able to handle these cases.

- Efficient abductive reasoning requires avoiding the generation of all possible consistent explaining sets. We have developed a greedy, hill-climbing abductive reasoning algorithm to ensure efficient reasoning. Experiments detailed later indicate that this is a good approach; however, it might fail to find a complete explaining set, rendering some evidence anomalous. Metareasoning might be able to repair the deficient explaining set by taking back incompatible hypothesis as suggested by the MetaIncompatHyp meta-hypothesis.

- As will be shown experimentally, noise detection is enhanced with metareasoning. Without abductive metareasoning, all anomalous reports are considered to be noisy reports. With abductive metareasoning, the correct explanation might be found for true reports so that anomalies that still remain after metareasoning are more likely to be actual noise.

Even though metareasoning adds value, the combined abductive reasoning and abductive metareasoning system is still just one cognitive system. Labeling one aspect of its reasoning process as a "meta" process is almost just a matter of semantics. However, the same could be said of all software architecture: though a software system may be designed and implemented in terms of complex class/type hierarchies and interfaces, the end product is still just one compiled sequence of instructions. But, as every software engineer knows, architecture matters both for efficiently implementing a system that meets specifications and, more importantly for our purposes, understanding the system's behavior. We have explicitly designed a two-level abductive reasoning architecture in order to identify exactly what capabilities metareasoning provides beyond the base-level reasoner.

Chapter 5: Prior work in metareasoning

The ideas of *metareasoning*, *meta-level knowledge*, *meta-rules*, and so on have been explored for much of the history of the field of Artificial Intelligence. After all, as soon as one designs and implements a system that encodes and reasons with expert knowledge (i.e., an expert system), one naturally thinks about better, more flexible ways to represent and utilize this knowledge. Davis and Buchanan (1984) illustrate meta-level knowledge with four examples from the TEIRESIAS expert system (Davis, 1976). They show how schemata and templates for knowledge representation and meta-rules for control of how knowledge is used give the system greater flexibility and self-reflection capabilities. These features are useful for, e.g., the system to perform self-diagnosis by finding out that it is missing necessary facts or rules to answer a query. They argue that meta-knowledge may also be used to support *knowledge transfer* between the expert system and the expert who is trying to use and improve the system. If the system is able to explain how it arrives at a conclusion, or that it is missing key information to complete a task, then it can communicate this information to the expert user. The user can add new rules or facts, and the system can then explain back to the user how it *interprets* this new information, e.g., how this new information would impact existing rules and how it might place this information in particular schemata or templates, etc.

Genesereth (1983) identifies the *meta-level* of system design as the level that controls base-level actions. For example, the base-level actions for a robot arm include grabbing, dropping, moving, etc. Meta-level control is essentially the program that decides what the robot arm is supposed to do. He gives examples of meta-level architecture from a variety of

high-level tasks. For example, he describes how one can write meta-level rules to control search (e.g., changing from depth-first to breadth-first search under certain circumstances). He also illustrates how planning systems operate at the meta-level since they arrange a sequence of base-level actions. Genesereth's distinction between base-level and meta-level is interesting because it shows how pervasive *meta-level* can be if one is "looking for it." Even commands to a robot arm, such as *grab(ball)*, may be considered at the meta-level if the base-level is equated with the actual sequence of primitive motor actions necessary to grab the ball. Explicitly identifying base- and meta-level features of a larger system architecture is a matter of pragmatics, not essentialness. There is nothing explicitly *meta* about a search algorithm, for example, but it might be used in a *meta* context, such as searching for the next applicable operator in a planning task. It is important to keep in mind that metareasoning, meta-knowledge, and so on, are distinctions made in context; there must be some kind of *base-reasoning*, *base-knowledge*, and so on to differentiate it from *meta* versions of the same, and to be the *subject* of meta-level concerns.

Metareasoning has enjoyed renewed interest since the work of the 1970's and 1980's. Some of this work may be found in a recent volume edited by Cox and Raja (2011b). Good summaries of metareasoning, or *metacognition* in general, include reviews by Cox (2005) and Anderson and Oates (2007). These reviews show that metacognition may be found in mathematics, psychology, artificial intelligence, and philosophy. Its varieties are too broad and numerous to cover here, but we note that metacognition and metareasoning are not only of interest to artificial intelligence researchers.

The remainder of this review of prior work in metareasoning is divided into two parts. The first addresses anomaly-driven metareasoning found in various cognitive systems and cognitive architectures. The second part gives an overview of *belief revision*, or what we have called previously *strict belief revision* (to differentiate it from Bayesian belief revision). Belief revision has primarily been studied in logical terms rather than computational

terms. Nevertheless, our abductive metareasoning system realizes a kind of belief revision, and the purpose of our review of belief revision is to compare and contrast our approach with those from the traditional logical approaches.

5.1 Anomaly-driven metareasoning

An early example of anomaly-driven metareasoning was developed by Karp (1989). That system 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. Much more recently, Bridewell (2004) has described 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 in that work 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. They experimented with a single simulated object tracking domain example, which leaves unanswered the question of whether their approach is domain-general.

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, rep-

resented 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.

Another metareasoning system that responds to anomalies, or expectation failures, is Meta-AQUA (Cox and Ram, 1999). AQUA (Ram, 1991), which stands for "Asking Questions and Understanding Answers," is a question-driven story understanding program. This means that it formulates questions about the story and then reads the story to find answers. In this way, it builds a causal and motivational model of events and actors in the story. Interestingly, AQUA would reread a story differently than it read it the first time since, on the second reading, it has some existing model of the story and therefore different questions. Meta-AQUA extends AQUA by monitoring for expectation failures. For example, if AQUA reads a story and predicts that X caused Y, but then later reads that in fact X never occurred, then Meta-AQUA notices the expectation failure and attempts to explain it. The metareasoning system obtains a trace of AQUA reasoning that led to the failure. The metareasoner of Meta-AQUA is a case-based reasoning system. As such, it attempts to find a similar case of expectation failure found in the reasoning trace. The result is a set of learning goals that are processed by a nonlinear planner to find a learning plan that ultimately updates the base-level reasoner's model of the story and resolves the anomaly.

Cox et al. (2011) have begun work integrating the MCL and Meta-AQUA architectures into a Metacognitive Integrated Dual-Cycle Architecture (MIDCA). The purpose of MIDCA is to address the shortcomings in MCL and Meta-AQUA. In the case of MCL, it

"has a weak model of perception and no model of high-level understanding and interpretation" (Cox et al., 2011). On the other hand, Meta-AQUA "is essentially a disembodied agent, lacking a model of action and personal agency." Furthermore,

> [...] neither MCL nor Meta-AQUA has an explicit model of self. The systems do not have a model of the contents of their background knowledge for example, and thus they cannot answer questions such as what kinds of tasks are they expert at. They have no feelings of confidence as they perform a cognitive task, and thus they cannot decide whether or not they are getting close to an answer (Cox et al., 2011).

Their solution is a new architecture "that addresses metacognition from the start." Though preliminary at the time of writing, the MIDCA architecture is designed to support two kinds of meta-level influence: (1) the meta-level can act as an "executive" that decides when to switch between cognitive processes (planning, acting, etc.), when to change goal priorities, and how to distribute resources between cognitive processes; and (2) the meta-level can change the representations used for reasoning as well as object-level goals, strategies, and knowledge (i.e., instigate learning).

It is clear that Meta-AQUA and MIDCA are larger, more inclusive metacognitive architectures than the architecture we have proposed (Figure 2.9). We are not attempting to address as broad a range of meta-level issues as the authors of those systems. Rather, we are focused narrowly on the design and utility of an abductive metareasoning system that is designed to work exclusively on a self-similar base-level abductive reasoning system. To our knowledge, no other prior work has investigated self-similar metareasoning, and only Bharathan (2010) has looked at abductive metareasoning. Furthermore, the present work appears to be unique in its experimental methodology. We experimentally evaluate our system in a variety of domains, without modifying any code of the reasoning or metar-

easoning components. Furthermore, our experimental analysis allows us to isolate exactly which features of the overall system are contributing to observed performance.

5.2 Belief revision as metareasoning

The term *belief revision* is often used to refer to what we describe as logical or *strict* belief revision. In this case, an agent holds a consistent and possibly infinite set of propositional beliefs. Suppose p, $p \rightarrow q$, and q are some of these beliefs. Now, if the agent learns $\neg q$ is the case, then the question is which beliefs should be taken back (contracted), and which beliefs should remain? Philosophers have come up with a variety of desiderata for belief revision. For example, the agent should hold on to as many beliefs as possible and only contract those necessary to accommodate the new belief. In our example, to accommodate $\neg q$, either p or $p \rightarrow q$ must be contracted. But unrelated beliefs should remain. We see this question, "which beliefs to contract?" as a meta-level question, and as such consider it to involve a kind of metareasoning. Chapter 1 explained how the present work differs from strict belief revision. The remainder of this section provides some details about strict belief revision and further points of comparison. Note that henceforth we will drop the modifier "strict" when referring to strict belief revision.

Several frameworks for belief revision (Alchourrón et al., 1985; Darwiche and Pearl, 1997; Jin and Thielscher, 2007; Williams, 1995, 1997) satisfy certain desiderata concerning ideal belief revisions, known as "the AGM postulates" from Alchourrón, Gärdenfors, and Makinson (1985). These frameworks represent an agent's doxastic state as a belief set (a closed theory) plus an entrenchment ordering on those beliefs (a belief μ is more entrenched than a belief ψ if the agent is more willing to give up ψ than μ, all else being equal). The AGM postulates cover the three types of belief change: expansion (adding a belief without taking any beliefs away), contraction (taking away beliefs without adding

any), and revision (adding and taking away beliefs). An expansion, contraction, or revision operation always occurs with respect to some input μ; thus, we say that a belief set is revised (or expanded or contracted) by μ. The belief set that results is unique, so the operations (expansion/contraction/revision) can be thought of as functions, with domain equal to the Cartesian product of the universes of doxastic states and propositions and range equal to the universe of doxastic states. If a belief revision framework satisfies the AGM postulates, then the Levi identity (Gärdenfors, 1981; Levi, 1977) holds that any revision can be modeled as a contraction followed by an expansion; thus, any of the three operations can be constructed from the other two. Revision is the most interesting case, since it may involve both adding and taking away beliefs, so we will only review the postulates for revision.

We will only present summaries of the postulates most relevant for this discussion. The presentation that follows should not be considered formal.

AGM-1. Revising a doxastic state with an input yields another doxastic state.

AGM-2. The doxastic state resulting from revision with μ has it that μ is believed. Note, this can get complicated if the prior doxastic state disbelieves μ (believes $\neg\mu$). In order to believe μ, the causes for belief in μ (e.g., believing both ϕ and $\phi \rightarrow \mu$) must be taken away.

AGM-3. The resulting doxastic state must be consistent, unless the input μ that the doxastic state was revised with is itself inconsistent.

Because a revision with μ might require taking away some beliefs (if those belies imply $\neg\mu$), the entrenchment ordering among beliefs may be consulted to determine which beliefs to remove. If the agent simply forgets everything it believed previously, and only believes μ when it learns μ (and whatever is logically implied by μ), the revision would technically satisfy the AGM postulates. But of course, this is belief revision of the most

useless sort; clearly some kind of *minimal revision* is desired, wherein the agent keeps any beliefs that are not related to or at least not incompatible with μ. Various researchers have devised entrenchment relations and revision functions with the goal of producing *minimal* revisions (it turns out that an entrenchment relation uniquely determines the revision function).

However, Tennant (2006) proved that the AGM postulates are fatally deficient as rationality postulates for belief revision. Tennant demonstrates that a revision function can be constructed such that the revision produces any "bizarre" belief set one wishes. Thus, the AGM postulates are not a *gold standard* for belief revision frameworks.

Darwiche and Pearl (1997) have provided additional postulates that attempt to ensure more rational revision for sequences of inputs, one presented after the other, with revisions possibly occurring after each input. Their additional postulates are intended to ensure that the entrenchment relations among beliefs are not severely altered. In particular,

DP-1. If μ is received (and one's doxastic state is revised to accept μ), and then α is received (and revision takes place), and it turns out that α logically implies μ, then the resulting doxastic state must be equivalent to revising just with α.

DP-2. If α logically implies $\neg\mu$, then revising first by μ and then by α must be the same as revising only by α. In other words, when an agent learns contradictory information, the second bit of information learned is what the agent ultimately believes.

DP-3. If learning α would imply that μ must also be true, then if the agent learned μ first, and subsequently learned α, then the agent continues to believe that μ must also be true.

DP-4. If learning α does not give the agent reason to believe that μ is false, then learning μ first, followed by α, still does not give the agent reason to believe that μ is false.

In Darwiche and Pearl's words, "no [input] can contribute to its own demise."

Darwiche and Pearl provide two additional postulates in order to further minimize changes during a belief revision operation, but the above descriptions of their first four postulates give the flavor of their work. Stalnaker (2009) gives examples that show some counter-intuitive restrictions in the DP postulates. In particular, in some cases μ may be contradicted by α but learning μ first followed by α should not always be equivalent to only learning α (DP-2). After all, μ may have contained much information not contradicted by α or even related to α, and therefore the most rational result would be to believe those parts of μ that are still valid, as well as believing α.

The AGM and DP postulates do not directly apply to the kind of belief revision enacted by our abductive metareasoning system. However, it is useful to compare the *spirit* of AGM+DP strict belief revision and our approach. First, abductive metareasoning does not necessarily accept the validity of the report, whereas the belief revision literature typically assumes that the incoming report (belief) must be accepted. We look at an alternative, non-prioritized belief revision, below. But for the remainder of this quick discussion, we will assume the incoming report is to be believed, and some explanation must be found. Finding an explanation might not be possible, triggering abductive metareasoning. The metareasoning procedure might take back some accepted hypotheses and execute the abductive reasoning process again, thus resulting in different beliefs than before (which might include explanations for the report that triggered the revisions). Thus, our approach agrees with the spirit of AGM-1: the result of a revision should be another doxastic state. AGM-2 states that the revised doxastic state should include a belief in the report that caused the revision. As stated previously, this point is not necessarily true in our framework, but we will address that again later. AGM-3 states that the resulting doxastic state should be consistent. We agree here and our the abductive reasoning algorithm will never produce an

inconsistent doxastic state.

DP-1 through DP-4 try to put limitations on how revisions interact. Abductive metareasoning agrees with the spirit of these postulates. We will not address each one in detail. Rather, we will just look at DP-2 as a case-in-point. Suppose μ is accepted to explain some report, and later α, which is incompatible with μ, is needed to explain a new report. But then α cannot be accepted, leaving the new report anomalous. Metareasoning is activated, which finds that by rejecting μ, α is freed up to explain. The result is that μ is disbelieved and α is believed. DP-2 states that the result should be the same as never observing the report that was explained by μ. In our system, the result is equivalent because upon accepting α, μ would be rejected.

We mentioned above that abductive metareasoning does not necessarily accept a report that cannot be explained, i.e., a report that might cause a belief revision. Strategies for strict belief revision that satisfy the AGM+DP postulates, on the other hand, assume that the incoming belief will be accepted (it is postulate 2 for AGM, after all). But there has been work on so-called *non-prioritized* belief revision which relaxes this requirement. In non-prioritized strategies, the incoming statement may be believed as is, modified before being integrated into the doxastic state, or rejected outright. There are a variety of approaches to non-prioritized belief revision, surveyed by Hansson (1999). Eloranta et al. (2008) explore the space of non-prioritized belief revision functions, specifically those functions that are able to rewrite the input given existing beliefs. Clearly, our work in abductive metareasoning is more similar to non-prioritized belief revision than traditional AGM+DP belief revision.

Logical varieties of abductive reasoning, where explanations logically entail what they explain, and belief revision have been shown to be closely related. Abduction is a way to do belief revision (Aliseda, 2000; Boutilier and Becher, 1995; Paglieri, 2003; Pagnucco, 1996). Furthermore, truth maintenance systems (Doyle, 1979) essentially combine abduc-

tion and belief revision. Dixon and Foo (1993) has shown how to simulate an assumption-based truth maintenance system with AGM-compatible belief revision functions.

Unlike strict belief revision, and more in-line with metareasoning in cognitive systems, reviewed previously, we believe that metacognition involves more than just finding a "minimal" revision to one's beliefs in order to accommodate new evidence. The fact that new evidence is inconsistent with existing beliefs is cognitively significant. If the trigger for abduction may be called a "cognitive irritant" (Garcez et al., 2007), then finding no plausible consistent explanation must be especially irritating. The fact that no explanation can be found itself seems to require explanation. This is the approach we have taken with *abductive* metareasoning. It turns out that there is evidence that humans do the same. Suppose each of these two sentences are believed:

- If it is raining, then the grass is wet.
- It is raining.

Deductive inference tells us that the grass is wet. Now, suppose the agent observes that the grass is not wet. Strict belief revision in the AGM style might find two possible *minimal* revisions: take back "it is raining" or take back the conditional. Note that an AGM strategy would define "it is raining" to be less *entrenched* and thus the conditional statement (i.e., the law or generalization) is retained rather than the case, producing a unique revision.

In any event, a different perspective on the issue comes from Khemlani and Johnson-Laird (2011), who state,

> In our view, however, the presupposition [that minimal revisions are ideal] has no warrant. In daily life, when an inconsistency arises because a fact collides with the consequences of your beliefs, your primary goal is to understand how

> the inconsistency could have occurred in the first place [by explaining the inconsistency], because its origins are likely to have consequences for how you should act. [...] [T]he process of reasoning to the best explanation is a hallmark of rationality, because it is a prerequisite for sensible action. A mere revision to beliefs, whether minimal or not, is not so useful a guide (Khemlani and Johnson-Laird, 2011).

They further state that an alternative to traditional belief revision, which brings with it the minimality criteria, is explaining away the inconsistency. They call this the *explanatory hypothesis*, which postulates that "the first goal in coping with an inconsistency is to explain its origin." Furthermore, "a plausible explanation is likely to imply changes to beliefs," but these changes may, in fact, not be minimal. The explanatory hypothesis and the minimality criteria produce different revisions.

The authors cite a series of studies by Elio and Pelletier (1997) that show humans are more likely to give up the conditional ("if it is raining, then the grass is wet") rather than the case ("it is raining"), and even more likely when the subjects are faced with an inconsistent natural language story as compared to a symbolic representation. Khemlani and Johnson-Laird's own studies "sought to establish whether reasoners spontaneously create explanations that resolve inconsistencies or instead revise the assertions giving rise to them, perhaps in a minimal way" (2011). Their results include:

- "Most individuals propose explanations that indirectly refute generalizations and that are far from minimal changes. Such explanations are often what psychologists refer to as 'disabling conditions,' which provide cases in which the generalization fails."
- "Because of their propensity to envisage disabling conditions, their explanations are indeed more likely to invoke such conditions than to imply that a proposition about

a specific individual or entity is wrong."

- "Participants tended to select such explanations as the most probable."
- "Participants tended to evaluate them as having the highest rank of probability."
- "Participants tended to assign them the highest probability."

They note that,

> Of course, our results leave open the possibility that minimalism is a normative theory, in which case, they show that untrained individuals depart from a canon of rationality (Khemlani and Johnson-Laird, 2011).

We do not attempt to answer the question of whether minimalism is a normative theory, nor whether abductive metareasoning as specified in Chapter 4 is normative. Instead, we opt to experimentally verify that abductive metareasoning actually boosts accuracy and noise identification in both randomly generated and real abductive reasoning tasks. The following three chapters do exactly that.

Chapter 6: Simulated tracking domain

The first experimental domain is a simulated object tracking domain. In this domain, the reasoning system is tasked with making sense of a virtual world of moving objects in a 10x10 discrete grid-space. This grid constitutes the world, and is fully observable, with one caveat described below. Each object occupies one full grid cell (so all objects have the same size and shape). Each object has a unique color, and no other properties. The object color stands in for more realistic object properties that might be present in actual surveillance systems, e.g., object size, color distributions or covariances, etc. The objects movements are random walks. At each time step, each object makes a constant number of mostly random 1-step movements (diagonals not allowed). No two objects are allowed to occupy the same grid cell at the completion of their walk, so their otherwise random movements are restricted. If at the end of an object's random walk, two objects occupy the same cell, a new random walk is generated. There is no need to handle merges and splits. Simulated sensors report the final location of each object's walk at the end of the time step. The reasoning system is tasked with explaining reports by describing which object's movements connect two reports from time step t to $t+1$.

The simulated object tracking domain is intentionally simple. The system has very little knowledge that can be brought to bear during the generation and scoring of hypotheses. Objects only have a single color and no other properties. Objects move mostly randomly, so Kalman filters (Welch and Bishop, 1995) and other common tracking techniques are not helpful. My intention with such a simple task domain is to be sure that we can identify features of the *reasoning process* that is responsible for the system's observed performance.

We want to be sure that performance is not primarily due to substantial domain knowledge or special-purpose preprocessing that reduces the computational task before abductive reasoning and abductive metareasoning occur. In other words, knowledge is *not* power (c.f., Ed Feigenbaum) when we want to study general-purpose reasoning systems rather than special-purpose deployed systems.

In Chapter 7, we look at an aerial tracking domain with real video data. The purpose of the aerial domain is to ensure that the reasoning system works in both simplistic and realistic domains. Object tracking domains (both simulated and aerial) are useful domains for studying abductive reasoning and metareasoning for the following reasons.

- The task is easily framed in explicitly abductive terms, in which object detections make up reports and object movements serve as the explanatory hypotheses.
- As will be shown, it is helpful to establish a minimum plausibility η for movement hypotheses, though anomalies due to implausible hypotheses might result. In the simulated domain, it is also useful to establish a minimum decisiveness δ.
- In the simulated 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.
- In the simulated domain, the plausibility of future movement hypotheses depend on the system's current estimate of the situation, i.e. its beliefs. Long-distance movements are generally less plausible, and very low-plausibility hypotheses might be rejected. Consequently, false beliefs might cause order dependency anomalies.

Abductive reasoning and metareasoning are not only applicable to object tracking tasks. Chapter 8 looks at a generic Bayesian network inference task. Other interesting

domains that have not been implemented or evaluated at this time are considered in Chapter 10.

6.1 Gray area

As stated, each object bears a color unique to that object. Objects can be identified by their color, and tracking is trivial (assuming all reports are factual). However, the center 60% 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 40% is watched by sensors that do report color. When objects move into this outer area, they can be uniquely identified. The presence of the gray area is an essential feature of the simulated tracking domain. Because of this gray area, objects might be mistracked when they move in close range in the gray area. Yet, when the objects emerge from the gray area, these mistracks might render the new reports unexplainable. Hence, the gray area may produce anomalies. This possibility is shown in the top diagram of Figure 6.1. In Section 6.5 we experimentally measure how the gray area influences the presence of anomalies.

6.2 Noise

Noisy reports are simulated by introducing false reports which describe non-existent objects, and by distorting (randomly modifying) and deleting reports about actual objects. Each report has a small chance of manipulation in one of four kinds, which in our experiments ranged from 0% to 20%. Specifically, the four kinds of noise are as follows.

Distortion noise: A true report is randomly modified to report the object as at a different location and/or with a different color.

Duplication noise: A true report is duplicated and modified to report a random location.

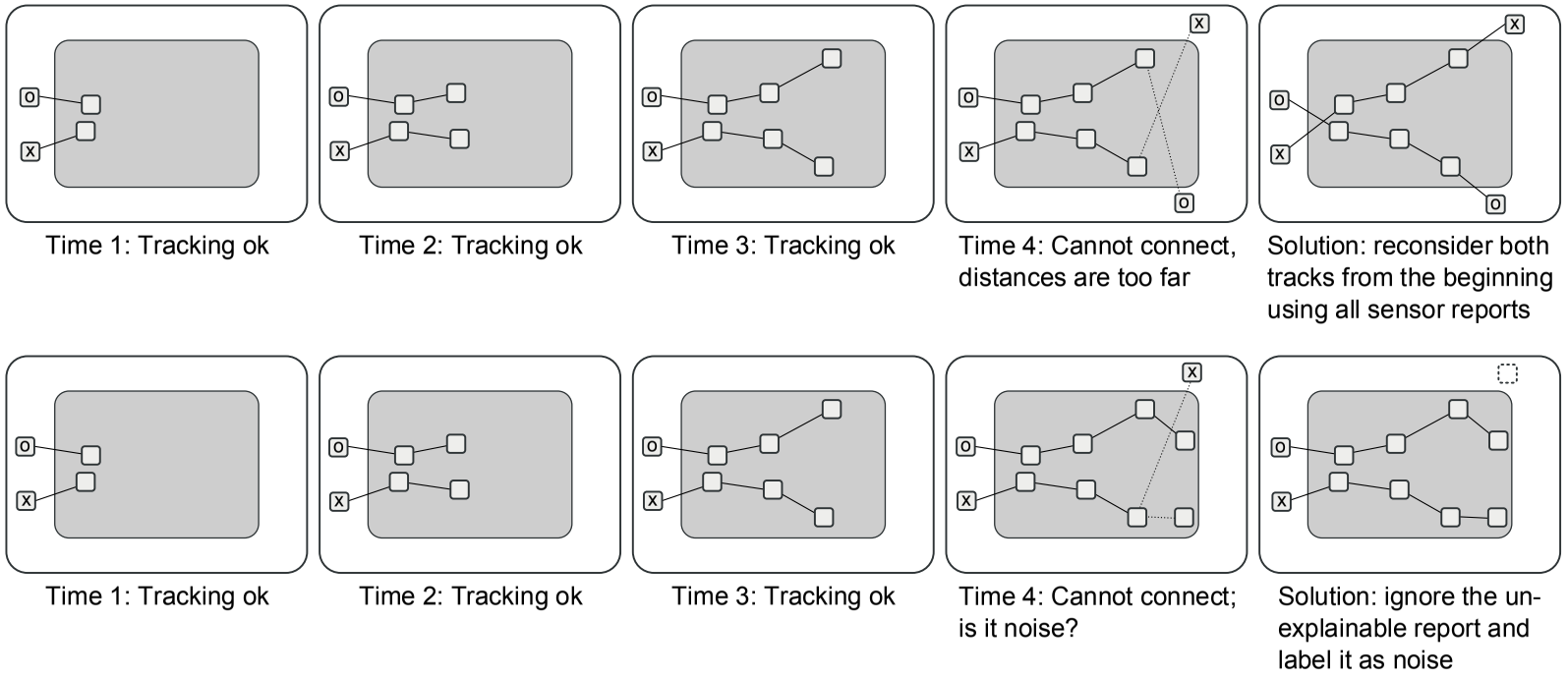


Figure 6.1. Anomalies due to order dependency (top) and noise (bottom) in the simulated object tracking domain. Reports for objects in the middle gray area do not specify the object's color (here represented by 'x' and 'o'), thus any object within a short range could possibly account for the detection.

Only the color of the object is retained in the duplicate report (though that color might be reported as "gray").

Insertion noise: A non-existent object is reported at some random location and with a random color.

Deletion noise: A true report is simply not reported.

Note that in each case excepting *deletion noise*, the reasoning system obtains a false report. If the system ultimately refuses to find or accept an explanation for such a false report, we say that it has identified the noisy report. However, deletion noise cannot be identified in this manner (there is no report to leave unexplained), so deletion noise is not represented in noise identification metrics.

Noise may produce anomalies. The bottom diagram of Figure 6.1 shows an example of duplication noise and the resulting anomaly. In Section 6.5 we experimentally measure how noise influences the presence of anomalies.

6.3 Definition as an abductive reasoning problem

Reports generated by the OBSERVE function take the form, "an object was detected at location x, y at time t with color c." Reports are assigned random plausibility scores such that noisy reports typically score low and true reports score high. Specifically, given a report r, its plausibility is $Pl(r) = \text{REPORTPLAUSIBILITY}(r)$, defined by Algorithm 6.1.

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 posit that two different objects moved into the same location or out of the same location at corresponding times (the domain does not handle merges and splits), or that the same object (identified by a color that is not "gray") is in two different locations

Algorithm 6.1 Algorithm for computing report plausibilities.

```
function REPORTPLAUSIBIILTY(r)
    p ← random number in [0,1]
    if (r is a true report and p < 0.9) or (r is a false report and p ≥ 0.9) then
        return GAUSSIAN(μ = 0.8, σ = 0.1)
    else
        return GAUSSIAN(μ = 0.4, σ = 0.1)
    end if
end function
```

at the same time. Figure 6.2 shows a simple example of reports and their hypotheses for the simulated object tracking domain.

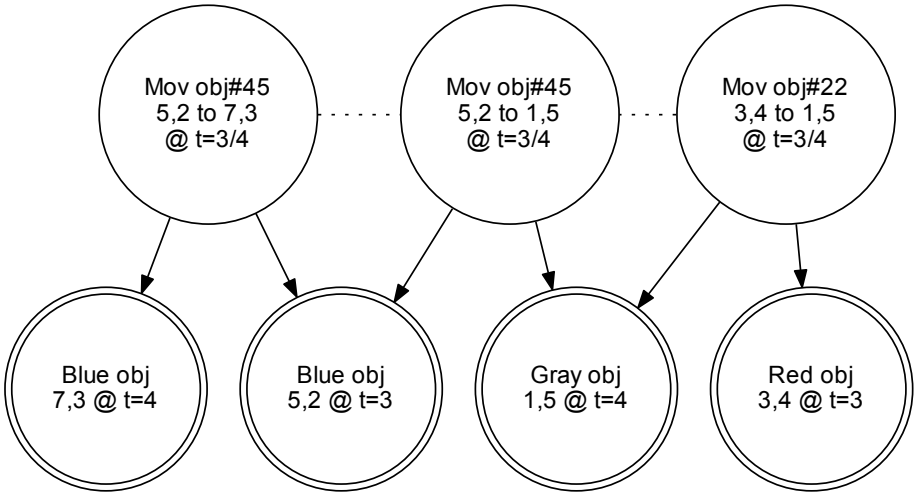


Figure 6.2. Example of an explanation graph for the simulated tracking domain.

Movement hypotheses are scored on the basis of the distance of the movement. The system is trained on 1,000 examples of object movements and builds a model of the probability of single time-step movements of various distances. Specifically, the plausibility of a movement hypothesis h is,

$$Pl(h) = G * \frac{1 + F(d)}{2 + \|\mathscr{T}\|}, \tag{6.1}$$

where h posits an object movement of distance d, and $F(d)$ is the frequency (count) of object movements of size d in the training examples, and $\|\mathscr{T}\|$ is the count of training examples (1,000). The extra 1 in the numerator and 2 in the denominator are added as Laplacian smoothing. G is a modifier equal to 1.0 if neither of the two reports that make up the movement describe a gray object, 0.75 if one report describes a gray object, or 0.50 if both reports describe a gray object. This modifier lowers the plausibility of movement hypotheses that explain gray reports, since such reports are not as certain as reports that explain reports of objects with colors. Experiments (not reported in this work) show that there is a small advantage to including the G modifier in overall tracking accuracy.

6.4 Experimental methodology

Each simulation involves 10 time steps and either two, four, six, eight, or ten different objects moving about. All experimental results are averaged across experiments with these different object counts. At each time step in a single experiment, each object takes a random walk of six grid steps (diagonals not allowed). Reports which remain unexplained after 10 time steps are called *noise claims*. The final doxastic state is evaluated according to the

following metrics, where $\|X\|$ denotes the cardinality of set X.

$$\begin{aligned}
\text{Precision} &= \frac{\|\text{Actual movements} \cap \text{Accepted movement hypotheses}\|}{\|\text{Accepted movement hypotheses}\|} \\
\text{Recall} &= \frac{\|\text{Actual movements} \cap \text{Accepted movement hypotheses}\|}{\|\text{Actual movements}\|} \\
\text{F1} &= \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \\
\text{Noise Precision} &= \frac{\|\text{Actual noisy reports} \cap \text{Noise claims}\|}{\|\text{Noise claims}\|} \\
\text{Noise Recall} &= \frac{\|\text{Actual noisy reports} \cap \text{Noise claims}\|}{\|\text{Actual noisy reports}\|} \\
\text{Noise F1} &= \frac{2 * \text{Noise Precision} * \text{Noise Recall}}{\text{Noise Precision} + \text{Noise Recall}}
\end{aligned}$$

The remainder of this chapter evaluates abductive reasoning and metareasoning in two stages. First, in Section 6.5, we look at experiments that validate the simulated object tracking domain in terms of its appropriateness as an abductive reasoning task. This validation is achieved by confirming, through a wide variety of experiments, that abductive reasoning, as applied to simulated object tracking, yields the kinds of results one would expect. These expectations are enumerated as hypotheses and each is confirmed in turn.

The second set of experiments in Section 6.6 evaluates the effectiveness of abductive metareasoning. We consider abductive metareasoning second to ensure that the base-level abductive reasoning task is not artificially handicapped just so that abductive metareasoning can be shown to increase accuracy. Rather, the first set of experiments show that abductive reasoning is appropriate and effective for the simulated object tracking task, and the second set of experiments show that abductive metareasoning brings an extra boost that could not have been achieved with just abductive reasoning.

6.5 Domain validation experiments

The following experimental hypotheses are labeled with the format "S-V-#" to indicate that these hypotheses are regarding simulated object tracking validation experiments.

Hypothesis S-V-1: EFLI abduction (Algorithm 2.3) yields significantly greater accuracy than arbitrary abduction. We expect this to be the case because EFLI prefers more plausible, more decisive hypotheses.

Hypothesis S-V-2: As the gray area increases (see Section 6.1), accuracy suffers. This is because objects are more often confusable. We also expect more anomalies when the gray area is neither absent (0%) nor total (100%). We expect this outcome because if the gray area is absent, all objects are uniquely identifiable, and if the gray area is total, then while mistracks are possible, anomalies are not.

Hypothesis S-V-3: As the noise level increases, anomalies also increase. This is to be expected because false reports should, in the usual case, have no plausible and consistent explanation. Accuracy should decrease, because noisy reports for which explanations are found usually have false explanations.

Hypothesis S-V-4: According to the discussion of the completeness–confidence trade-off (Section 2.8), we expect the following outcomes. (1) A minimum plausibility threshold $\eta > 0$ is expected to produce greater accuracy than $\eta = 0$, thus demonstrating the usefulness of this parameter. (2) A decisiveness threshold $\delta > 0$ is expected to produce greater accuracy than $\delta = 0$, thus demonstrating the usefulness of this parameter. However, we also expect that (3) when $\eta > 0$, more anomalies occur because hypotheses are more often rejected due to not meeting the minimum plausibility threshold.

Hypothesis S-V-5: When hypothesis plausibility estimates are hidden (or, equivalently, $Pl(h) = 1.0$ for all hypotheses h), abductive reasoning suffers. This outcome would show that plausibilities are used and useful. However, we also expect that only a small value for the plausibility precision (Definition 2.2.2), say, precision ≤ 7, is required to produce accuracy on par with very precise plausibilities. This is expected because in cases where two or more competing hypotheses have very small plausibility deltas, not enough information is available to make a confident decision. The same reasoning explains why we expect the decisiveness threshold $\delta > 0$ to yield better accuracy than $\delta = 0$.

In the remainder of this section, each domain validation hypothesis is addressed in turn.

Hypothesis S-V-1

Table 6.1 shows a comparison between EFLI and the arbitrary abduction algorithm, under noise-free conditions and $\eta = \delta = 0$. As expected, EFLI yields significantly better accuracy. We also see that EFLI and arbitrary abduction perform just as well when there was no gray area (gray = 0%). In such a configuration, all objects are uniquely identifiable, so both abduction algorithms (arbitrary and EFLI) should arrive at the same conclusions.

Hypothesis S-V-2

The gray area allows for object misidentification. Outside the gray area, objects are uniquely identifiable. Thus, we would expect that, if the gray area is entirely absent, every object is identifiable and perfect accuracy is achieved (assuming noise-free conditions). Additionally, we would expect no anomalies in this case; all reports should be explainable. On the other hand, if the entire grid is gray, then no objects can be uniquely identified, ever; all objects appear exactly the same. Accuracy should suffer and anomalies are impossible

Gray (%)	Precision	Recall	F1
0	0.000	0.000	0.000
20	+0.326 ***	+0.427 ***	+0.383 ***
40	+0.457 ***	+0.517 ***	+0.490 ***
60	+0.472 ***	+0.502 ***	+0.488 ***
80	+0.516 ***	+0.529 ***	+0.523 ***
100	+0.485 ***	+0.497 ***	+0.491 ***

Table 6.1. EFLI vs. arbitrary abduction for different sizes of the gray area, under no noise conditions and $\eta = \delta = 0$, supporting Hypothesis S-V-1. A value of 0.000 means that EFLI and arbitrary abduction produced equivalent accuracy. A value +0.326 for some metric means that EFLI produced, on average, 0.326 higher on that metric. Statistical significance is indicated by asterisks: *** indicates $p < 0.001$.

(assuming no noise and $\eta = 0$). An example of an anomaly resulting from partial gray area and no noise is shown in the top diagram of Figure 6.1.

We see in Figure 6.3 that when the gray area is absent (0%), accuracy is maximal, and accuracy generally decreases as the gray area increases. However, an interesting phenomenon occurs when the gray area is 100%. In this case, no objects bear any distinguishing characteristics, so objects are tracking based solely on distance. Accuracy is better in this case than when the gray area is 60% or 80%. This is because at 60% and 80% gray area, more anomalies are present (which we will see in Figure 6.5) and thus more mistracks result simply because tracks are lost.

Figure 6.4 illustrates that Plausibility errors become more numerous as the gray area increases. This is because plausibilities are more often calculated based entirely on distance and less often on object identifiability when the gray area is larger. NoExplOffered (no explanation offered) errors are most numerous when the gray area is 60%. Recall from Section 2.10 that a NoExplOffered error is defined as the acceptance of a false hypothesis because the true hypothesis was never offered. This can only occur, in the simulated

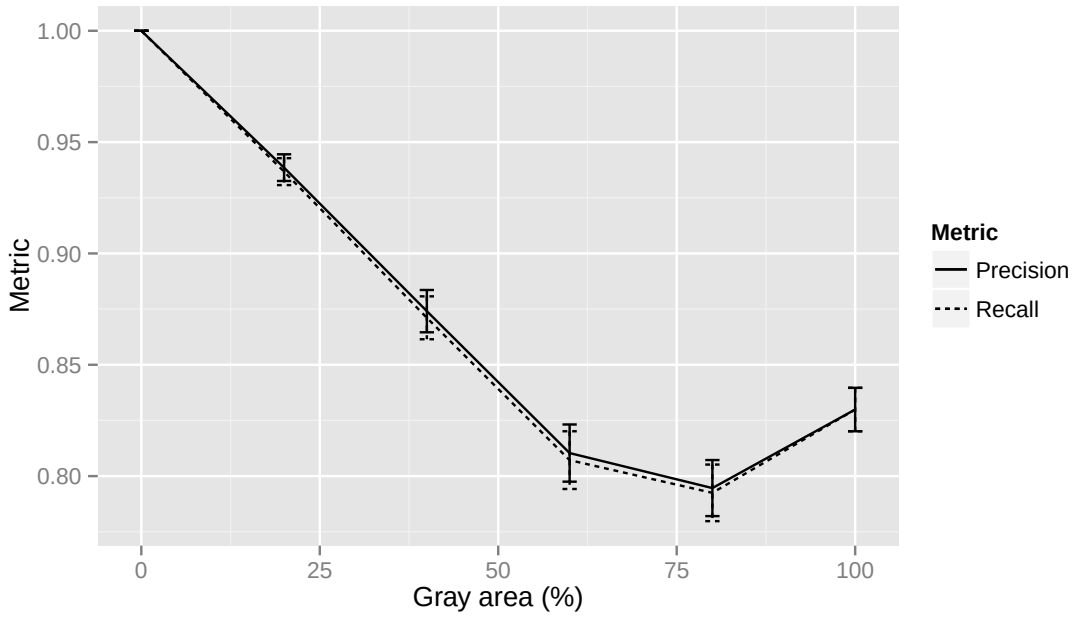


Figure 6.3. Impact of the gray area on accuracy, supporting Hypothesis S-V-2. $\eta = \delta = 0$ and there was no noise.

tracking domain, when an object is mistracked in the gray area but then cannot be linked to a report outside the gray area (i.e., a report indicating the object's color) because the colored report is simply too far from the object's prior believed (but incorrect) location. This scenario is demonstrated in the top diagram of Figure 6.1. Only when the gray area is neither absent nor total can this occur.

Finally, Figure 6.5 shows the impact of the gray area on the occurrence of anomalies. As expected, when the gray area is absent (0%) or total (100%), no anomalies occur. At 60% gray area, anomalies are maximized, though still relatively uncommon: only about 0.3% of reports are anomalous. Note that in these experiments, the gray area is a contiguous block in the middle of the tracking area. We have not experimented randomly-distributed gray cells rather than a single contiguous block.

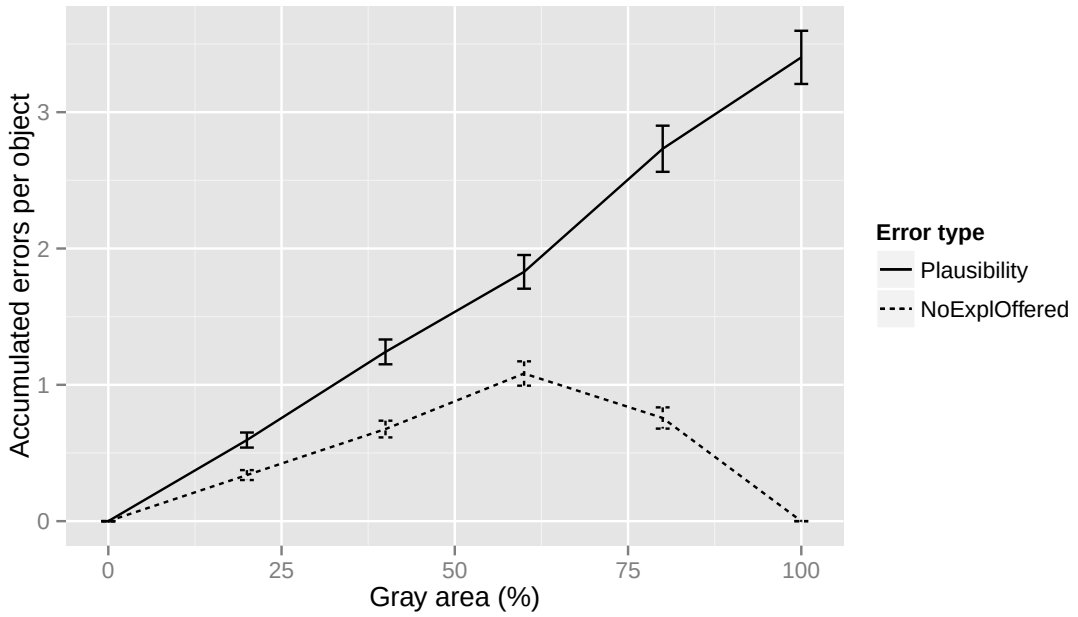


Figure 6.4. Impact of the gray area on errors (see Section 2.10), supporting Hypothesis S-V-2. Only Plausibility and NoExplOffered errors are shown because the occurrences of other kinds of errors were not affected by the size of the gray area. $\eta = \delta = 0$ and there was no noise.

Because we wish to investigate metareasoning, which responds to the presence of anomalies, further experiments will be conducted with a gray area of 60%. Accuracy for these cases is also quite low, as shown in Figure 6.3. This leaves wide room for improvement, which might be realized by metareasoning.

Hypothesis S-V-3

We see in Figure 6.6 that the presence of anomalies increases when the noise level increases. Recall from Section 6.2 that a noise level of $n\%$ indicates that each report has an $n\%$ chance of being modified or deleted and that for each report, there is an $n\%$ chance of a new false report being introduced. Thus, even a noise level of 20% introduces a significant

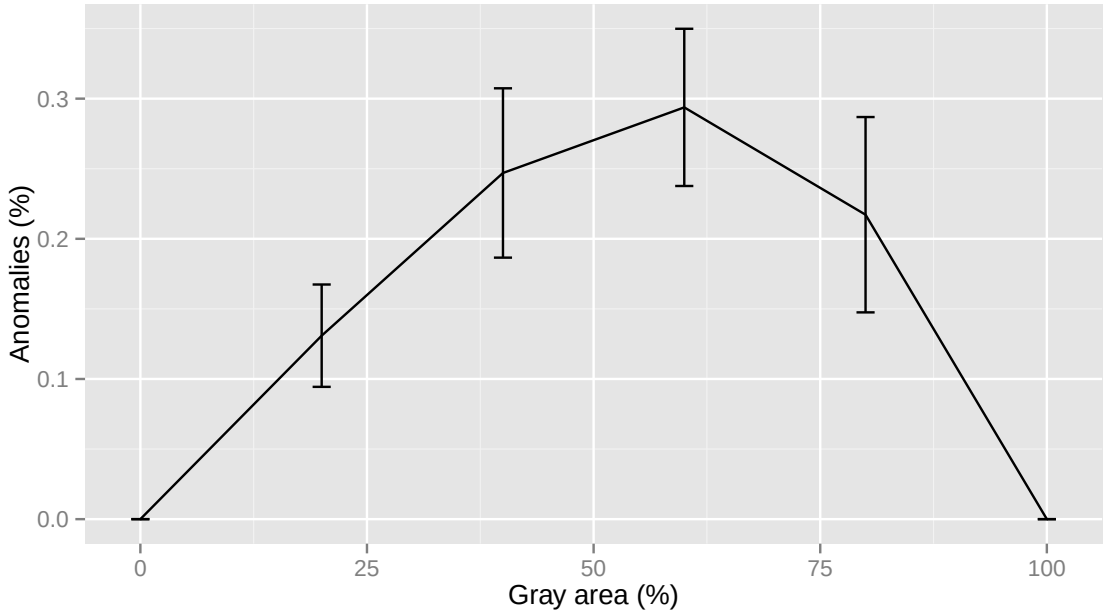


Figure 6.5. Impact of the gray area on the occurrence of anomalies, supporting Hypothesis S-V-2. $\eta = \delta = 0$ and there was no noise.

number of false and missing reports. As expected, when more false reports are introduced and more true reports are deleted, a larger percentage of reports are unexplainable. Furthermore, as Figure 6.7 shows, accuracy also suffers. Future experiments will typically involve noise levels at 0%, 10%, and 20%.

Hypothesis S-V-4

We expect that a minimum plausibility threshold $\eta > 0$ and a minimum decisiveness threshold $\delta > 0$ yield better accuracy in both noise-free and noisy conditions. Figure 6.8 shows the impact of δ on accuracy and noise detection for different values of η. We see that for each value of η, best performance is obtained with δ equal to about 0.20. Furthermore, $\eta = 0.10$ gives better performance than $\eta = 0$ or $\eta = 0.20$. Noise identification is also

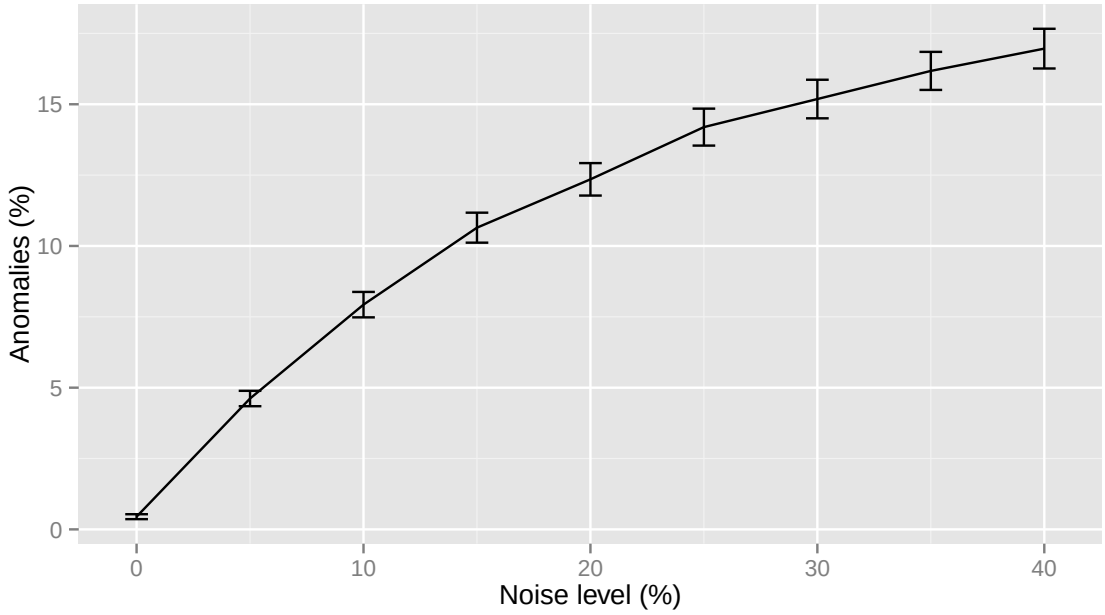


Figure 6.6. Impact of noise level on the occurrence of anomalies, supporting Hypothesis S-V-3. $\eta = \delta = 0$.

maximized at these same parameters.

In order to more carefully investigate the impact of η, we experimented with $\delta = 0.20$ and various η values. Figure 6.9 shows the impact of η on frequencies of errors and anomalies. We see in the top graph that increasing η produces more MinPlausibility errors but fewer of each other kind of error. This is to be expected, since more hypotheses are incorrectly rejected as η increases, but noisy reports and errors due to inaccurate plausibilities are reduced since fewer hypotheses are accepted. The bottom figure shows that MinPlausibility anomalies increase as η increases. This trend is also not surprising.

Unless stated otherwise, for the remainder of the simulated object tracking experiments, we set $\eta = 0.10$ and $\delta = 0.20$.

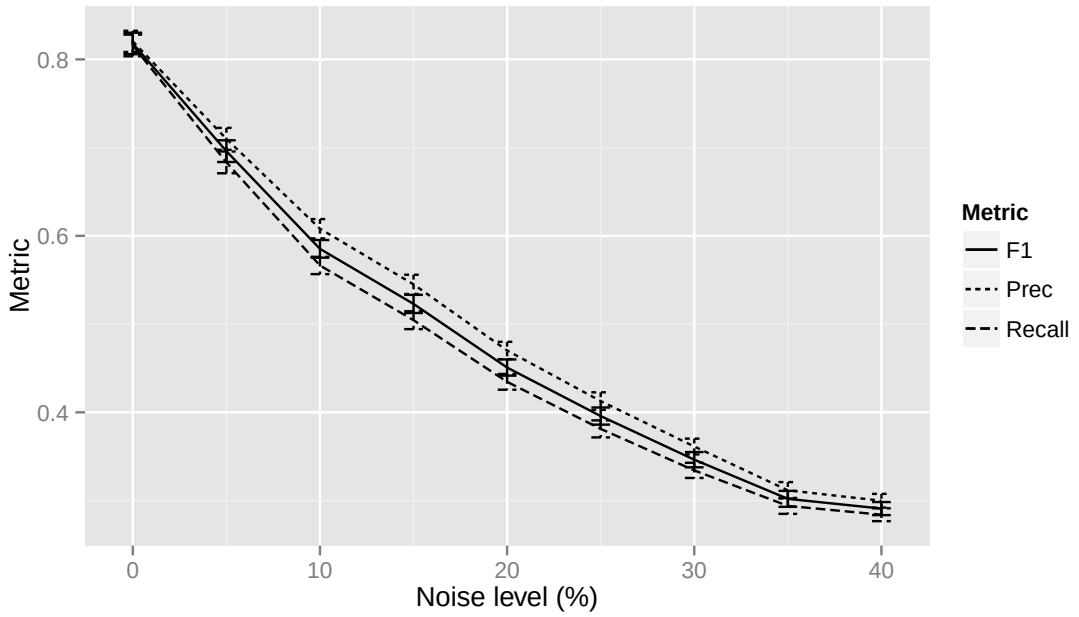


Figure 6.7. Impact of noise level on accuracy, relating to Hypothesis S-V-3. $\eta = \delta = 0$.

Hypothesis S-V-5

The first claim of Hypothesis S-V-5 states that hypothesis plausibility estimates are used and useful. We can simulate abductive reasoning with no information about plausibilities by setting $Pl(h) = 1.0$ for all hypotheses h. In this case, EFLI is unable to prefer hypotheses based on their plausibilities. When a contrast set contains two or more hypotheses, they also cannot be differentiated according to decisiveness. However, essential explainers are still preferred over non-essential explainers. Table 6.2 shows the increase in accuracy when scores are present. As expected, this increase is large and significant.

The second claim of Hypothesis S-V-5 states that only a small value for the plausibility precision (say, precision ≤ 7) is required to obtain accuracy on par with full plausibility precision (i.e., double-precision floating-point numbers). Figure 6.10 shows the impact of

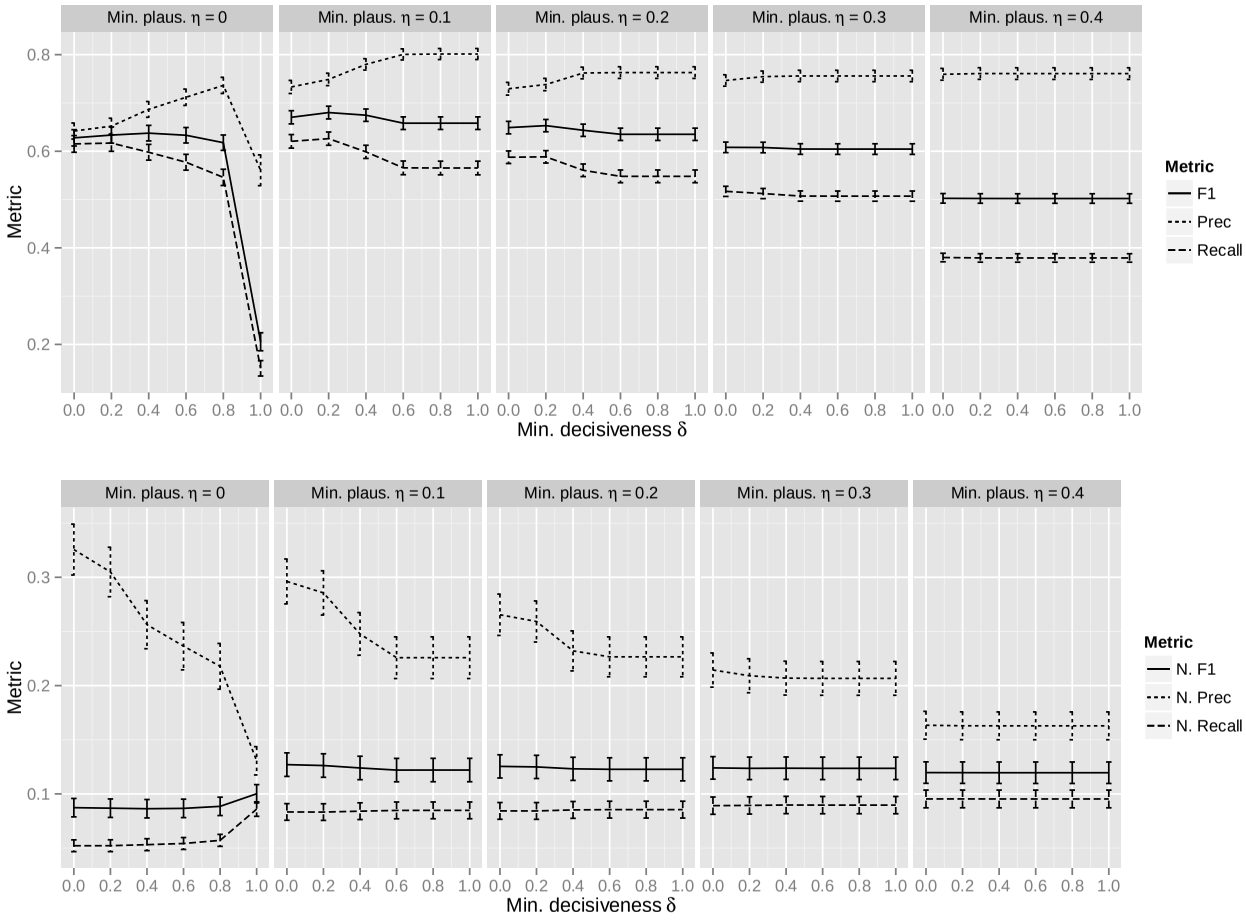


Figure 6.8. Impact of the decisiveness threshold δ on accuracy and noise identification for several values of the minimum plausibility η, and averaged across noise level values 0%, 10%, and 20%. "N. Prec." means "Noise Precision," etc. Supports Hypothesis S-V-4.

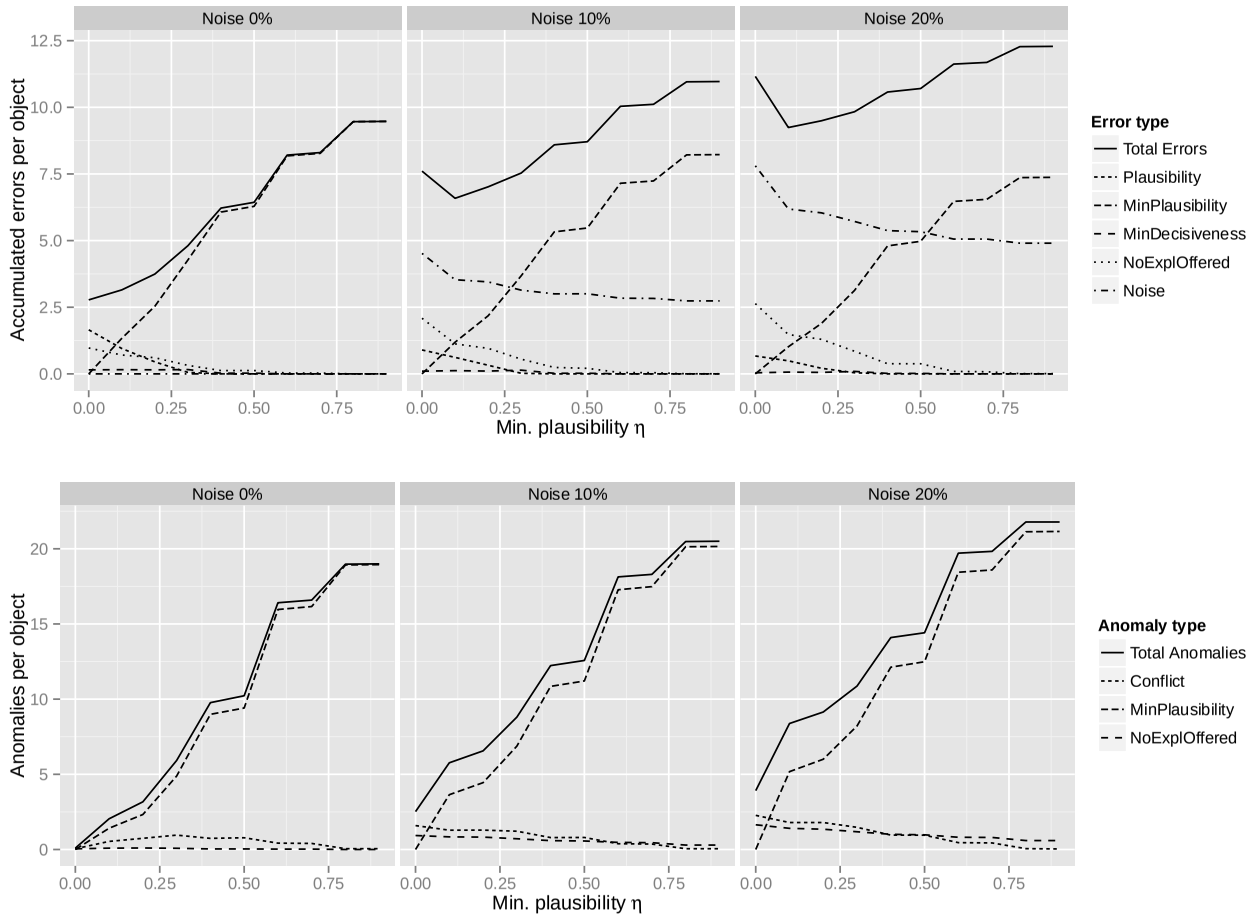


Figure 6.9. Impact of minimum plausibility η on frequencies of errors and anomalies, with $\delta = 0.20$, and averaged across noise level values 0%, 10%, and 20%. Supports Hypothesis S-V-4.

Noise	Prec.	Recall	F1	N. Prec.	N. Recall	N. F1
0	+0.420 ***	+0.429 ***	+0.425 ***			
10	+0.326 ***	+0.311 ***	+0.318 ***	+0.161 ***	+0.014 ***	+0.025 ***
20	+0.246 ***	+0.231 ***	+0.238 ***	+0.085 ***	+0.010 ***	+0.017 ***

Table 6.2. Impact of having no plausibility information, supporting Hypothesis S-V-5. $\eta = \delta = 0$. This experiment compares abductive reasoning with no plausibility information and abductive reasoning with normal plausibility information. In both cases with and without plausibility information, results are averaged across noise levels of 0%, 10%, and 20%. In the table, a Noise value of 10 indicates 10% noise level. "N. Prec." etc. refer to the Noise Precision metric, etc. A value $+0.432$ for some metric means that EFLI produced, averaged across individual cases, 0.432 higher on that metric. Statistical significance is indicated by asterisks: *** indicates $p < 0.001$.

plausibility precision on accuracy. We see that at plausibility precision $n = 5$, accuracy is virtually equivalent to full plausibility precision. Even so, in the remaining experiments, we use full plausibility precision.

At this point, all domain validation hypotheses for the simulated object tracking domain have been confirmed. Thus, we have demonstrated the usefulness and appropriateness for reasoning about the simulated object tracking domain with abductive reasoning. Furthermore, we have identified optimal parameters ($\eta = 0.10, \delta = 0.20$) for EFLI-based abductive reasoning, and have chosen to focus further experiments on a gray area size of 60% in order to maximize the likelihood of anomalies due to mistracking objects in the gray area. Next, we will examine how metareasoning can boost accuracy when applied to this domain.

6.6 Metareasoning experiments

The following experimental hypotheses are labeled with the format "S-M-#" to indicate that these hypotheses are regarding simulated object tracking metareasoning experiments.

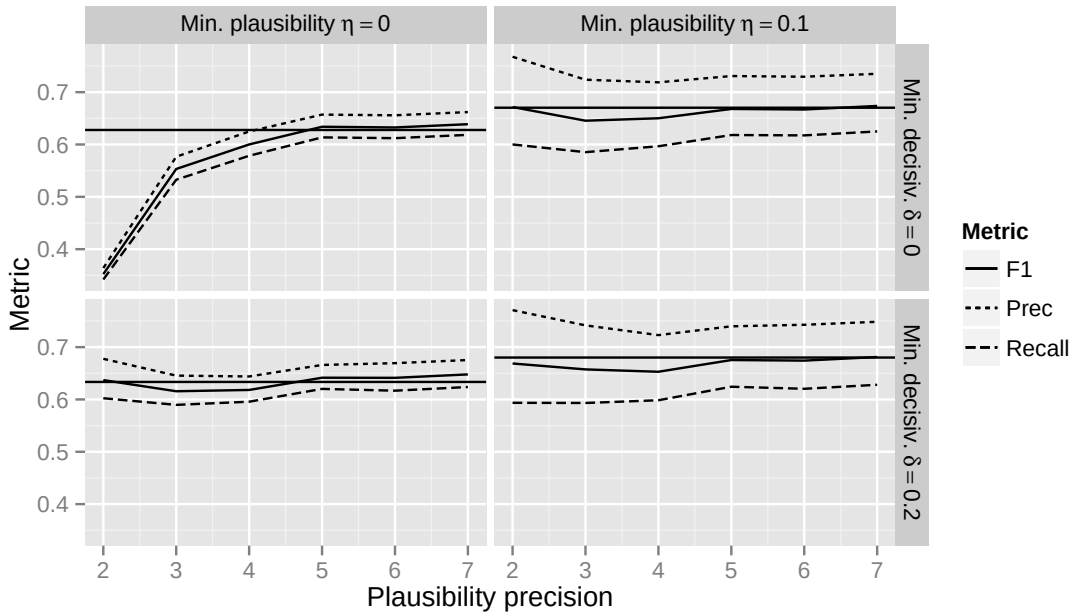


Figure 6.10. Impact of plausibility precision on accuracy for different η and δ values, supporting Hypothesis S-V-5. Results are averaged across noise levels of 0%, 10%, and 20%. The horizontal line marks F1 at maximum plausibility precision (double-precision floating-point numbers).

Hypothesis S-M-1: Analogous to Hypothesis S-V-4, we expect that $\eta_{\text{meta}} > 0$ and $\delta_{\text{meta}} > 0$ yield greatest accuracy when abductive metareasoning is activated. We expect this because if $\eta_{\text{meta}} = 0$, then even very implausible meta-hypotheses may be accepted and the system may thereby acquire false beliefs or inappropriately alter existing true beliefs. If $\delta_{\text{meta}} = 0$, then some arbitrary meta-hypothesis will be accepted when two or more are applicable as possible explanations for some anomaly. Generally, an anomaly has one cause (though this is not always the case), and assuming the anomaly is not a noisy report, only one meta-hypothesis of the competing meta-hypotheses is correct. But if the meta-hypotheses have the same plausibilities, then the system will not be able to ensure the correct one is accepted.

Hypothesis S-M-2: Abductive metareasoning gives better performance than no metareasoning, at the values of η_{meta} and δ_{meta} determined to be best by experiments related to Hypothesis S-M-1. Furthermore, even when we vary η and δ, overall maximum accuracy is achieved for some $\eta, \eta_{\text{meta}}, \delta$, and δ_{meta} when the system supports abductive metareasoning.

Hypothesis S-M-3: Abductive metareasoning gives better performance than no metareasoning even when report plausibilities are unknown (using η_{meta} and δ_{meta} established earlier).

Hypothesis S-M-4: Performance is maximized when each of the MetaImplHyp, MetaIncompatHyp, MetaOrderDep meta-hypotheses is available as a possible explainer of anomalies. This can be tested by ablation experiments in which the various subsets of meta-hypotheses are disabled.

In the remainder of this section, each metareasoning experimental hypothesis is addressed in turn.

Hypothesis S-M-1

Given $\eta = 0.10$ and $\delta = 0.20$, found earlier to be optimal for base-level abductive reasoning, Figure 6.11 shows that $\eta_{\text{meta}} > 0$ and $\delta_{\text{meta}} > 0$ yield best accuracy when abductive metareasoning is supported. In particular, $\eta_{\text{meta}} = 0.60$ and $\delta_{\text{meta}} = 0.10$ appear to be best.

Hypothesis S-M-2

Table 6.3 shows the impact of abductive metareasoning for $\eta = 0.10$, $\delta = 0.20$, $\eta_{\text{meta}} = 0.60$, and $\delta_{\text{meta}} = 0.10$. We see that in cases of no noise, abductive metareasoning sig-

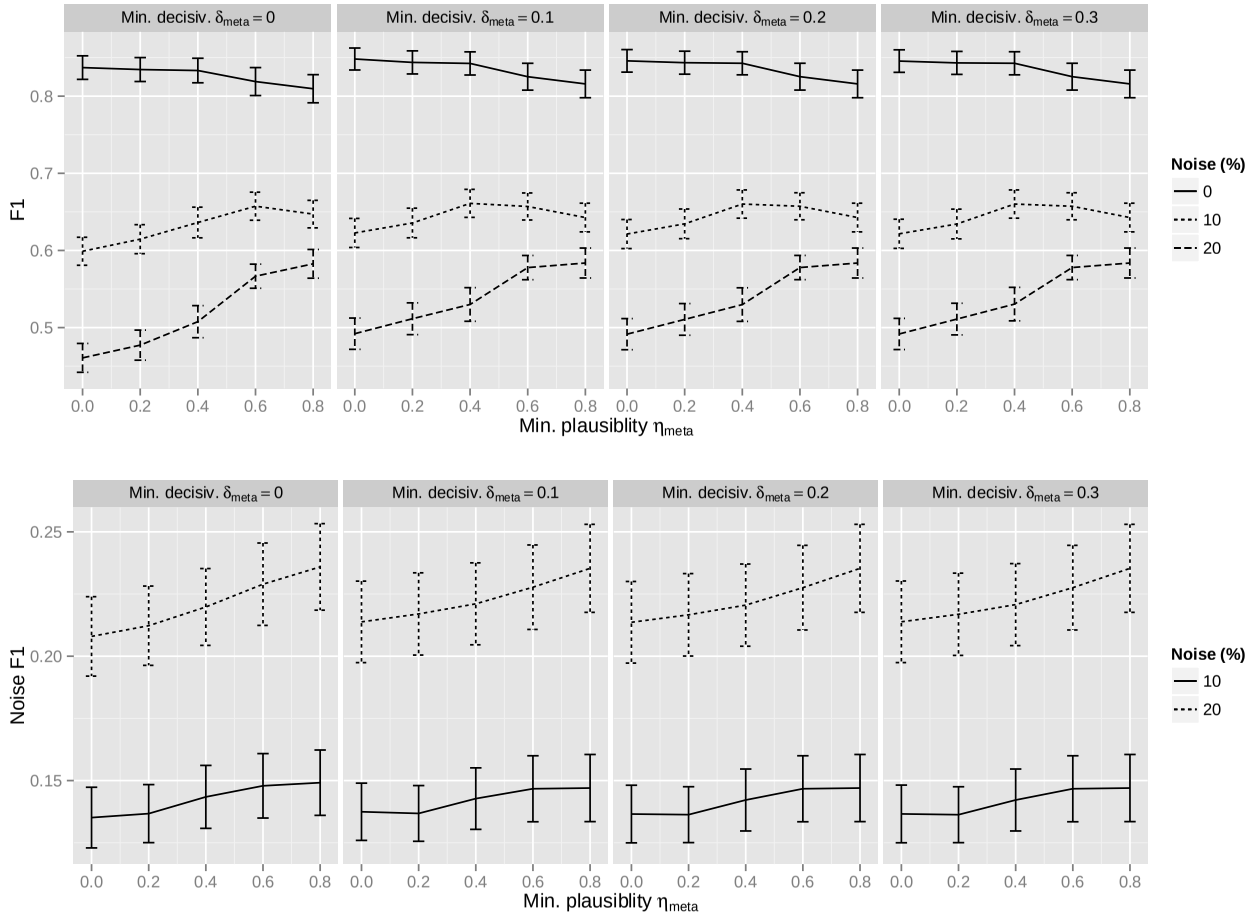


Figure 6.11. Impact of metareasoning parameters η_{meta} and δ_{meta} on accuracy. In every case, $\eta = 0.10$ and $\delta = 0.20$.

nificantly but only slightly improves accuracy. In cases when noise is present, abductive metareasoning significantly improves only Noise Precision.

Noise	Prec.	Recall	F1	N. Prec.	N. Recall	Noise F1
0	+0.024 ***	+0.058 ***	+0.042 ***			
10	+0.008	+0.004	+0.003	+0.030 ***	+0.001	+0.002
20	+0.012	+0.010	+0.008	+0.029 ***	0.000	+0.002

Table 6.3. Results from comparative experiments, supporting Hypothesis S-M-2, with abductive metareasoning and no metareasoning, for different noise levels. In the table, a Noise value of 10 indicates 10% noise level. "N. Prec." etc. refer to the Noise Precision metric, etc. A metric value $+0.021$ indicates that abductive metareasoning increased that metric on average by 0.021 compared to no metareasoning. Statistical significance is indicated by asterisks: *** indicates $p < 0.001$.

These results indicate that abductive metareasoning improves accuracy slightly but does not dramatically change the overall reasoning accuracy of the system. This outcome might seem surprising; perhaps we should expect abductive metareasoning to yield greater improvements. However, it is important to keep in mind that these experiments have focused on especially difficult object tracking tasks with a large gray area (60%), random object movements and, in some cases, significant noise (with noise level set to 10% or 20%). Given such a dearth of information that might help distinguish objects and identify noise, it should not be surprising that abductive metareasoning offers little improvement. There is very little information available to the abductive metareasoning system that can be used to explain anomalies. The base-level abductive reasoning system was not artificially handicapped (the EFLI algorithm and η and δ were defined to maximize accuracy before metareasoning was activated), so whatever knowledge could have been brought to bear on the task would have already been used in the base-level reasoning system. The base-level reasoning system fails, in some cases, to correctly track the objects and identify noise due to either an inappropriate η value for one or more true hypotheses (handled by

MetaImplHyp meta-hypotheses), failure to prevent an anomaly due to conflicting hypotheses and EFLI's greedy algorithm (handled by MetaIncompatHyp), or failure to explain a report due to mistracking an object at an earlier time step (handled by MetaOrderDep). Abductive metareasoning is designed to handle these contingencies, even though they are rare.

It is important to verify that abductive metareasoning is optimal for the simulated object tracking task. Table 6.3 only shows that abductive metareasoning improves accuracy given $\eta = 0.10$, $\delta = 0.20$, $\eta_{\text{meta}} = 0.60$ and $\delta_{\text{meta}} = 0.10$. However, one might wonder if perhaps for some other values of η and δ, accuracy can be maximized without abductive metareasoning. Figures 6.12 and 6.13 show results for various values of η, δ, and η_{meta} (experiments not reported here indicate that $\delta_{\text{meta}} = 0.10$ is best in all cases). We see that maximum accuracy is achieved by including abductive metareasoning.

Further experiments show that metareasoning improves accuracy across various sizes of the gray area. Figures 6.14 and 6.15 show tracking accuracy and noise identification accuracy, respectively, for abductive reasoning with and without metareasoning for different noise levels. The parameters η, η_{meta}, δ, and δ_{meta} are held constant at optimal values found earlier. We see that abductive metareasoning almost always increases accuracy over the base-level system. Only in cases where noise is present and the gray area is very small or non-existent do Recall and Noise Precision suffer. These metrics are correlated. When more true movements are left unexplained, and thus claimed to be noise, both Recall and Noise Precision decline. When the gray area is very small, most noise will take the form of a false report describing a *colored* object. This is because the false report will likely occur outside of the gray area. Movement hypotheses that connect this false report to other reports will more often connect two colored reports, so their plausibility estimates will on average be higher (recall from Section 6.3 and Equation 6.1 that movement hypothesis plausibilities are penalized if one or both of the reports are *gray* reports). The effect is that

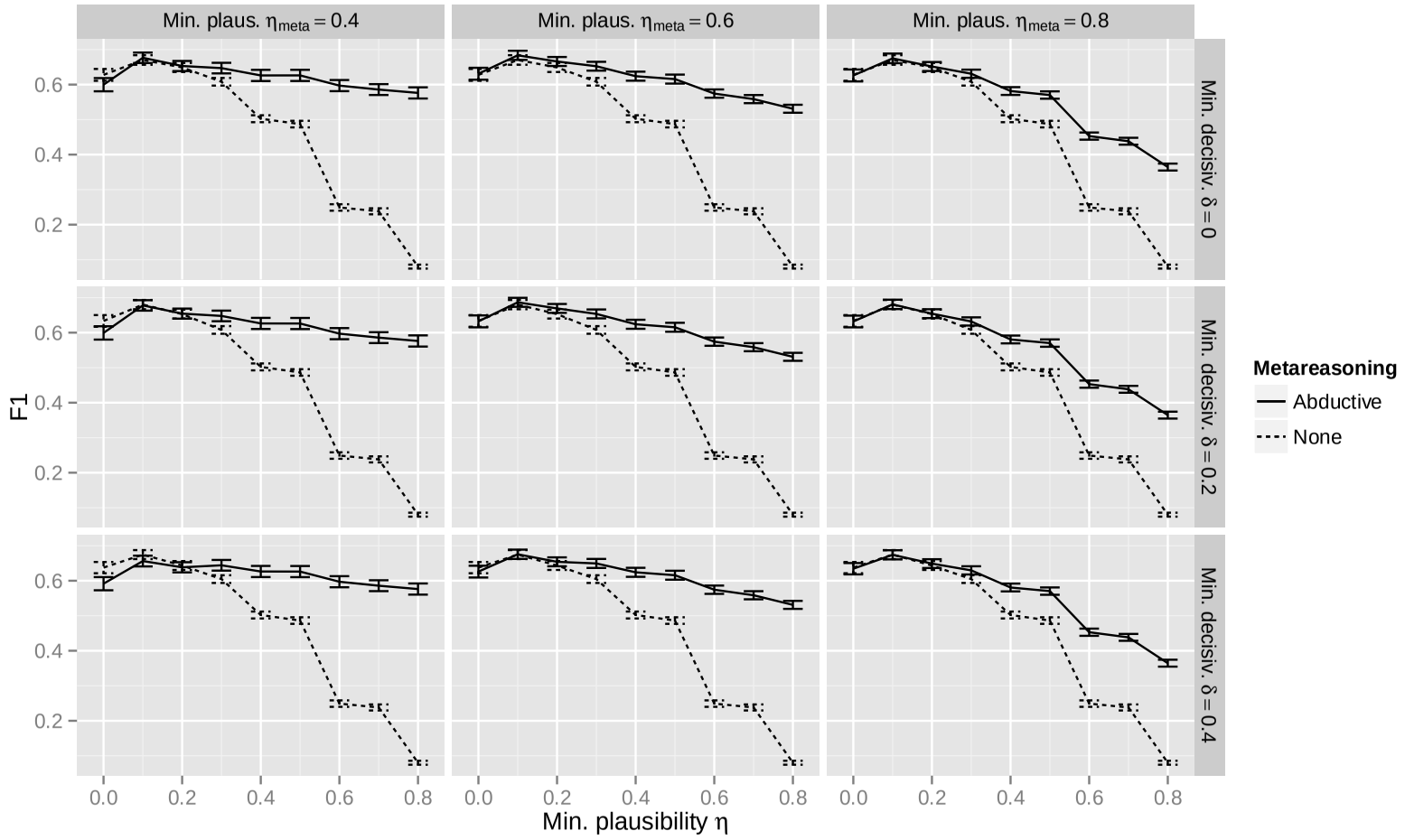


Figure 6.12. Accuracy (F1) for various values of η, δ, and η_{meta}, averaged across 0%, 10%, and 20% noise. Supports Hypothesis S-M-2. In all cases, $\delta_{\text{meta}} = 0.10$.

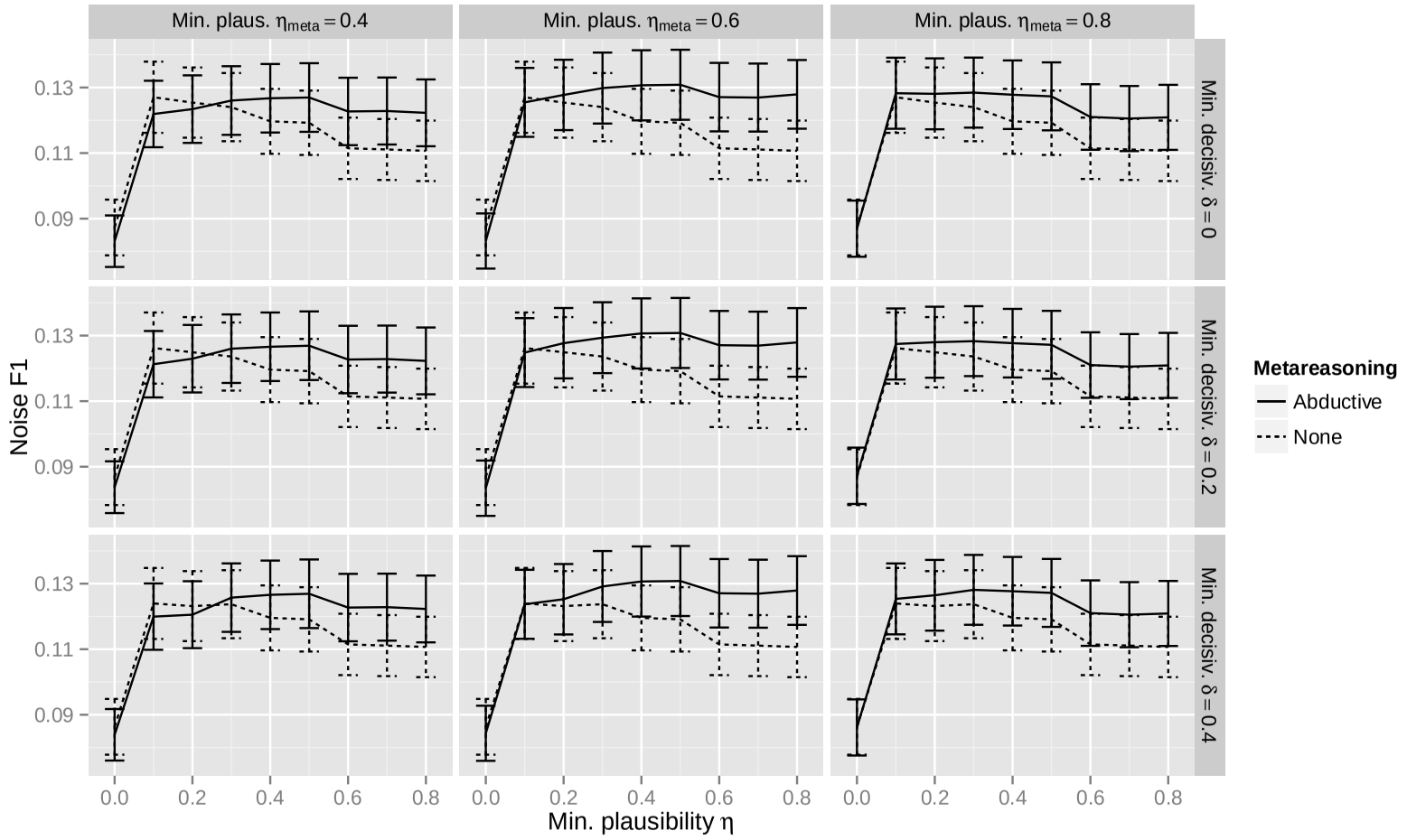


Figure 6.13. Noise identification accuracy (Noise F1) for various values of η, δ, and η_{meta}, averaged across 0%, 10$, and 20% noise. Supports Hypothesis S-M-2. In all cases, $\delta_{\text{meta}} = 0.10$.

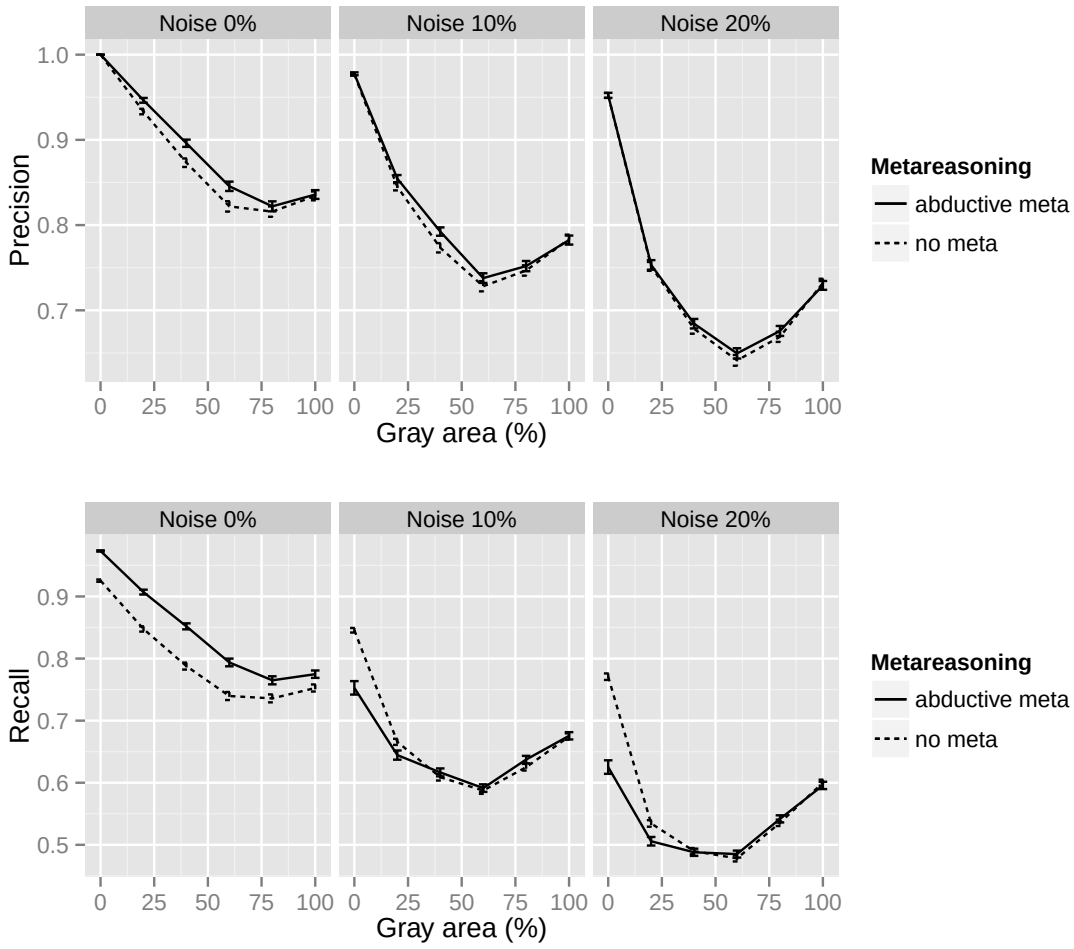


Figure 6.14. Impact of gray area on metareasoning accuracy, for various noise levels. Supports Hypothesis S-M-2. In all cases, $\eta = 0.10$, $\eta_{\text{meta}} = 0.60$, $\delta = 0.20$, $\delta_{\text{meta}} = 0.10$.

false movement hypotheses will generally score higher in cases with small or non-existent gray areas than in cases with large gray areas. We suspect, therefore, that abductive metareasoning will continue to boost accuracy over the base-level system if the parameters η, η_{meta}, δ, and δ_{meta} are appropriately adjusted with respect to the size of the gray area.

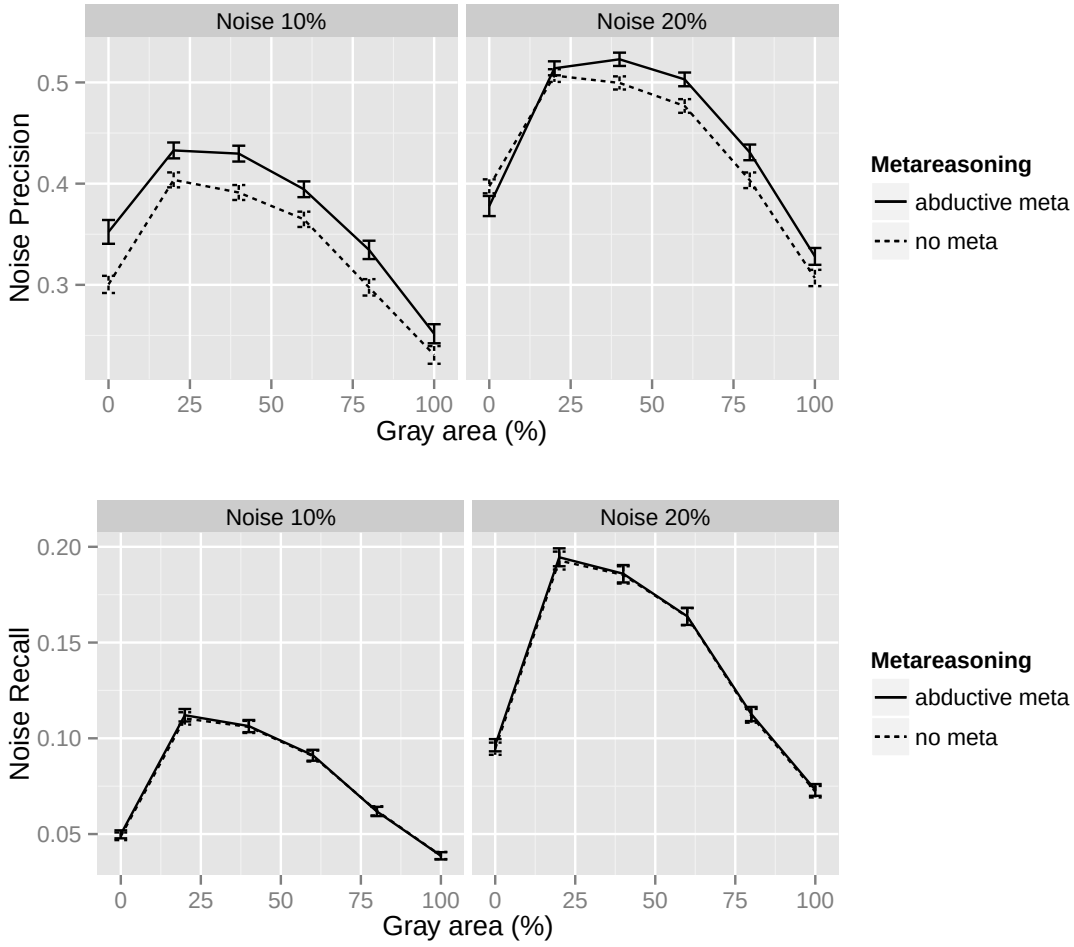


Figure 6.15. Impact of gray area on metareasoning for noise identification, for two noise levels. Supports Hypothesis S-M-2. In all cases, $\eta = 0.10$, $\eta_{\text{meta}} = 0.60$, $\delta = 0.20$, $\delta_{\text{meta}} = 0.10$.

Hypothesis S-M-3

Recall that the plausibility function for meta-hypotheses (Section 4.5) is simply the average of the plausibilities of the anomalies that a meta-hypothesis is capable of explaining. In the case of the simulated object tracking domain, only sensor reports require explanation, so only sensor reports may be anomalous. Thus, in the simulated tracking domain, the plausi-

bility of meta-hypotheses depends directly on the plausibility of sensor reports. Therefore, we feel that it is important to understand how much abductive metareasoning depends on the plausibilities of anomalous sensor reports, and how much abductive metareasoning depends on other factors such as the determination of which meta-hypotheses are applicable in a given situation (detailed in Section 4.1).

In order to investigate this issue, we tested abductive metareasoning in situations where all sensor reports have plausibilities equal to 1.0. Thus, all meta-hypotheses likewise have plausibilities equal to 1.0 and therefore cannot be compared according to their plausibilities. Rather, in these experiments, meta-hypotheses are accepted, and their corresponding belief revisions applied, simply on the basis of whether they can explain the anomalies. We retained $\delta_{\text{meta}} = 0.10$, so meta-hypotheses were accepted only if they were the only possible explanation for an anomaly.

Noise	Prec.	Recall	F1	N. Prec.	N. Recall	Noise F1
0	+0.035 ***	+0.078 ***	+0.058 ***			
10	+0.019 *	-0.087 ***	-0.067 ***	-0.055 ***	-0.001	-0.011 ***
20	+0.008	-0.132 ***	-0.114 ***	-0.093 ***	+0.005	-0.015 ***

Table 6.4. Results from comparative experiments, supporting Hypothesis S-M-3, in which report plausibilities are unknown (i.e., constantly 1.0). Abductive metareasoning is compared to no metareasoning for different noise levels. In the table, a Noise value of 20 indicates 20% noise level. "N. Prec." etc. refer to the Noise Precision metric, etc. A metric value $+0.021$ indicates that abductive metareasoning increased that metric on average by 0.021 compared to no metareasoning. Statistical significance is indicated by asterisks: * indicates $p < 0.05$, *** indicates $p < 0.001$.

Results are shown in Table 6.4. We see that Recall and Noise Precision suffer when noise is present. These two metrics are related: if abductive metareasoning is more reluctant to accept meta-hypotheses, then more anomalies will remain unexplained, and ultimately be labeled as "noise." Thus, some true reports will have no explanation (reducing Recall) and subsequently will be labeled as noise (reducing Noise Precision). As described

above, it is indeed the case that if all meta-hypotheses have equal plausibilities (1.0) and $\delta_{\text{meta}} > 0$, then fewer meta-hypotheses are accepted. In conclusion, we find that abductive metareasoning does depend crucially on the presence of (reasonably accurate) sensor report plausibility estimates.

Hypothesis S-M-4

Finally, Hypothesis S-M-4 states that each kind of meta-hypothesis (MetaImplHyp, MetaIncompatHyp, MetaOrderDep; refer to Section 4.1) plays an important role in abductive metareasoning. We can test this claim by conducting ablation experiments, in which only a subset of the three meta-hypotheses is available for abductive metareasoning. Each subset is tested in turn (excluding the case in which no meta-hypotheses are available, since we essentially tested that case earlier in regards to Hypothesis S-M-2).

Table 6.5 summarizes the results. We see that in noise levels 10% and 20%, abductive metareasoning with all three kinds of meta-hypotheses gives best results. However, when no noise is present, leaving out MetaIncompatHyp gives best results. The difference is minute, but nevertheless this outcome is interesting. It suggests that, in some (rare) cases, a false MetaIncompatHyp is accepted and an inappropriate belief revision is committed. It is unclear at this time how to identify and avoid these cases.

6.7 Conclusions

The simulated object tracking domain supported our first experimental look into abductive reasoning and abductive metareasoning. Each of the experimental hypotheses was confirmed. Thus, we now have strong evidence that abductive reasoning, and the EFLI algorithm in particular, is an appropriate reasoning strategy for making sense of reports about simulated object movements, and that abductive metareasoning is effective at find-

Noise	Impl	Incompat	OD	Precision	Recall	F1	N. Precision	N. Recall	N. F1
0	X	X	X	0.832 ± 0.010	0.773 ± 0.011	0.801 ± 0.010			
0	X	X		0.811 ± 0.011	0.751 ± 0.013	0.780 ± 0.012			
0	**X**		**X**	**0.836 ± 0.010**	**0.777 ± 0.011**	**0.805 ± 0.010**			
0	X			0.814 ± 0.012	0.753 ± 0.014	0.782 ± 0.013			
0		X	X	0.811 ± 0.010	0.719 ± 0.011	0.761 ± 0.010			
0		X		0.767 ± 0.012	0.671 ± 0.013	0.715 ± 0.012			
0			X	0.797 ± 0.014	0.711 ± 0.014	0.751 ± 0.014			
10	**X**	**X**	**X**	**0.735 ± 0.011**	**0.590 ± 0.011**	**0.650 ± 0.010**	**0.389 ± 0.013**	**0.090 ± 0.005**	**0.143 ± 0.007**
10	X	X		0.709 ± 0.012	0.592 ± 0.011	0.644 ± 0.011	0.399 ± 0.013	0.089 ± 0.005	0.143 ± 0.007
10	X		X	0.737 ± 0.011	0.578 ± 0.012	0.640 ± 0.011	0.354 ± 0.014	0.083 ± 0.005	0.131 ± 0.007
10	X			0.700 ± 0.013	0.588 ± 0.012	0.638 ± 0.012	0.358 ± 0.015	0.075 ± 0.005	0.123 ± 0.007
10		X	X	0.725 ± 0.010	0.562 ± 0.010	0.630 ± 0.009	0.347 ± 0.012	0.090 ± 0.005	0.140 ± 0.007
10		X		0.693 ± 0.011	0.557 ± 0.010	0.616 ± 0.010	0.346 ± 0.013	0.087 ± 0.005	0.136 ± 0.007
10			X	0.739 ± 0.013	0.543 ± 0.015	0.616 ± 0.013	0.319 ± 0.015	0.086 ± 0.006	0.131 ± 0.008
20	**X**	**X**	**X**	**0.656 ± 0.011**	**0.491 ± 0.010**	**0.556 ± 0.010**	**0.499 ± 0.013**	**0.163 ± 0.009**	**0.239 ± 0.010**
20	X	X		0.631 ± 0.012	0.488 ± 0.010	0.548 ± 0.010	0.500 ± 0.013	0.160 ± 0.008	0.237 ± 0.010
20	X		X	0.654 ± 0.012	0.469 ± 0.012	0.538 ± 0.011	0.466 ± 0.014	0.155 ± 0.009	0.226 ± 0.011
20	X			0.619 ± 0.013	0.485 ± 0.011	0.543 ± 0.012	0.476 ± 0.014	0.144 ± 0.009	0.216 ± 0.011
20		X	X	0.654 ± 0.011	0.472 ± 0.010	0.543 ± 0.010	0.468 ± 0.012	0.165 ± 0.009	0.238 ± 0.010
20		X		0.625 ± 0.012	0.464 ± 0.010	0.530 ± 0.010	0.470 ± 0.012	0.162 ± 0.008	0.236 ± 0.010
20			X	0.664 ± 0.015	0.439 ± 0.014	0.518 ± 0.013	0.422 ± 0.016	0.160 ± 0.012	0.225 ± 0.014

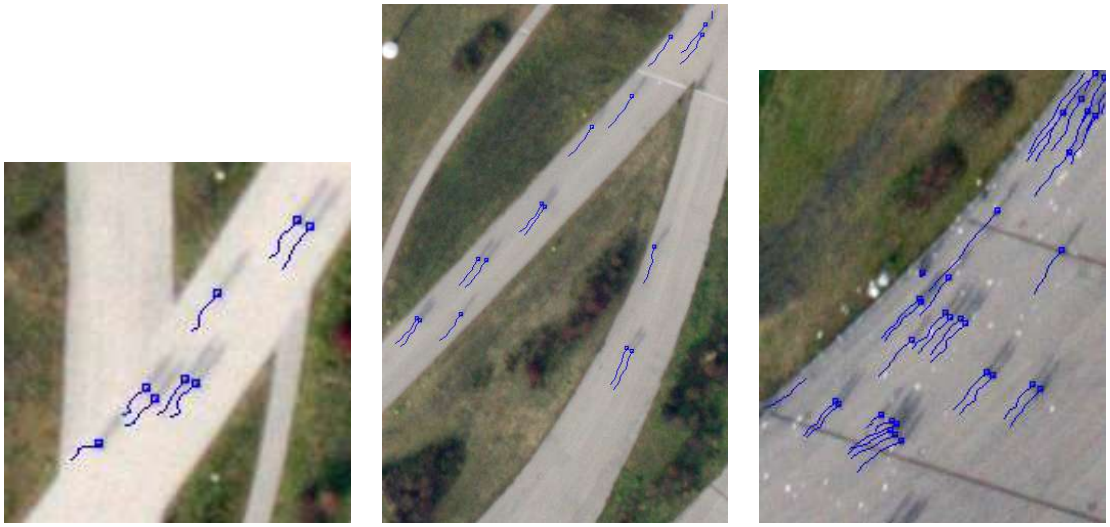
Table 6.5. Results from ablation experiments, supporting Hypothesis S-M-4, in which various meta-hypotheses are disabled. A Noise value of 20 indicates 20% noise level. Columns "Impl," "Incompat," "OD" refer to presence ("X") or absence of support for each meta-hypothesis: MetaImplHyp, MetaIncompatHyp, MetaOrderDep. Each value for each metric is shown as average ± standard error. Rows with best performance for each of the three noise levels are highlighted in bold.

ing appropriate belief revisions that increase tracking accuracy and at identifying noise.

However, results with just the simulated object tracking domain are not enough to prove that abductive reasoning and metareasoning are *generally* appropriate strategies for making sense of the world. More evidence is needed from other task domains. To this end, we next will consider an aerial tracking domain which, while in the same style as simulated tracking, brings another perspective by employing sensor reports from real-world aerial video surveillance. Then, we will look at a completely different set of domains that are abstract in nature and defined by arbitrary Bayesian networks. When results from all three domains are combined, we will possess strong evidence that abductive reasoning combined with abductive metareasoning is a very effective manner of reasoning.

Chapter 7: Aerial tracking domain

In addition to the simulated object tracking domain, we experimented with object tracking from aerial imagery using the KIT AIS Data Set (Schmidt and Hinz, 2011). See Figure 7.1 for examples of this dataset. In each frame, hundreds of very small (4x4 pixel) person-like objects are detected by a *Gentle AdaBoost* classifier (Friedman et al., 2000); we take these detections to be reports that need to be explained. Each detection is assigned a confidence score which we take to be the report's plausibility. Details may be found in Schmidt and Hinz (2011). Once reports are obtained, object tracking is performed in the same manner as the simulated tracking domain.



AA_Easy_01 AA_Easy_02 AA_Walking_01

Figure 7.1. Examples of aerial tracking datasets. Only the human-labeled ground-truth tracks are shown, though each scenario includes abundant false detections.

Our goal with the aerial tracking domain is to show that abductive reasoning and

metareasoning work well on realistic data. However, there are also drawbacks to using realistic data: we cannot *generate* alternative cases that exhibit interesting behavior, and we cannot *objectively* measure the system's performance; rather, we can only compare the system's performance with human performance as defined by the human-labeled ground-truth. However, we feel that the inclusion of the aerial tracking domain can only increase insights about abductive reasoning and metareasoning.

7.1 Definition as an abductive reasoning problem

The aerial tracking domain is implemented nearly equivalently to the simulated object tracking domain. Refer to Chapter 6 for details. There are two important differences between the two domains. (1) In the aerial domain, detections never report a unique identifying property. The simulated object tracking domain, on the other hand, sometimes reports object colors, which are uniquely identifying. (2) The second difference is that the aerial domain already includes noisy detections, so noise is not simulated (i.e., reports are not randomly modified or generated in order to create confusion).

Movement hypotheses connecting two reports are scored on the basis of the distance of the movement, just as in the simulated tracking domain. The system is trained in five frames of video and provided the true object movements for those frames. A model of movement distances (in terms of pixels) is established on the basis of these examples. The video footage is already orthorectified with a fixed ground sampling distance so pixels are a reasonable stand-in for actual travel distance. The GENERATEHYPOTHESES function only generates movement hypotheses for reports within $2 * d_{\text{avg}}$ distance, where d_{avg} is the average movement distance in the training examples. Thus, in some cases, reports are anomalous because no other report is near enough to generate a movement hypothesis. This hard limit for maximum movement distance is used so that the system does not generate all

Dataset	True reports	False reports	Avg. true plaus.	Avg. false plaus.
AA_Easy_01	75	204	0.660	0.407
AA_Easy_02	137	2005	0.623	0.402
AA_Walking_01	296	1060	0.575	0.443

Table 7.1. Noise per dataset. "Avg. true plaus." means "Average plausibility of true reports."

$O(n^2)$ possible movement hypotheses, where n is the number of reports for a single video frame.

7.2 Experimental methodology

The original datasets contain hundreds of "detections" for each frame. As mentioned, the AdaBoost classifier assigns each detection a confidence score (between 0.0 and 1.0). We take this confidence score to be the plausibility of the detection (a.k.a., report). However, we have found that in nearly all cases, very-low confidence detections are false detections. Thus, we apply a confidence threshold to the detections: any detection with confidence < 0.35 is eliminated and not converted into a sensor report. This threshold is applied to each of the datasets. We do so only to reduce the number of useless reports and thereby speed up our experiments. Even so, Table 7.1 shows that many false detections are still reported and true and false reports are not well-separated according to plausibility.

As in the simulated object tracking domain, we measure the accuracy of beliefs in the final doxastic state according to the following metrics, where $\|X\|$ denotes the cardinality of set X.

$$
\begin{aligned}
\text{Precision} &= \frac{\|\text{Actual movements} \cap \text{Accepted movement hypotheses}\|}{\|\text{Accepted movement hypotheses}\|} \\
\text{Recall} &= \frac{\|\text{Actual movements} \cap \text{Accepted movement hypotheses}\|}{\|\text{Actual movements}\|} \\
\text{F1} &= \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \\
\text{Noise Precision} &= \frac{\|\text{Actual noisy reports} \cap \text{Noise claims}\|}{\|\text{Noise claims}\|} \\
\text{Noise Recall} &= \frac{\|\text{Actual noisy reports} \cap \text{Noise claims}\|}{\|\text{Actual noisy reports}\|} \\
\text{Noise F1} &= \frac{2 * \text{Noise Precision} * \text{Noise Recall}}{\text{Noise Precision} + \text{Noise Recall}}
\end{aligned}
$$

It is worth noting that each dataset (see Figure 7.1) is tested once for each parameter specification, rather than multiple times with different random object movements as in the simulated tracking domain. All the detections in the dataset are processed in each experiment. Thus, there is no need to consider standard error or statistical significance of experimental results.

The remainder of this chapter evaluates abductive reasoning and metareasoning in two stages. First, in Section 7.3, we look at experiments that validate the aerial tracking domain in terms of its appropriateness as an abductive reasoning task. This validation is achieved by confirming, through a wide variety of experiments, that abductive reasoning, as applied to aerial tracking, yields the kinds of results one would expect. These expectations are enumerated as hypotheses and each is confirmed in turn.

The second set of experiments in Section 7.4 evaluates the effectiveness of abductive metareasoning. We consider abductive metareasoning second to ensure that the base-level abductive reasoning task is not artificially handicapped just so that abductive metareasoning can be shown to increase accuracy. Rather, the first set of experiments show that abductive

reasoning is appropriate and effective for the aerial tracking task, and the second set of experiments show that abductive metareasoning brings an extra boost that could not have been achieved with just abductive reasoning. We note in passing that the computational cost of abductive metareasoning is not being considered here. This issue is address more in Section 9.1.

7.3 Domain validation experiments

The following experimental hypotheses are labeled with the format "A-V-#" to indicate that these hypotheses are regarding aerial tracking validation experiments. The experimental hypotheses should be taken implicitly to refer to each of the three datasets equally.

Hypothesis A-V-1: EFLI abduction (Algorithm 2.3) yields substantially greater accuracy than arbitrary abduction. We expect this to be the case because EFLI prefers more plausible, more decisive hypotheses.

Hypothesis A-V-2: According to the discussion of the completeness–confidence trade-off (Figure 2.10), we expect the following outcomes. (1) A minimum plausibility threshold $\eta > 0$ is expected to produce greater accuracy than $\eta = 0$, thus demonstrating the usefulness of this parameter. (2) A decisiveness threshold $\delta > 0$ is expected to produce greater accuracy than $\delta = 0$, thus demonstrating the usefulness of this parameter. However, we also expect that (3) when $\eta > 0$, more anomalies occur because hypotheses are more often rejected due to not meeting the minimum plausibility threshold.

Hypothesis A-V-3: When hypothesis plausibility estimates are hidden (or, equivalently, $Pl(h) = 1.0$ for all hypotheses h), abductive reasoning suffers. This outcome would show that plausibilities are used and useful. However, we also expect that only a

small value for the plausibility precision (Definition 2.2.2), say, precision ≤ 7, is required to produce accuracy on par with very precise plausibilities. This is expected because in cases where two or more competing hypotheses have very small plausibility deltas, not enough information is available to make a confident decision. The same reasoning explains why we expect the decisiveness threshold $\delta > 0$ to yield better accuracy than $\delta = 0$.

In the remainder of this section, each domain validation hypothesis is addressed in turn.

Hypothesis A-V-1

Table 7.2 shows a comparison between EFLI and the arbitrary abduction algorithm, with $\eta = \delta = 0$. EFLI increases Recall but slightly lowers Noise Recall. To see why, notice that EFLI also makes fewer Noise Claims (i.e., leaves fewer reports unexplained). Thus, EFLI is able to explain more, and most of these explanations are true (otherwise Precision would suffer). However, in some rare cases, those explanations were false and explained false reports, thus lowering Noise Recall. In summary, Hypothesis A-V-1 is mostly true, but the case is not as clear as it was in the simulated tracking domain.

Dataset	**Prec.**	**Recall**	**F1**	**N. Prec.**	**N. Recall**	**N. F1**	**N. Claims**
AA_Easy_01	+0.058	+0.150	+0.084	+0.016	-0.014	-0.013	-6
AA_Easy_02	+0.009	+0.101	+0.017	+0.005	-0.028	-0.031	-64
AA_Walking_01	+0.010	+0.038	+0.016	+0.013	-0.027	-0.036	-46

Table 7.2. EFLI vs. arbitrary abduction for different datasets and $\eta = \delta = 0$, supporting Hypothesis A-V-1. A value of 0.000 means that EFLI and arbitrary abduction produced equivalent accuracy. "N. Prec." etc. refer to the Noise Precision metric, etc. A value $+0.058$ for some metric means that EFLI produced 0.058 higher on that metric.

Hypothesis A-V-2

Figure 7.2 shows the impact of η on accuracy for $\delta = 0$. For tracking accuracy (F1, Prec, Recall), we see that $\eta = 0.80$ consistently gives the best results. Noise identification is also nearly best at $\eta = 0.80$, but Noise Recall is greater with even larger η. However, at higher η, Noise Precision suffers. These results match expectations about the impact of η on precision and recall in general.

Accuracy in these datasets, particularly AA_Easy_02 and AA_Walking_01, is very low, in an absolute sense. This might indicate that either (1) abductive reasoning, as implemented, is a poor choice for tracking people in aerial surveillance, or that (2) our hypothesis plausibility function, which is based on movement distance, is a poor choice for these datasets, or (3) both choices are inappropriate. In any event, our goal in this work is not to produce a great aerial tracking system, but to investigate abductive reasoning and abductive metareasoning. Surely, we can build a better tracker for the aerial domain by using Kalman filters (Welch and Bishop, 1995), since the movements of people in these datasets mostly follow smooth trajectories. However, we emphasize again that the goal with this work is not to maximize performance on particular tasks but rather to understand the behavior and trade-offs of abductive reasoning and abductive metareasoning.

Figure 7.3 shows the impact of η on frequencies of errors and anomalies for $\delta = 0$. We see in the top graph that increasing η only slightly increases MinPlausibility errors but decreases Noise errors. Note that Noise errors are the vast majority of errors in the aerial domain. This is because so many reported detections are noise.

The bottom figure shows that MinPlausibility anomalies increase as η increases. This is to be expected. Anomalies due to conflicts among hypotheses are reduced when η increases. This is likely due to the fact that fewer hypotheses are accepted as η increases, thus fewer hypotheses are rejected due to incompatibility with accepted hypotheses. Also

note that the number of NoExplOffered (no explanation offered) anomalies remains constant across the η values. These anomalies are reports that are not near enough to other reports to be considered part of a movement. Refer to Section 7.1 for details.

Figure 7.4 shows the impact of δ on accuracy and noise detection for $\eta = 0.80$. There is a very slight boost in noise identification accuracy for $\delta = 0.10$, and a slight loss in tracking accuracy at the same setting. This is because Recall decreases at $\delta = 0.10$ (and thus Noise Recall increases) since some hypotheses are not accepted that otherwise would have been under $\delta = 0$. These differences are so minor because, in this domain, between 0.98 and 1.02 hypotheses on average (across the various datasets) compete to explain the same report. Because these differences are so minor, however, further experiments will have $\delta = 0$.

Hypothesis A-V-3

The first claim of Hypothesis A-V-3 states that hypothesis plausibility estimates are used and useful. We can simulate abductive reasoning with no information about plausibilities by setting $Pl(h) = 1.0$ for all hypotheses h. In this case, EFLI is unable to prefer hypotheses based on their plausibilities. When a contrast set contains two or more hypotheses, they also cannot be differentiated according to decisiveness. However, essential explainers are still preferred over non-essential explainers. Table 7.3 shows the increase in accuracy when scores are present. As expected, there is an increase, but it is not large because $\eta = 0$ and most reports are noisy.

The second claim of Hypothesis A-V-3 states that only a small value for the plausibility precision (say, precision ≤ 7) is required in order to obtain accuracy on par with full plausibility precision (i.e., double-precision floating-point numbers). Figure 7.5 shows the impact of plausibility precision on accuracy. The impact is not at all consistent and

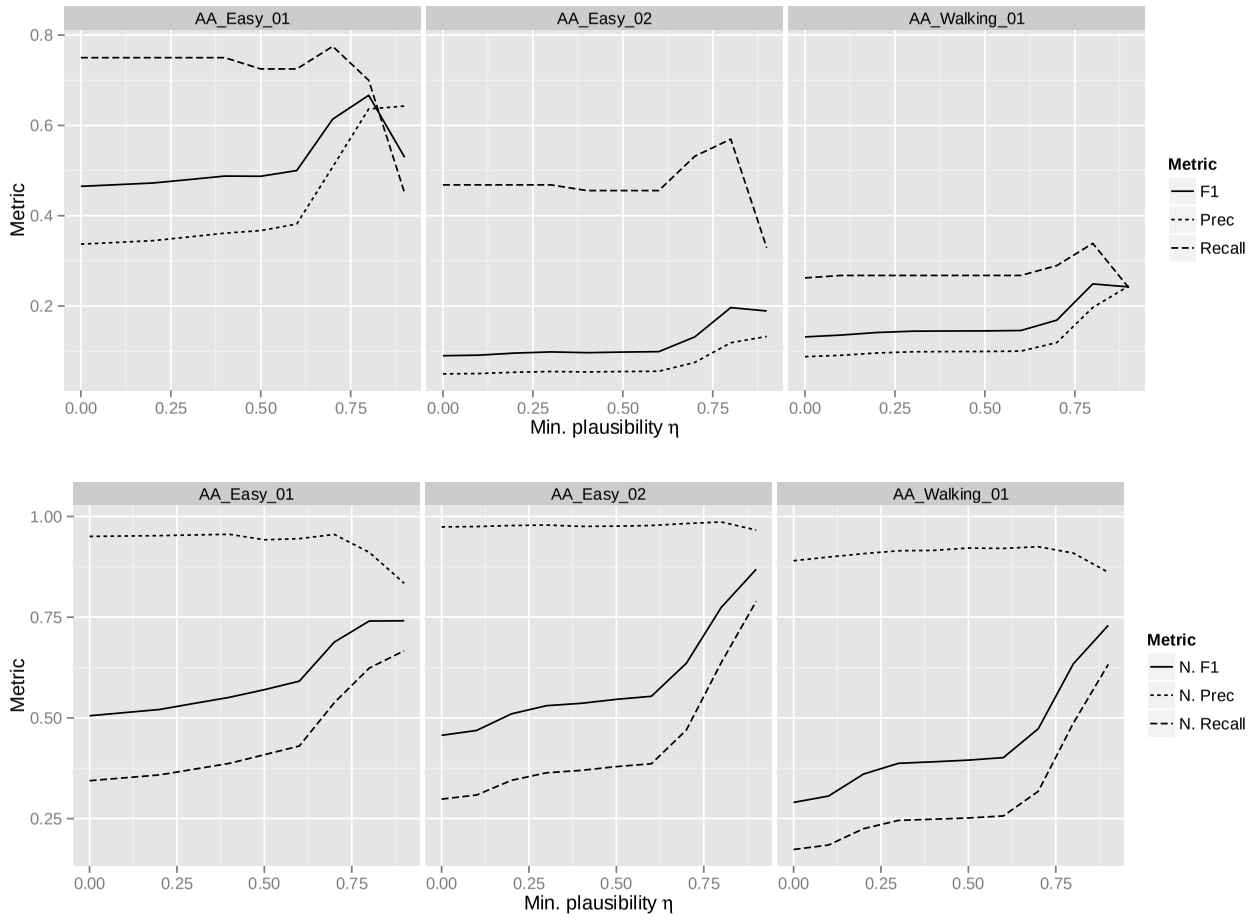


Figure 7.2. Impact of the minimum plausibility threshold η on accuracy and noise identification, with $\delta = 0$. "N. Prec." means "Noise Precision," etc. Supports Hypothesis A-V-2.

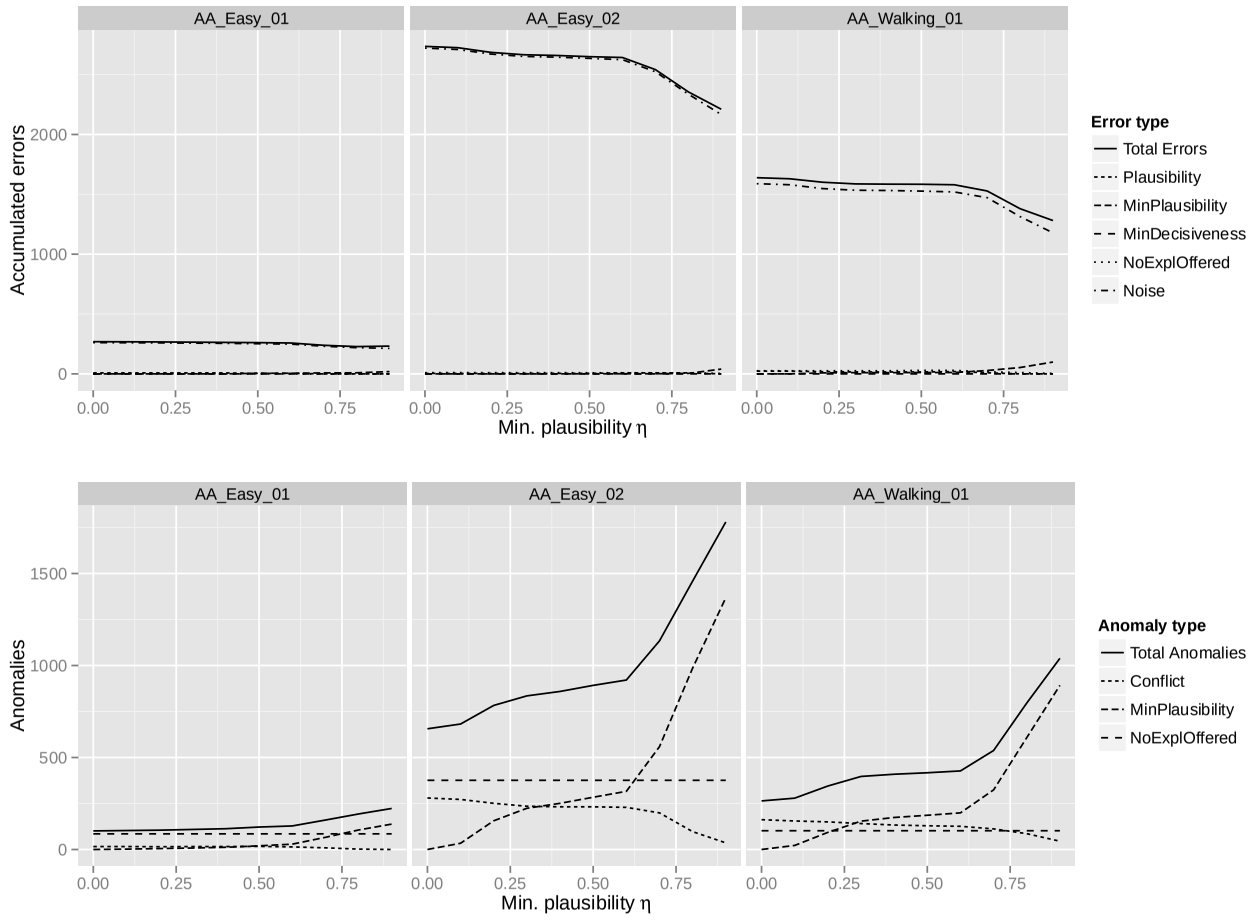


Figure 7.3. Impact of the minimum plausibility threshold η on frequencies of errors and anomalies, with $\delta = 0$. Supports Hypothesis A-V-2.

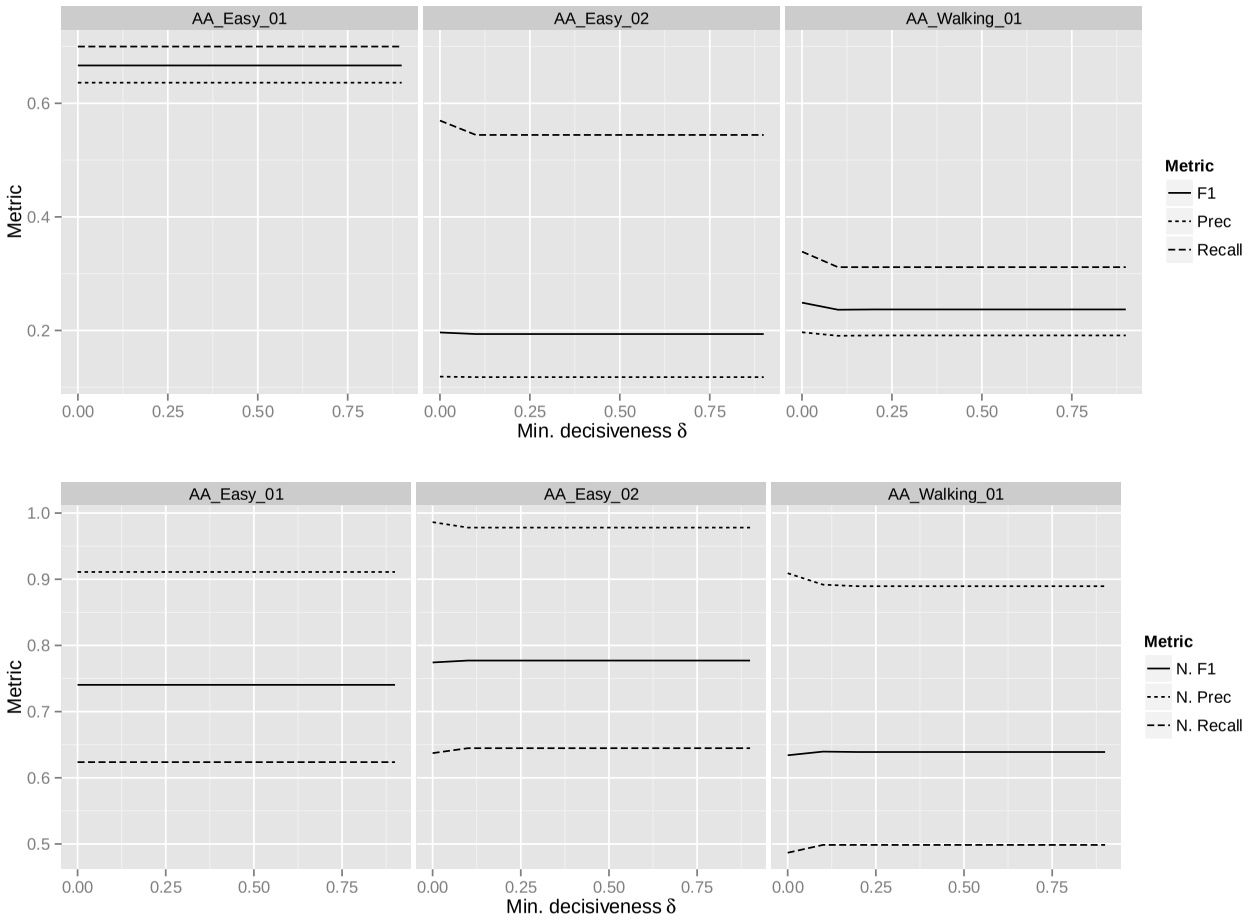


Figure 7.4. Impact of the minimum decisiveness threshold δ on accuracy and noise identification, with $\eta = 0.80$. "N. Prec." means "Noise Precision," etc. Supports Hypothesis A-V-2.

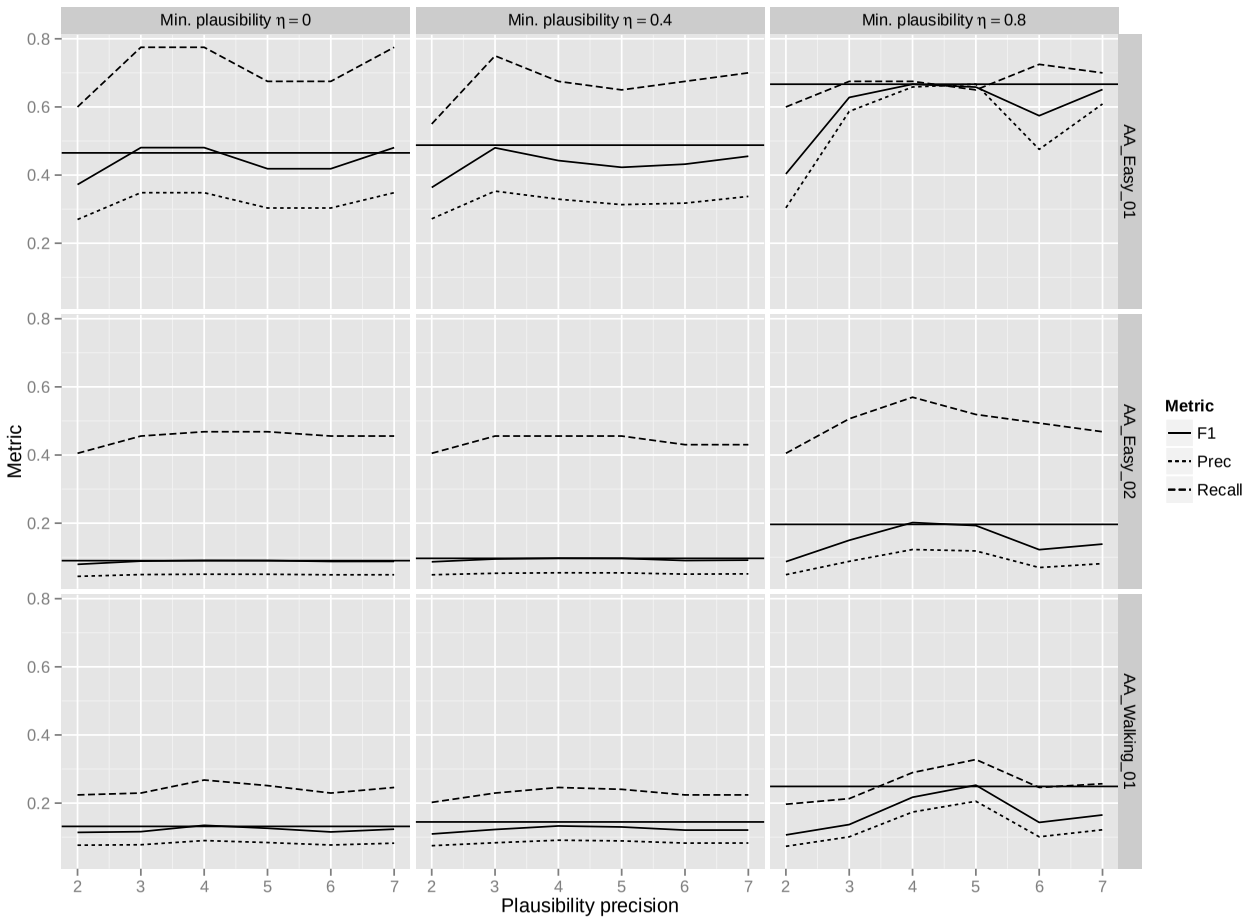


Figure 7.5. Impact of plausibility precision on accuracy for different η values, supporting Hypothesis A-V-3. The horizontal line marks F1 at maximum plausibility precision (double-precision floating-point numbers).

Dataset	Prec.	Recall	F1	N. Prec.	N. Recall	N. F1
AA_Easy_01	+0.056	+0.125	+0.078	+0.010	+0.004	+0.005
AA_Easy_02	+0.005	+0.051	+0.009	-0.001	+0.006	0.000
AA_Walking_01	+0.018	+0.055	+0.027	+0.047	+0.007	+0.012

Table 7.3. Impact of having no plausibility information, supporting Hypothesis A-V-3. $\eta = \delta = 0$. This experiment compares abductive reasoning with no plausibility information and abductive reasoning with normal plausibility information. "N. Prec." etc. refer to the Noise Precision metric, etc. A value $+0.056$ for some metric means that EFLI produced, averaged across individual cases, 0.056 higher on that metric.

we see no obvious trend. These results do not match previous findings in the simulated tracking domain (Figure 6.10). The reason seems to be that, in the aerial domain, hypothesis plausibility estimates generally cover a very small range. Thus, even with plausibility precision is limited, plausibility estimates still cover a similar range. Table 7.4 shows the average plausibility estimates for true and false movement hypotheses under different plausibility precisions. We see that the differences in true and false plausibilities does not vary much across different plausibility precisions. Thus, accuracy does not differ much either, as reflected in the graphs.

The first, but not second, claim of Hypothesis A-V-3 has been confirmed. We have also identified optimal parameters ($\eta = 0.80, \delta = 0$) for EFLI-based abductive reasoning. Next, we will examine how metareasoning can boost accuracy when applied to this domain.

7.4 Metareasoning experiments

The following experimental hypotheses are labeled with the format "A-M-#" to indicate that these hypotheses are regarding aerial tracking metareasoning experiments.

Hypothesis A-M-1: Metareasoning gives better tracking accuracy and noise identification than no metareasoning, for certain values of $\eta, \delta, \eta_{\text{meta}}, \delta_{\text{meta}}$.

Dataset	Plaus. prec.	Avg. true plaus.	Avg. false plaus.
AA_Easy_01	2	0.94	0.89
AA_Easy_01	3	0.90	0.60
AA_Easy_01	4	0.87	0.66
AA_Easy_01	5	0.87	0.71
AA_Easy_01	6	0.87	0.68
AA_Easy_01	7	0.86	0.67
AA_Easy_01	(full)	0.86	0.68
AA_Easy_02	2	0.98	0.79
AA_Easy_02	3	0.97	0.61
AA_Easy_02	4	0.92	0.64
AA_Easy_02	5	0.92	0.67
AA_Easy_02	6	0.90	0.65
AA_Easy_02	7	0.89	0.65
AA_Easy_02	(full)	0.89	0.65
AA_Walking_01	2	0.95	0.81
AA_Walking_01	3	0.90	0.64
AA_Walking_01	4	0.85	0.65
AA_Walking_01	5	0.86	0.68
AA_Walking_01	6	0.84	0.67
AA_Walking_01	7	0.85	0.66
AA_Walking_01	(full)	0.84	0.67

Table 7.4. Average true and false movement hypothesis plausibilities for different datasets and plausibility precision. “(full)” plausibility precision means that the plausibilities are not limited, and plausibilities are represented as full double-precision floating point values.

Folder	Prec.	Recall	F1	N. Prec.	N. Recall	Noise F1
AA_Easy_01	+0.024	+0.125	+0.066	+0.061	0.000	+0.020
AA_Easy_02	+0.004	+0.025	+0.007	+0.002	0.000	+0.001
AA_Walking_01	+0.016	+0.044	+0.025	+0.030	+0.002	+0.005

Table 7.5. Results from comparative experiments, supporting Hypothesis A-M-1, with abductive metareasoning and no metareasoning, for different datasets. $\eta = 0.80$, $\eta_{\text{meta}} = 0.60$, and $\delta = \delta_{\text{meta}} = 0$. A metric value $+0.024$ indicates that abductive metareasoning increased that metric by 0.024 compared to no metareasoning.

Hypothesis A-M-2: Performance is maximized when each of the MetaImplHyp, MetaIncompatHyp, MetaOrderDep meta-hypotheses are available as a possible explainer of anomalies. This can be tested by ablation experiments in which the various subsets of meta-hypotheses are disabled.

In the remainder of this section, each metareasoning experimental hypothesis is addressed in turn.

Hypothesis A-M-1

Figures 7.6 and 7.7 show tracking accuracy and noise identification, respectively, under both no metareasoning and abductive metareasoning and various values of η and η_{meta}. Experiments not reported here indicate that $\delta_{\text{meta}} = 0$ is the optimal value for that parameter, and we established earlier that $\delta = 0$ also yielded best performance. We see from the figures that abductive metareasoning increases tracking accuracy and at $\eta = 0.80$ and $\eta_{\text{meta}} = 0.60$, F1 is maximized (for each dataset). Not surprisingly, Noise F1 is maximized with $\eta = 0.90$ rather than $\eta = 0.80$, because Noise Recall increases since more reports are left unexplained. However, Noise F1 at $\eta = 0.80$ is nearly as good. Additionally, in most cases, abductive metareasoning yields equal or better Noise F1 than no metareasoning.

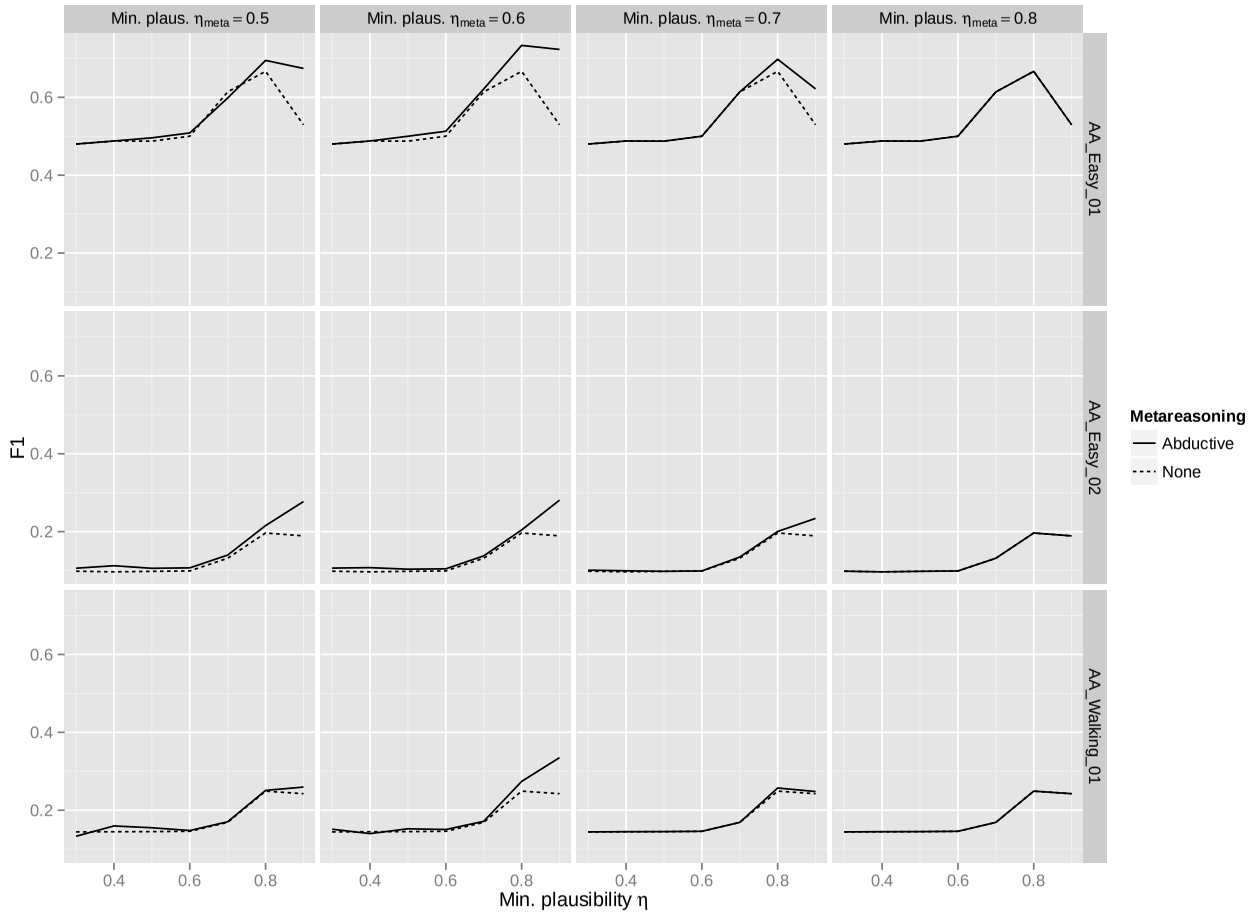


Figure 7.6. Accuracy (F1) for various values of η and η_{meta}. Supports Hypothesis A-M-1. In all cases, $\delta = \delta_{\text{meta}} = 0$.

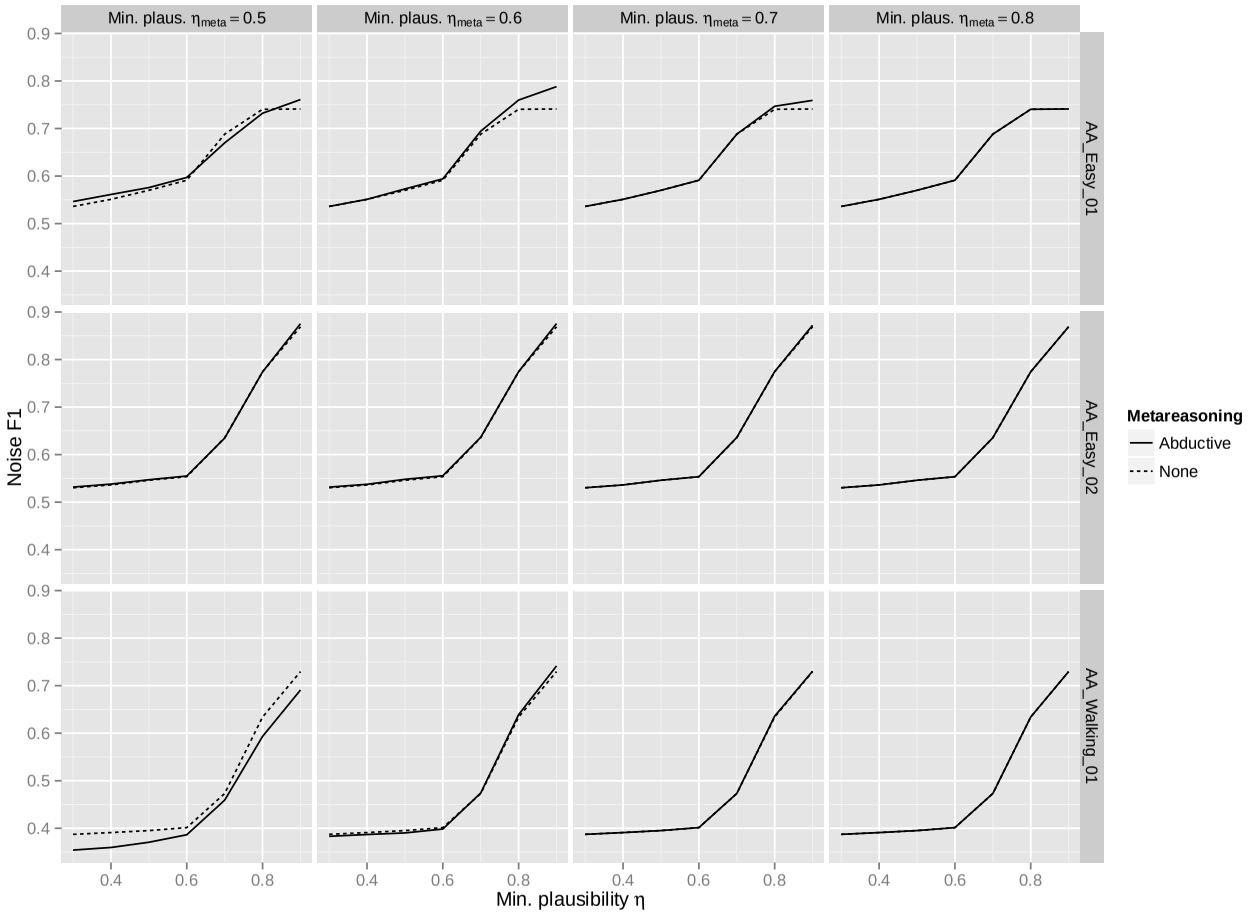


Figure 7.7. Noise identification accuracy (Noise F1) for various values of η and η_{meta}. Supports Hypothesis A-M-1. In all cases, $\delta = \delta_{\text{meta}} = 0$.

Table 7.5 highlights the improvement of abductive metareasoning over no metareasoning at $\eta = 0.80, \eta_{\text{meta}} = 0.60$, which we have identified from Figures 7.6 and 7.7 as the best performance overall. Thus, Hypothesis A-M-1 is confirmed.

Hypothesis A-M-2

Hypothesis A-M-2 states that each kind of meta-hypothesis (MetaImplHyp, MetaIncompatHyp, MetaOrderDep; refer to Section 4.1) plays an important role in abductive metareasoning. Table 7.6 summarizes the results. We see that the AA_Easy_01 dataset benefits from MetaImplHyp but not any other meta-hypothesis. All rows for AA_Easy_01 where MetaImplHyp was ablated are not shown in the table because no meta-hypotheses were accepted (thus tracking accuracy and noise identification were equal to performance with no metareasoning), and the inclusion of MetaIncompatHyp and MetaOrderDep did not change the results as long as MetaImplHyp was available. In fact, MetaIncompatHyp and MetaOrderDep meta-hypotheses were never accepted to explain any anomalies for AA_Easy_01. Thus, we highlighted the row for AA_Easy_01 where only MetaImplHyp was available since the other two types of meta-hypotheses offered no additional benefits.

For AA_Easy_02, best performance was obtained when only MetaIncompatHyp was available. This outcome is quite different than the state of affairs for AA_Easy_01. The difference is small compared to the case where all meta-hypotheses were available. Additionally, no MetaOrderDep was accepted in any case for AA_Easy_02. On all cases, except for the last two rows for AA_Easy_02, including MetaOrderDep did not affect performance. The last two rows compare MetaOrderDep+MetaIncompatHyp and MetaIncompatHyp cases. The accuracy differed even though no MetaOrderDep meta-hypotheses were accepted. This difference is due to the fact that some MetaIncompatHyp meta-hypotheses were accepted in a different order across the two experiments and as a result produced

different answers. The order of acceptance differed because, when MetaOrderDep meta-hypotheses were competing to explain, the decisiveness of at least one MetaIncompatHyp meta-hypotheses was reduced, causing a different MetaIncompatHyp to be accepted first (note that there was no minimum decisiveness threshold, i.e., $\delta_{\text{meta}} = 0$ in these experiments). This outcome demonstrates that the order that meta-hypotheses are accepted (i.e., the order that belief revisions are applied) might affect overall accuracy.

In summary, Table 7.6 shows that the aerial tracking domain does not benefit from all meta-hypotheses in the same way that the simulated tracking domain (mostly) does. For the most part, including all meta-hypotheses was harmless, but the case of AA_Easy_02 shows that including more types of meta-hypotheses might impact decisiveness and thus impact the order that meta-hypotheses are accepted. We have no reason to believe at this time that the impact of this alternative acceptance ordering is always a negative impact; in fact, we only have one such example. MetaOrderDep and MetaIncompatHyp meta-hypotheses were present in some cases of AA_Easy_01 yet accuracy was not affected. In any event, we do not yet have a good understanding of which meta-hypotheses should and should not be included for a particular problem domain, nor do we know exactly how the presence of more or fewer meta-hypotheses affects overall accuracy. We would not be surprised to learn, however, that there are no domain-general answers to these questions.

7.5 Prior work

The producers of the KIT AIS Data Set (Schmidt and Hinz, 2011) evaluated an iterative Bayesian tracking approach that uses information about object position, velocity, and color. They also include information about optical flow throughout the entire scene. Recall that our approach uses a simple distance-based tracker that only considers object position. We chose to do so for the sake of simplicity, to ensure the "power" was in the abductive rea-

Dataset	Impl	Incompat	OD	Prec.	Recall	F1	N. Prec.	N. Recall	N. F1
AA_Easy_01	X	X	X	0.660	0.825	0.733	0.972	0.624	0.760
AA_Easy_01	X	X		0.660	0.825	0.733	0.972	0.624	0.760
AA_Easy_01	X		X	0.660	0.825	0.733	0.972	0.624	0.760
AA_Easy_01	**X**			**0.660**	**0.825**	**0.733**	**0.972**	**0.624**	**0.760**
AA_Easy_02	X	X	X	0.123	0.595	0.204	0.988	0.637	0.775
AA_Easy_02	X	X		0.123	0.595	0.204	0.988	0.637	0.775
AA_Easy_02	X		X	0.118	0.570	0.196	0.988	0.636	0.774
AA_Easy_02	X			0.118	0.570	0.196	0.988	0.636	0.774
AA_Easy_02		X	X	0.124	0.595	0.205	0.987	0.638	0.775
AA_Easy_02		**X**		**0.127**	**0.608**	**0.210**	**0.987**	**0.638**	**0.775**
AA_Walking_01	**X**	**X**	**X**	**0.213**	**0.383**	**0.274**	**0.939**	**0.485**	**0.639**
AA_Walking_01	X	X		0.213	0.383	0.274	0.939	0.485	0.639
AA_Walking_01	X		X	0.204	0.366	0.262	0.929	0.479	0.632
AA_Walking_01	X			0.204	0.366	0.262	0.929	0.479	0.632
AA_Walking_01		X	X	0.210	0.361	0.265	0.919	0.492	0.641
AA_Walking_01		X		0.206	0.355	0.261	0.920	0.493	0.642

Table 7.6. Results from ablation experiments, supporting Hypothesis A-M-2, in which various meta-hypotheses are disabled. Columns "Impl," "Incompat," "OD" refer to presence ("X") or absence of support for each meta-hypothesis: MetaImplHyp, MetaIncompatHyp, MetaOrderDep. Missing rows for certain combinations of meta-hypotheses indicates that no meta-hypotheses were accepted in those cases. Rows with best performance for each dataset are highlighted in bold.

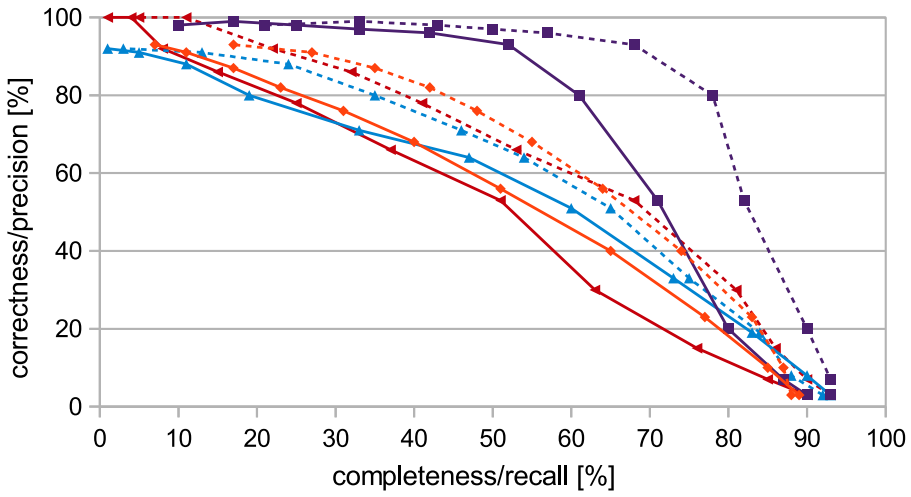


Figure 7.8. Results reproduced from Schmidt and Hinz (2011) showing precision/recall for different datasets and confidence thresholds. Compare our completeness–confidence diagram in Figure 2.10.

soning and metareasoning systems rather than in, say, a Bayesian tracker. They used performance metrics equivalent to our uses of Precision and Recall. They did not attempt to explicitly identify noise. Each possible track found by their algorithm bears a confidence score. By varying a threshold of confidence, they obtained precision/recall graphs (which they also call correctness/completeness graphs) very similar to our completeness–confidence diagram (Figure 2.10). One such graph is reproduced here as Figure 7.8. However, their results and our results are not comparable in any meaningful way since they utilize substantially more information about object movements to track the objects. A critical discussion of our methodology, specifically regarding our decision to model only very simple aspects of each domain, may be found in Section 9.1.

7.6 Conclusions

The aerial tracking domain is the only realistic problem domain that we investigate in this work. We state in Section 10.7 that our planned future work includes experiments with additional realistic problem domains, such as plan recognition, speech recognition, and robotics applications. We learned from the aerial domain that abductive reasoning is suitable for the task, though it might benefit from being able to utilize more information about object velocity, color, and so on, in order to produce more accurate plausibility estimates for the movement hypotheses. We also saw that the η parameter impacts abductive reasoning performance in the ways we would expect. However, we saw that $\delta = 0$ is better than $\delta > 0$. This result is likely due to the dearth of competing explainers found when we set a high η threshold. We demonstrated that abductive metareasoning maximizes accuracy and nearly maximizes noise identification across all tested configurations (both with and without metareasoning) when we choose $\eta = 0.80$ and $\eta_{\text{meta}} = 0.60$. These results confirm that abductive reasoning and abductive metareasoning are a powerful combination.

Chapter 8: Bayesian network domains

Each world estimation task has unique properties. For example, medical diagnosis and accident investigation differ in some ways. Medical diagnosis often involves elaborate tests and diagnosis by attempting to 'repair' the problem with medicines. Accident investigation, on the other hand, typically must make do with whatever evidence was gathered at the scene. Many properties are shared, for example, finding explanations for the evidence, updating a world estimate or doxastic state as evidence is gathered and processed in a particular order, and evaluating whether evidence is factual or not. We have built algorithms that generate random task domains that are represented as Bayesian networks (Pearl, 1988). These algorithms attempt to model the shared properties of medical diagnosis, accident investigation, and other domains while abstracting away from the specifics that differentiate domains.

8.1 Notation

Bayesian networks are used here to represent causal and incompatibility 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 function $\mathscr{S}(\mathbf{X})$ gives all possible state assignments to every variable in $\mathbf{X}$. The set $\overline{\mathbf{x}}$ 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 vari-

able 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. For convenience, we denote the beliefs of a doxastic state as the set of variable states **b**.

Certain states of certain variables may be incompatible with certain states of other variables. A set of states **x** is incompatible with another set **y**, for our purposes, if and only if some $x \in \mathbf{x}$ is incompatible with some $y \in \mathbf{y}$. When **b** does not include both states of an incompatible pair, we say **b** is 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 shown in Table 8.1.

	$\overline{x},\overline{y}$	$\overline{x},y$	$x,\overline{y}$	x,y
c_{xy}	1.0	1.0	1.0	0.0
$\overline{c_{xy}}$	0.0	0.0	0.0	1.0

Table 8.1. Conditional probability table for a constraint variable c_{xy} with parents x and y. Note that c_{xy} is always "observed."

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 et al. (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 (ancestors) is less likely than if x and y were not incompatible states.

A Bayesian network is randomly generated in order to produce a generic causal domain with which to perform abductive reasoning and metareasoning. The algorithms guiding this generation are detailed in Section 8.3. The variables and their states, and indeed the incompatibility relationships, have no "meaning" as they do not represent any particular knowledge base. However, our experiments with randomly-generated Bayesian networks are intended to shed light on how abductive reasoning and metareasoning perform in a variety of task domains.

8.2 Definition as an abductive reasoning problem

Unlike the tracking domains, in the Bayesian network domains we must first define what constitutes the *ground truth* for a particular network. This ground truth is to be inferred by the reasoning system. It is generated by taking a direct sample of the whole network, 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 become reports. The reasoning system's goal is to explain the reports. Note that the same Bayesian network may yield different ground truth configurations. Figure 8.2 shows an example Bayesian network and an example of its ground truth.

Furthermore, we say that a belief $X = x$ requires explanation if it has parents in the network. Possible explanations range across the various combinations of states of its parents. More formally,

Definition 8.2.1. Suppose X is believed (or observed) to have state x. Then a *possible*

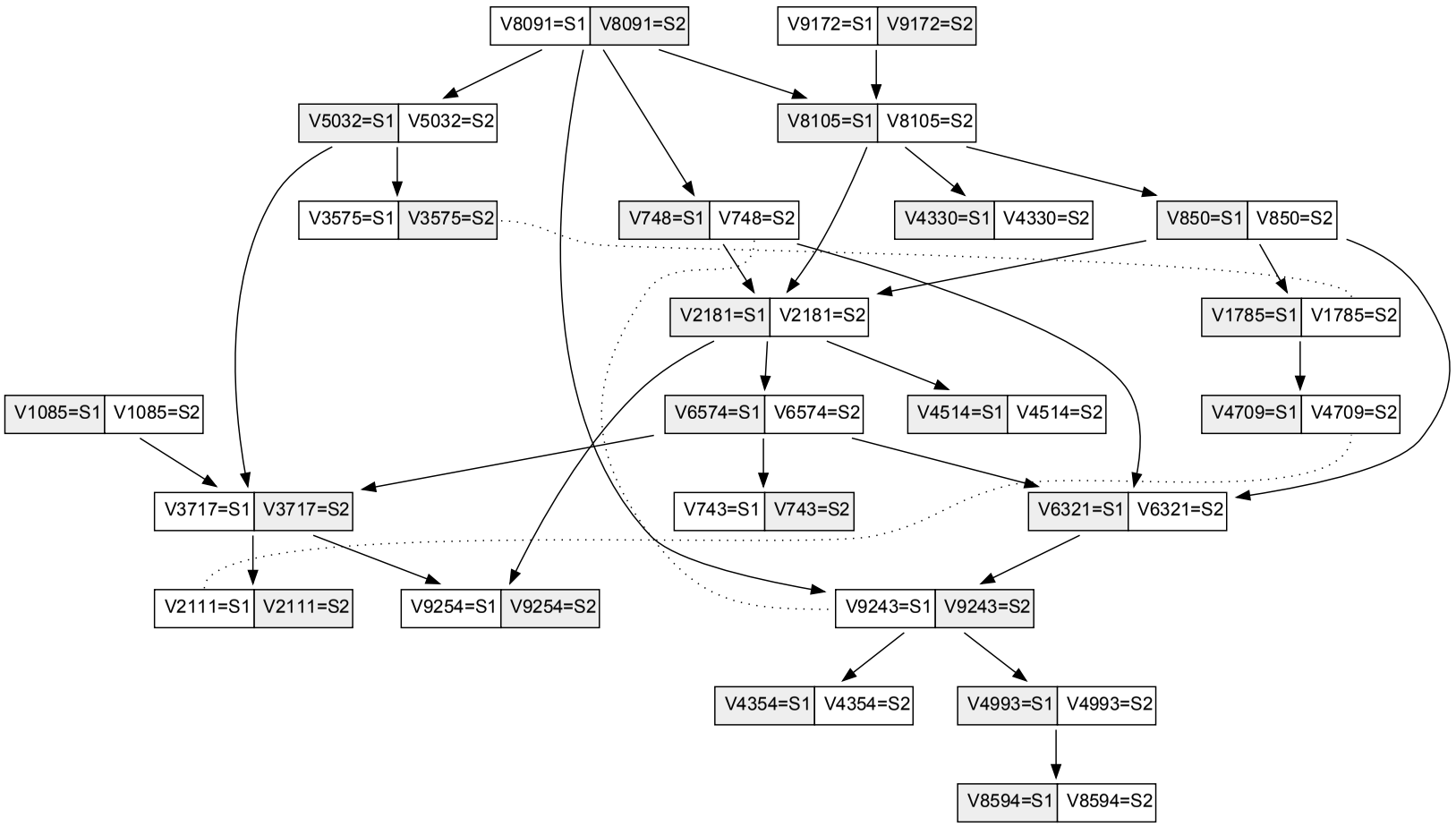


Figure 8.1. An example network randomly generated for the Bayesian network domain. Each vertex can have two distinct states. Arrows represent causal, i.e., explanatory relationships. Dotted-lines represent incompatibilities among two states; it cannot be the case that both states in an incompatible pair are true. A sampling of the Bayesian network for a particular simulation is shown with gray vertex states. The sampling will always respect incompatibility relationships.

explanation of x is a partial instantiation $\mathbf{y} \subseteq \mathbf{Y}, \mathbf{y} \neq \emptyset$ of parents $\mathbf{Y}$ of X such that $\mathbf{b} \cup \mathbf{y}$ is consistent.

As an example, referring to Figure 8.2, if V9243=S2 were believed, then its possible explanations are:

V8091=S1	V8091=S2
V6321=S1	V6321=S2
V8091=S1, V6321=S1	V8091=S1, V6321=S2
V8091=S2, V6321=S2	V8091=S2, V6321=S2

Supposing V8091=S2, V6321=S2 were accepted, then both V8091=S2 and V6321=S2 would require explanation. The plausibility of a possible explanation $\mathbf{y}$ is simply the posterior (where $\setminus$ denotes set difference):

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

8.3 Network generation algorithms

The four algorithms below define how random networks are generated. The entry-point is Algorithm 8.1. The function PICKRANDOM(S,n) is referred to in several instances, but never defined. This function simply picks n random elements from set S.

Algorithm 8.1 Algorithm for generating a random network.

```
function RANDOMNETWORK
    Reset random seed to a random integer in [0, 1000]
    E ← GENERATEEXPLAINSLINKS                    ▷ See Algorithm 8.2
    I ← ADDINCOMPATIBILITIES(E)                  ▷ See Algorithm 8.4
    P ← ASSIGNRANDOMPROBABILITIES(E)             ▷ See Algorithm 8.5
    return (E, I, P)
end function
```

Algorithm 8.2 Algorithm for generating edges.

function GENERATEEXPLAINSLINKS
  $\mathbf{U} \leftarrow$ Generate 10 random variables ▷ $\mathbf{U}$ contains variables with no parents
  $\mathbf{P} \leftarrow \{\}$ ▷ $\mathbf{P}$ contains generated parent variables
  $E \leftarrow \{\}$ ▷ E contains parent-child edges
  while $\|E\| < 40$ **do**
    $E' \leftarrow \bigcup_{V \in \mathbf{U}} \text{GENERATEPARENTEDGES}(V, \mathbf{P})$ ▷ See Algorithm 8.3
    $E \leftarrow E \cup E'$
    $\mathbf{V}_{\text{exp}} \leftarrow \{V | (V', V) \in E'\}$ ▷ Find the newly explained
    $\mathbf{V}_{\text{unexp}} \leftarrow \{V' | (V', V) \in E'\}$ ▷ Find the newly unexplained
    $\mathbf{U} \leftarrow (\mathbf{U} \setminus \mathbf{V}_{\text{exp}}) \cup \mathbf{V}_{\text{unexp}}$
    $\mathbf{P} \leftarrow \mathbf{P} \cup \mathbf{V}_{\text{unexp}}$
  end while
  return E
end function

Algorithm 8.3 Algorithm for generating parents for some variable.

function GENERATEPARENTEDGES($V, \mathbf{P}$)
  $\hat{V} \leftarrow$ Generate a single random variable
  $\mathbf{P}' \leftarrow \text{PICKRANDOM}(\mathbf{P}, 10) \cup \{\hat{V}\}$
  return $\{(V', V) | V' \in \text{PICKRANDOM}(\mathbf{P}', 3)\}$
end function

Algorithm 8.4 Algorithm for generating incompatible pairs.

function ADDINCOMPATIBILITIES(E)
  $\mathbf{V} \leftarrow \{V | (V', V) \in E\} \cup \{V' | (V', V) \in E\}$ ▷ Get all variables
  $\mathbf{v} \leftarrow \{(V, s) | V \in \mathbf{V}\} \cup \{(V, \overline{s}) | V \in \mathbf{V}\}$ ▷ Get all variable–state pairs (states are $s, \overline{s}$)
  $C \leftarrow \{(V', v', V, v) | (V', v') \in \mathbf{v}, (V, v) \in \mathbf{V}\}$ ▷ Get all combinations of variable–state pairs
  $I \leftarrow \{(V', v', V, v) | (V', v', V, v) \in C \wedge v \neq v' \wedge (V, V') \notin E \wedge (V', V) \notin E\}$ ▷ Keep valid pairs
  return $\text{PICKRANDOM}(I, 10)$ ▷ Only keep at most 10 incompatible pairs
end function

Algorithm 8.5 Algorithm for assigning probabilities.

```
function AssignRandomProbabilities(E)
    Establish structure P to record probabilities
    V ← {V|(V′,V) ∈ E} ∪ {V′|(V′,V) ∈ E}                    ▷ Get all variables
    for all V ∈ V do
        P ← {P|(P,V) ∈ E}                                   ▷ Get parents of V
        𝒫 ← All combinations of parents
        for all P′ ∈ 𝒫 do          ▷ Generate probabilities for each parent combination
            p ← uniform-random(0,1)
            Record probability P(V = v|P′) = p into structure P
        end for
    end for
    return P
end function
```

8.4 Noise

Over the course of an experiment, the OBSERVE function randomly picks 10 true variable states from anywhere in the network and generates reports for these variable states. 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 include four types of noise.

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

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

Insertion noise: A random unobserved variable Y, with true state y, is reported to have state $\overline{y}$.

Deletion noise: A true report is simply deleted from the set of variable states that would have otherwise been reported.

Note that in each case excepting *deletion noise*, the reasoning system obtains a false report. If the system ultimately refuses to find or accept an explanation for such a false report, we say that it has identified the noisy report. However, deletion noise cannot be

identified in this manner (there is no report to leave unexplained), so deletion noise is not represented in noise identification metrics.

Noise may produce anomalies. In Section 8.6 we experimentally measure how noise influences the presence of anomalies.

8.5 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 1000 randomly generated networks, the networks have 22.7 variables (excluding constraint variables), 1.5 parent variables, a depth of 4.5, and contains 9.8 incompatible variable state pairs.

At each time step, a small number (≤ 3) of variable states are reported (these reports might contain noise) from the OBSERVE function (see Section 2.6). The GENERATEHYPOTHESES function is called to generate new hypotheses for the reports or other believed variable states that themselves need explanation. After each hypothesis is accepted, GENERATEHYPOTHESES is called again to ensure the hypothesis plausibility estimates, which are posteriors given the current beliefs, are continually up-to-date.

We measure performance according to the following metrics. Let $\mathbf{r}$ be reported variable states (including noisy reports), $\mathbf{B}$ be believed variables (not including reports), $\mathbf{b}$ be the believed variable states, $\mathbf{E} \subseteq (\mathbf{B} \cup \mathbf{r})$ be reports and believed variables that require explanation, $\mathbf{U} \subseteq \mathbf{E}$ be those that are unexplained, and $\mathbf{t}$ be the ground truth variable states. $\|X\|$ represents the cardinality of the set X, and $P(\cdot)$ is the normal probability function.

$$
\begin{aligned}
\text{Accuracy} &= \frac{\|\mathbf{t} \cap \mathbf{b}\|}{\|\mathbf{b}\|} \\
\text{Coverage} &= 1.0 - \frac{\|\mathbf{U}\|}{\|\mathbf{E}\|} \\
\text{AccCov} &= \frac{2 * \text{Accuracy} * \text{Coverage}}{\text{Accuracy} + \text{Coverage}} \\
\text{MPE} &= \max_{\mathbf{v} \in \mathscr{S}(\mathbf{B})} P(\mathbf{v}|\mathbf{r}) \\
\text{MPEAccuracy} &= \frac{\|\mathbf{t} \cap \text{MPE}\|}{\|\text{MPE}\|}
\end{aligned}
$$

Note that MPE means "most probable explanation," and here we define it as the most probable assignment of states $\mathbf{v} \in \mathscr{S}(\mathbf{B})$ to believed variables $\mathbf{B}$, given the reported states (which might include noise). It is possible that MPE $\neq \mathbf{b}$, i.e., the most probable variable states for believed variables might not equal the actual believed variable states (arrived at via abductive reasoning/metareasoning). Additionally, as in the simulated and aerial tracking domains, we measure noise identification accuracy according to the following metrics.

$$
\begin{aligned}
\text{Noise Precision} &= \frac{\|\text{Actual noisy reports} \cap \text{Noise claims}\|}{\|\text{Noise claims}\|} \\
\text{Noise Recall} &= \frac{\|\text{Actual noisy reports} \cap \text{Noise claims}\|}{\|\text{Actual noisy reports}\|} \\
\text{Noise F1} &= \frac{2 * \text{Noise Precision} * \text{Noise Recall}}{\text{Noise Precision} + \text{Noise Recall}}
\end{aligned}
$$

8.6 Domain validation experiments

The following experimental hypotheses are labeled with the format "B-V-#" to indicate that these hypotheses are regarding the Bayesian network domains' validation experiments.

Hypothesis B-V-1: EFLI abduction (Algorithm 2.3) yields significantly greater accuracy than arbitrary abduction. We would expect this to be the case because EFLI prefers more plausible, more decisive hypotheses.

Hypothesis B-V-2: As the noise level increases, anomalies also increase. This is to be expected because false reports should, in the usual case, have no plausible and consistent explanation. Accuracy should decrease, because noisy reports for which explanations are found usually have false explanations.

Hypothesis B-V-3: According to the discussion of the completeness–confidence trade-off (Section 2.8), we expect the following outcomes. (1) A minimum plausibility threshold $\eta > 0$ is expected to produce greater accuracy than $\eta = 0$, thus demonstrating the usefulness of this parameter. (2) A decisiveness threshold $\delta > 0$ is expected to produce greater accuracy than $\delta = 0$, thus demonstrating the usefulness of this parameter. However, we also expect that (3) when $\eta > 0$, more anomalies occur because hypotheses are more often rejected for not meeting the minimum plausibility threshold.

Hypothesis B-V-4: In noise-free scenarios, we expect that the MPE is no less accurate than beliefs acquired via abductive reasoning. This is expected because the MPE is provably the most probable configuration of variable states given the reports (which are true reports, by assumption). However, in noisy scenarios, in which some reports might be false, abductive reasoning should yield more accurate beliefs than the MPE. This is expected because the MPE calculates the most probable explanation "all at once," taking the noisy reports to be true; while the abductive reasoning system that we have described is iterative, explaining one report or belief at a time and regenerating hypotheses and re-estimating their plausibilities after each acceptance. The

impact of noise is more transient with abduction, and the noise is effectively "washed out" as explainers at higher levels in the network are considered.

Hypothesis B-V-5: When plausibility estimates are missing or, equivalently, $Pl(\mathbf{y}) = 1.0$ for every possible explanation $\mathbf{y}$, performance is significantly and strongly degraded. This outcome would show that performance depends on information about probabilities and not just network structure. Furthermore, we expect that only a small value for the plausibility precision (Definition 2.2.2), say, precision ≤ 7, is required to produce accuracy on par with very precise plausibilities. This is expected because in cases where two or more competing hypotheses have very small plausibility deltas, not enough information is available to make a confident decision. The same reasoning explains why we expect the decisiveness threshold $\delta > 0$ to yield better accuracy than $\delta = 0$.

Hypothesis B-V-1

At $\eta = \delta = 0$ and no noise, the Accuracy metric increases by 0.095 on average ($p < 0.001$) when EFLI abductive reasoning is employed compared to arbitrary abductive reasoning. Coverage is not affected because $\eta = \delta = 0$, so no hypotheses are rejected due to not meeting minimum plausibility or not accepted due to inadequate decisiveness. So Hypothesis B-V-1 is confirmed.

Hypothesis B-V-2

Figure 8.2 shows how the noise level affects the occurrence of anomalies in the Bayesian network domains. Interestingly, the impact of noise on anomalies is minimal and not as expected. The reason seems to be that false reports typically still have plausible explainers: even for false reports, there still exists some plausible combination of parent variable states.

But these explainers for the false reports are often false. We see this in Figure 8.3, although again the impact is minor. The reason might be that a false observation in a Bayesian network does not strongly impact the rest of the network. It is possible that the network structure influences this result. Perhaps networks with greater connectedness are better able to "absorb" random false reports.

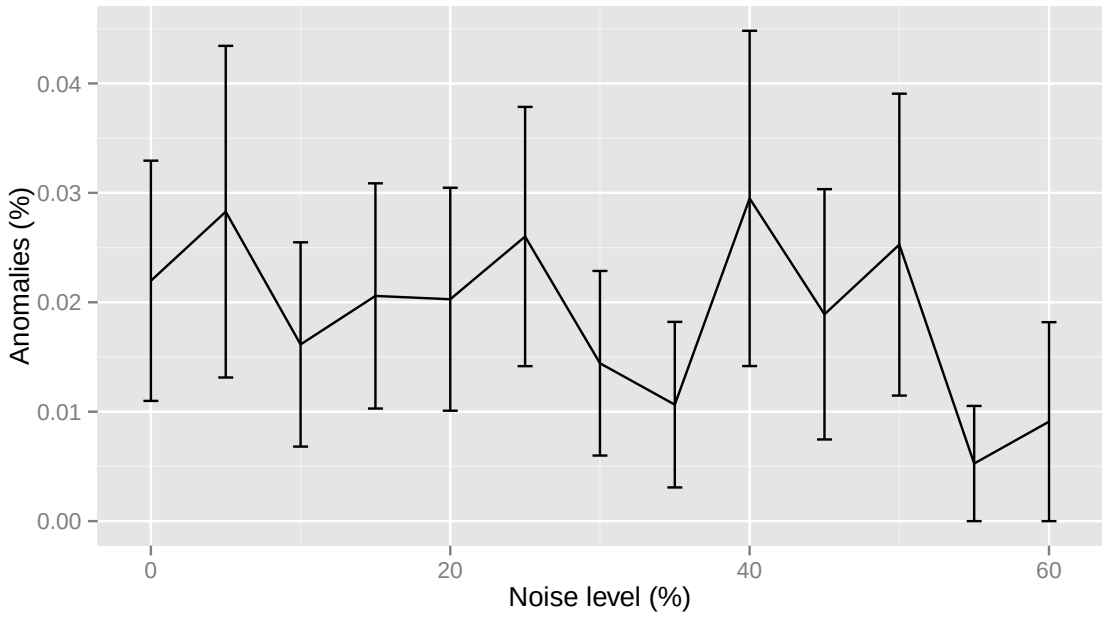


Figure 8.2. Impact of noise level on anomalies, supporting Hypothesis B-V-2. $\eta = \delta = 0$.

Hypothesis B-V-3

Figure 8.4 shows the impact of δ on accuracy and noise detection for different values of η. Counter to our expectations, $\delta = 0$ yields the best Accuracy/Coverage trade-off represented by the AccCov metric. Though these results do show that Accuracy improves with $\delta = 0.20$, there is a strong loss of Coverage. In the Bayesian networks, plausibilities

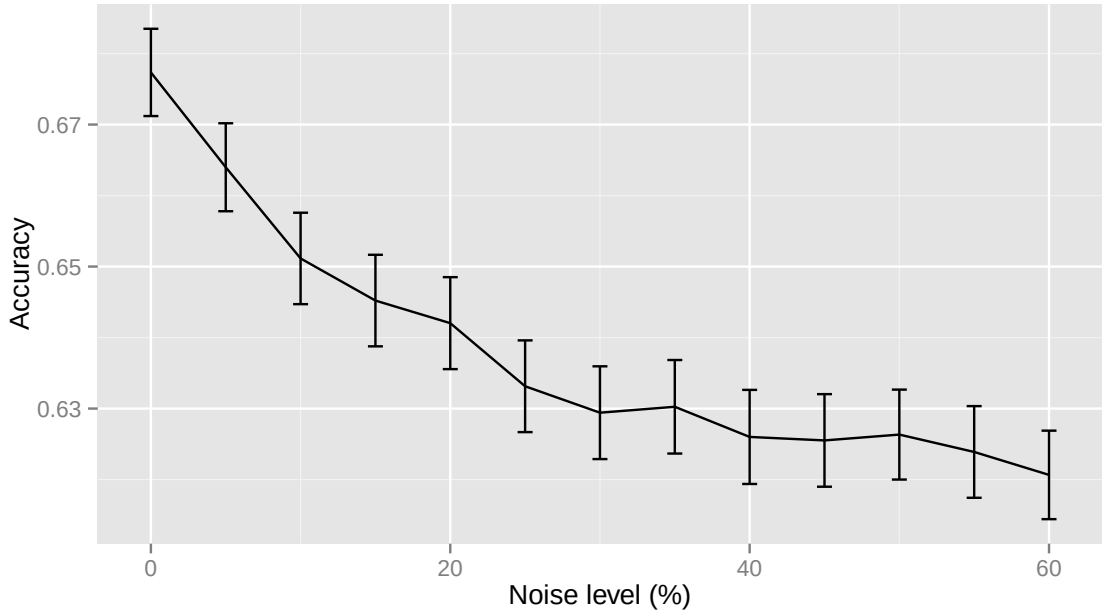


Figure 8.3. Impact of noise level on accuracy, relating to Hypothesis B-V-2. $\eta = 0 = \delta = 0$.

of competing hypotheses are often close, so even a small minimum decisiveness threshold causes many reports and other beliefs to remain unexplained, which reduces Coverage.

Noise identification improves with greater δ, apparently because noisy reports generally do not have decisive explanations. Noise identification also increases with greater minimum plausibility η threshold, since noisy reports also generally have less-plausible possible explanations. However, Accuracy and Coverage are reduced with a large η threshold.

The right choices of δ and η depend on whether one wishes to increase Accuracy and Coverage or increase Noise F1, since there appears to be a trade-off between these metrics. We will often vary η to see how abductive metareasoning performs under different choices in this trade-off. Recall that $\eta > 0$ might produce anomalies. Since $\delta > 0$ does not produce

anomalies, there are no insights to gain about metareasoning by varying δ. We will leave $\delta = 0$ unless stated otherwise.

Figure 8.5 shows the impact of η on frequencies of errors and anomalies. Note that $\delta = 0$. We see in the top graph that increasing η produces more MinPlausibility errors but fewer of each other kind of error, especially Plausibility errors. NoExplOffered errors are virtually non-existent because the network structure almost always provides possible explanations; only about 20 cases among the 3000 summarized in these graphs had just one or two occurrences of NoExplOffered errors. These errors only manifest when believed variable states in the network are incompatible with all possible explanations of some report or belief. Finally, MinDecisiveness errors are not possible since $\delta = 0$.

The bottom figure shows that MinPlausibility anomalies increase as η increases. They account for virtually all cases of anomalies in the Bayesian network domains. However, there do exist a handful of cases in which anomalies are caused by Conflict or NoExplOffered.

Hypothesis B-V-4

Hypothesis B-V-4 states that, in noise-free scenarios, we expect that the MPE is no less accurate than beliefs acquired via abductive reasoning (that is to say, EFLI-based abductive reasoning). On the other hand, in noisy scenarios, we expect the MPE to be less accurate than beliefs acquired by abductive reasoning since the MPE does not iteratively build its beliefs. EFLI-based abductive reasoning accepts one hypothesis (vertex–value pair) at a time, then generates new hypotheses and re-estimates the plausibilities of existing hypotheses. Noise is better handled in this way because hypotheses further separated from the noisy reports are less affected by the noise in terms of their plausibility estimates. The MPE, on the other hand, considers the probabilities of all the relevant vertices at once, so

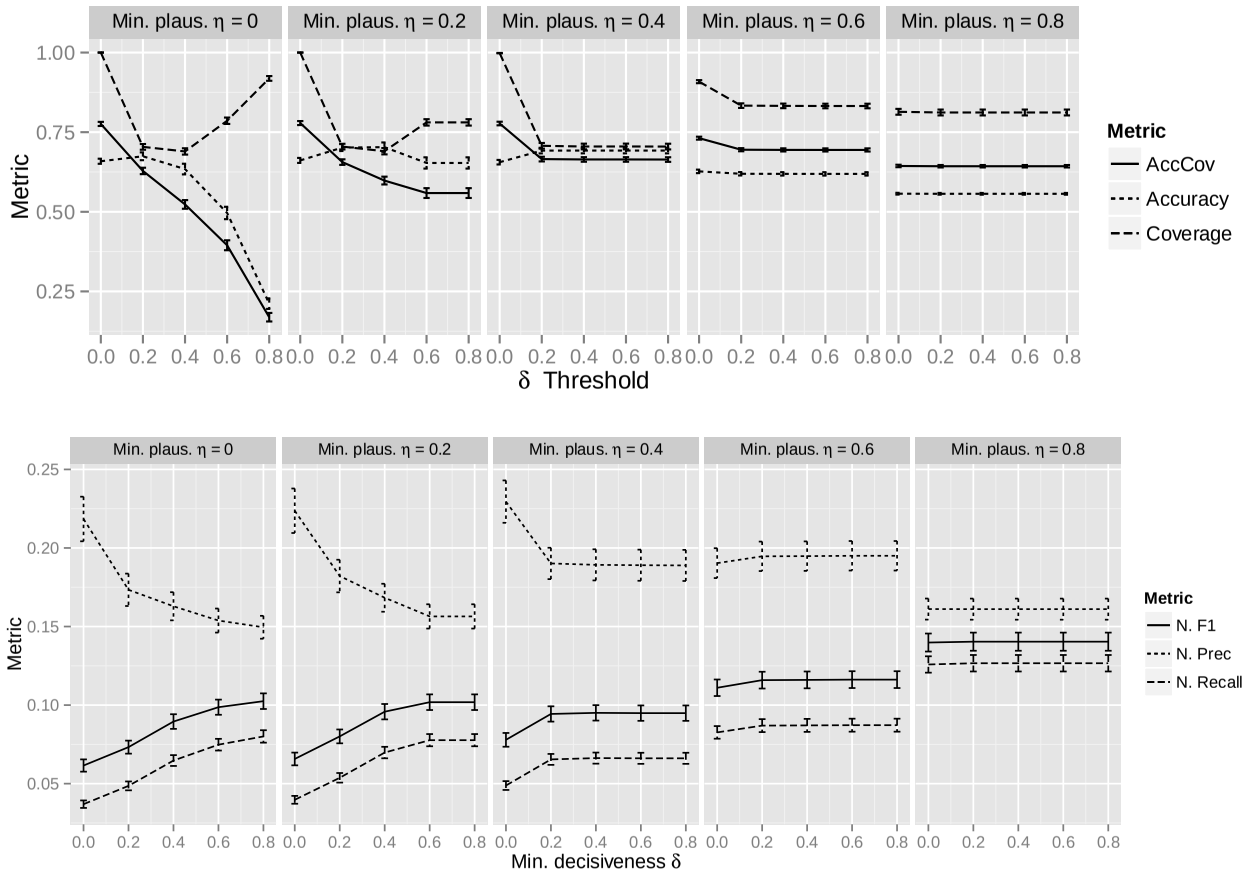


Figure 8.4. Impact of the decisiveness threshold δ on accuracy and noise identification for several values of the minimum plausibility η, and averaged across noise level values 0%, 10%, and 20%. "N. Prec." means "Noise Precision," etc. Supports Hypothesis B-V-3.

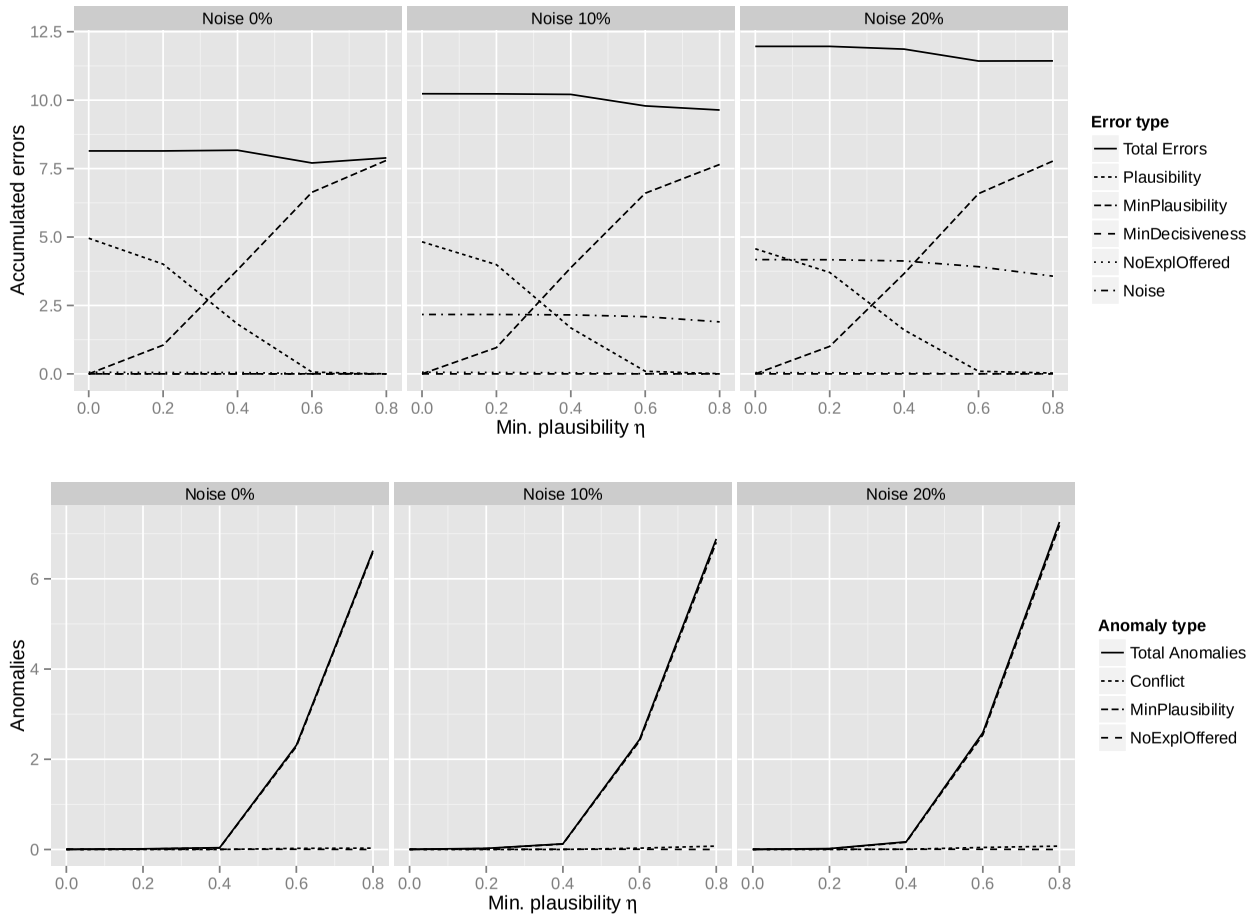


Figure 8.5. Impact of minimum plausibility η on frequencies of errors and anomalies, with $\delta = 0$, and averaged across noise level values 0%, 10%, and 20%. Supports Hypothesis B-V-3.

the noisy reports are never "washed out" as they are by the iterative abduction process.

The MPE is measured here only for those vertices that our abduction system has beliefs about (those vertices that have states that were accepted in order to explain). In this usage, the MPE is synonymous with the MAP (*maximum a posteriori*) for the subset of vertices that have believed states from our abductive reasoning process. Thus, there is no difference in Coverage or number of vertex–value pairs that make up the beliefs found by our abduction process and vertex–value pairs found by the MPE.

Figures 8.6, 8.7, and 8.8 compare Accuracy and MPEAccuracy for different kinds of noise at various noise levels. Arbitrary and EFLI abduction are separated to indicate that EFLI is generally responsible for the dominance of Accuracy over MPEAccuracy at higher noise levels.

In the case of Figure 8.6, we see two interesting phenomena. First, arbitrary abduction is not greatly affected by distortion noise. Recall that distortion noise means randomly modifying a reported vertex value to the vertex's other value. The same vertex is reported, so the same parent nodes are examined as possible explainers. Only the plausibility estimates of the possible explainers are impacted by distortion noise. Arbitrary abduction ignores the plausibility estimates of hypotheses, so distortion noise has no impact on accuracy with arbitrary abduction. The MPE is directly affected by distortion noise because the vertex values affect the posteriors of all other vertices in the network, and thus impact the MPE.

Duplication noise is shown in Figure 8.7. Duplication noise means reporting both vertex values. In this case, the MPE only accepts a random choice for the duplicate reports, since it cannot accept both; thus, the MPE behaves in a similar manner as in distortion noise, just there is a 50% chance it accepts the correct value of the pair under duplication noise.

More interestingly, EFLI-based abduction performs better than MPE for higher dis-

tortion and duplication noise levels. The reason is as described above: that the iterative abductive reasoning algorithm effectively "washes out" the impact of the noisy reports because new hypotheses are generated or plausibility estimates are recalculated after each hypothesis is accepted. The MPE, on the other hand, considers the probability of an entire composite explanation all at once, so the noise is never "washed out." Insertion noise, as shown in Figure 8.8, does not strongly affect the accuracy of either abduction or the MPE.

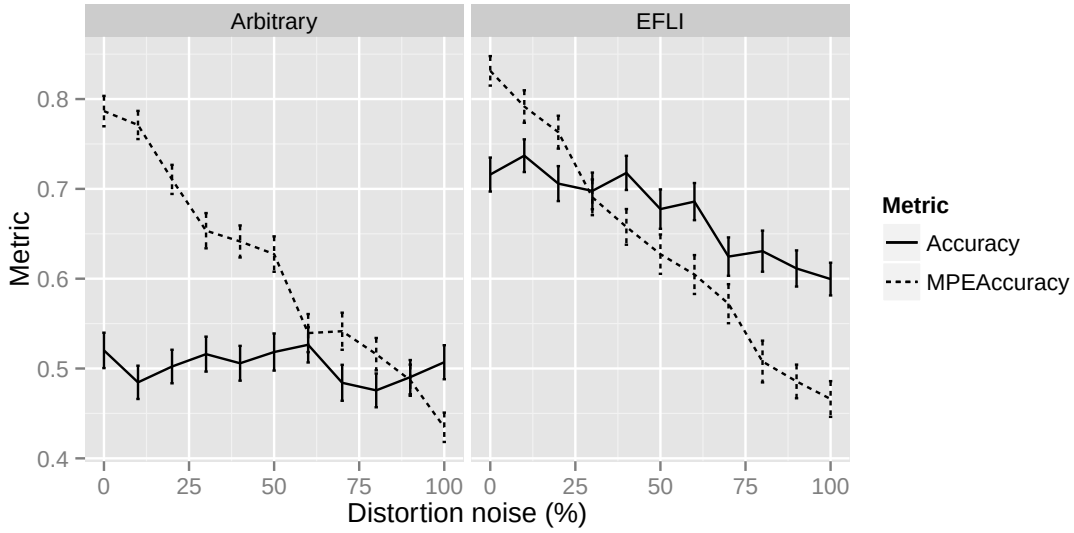


Figure 8.6. Comparison of Accuracy and MPEAccuracy for different levels of distortion noise, supporting Hypothesis B-V-4. $\eta = \delta = 0$ and reports were obtained all at once rather than sequentially. Arbitrary contrast set preference and EFLI contrast set preference are distinguished.

Hypothesis B-V-5

The first claim of Hypothesis B-V-5 states that hypothesis plausibility estimates are used and useful. We can simulate abductive reasoning with no information about plausibilities by setting $Pl(\mathbf{y}) = 1.0$ for all possible explanations $\mathbf{y}$. In this case, EFLI is unable to prefer

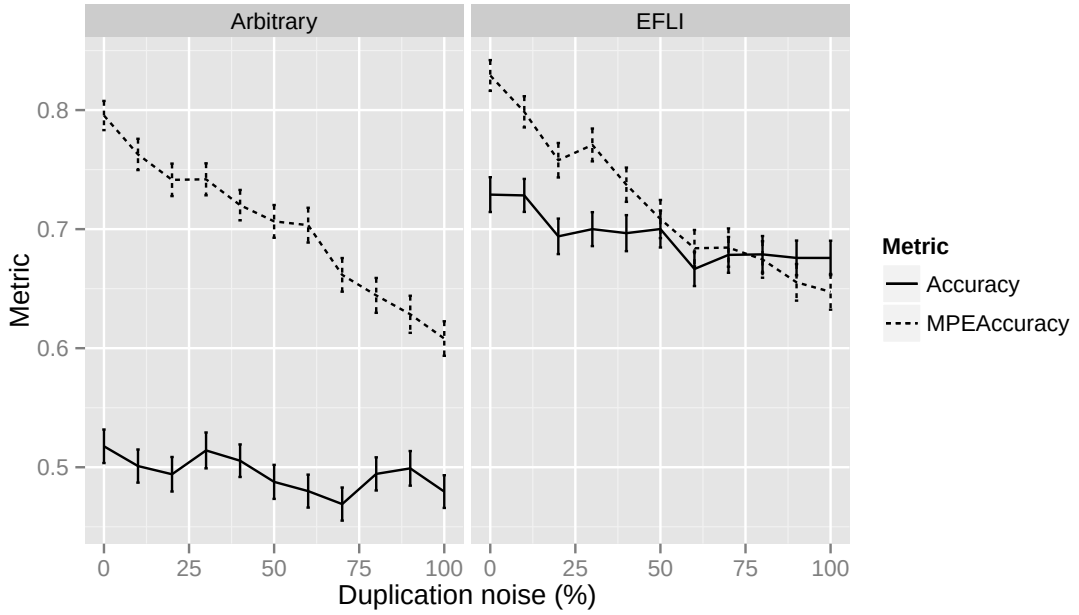


Figure 8.7. Comparison of Accuracy and MPEAccuracy for different levels of duplication noise, supporting Hypothesis B-V-4. $\eta = \delta = 0$ and reports were obtained all at once rather than sequentially. Arbitrary contrast set preference and EFLI contrast set preference are distinguished.

hypotheses based on their plausibilities. When a contrast set contains two or more hypotheses, they also cannot be differentiated according to decisiveness. However, essential explainers are still preferred over non-essential explainers. Table 8.2 shows the increase in accuracy when scores are present. Accuracy and noise identification are significantly and strongly increased. Coverage is not affected since $\eta = \delta = 0$ and Coverage$= 1.0$ in all cases.

The second claim of Hypothesis B-V-5 states that only a small value for the plausibility precision (say, precision ≤ 7) is required to obtain accuracy on par with full plausibility precision (i.e., double-precision floating-point numbers). Figure 8.9 shows the impact of plausibility precision on Accuracy. We see that when $\eta = 0$, plausibility precision at about

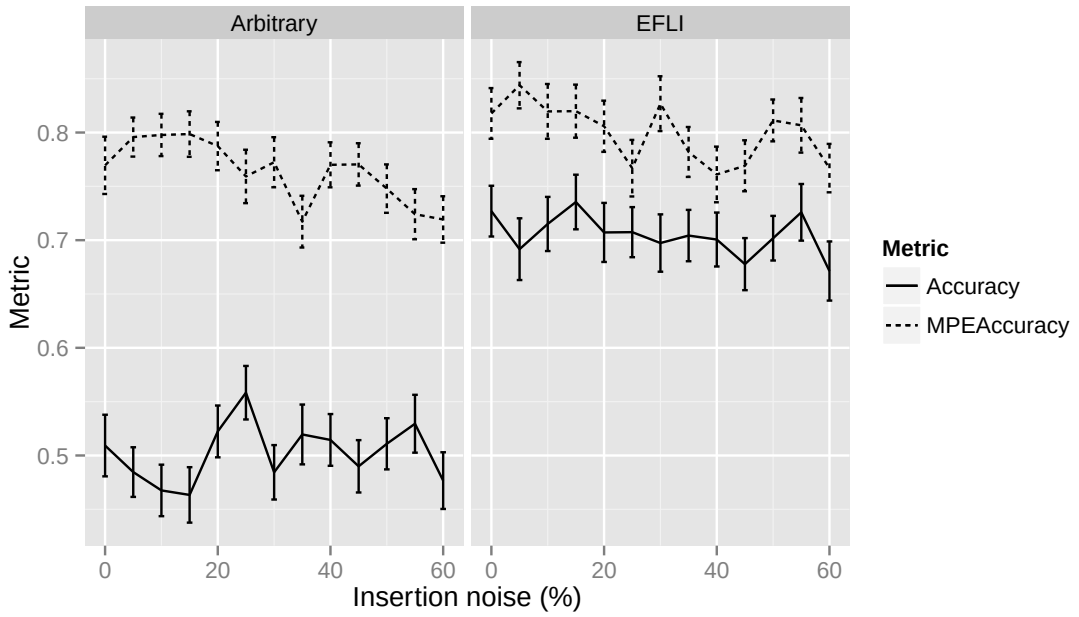


Figure 8.8. Comparison of Accuracy and MPEAccuracy for different levels of insertion noise, supporting Hypothesis B-V-4. $\eta = \delta = 0$ and reports were obtained all at once rather than sequentially. Arbitrary contrast set preference and EFLI contrast set preference are distinguished.

7 or 10 is nearly as good as full plausibility precision. This result agrees with previous findings regarding a Bayesian network's sensitivity to plausibility precision (Pradhan et al., 1996). When $\eta = 0.6$, low plausibility precision has even less of an impact.

There is curious behavior at plausibility precision 3. In this case, the plausibility scores are all one of $[0, 0.50, 1.0]$. The original plausibility score for each report or hypothesis is replaced with the closest of these three scores. For example, 0.35 is replaced with 0.50. As it happens, in the Bayesian network domains, 0.50 is almost always the closest such plausibility score, so virtually all reports and hypotheses are assigned the same plausibility. This results in a scenario much like having no plausibility estimates at all (addressed earlier). Yet in cases where there are two possible plausibilities (0 or 1.0), the true reports

Noise	Accuracy	Coverage	N. Prec.	N. Recall	N. F1
0	+0.152 ***	0.000			
10	+0.141 ***	0.000	+0.056 *	+0.007	+0.012 *
20	+0.134 ***	0.000	+0.123 ***	+0.013 **	+0.025 **

Table 8.2. Impact of having no plausibility information, supporting Hypothesis B-V-5. $\eta = \delta = 0$. This experiment compares abductive reasoning with no plausibility information and abductive reasoning with normal plausibility information. In both cases with and without plausibility information, results are averaged across noise levels of 0%, 10%, and 20%. In the table, a Noise value of 10 indicates 10% noise level. "N. Prec." etc. refer to the Noise Precision metric, etc. A value $+0.152$ for some metric means that EFLI produced, averaged across individual cases, 0.152 higher on that metric. Statistical significance is indicated by asterisks: * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$.

and hypotheses are more often assigned the plausibility 1.0 and the false 0.0. In cases of plausibility precision of four (0, 0.33, 0.66, 1.0), again true reports and hypotheses generally score higher than false ones. Thus, a plausibility precision of 3 performs worse than 2 or 4.

8.7 Metareasoning experiments

The following experimental hypotheses are labeled with the format "B-M-#" to indicate that these hypotheses are regarding the Bayesian network domains' metareasoning experiments.

Hypothesis B-M-1: Metareasoning gives better Accuracy, Coverage, and Noise identification than no metareasoning, for certain values of $\eta, \delta, \eta_{\text{meta}}, \delta_{\text{meta}}$.

Hypothesis B-M-2: Metareasoning gives better performance than no metareasoning even when report plausibilities are unknown.

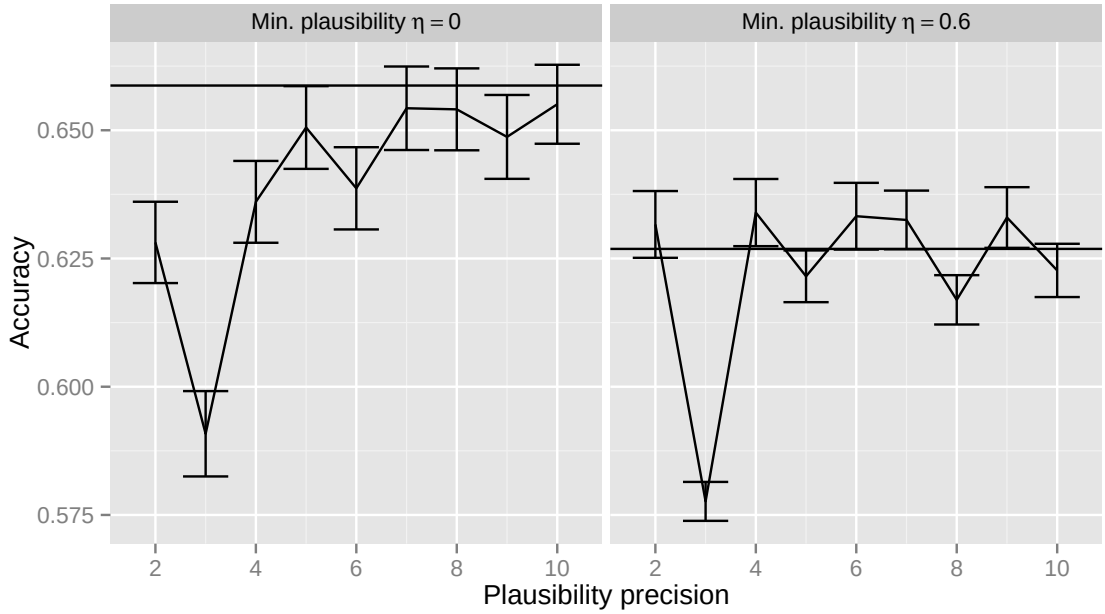


Figure 8.9. Impact of plausibility precision on accuracy for different both $\eta = 0$ and $\eta = 0.60$, and $\delta = 0$, supporting Hypothesis B-V-5. Results are averaged across noise levels of 0%, 10%, and 20%. The horizontal line marks Accuracy at maximum plausibility precision (double-precision floating-point numbers).

Hypothesis B-M-3: Performance is maximized when each of MetaImplHyp, MetaIncompatHyp, MetaOrderDep meta-hypotheses is available as a possible explainer of anomalies. This can be tested with ablation experiments in which various combinations of meta-hypotheses are supported.

Hypothesis B-M-1

We see from Figures 8.10 and 8.11 that abductive metareasoning almost always increases Accuracy, Coverage, and AccCov for all values of η. We found that $\delta = \delta_{\text{meta}} = 0$ gave best performance in all cases. Noise identification, shown in the bottom of Figure 8.11, suffers under abductive metareasoning for certain values of η and η_{meta}. This outcome

seems to be due to a decrease in Noise Recall, although Noise Precision increases. We see this in Table 8.3. The reason is that abductive metareasoning often finds ways to explain more reports, both noisy and true reports, by lowering the minimum plausibility threshold η in certain cases. Although Accuracy and Noise Precision show an increase, Noise Recall decreases because in some rare cases, abductive metareasoning finds explanations for false reports.

A higher η value is the primary cause of anomalies in the Bayesian network domains. We saw this earlier in Figure 8.5. Table 8.3 reinforces this result. The table includes a "Cases" column which shows how many cases, among 15,000 random scenarios, contained anomalies and thus activated abductive metareasoning. When $\eta < 0.6$ or so, anomalies are virtually non-existent. Even so, abductive metareasoning almost always brings an increase in Accuracy, Coverage, and Noise Precision.

Hypothesis B-M-2

As described in Section 6.6, with regards to Hypothesis S-M-3 of the simulated tracking domain, the plausibility function for meta-hypotheses (Section 4.5) is simply the average of the plausibilities of the anomalies that a meta-hypothesis is capable of explaining. It is important to investigate how abductive metareasoning performs when report plausibilities are unknown (equivalently 1.0). Table 8.4 shows these results. Comparing with Table 8.3, which shows results for experimental cases that do include report plausibilities, we see that only Noise Recall is significantly lowered when report plausibilities are obscured. This outcome is due to the fact that noise identification depends crucially on having accuracy report plausibilities because noise is detected only when it remains unexplained. Usually, very implausible reports are left unexplained by abductive metareasoning due to the η_{meta} threshold. However, when all reports have plausibility 1.0, they seem to be highly plausible

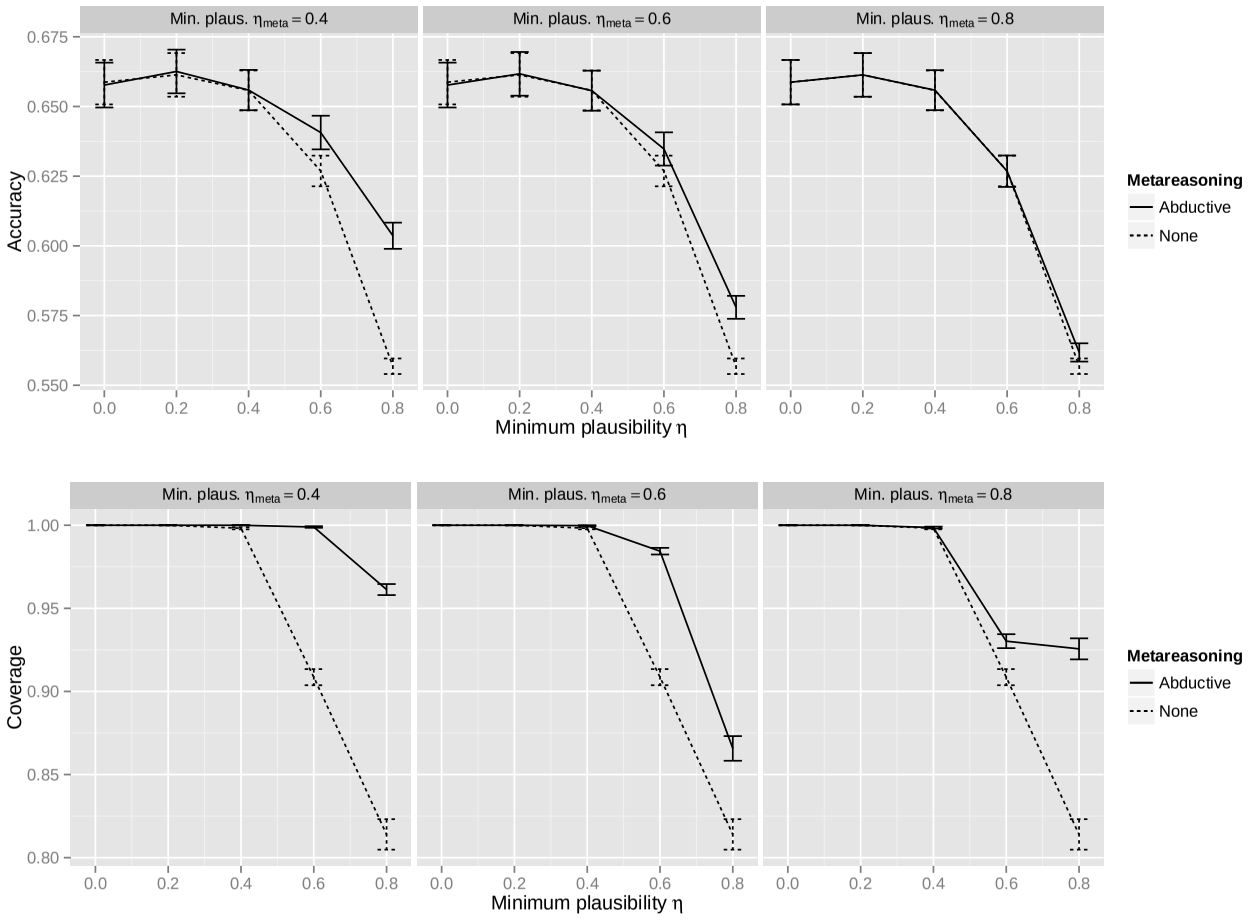


Figure 8.10. Accuracy and Coverage for various values of η and η_{meta}, averaged across 0%, 10%, and 20% noise. Supports Hypothesis B-M-1. In all cases, $\delta = \delta_{\text{meta}} = 0$.

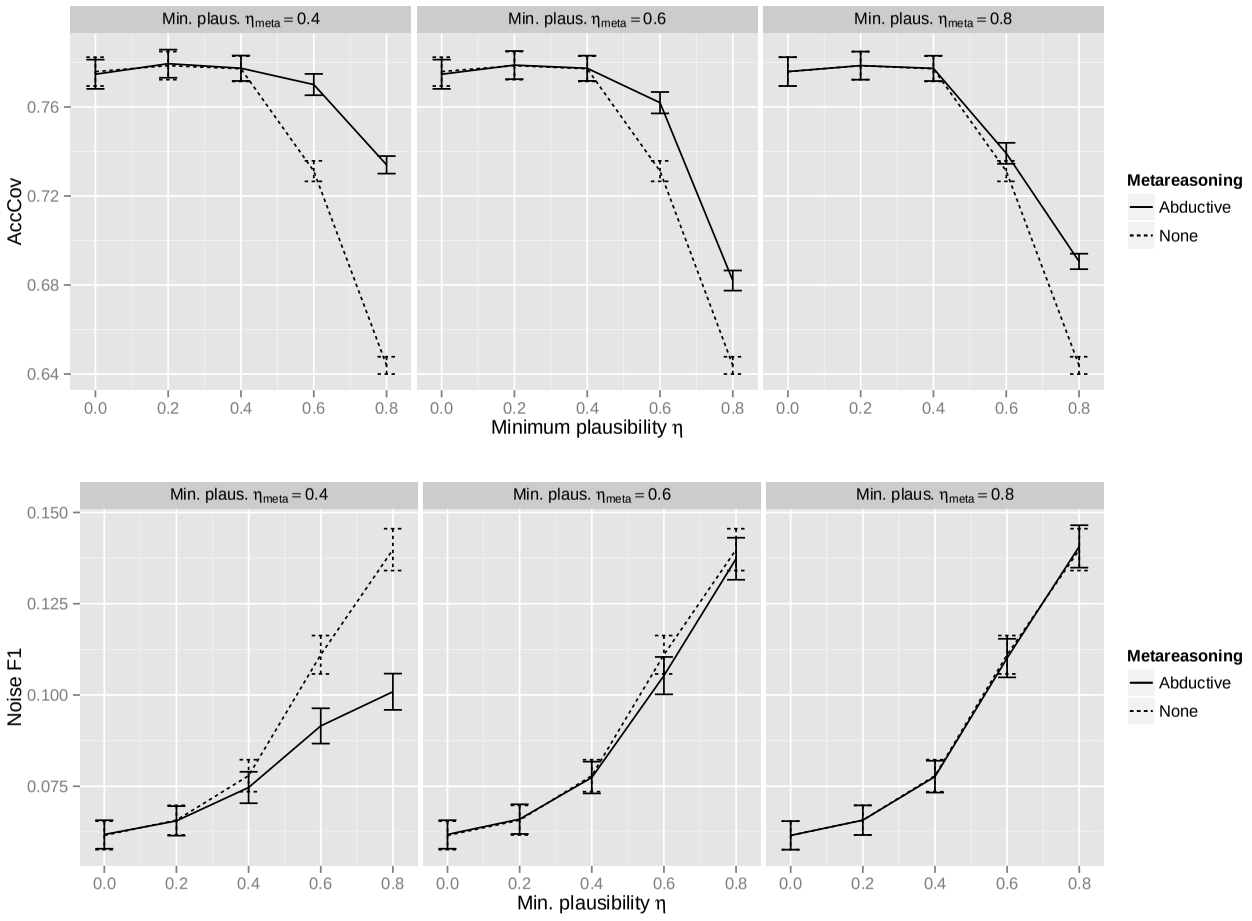


Figure 8.11. AccCov and Noise F1 for various values of η and η_{meta}, averaged across 0%, 10%, and 20% noise. Supports Hypothesis B-M-1. In all cases, $\delta = \delta_{\text{meta}} = 0$.

η	Noise	Accuracy	Coverage	N. Prec.	N. Recall	N. F1	Cases
0.00	0	+0.016	0.000				3
0.00	10	-0.082	0.000	+0.167	+0.050	+0.077	2
0.20	0	+0.054	+0.065 *				8
0.20	10	+0.018	+0.063	+0.278 *	+0.017	+0.034	6
0.20	20	-0.015	+0.051	+0.067	0.000	+0.003	5
0.40	0	+0.026	+0.056 **				21
0.40	10	+0.018	+0.071 ***	+0.116 **	-0.004	+0.003	21
0.40	20	-0.010	+0.048 **	+0.088	-0.011	-0.009	22
0.60	0	+0.014 ***	+0.117 ***				701
0.60	10	+0.007 *	+0.122 ***	+0.048 ***	-0.009 ***	-0.002	662
0.60	20	+0.007	+0.128 ***	+0.052 ***	-0.017 ***	-0.008 **	604
0.80	0	+0.025 ***	+0.067 ***				942
0.80	10	+0.016 ***	+0.074 ***	+0.038 ***	-0.015 ***	-0.001	908
0.80	20	+0.011 ***	+0.078 ***	+0.037 ***	-0.023 ***	-0.006 *	873

Table 8.3. Results from comparative experiments, supporting Hypothesis B-M-1, with abductive metareasoning and no metareasoning, for different minimum plausibility η thresholds and noise levels. $\eta_{\text{meta}} = 0.60, \delta = \delta_{\text{meta}} = 0$. In the table, a Noise value of 10 indicates 10% noise level. A metric value $+0.016$ indicates that abductive metareasoning increased that metric on average by 0.016 compared to no metareasoning. The "Cases" column indicates the number of cases, among 15,000 random scenarios, when anomalies were present and abductive metareasoning as activated. Statistical significance is indicated by asterisks: * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$.

reports, and explanations are more often found, thus reducing Noise Recall. Note that, unlike the tracking domains, non-reported variable states in the Bayesian network may become anomalies just as reports may. The plausibility of those anomalies is calculated in the usual way (using posteriors, given prior beliefs and reports) in these experiments.

Hypothesis B-M-3

Finally, Hypothesis B-M-3 states that each kind of meta-hypothesis (MetaImplHyp, MetaIncompatHyp, MetaOrderDep; refer to Section 4.1) plays an important role in abductive

η	Noise	Accuracy	Coverage	N. Prec.	N. Recall	N. F1	Cases
0.00	0	+0.009	0.000				4
0.00	10	+0.035	0.000	+0.139	+0.033	+0.055	3
0.00	20	-0.047	0.000	+0.083	+0.007	+0.017	4
0.20	0	+0.036	+0.040 *				17
0.20	10	+0.033	+0.022	+0.025	-0.015	-0.020	17
0.20	20	+0.001	+0.015	+0.117	-0.024	-0.027	17
0.40	0	+0.008	+0.022 *				47
0.40	10	+0.001	+0.027 ***	-0.021	-0.027 ***	-0.036 ***	63
0.40	20	-0.003	+0.015 **	-0.013	-0.040 ***	-0.052 ***	77
0.60	0	+0.023 ***	+0.096 ***				901
0.60	10	+0.015 ***	+0.097 ***	+0.041 ***	-0.034 ***	-0.032 ***	884
0.60	20	+0.015 ***	+0.094 ***	+0.055 ***	-0.056 ***	-0.052 ***	864
0.80	0	+0.051 ***	+0.136 ***				999
0.80	10	+0.040 ***	+0.132 ***	+0.039 ***	-0.072 ***	-0.054 ***	999
0.80	20	+0.031 ***	+0.131 ***	+0.039 ***	-0.120 ***	-0.094 ***	999

Table 8.4. Results from comparative experiments, supporting Hypothesis B-M-2, in which report plausibilities are unknown (i.e., constantly 1.0). Abductive metareasoning is compared to no metareasoning for different minimum plausibility η thresholds and different noise levels. $\eta_{\text{meta}} = 0.60, \delta = \delta_{\text{meta}} = 0$. In the table, a Noise value of 10 indicates 10% noise level. A metric value $+0.009$ indicates that abductive metareasoning increased that metric on average by 0.009 compared to no metareasoning. The "Cases" column indicates the number of cases, among 15,000 random scenarios, when anomalies were present and abductive metareasoning as activated. Statistical significance is indicated by asterisks: * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$.

metareasoning. We can test this claim by conducting ablation experiments, in which only a subset of the three meta-hypotheses is available for abductive metareasoning. Each subset is tested in turn (excluding the case in which no meta-hypotheses are available, since we essentially tested that case earlier in regards to Hypothesis B-M-1).

Table 8.5 summarizes the results. We see that in all cases, leaving out MetaIncompatHyp gives best results. A similar result was found in the simulated tracking domain. As in that domain, the difference here between including MetaIncompatHyp and not in-

cluding it is minute. We saw earlier (Figure 8.5 and Table 8.3) that most anomalies in the Bayesian network domains are due to a high η value. Thus, it is no surprise that Table 8.5 shows that MetaImplHyp meta-hypotheses are very important; leaving them out severely negatively impacts Accuracy and Coverage. However, not surprisingly, leaving out MetaImplHyp improves Noise Recall because more reports remain anomalous (since they are not explained by a MetaImplHyp meta-hypothesis) and thus are considered noise claims. Finally, MetaOrderDep on its own does not offer much benefit, but when included with MetaImplHyp, abductive metareasoning offers its maximum benefits.

8.8 Prior work

There exists a large body of work about inferencing with Bayesian networks. Randomly-generated Bayesian networks were used to evaluate abductive reasoning and metareasoning because their structure and probabilistic properties are familiar to many readers. However, our approach has some interesting differences with prior work. These differences can be categorized under two headings: the definition of a *relevant explanation* and the definition of the *best explanation*.

Relevant explanation

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, equating *explanation* with MPE or MAP introduces issues of relevance. 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,

Noise	Impl	Incompat	OD	Accuracy	Coverage	N. Precision	N. Recall	N. F1
0	X	X	X	0.654 ± 0.008	0.981 ± 0.003			
0	X	X		0.654 ± 0.008	0.980 ± 0.003			
0	**X**		**X**	$\mathbf{0.655 \pm 0.008}$	$\mathbf{0.983 \pm 0.003}$			
0	X			0.655 ± 0.008	0.982 ± 0.003			
0		X	X	0.586 ± 0.034	0.704 ± 0.057			
0		X		0.610 ± 0.032	0.665 ± 0.065			
0			X	0.488 ± 0.112	0.857 ± 0.000			
10	X	X	X	0.632 ± 0.008	0.979 ± 0.003	0.253 ± 0.014	0.085 ± 0.005	0.122 ± 0.006
10	X	X		0.633 ± 0.008	0.979 ± 0.003	0.251 ± 0.014	0.085 ± 0.005	0.121 ± 0.006
10	**X**		**X**	$\mathbf{0.633 \pm 0.008}$	$\mathbf{0.982 \pm 0.003}$	$\mathbf{0.253 \pm 0.014}$	$\mathbf{0.085 \pm 0.005}$	$\mathbf{0.121 \pm 0.006}$
10	X			0.633 ± 0.008	0.981 ± 0.003	0.250 ± 0.014	0.084 ± 0.005	0.120 ± 0.006
10		X	X	0.510 ± 0.052	0.777 ± 0.072	0.153 ± 0.060	0.077 ± 0.030	0.102 ± 0.040
10		X		0.529 ± 0.056	0.765 ± 0.082	0.139 ± 0.067	0.074 ± 0.035	0.096 ± 0.045
10			X	$0.375 \pm$ NA	$0.857 \pm$ NA	$0.250 \pm$ NA	$0.100 \pm$ NA	$0.143 \pm$ NA
20	X	X	X	0.607 ± 0.008	0.978 ± 0.003	0.372 ± 0.014	0.140 ± 0.006	0.196 ± 0.007
20	X	X		0.606 ± 0.008	0.979 ± 0.003	0.372 ± 0.014	0.140 ± 0.006	0.196 ± 0.007
20	**X**		**X**	$\mathbf{0.607 \pm 0.008}$	$\mathbf{0.981 \pm 0.003}$	$\mathbf{0.371 \pm 0.014}$	$\mathbf{0.140 \pm 0.006}$	$\mathbf{0.196 \pm 0.007}$
20	X			0.607 ± 0.008	0.981 ± 0.003	0.370 ± 0.014	0.140 ± 0.006	0.196 ± 0.007
20		X	X	0.509 ± 0.029	0.746 ± 0.077	0.214 ± 0.050	0.137 ± 0.041	0.165 ± 0.044
20		X		0.524 ± 0.028	0.733 ± 0.085	0.238 ± 0.049	0.153 ± 0.042	0.183 ± 0.045
20			X	$0.375 \pm$ NA	$0.857 \pm$ NA	$0.000 \pm$ NA	$0.000 \pm$ NA	$0.000 \pm$ NA

Table 8.5. Results from ablation experiments, supporting Hypothesis B-M-3, in which various meta-hypotheses are disabled. A Noise value of 20 indicates 20% noise level. Columns "Impl," "Incompat," "OD" refer to presence ("X") or absence of support for each meta-hypothesis: MetaImplHyp, MetaIncompatHyp, MetaOrderDep. Each value for each metric is shown as average $\pm$ standard error, except in rows where meta-hypotheses were only accepted in one experimental case. Best performance for each of the noise levels are highlighted with bold text. $\eta = \eta_{\text{meta}} = 0.60$, $\delta = \delta_{\text{meta}} = 0$.

e.g., downstream effects of the observations, are included in the explanation.

Chajewska and Halpern (1997) introduce a new definition of explanation, attributed to Gärdenfors (1988), which requires that $P(e|x) > P(e)$, where x is a possible explanation of e, and that $P(x) < 1$. Chajewska and Halpern show that this definition is not sufficiently limiting, as for any f where $P(f) < 1$, $e \wedge f$ will be considered an explanation of e. They go on to offer a "synthesis" of Gärdenfors' approach and MAP-based explanation. Their synthesis requires the specification of a "causal structure" that gives the *relevant* causal relations for e. Again, it seems that the relevant variables in the network must be determined *a priori*.

Yuan and Lu (2007) develop an approach they call the *Most Relevant Explanation* (MRE) and refine it in subsequent work (Yuan et al., 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 probabilistic 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 criteria. These relevance criteria are encoded in the EFLI algorithm (Algorithm 2.3), which prefers hypotheses that are both plausible and decisive, and capable of explaining unexplained reports or beliefs. Yuan and Lu's MRE differs from EFLI-based abduction, however, in that the target variables for the MRE must be established ahead of time, while our approach automatically seeks explanations for unexplained reports and beliefs.

Flores et al. (2005) describe a unique approach to finding relevant explanations. They build an *explanation tree* in which a node is a variable in the explanation and every branch is a particular state for that variable. Leaf nodes store the probability of the variable state assignments that are found along the path from the leaf to the root. The tree is not neces-

sarily balanced or symmetric; thus, different leaf nodes might have different depths (i.e., different number of variables in the corresponding explanation). The tree is built according to criteria involving probabilities and information entropy that aim to ensure all possible explanations represented in the tree are relevant. Once the tree is built, the best relevant explanation may be found by picking out the most probable leaf node.

This 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 relevant explanations. In the work presented here, all combinations (of all sizes) of parents of each unexplained variable are generated and scored according to their posteriors. The abductive reasoning process determines which hypotheses are best according to plausibility and decisiveness, and then rejects incompatible hypotheses (such as those that posit different states for variables that it has in common with the accepted hypothesis). "Relevant" explanations need not be determined *a priori*; instead, abductive reasoning decides what's relevant according to whether or not it is plausible, decisive, and explanatory. While we do not offer any way of measuring which approach (EFLI, MAP, MPE, MRE, etc.) actually produces more or less relevant explanations, it is worth noting the differences among these approaches.

Best explanation

What constitutes the *best explanation* is also contentious. For the most probable explanation (MPE), "best" is equal to "most probable." However, Glass (2007) has shown that this equivocation is problematic:

> A problem with this approach is that a hypothesis could turn out to be the best explanation even if the evidence is extremely unlikely given the hypothesis. For example, in many cases a hypothesis which lowers the probability of

> the evidence relative to the unconditional case, i.e., $P(e|c_i) < P(e)$, will turn out to be the best explanation according to the MPE approach because it has a high prior probability. While it may not be desirable to rule out the possibility that the evidence could be negatively dependent on the best explanation, it still seems reasonable to say that such dependence should count against the hypothesis more than it does in the MPE approach. (Glass, 2007)

Glass (2009) explores seven definitions for *best explanation* and experimentally evaluates the accuracy of the various approaches. There is also a significant body of work comparing and contrasting "inference to the best explanation" (IBE) and Bayesian abduction (Bartelborth, 2006; Douven, 1999; Iranzo, 2008; Lipton, 2004; Psillos, 2004; Weisberg, 2009).

8.9 Conclusions

Each of our experimental hypotheses is confirmed by this work. They tell us that abductive reasoning and metareasoning is a very effective strategy for using Bayesian networks to infer the true 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.

Chapter 9: Conclusions

The goals of this research were two-fold: (1) design, implement, and analyze an effective, domain-general abductive reasoning procedure; and (2) design, implement, and analyze an effective, domain-general metareasoning procedure that boosts accuracy over the base-level abductive reasoning system. These goals were met, and the second goal was even exceeded because we were able to design a metareasoning procedure that is self-similar with the base-level reasoning system. This self-similarity, in which both the base-level reasoning system and the metareasoning system utilize identical abductive reasoning algorithms, is attractive from a system design standpoint and is cognitively plausible, as argued in Section 4.10.

The remainder of this chapter addresses some concerns and issues about the present work, and hints at plans for future work. More extensive discussion of future work may be found in Chapter 10.

9.1 Methodology

There are a few important limitations to discuss about our methodology. First, we will address a variety of idealizations. In our definition of a doxastic state (Definition 2.2.1), we idealize *explanation* as a binary relation. Explanations cannot be partial. Hypotheses are also either accepted, rejected, or undetermined; they cannot be partially believed as you find, for example, in Bayesian networks. Of course, this latter idealization is important for our development of metareasoning, since beliefs need not be taken back or revised if they are not categorical but rather partial or probabilistic beliefs. We also assume that

plausibility estimates are non-malevolent, i.e., mostly veridical. We assume there is no evil deceiver whose aim it is to ensure the reasoning system acquires the wrong beliefs. However, we do plan in future work to address the issue of *deception* (see Section 10.5).

Another concern is that our experiments only cover a small number of limited problem domains. We have two object tracking domains and Bayesian network domains represented by a collection of randomly-generated Bayesian networks. None of these domains covers the range of properties one might find in real-world problem domains, such as: (1) thousands or hundreds of thousands of reports and/or hypotheses, (2) very long-term reasoning tasks that span thousands of "time steps" (our experiments ran for 20 or fewer time steps), (3) strict time or memory resource limits such as might be required in robotics applications. Some of these issues are planned in future work (see Section 10.7). We also recognize that our results would be more convincing if we experimented with a well-known domain for which there are clearly-defined success criteria. For example, some datasets for speech recognition have been subjected to a wide variety of approaches. It would be clear that abductive reasoning and abductive metareasoning are a powerful combination if it could be demonstrated that our system beats the state-of-the-art for one or more of these datasets. However, we consider the present work simply an initial effort that exposes the fundamental features of our system. Future work will partly aim to demonstrate its wider applicability and better understand its performance trade-offs.

Finally, we chose not to analyze performance of our system in terms of computation time and memory use. This is a glaring omission because we hope for the system to be, above all, a pragmatic choice for cognitive agents. In fact, engaging in metareasoning is more expensive than not doing so, at least according to preliminary experiments. Optimizing the entire process is planned in future work (see Section 10.1). Nevertheless, our analysis of the computational complexity of abductive reasoning (Section 2.9) shows that it is polynomial-time and therefore not obviously impracticable.

9.2 Abductive reasoning

Our approach to abductive reasoning uses a greedy, hill-climbing (non-backtracking) algorithm that may not find a complete composite explanation that explains all the reports and unexplained beliefs. We chose to use a greedy algorithm because the problem of finding a complete and consistent explanation is NP-complete. Additionally, the greedy algorithm may not find the most plausible consistent and complete explanation, but that problem is NP-hard. See Section 2.4 for details. However, our algorithm does a reasonable job at efficiently finding complete or nearly-complete and highly plausible consistent composite explanations. Furthermore, the EFLI algorithm, which prefers hypotheses that are plausible and decisive, has been shown experimentally to boost performance over arbitrary hypothesis preferences.

9.3 Abductive metareasoning

The abductive metareasoning system that we have developed only responds to *anomalies*, i.e., unexplainable evidence. We have shown that this works considerably well for correcting errors in the doxastic state. However, anomalies as defined might not be the only candidates for abductive metareasoning. In other words, there might be other signals of trouble:

- the absence of a *decisive* explanation for some evidence; i.e., an explainer where its decisiveness is less than δ;
- a gradual decline of *confidence* in the doxastic state, where confidence is measured by average plausibility of accepted hypotheses, or average decisiveness, or some combination of these measures;

We saw in Section 4.8 that abductive metareasoning does not always properly handle anomalies with multiple causes. This limitation suggests that the meta-hypotheses should perhaps be more sophisticated to allow combinations that specify causes to the effect, "this anomaly is due to both implausible hypotheses and, should that be resolved, incompatible hypotheses." This may be achieved by creating more kinds of meta-hypotheses, or extending the core abductive reasoning algorithm to automatically merge hypotheses in some manner.

As mentioned earlier, metareasoning is a potentially costly operation. Abductive metareasoning involves at least one more execution of the abductive reasoning algorithm, and if any belief revisions are involved or a MetaOrderDep meta-hypothesis is accepted, then after the revisions are made, abductive reasoning is executed again on the revised doxastic state in order to *finalize* the state. Metareasoning might be activated yet again if more anomalies remain. This all adds up to more time spent on the abductive reasoning algorithm than just avoiding metareasoning altogether. We have measured *cost-benefit ratio* of metareasoning, in which cost is measured in milliseconds and benefit is measured in change in F1 or Accuracy, and discovered that, essentially, metareasoning is not worth it. However, in future work we wish to extend our experiments in a variety of ways. First, perhaps metareasoning is too costly when it is performed at every time step, but its cost is amortized effectively when it is performed only after a certain number of anomalies have accumulated or a certain amount of time has passed. We also want to identify if there are cases where metareasoning makes corrections early on, and these corrections reduce confusion later and result in more efficient abductive reasoning and fewer anomalies. Finally, our experiments that measured the cost-benefit ratio included problem domain computations in the cost measure. It might be more insightful to isolate the reasoning computations from the observing and hypothesis generation functions in order to find the true cost of metareasoning.

Finally, we saw in the overview of prior work in metareasoning, which includes belief revision (Chapter 5), that some metareasoning strategies make commitments to a language and logic for beliefs. For example, in traditional strict belief revision, beliefs are represented as propositions, in classical symbolic logic. These metareasoning systems are able to generate alternative explainers or perform very specific belief revisions in order to resolve anomalies. Our metareasoning system, on the other hand, does not have access to the *content* of the reports and hypotheses, so it is unable directly to generate alternative doxastic states. Instead, revisions are made to the acceptance status of existing hypotheses and abductive reasoning is activated again. This second run of abductive reasoning may result in new hypotheses or even new evidence, but the metareasoning system cannot know what those new hypotheses or evidence will be. This limitation means that the metareasoning system might not be as effective as a system that does have direct access to domain-specific information. However, we have chosen to design both the base-level and meta-level reasoning systems in this way in order to maximize generality.

9.4 Wider relevance

This is a study about high-level artificial cognitive systems, so it is reasonable to ask if there are any take-away lessons for all cognitive agents. In other words, are there any insights we can teach our children? We believe this work has illustrated a few insights about how to be a smart thinker. After each insight, we summarize the evidence that supports it.

> "If one has alternative ways to explain some evidence, it is more likely that the most plausible, most decisive explanation is true."

The EFLI algorithm was shown in each domain to yield greater accuracy than arbitrary abduction.

> "It is wise to avoid believing something implausible."

We saw in each domain that $\eta > 0$ yielded best accuracy.

> "It is wise to avoid believing that X explains Y if some $X' \neq X$ seems nearly as plausible an explainer of Y, assuming X and X' are different (possibly incompatible) stories of how Y came to be. In particular, it is wise to avoid believing X if $\neg X$ is almost as likely. Rather than believing X in these cases (or X'), it is better to remain uncommitted about X until further information is available."

We saw in the simulated tracking domain, $\delta > 0$ yielded best accuracy. However, the aerial tracking domain and Bayesian network domains did not exhibit the same property. Nevertheless, it seems intuitive that, under normal circumstances, $\delta > 0$ is a good idea. We do not expect that there is one single value for η or δ that is best in all scenarios, however. The completeness–confidence trade-off (Figure 2.10) illustrates that pragmatic concerns, such as the cost of being wrong and the cost of not explaining all the evidence, play an important role in determining η and δ.

> "If there is some fact or claim Y that cannot plausibly be explained in a way that is consistent with existing beliefs, then one has reason to be doubtful both of (a relevant subset) of existing beliefs and the veracity of Y."

Normally, one expects to find a plausible and consistent explanation for all reports, all observations, all statements made by other agents, etc. When this is not possible, both the evidence and one's existing beliefs that are somehow related to the evidence should be called into question. Note that calling beliefs and the evidence into question does not imply that one must be able to decide, at that moment, which beliefs are false or that the evidence is noise. Perhaps by waiting, more evidence will be acquired to resolve the

anomaly. The metareasoning system significantly boosted accuracy and noise identification when it behaved in exactly this way. Abductive metareasoning first considered whether some kind of belief revision was possible in order to find consistent and plausible explainers for the anomalies. If there was no such plausible meta-hypotheses, then the anomalies were left alone, and at the end of the experiment, any anomalies that were not later explained were considered to be labeled as noise.

We hope to find deeper insights into the nature of cognition and intelligence by extending this work in a variety of ways. These ways are addressed in the next chapter.

Chapter 10: Future work

The combined abductive reasoning and abductive metareasoning system has been demonstrated to be a very effective strategy for "making sense" in a variety of problem domains. The foundations have been built, and now a variety of deeper questions about cognition in general, and metareasoning in particular, can be explored.

10.1 Efficient metareasoning

It would be useful to characterize the cost–benefit trade-off of metareasoning. The system we have investigated performs metareasoning for every time step in which there are anomalies. Perhaps it is more computationally efficient to perform metareasoning every n time steps, or only after some number of anomalies have accumulated.

Consider the folk wisdom, "perfect is the enemy of good" (often attributed to Voltaire). Is it better to make commitments early and possibly detect and correct mistakes later, than to spend more time gathering more evidence and evaluating more hypotheses now? And will much of the confusion now simply work itself out later, due to the availability of new evidence and the hindsight that comes with time? Can a cognitive system reason about when it is best to perform metareasoning, in terms of the expected cost–benefit trade-off?

10.2 Bounded memory

How does *forgetting* affect abductive metareasoning? It is apparent that humans quickly forget low-level inputs (e.g., the sounds of a speaker's voice) but remember higher-level concepts (e.g., the words themselves, or the gist of what the person was saying) for longer

periods. Forgetting the original observations (what was picked up by the senses) might limit abductive metareasoning's ability to find alternative explanations for old data. Is it possible for an agent to reason that some data might be needed again at later time in order to eliminate future confusion? Are there "forgetting heuristics" that optimize the use of memory but simultaneously prepare for future demands for self-correction? Can an agent *plan* and *prepare* to perform abductive metareasoning in the future?

10.3 Dunning–Kruger effect

The so-called Dunning–Kruger effect (Kruger and Dunning, 1999) is the tendency for incompetent persons to strongly over-estimate their competence in a wide range of tasks, such as logic and math problems, identifying humor, and so on (Figure 10.1). Kruger and Dunning claim that incompetent subjects lack the "metacognitive ability" to identify their own errors and others'. In other words, their incompetence in the task results in their inability to be aware of their incompetence.

Suppose we gave a cognitive system some way to estimate its own competence in its task. An abductive reasoning system might measure its competence by taking the average of the decisiveness of accepted hypotheses. The intuition is that competence means the best explanation is usually better than the alternatives (very decisive). In any event, is it the case that incompetent systems (that is, those that perform poorly on the task because, for example, its plausibility estimates are poor, or it routinely fails to consider alternative explainers) also over-estimate their competence, i.e., the wrong explainers are often the most decisive? Furthermore, is it the case that precisely those incompetent systems fail to self-correct, because they lack the kind of knowledge required by abductive metareasoning to detect and correct false beliefs and noise? And is there some kind of fix that applies to artificial cognitive systems and can be shown experimentally to bring the incompetent

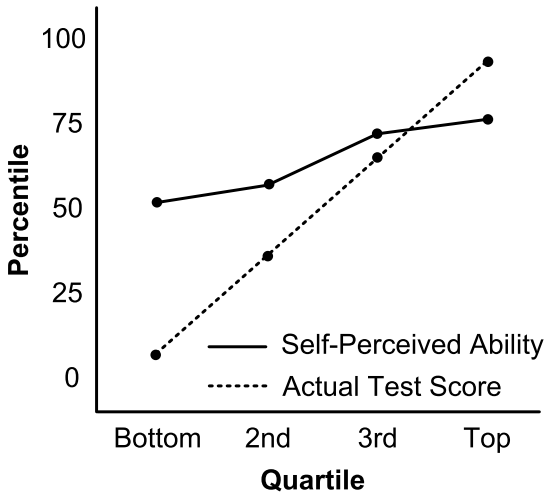


Figure 10.1. Dunning–Kruger effect. Recreated from Fig. 1 in Kruger and Dunning (1999).

agent out of this mess and allow it to perform some self-correction? Finally, will this trick work on humans as well?

10.4 Dogmatism and delusions

Are artificial cognitive systems susceptible to dogmatic or delusional beliefs just as humans are? We might say that a cognitive system is acting dogmatically if it has a preponderance of evidence that its beliefs are mistaken but fails to revise those beliefs appropriately. A delusional cognitive system might be described as one that holds highly implausible beliefs even in the face of significant counter-evidence (Bell et al., 2006). We also often think of a delusional person as one who has a mostly-consistent but completely wrong model of the world. Can reasoning systems be designed so that they can detect and repair dogmatic or delusional patterns of reasoning? Are there cues, like consistently implausible hypotheses or consistently few alternative explainers? And can anything be done to shed bad beliefs in

one fell swoop? Humans sometimes arrive at delusional beliefs after experiencing "abnormal data" (Coltheart et al., 2010). How do implausible or noisy inputs affect the likelihood of a cognitive system behaving dogmatically or delusively?

10.5 Identifying deception

When should putative observations not be trusted as plausibly noisy or deceptive? In other words, when should a cognitive system disbelieve the data? The approach we have taken is to disbelieve data that introduce significant confusion, and cannot be resolved by abductive metareasoning. This strategy seems to work well for noisy data. How can *deception* specifically be detected? Deception seems more difficult to detect because, if it is good deception, it is designed to go undetected as deceptive, but nevertheless lead the system to the wrong beliefs. Being wrong beliefs, there might eventually be some evidence that the beliefs are wrong. No deception is perfect (presumably). What are the signs that deception has been successful? Are beliefs that are arrived at in order to explain deceptive inputs often implausible without having obtained the report (i.e., are their priors implausible but their posteriors, conditioned on the report, not)? Or do false beliefs more often conflict with other beliefs? That is, are false beliefs less compatible in general? These and other cues might give a cognitive system some hope in identifying deceptive inputs.

10.6 Meta-metareasoning

It seems obvious that the abductive metareasoning system should support meta-metareasoning. A meta-metareasoning system would monitor and control the metareasoning system. Because our abductive reasoning and abductive metareasoning systems are self-similar, and utilize the same machinery to do their jobs, an abductive meta-metareasoning system should be easily devised that monitors for anomalies in the abductive metareasoning sys-

tem. The question is, what would the meta-metareasoning system hypothesize? What are its meta-meta-hypotheses? Why would the metareasoning system have anomalies, i.e., no plausible meta-hypotheses that explain base-level anomalies? If we use the same approach as the abductive metareasoner, then perhaps the meta-meta-hypotheses posit that the metareasoner discounted possible explainers because they were deemed too implausible, or that some anomaly had no explanation (at the meta-level) because some other meta-hypothesis conflicted with a possible explainer. A meta-meta-hypothesis that mirrors the MetaOrderDep meta-hypothesis may also be useful, as abductive metareasoning is iterative and new anomalies (meta-level reports) might appear as metareasoning proceeds. Meta-metareasoning is nearly as easy as flipping a switch, given the existing system architecture. We are not certain the exploration will yield interesting results, but it seems worth trying.

10.7 New problem domains

We have experimentally validated our system in two object tracking domains (one simulated, one using aerial surveillance data), and a collection of Bayesian network domains. But there are many more kinds of problem domains that we think would benefit from abductive reasoning and abductive metareasoning. More experiments in different domains will also provide further evidence that our approach is practical and correct.

Plan recognition

Plan recognition is a paradigmatic case of abductive reasoning. The task is to observe actions committed by one or more agents and infer the plan that the agents are following (and hence infer their goals). Under normal circumstances, most of the intermediate actions required to execute a plan are hidden from the observer; the observer cannot be sure the

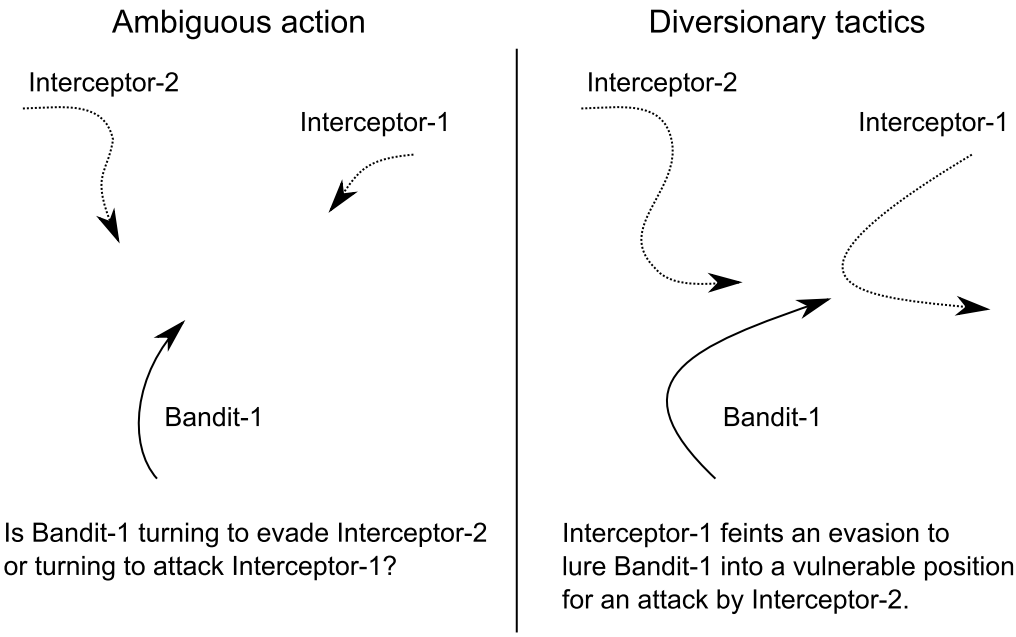


Figure 10.2. Ambiguity in a plan recognition task, from Heinze et al. (1999).

agents performed all the steps of the plan. There are usually more than one plausible plans that the agents are following at any particular time. The observed agents may carry out actions over time, so the most plausible plan given the observed actions may change over time. Plan recognition is rife with ambiguity. Consider the case shown in Figure 10.2. In the left diagram, there are two possibilities, indicated by the question posed in the diagram. In the right diagram, after some time has passed, it becomes more clear that Bandit-1 is chasing Interceptor-1. But rather than wait until sufficient evidence is available, strategists may wish to recognize Bandit-1's plan early and then respond (in this case, respond by luring Bandit-1 toward Interceptor-1).

It is clear that plan recognition has many of the properties that might make abductive reasoning an effective strategy. Previous work has explored various kinds of abductive reasoning procedures for plan recognition (Goldman et al., 1999; Paul, 1993). We expect that

we will be able to build a plan recognition problem domain and fit it into our existing system architecture. Furthermore, we will be able to test both realistic and vetted datasets as well as generate random cases. Once the plan recognition problem domain is implemented, we can activate metareasoning and see if it boosts accuracy. We expect that it will because, as described, there seem to be many scenarios where the most plausible plan turns out to be clearly wrong and anomalies appear, i.e., the next predicted actions are not observed, but other actions are observed that cannot be explained given the currently believed plan.

Speech recognition

Our work with the Bayesian network domains showed that both abductive reasoning and abductive metareasoning were very effective for inferring the true vertex states from reports. Furthermore, when certain kinds of noise were present, abductive reasoning yielded more accurate beliefs than the most probable explanation, which is the gold-standard in noise-free environments. For future work, we would like to extend these efforts into realistic and high-value problems. One such problem is speech recognition, in which domain knowledge is typically represented by Hidden Markov Models (Gales and Young, 2008; Juang and Rabiner, 1991) and Bayesian inference yields phonemes, words, and sentences from spectral vectors. We expect to find similar gains in accuracy in noisy conditions by using our abductive reasoning strategy on these problems. Furthermore, because inference at the word and sentence level crucially depends on having accurate "beliefs" about prior sentences and context of the sentence, we expect that abductive metareasoning, which is capable of belief revision given anomalous evidence, will improve accuracy beyond the base-level abductive reasoning system.

Speech recognition as a problem domain has the added benefit that there exist plentiful evaluation data and success is well-defined. However, it has the drawback that it is virtually

impossible to generate random cases (though one could simulate noise). Nevertheless, speech recognition is a good domain to explore in future work since the community of interested researchers and application developers is very large.

Robots

Finally, we believe it is always a good idea to try building a physical, bounded autonomous version of a cognitive system. Robots often have far fewer computational resources than system designers plan for in their initial developments. In our case, our cognitive system is very abstract and modularized to support swapping out the problem domain and to support careful evaluation of the base-level and meta-level reasoning systems. However, the entire system would require a redesign and reimplementation if it is to fit inside a small embedded system. Though not all robots are computationally-deficient, it is interesting to ask what are the *essential* features of the combined abductive reasoning and metareasoning system? And what can be left out?

We also must ask, what is the robot reasoning about and how might metareasoning help? We believe the way to answer this question is to ask what data structures, i.e., beliefs, the robot is constructing and updating over time, and how might this process be improved? Suppose that the robot is mapping a room, so it is building an internal map. Are there cases where this map is incorrect, and future observations seem not to make sense? These cases might be candidates for a metareasoning procedure to analyze the internal structures and the history of reasoning and find appropriate revisions.

The case of robots exposes an important point. We believe that abductive reasoning and abductive metareasoning are appropriate features of virtually any cognitive system. What makes a system cognitive is that it keeps internal representations of the world and reasons about them (cf. Brooks (1991)). These internal representations are likely estab-

lished and/or updated on the basis of evidence from the world (from sensors, reports from other agents, etc.). If these representations are inaccurate in some way, then evidence from the world might not make sense, leading the cognitive system to confusion. This confusion may be resolved with metareasoning. And we expect that, *ceteris paribus*, a less confused cognitive system is a more accurate one.

References

A. Abdelbar. Approximating cost-based abduction is NP-hard. *Artificial Intelligence*, 159 (1):231–239, 2004.

A. Abdelbar, E. Andrews, and D. Wunsch. Abductive reasoning with recurrent neural networks. *Neural Networks*, 16(5):665–673, 2003.

C. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of symbolic logic*, 50(2):510–530, 1985. ISSN 0022-4812.

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

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

M. L. Anderson and T. Oates. A review of recent research in metareasoning and metalearning. *AI Magazine*, 28(1):12, 2007.

T. Bartelborth. Is the best explaining theory the most probable one? *Grazer Philosophische Studien*, 70(1):1–23, 2006.

V. Bell, P. W. Halligan, and H. D. Ellis. Explaining delusions: a cognitive perspective. *Trends in cognitive sciences*, 10(5):219–226, 2006.

M. Ben-Bassat, R. Carlson, V. Puri, M. Davenport, J. Schriver, M. Latif, R. Smith, L. Portigal, E. Lipnick, and M. Weil. Pattern-based interactive diagnosis of multiple disorders:

The medas system. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2:148–160, 1980.

E. Beth. *Semantic entailment and formal derivability*. Noord-Hollandsche Uitg. Mij., 1961.

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

C. Boutilier and V. Becher. Abduction as belief revision. *Artificial intelligence*, 77(1): 43–94, 1995.

W. Bridewell. *Science as an Anomaly-Driven Enterprise: A Computational Approach to Generating Acceptable Theory Revisions in the Face of Anomalous Data*. PhD thesis, University of Pittsburgh, 2004.

R. A. Brooks. Intelligence without representation. *Artificial intelligence*, 47(1):139–159, 1991.

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

U. Chajewska and J. Halpern. Defining explanation in probabilistic systems. In *Proc. Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI'97)*, pages 62–71, 1997.

E. Charniak and S. Shimony. Probabilistic semantics for cost based abduction. In *Proceedings of the 8th National Conference on Artificial intelligence*, pages 106–111. AAAI Press, 1990.

E. Charniak and S. Shimony. Cost-based abduction and map explanation. *Artificial Intelligence*, 66(2):345–374, 1994.

M. Coltheart, P. Menzies, and J. Sutton. Abductive inference and delusional belief. *Cognitive Neuropsychiatry*, 15(1):261–287, 2010.

M. T. Cox. Metacognition in computation: A selected research review. *Artificial Intelligence*, 169(2):104–141, 2005.

M. T. Cox, T. Oates, and D. Perlis. Toward an integrated metacognitive architecture. In *2011 AAAI Fall Symposium Series*, 2011.

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

M. Cox and A. Raja, editors. *Metareasoning: Thinking about thinking*. MIT Press, 2011b.

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

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

M. D'Agostino. Are tableaux an improvement on truth-tables? *Journal of Logic, Language and Information*, 1(3):235–252, 1992.

A. Darwiche and J. Pearl. On the logic of iterated belief revision. *Artificial intelligence*, 89 (1-2):1–29, 1997.

R. Davis. Applications of meta-level knowledge to the construction, maintenance and use of large knowledge bases. HPP Memo 76-7, Stanford, July 1976.

R. Davis and B. G. Buchanan. Meta-level knowledge. In B. G. Buchanan and E. Shortliffe, editors, *Rulebased expert system: The MYCIN Experiments of the Stanford Heuristic Programming Project*, pages 507–530. Addison-Wesley, 1984.

J. de Kleer. An assumption-based TMS. *Artificial intelligence*, 28(2):127–162, 1986. ISSN 0004-3702.

S. Dixon and N. Foo. Connections between the atms and agm belief revision. In *International Joint Conference on Artificial Intelligence*, volume 13, pages 534–534, 1993.

S. Dixon. *Belief Revision: A Computational Approach*. PhD thesis, University of Sydney, 1994.

I. Douven. Inference to the best explanation made coherent. *Philosophy of Science*, 66(3): 424–435, 1999.

J. Doyle. A truth maintenance system. *Artificial intelligence*, 12(3):231–272, 1979.

J. Doyle. Reason maintenance and belief revision: Foundations vs. coherence theories. In P. Gärdenfors, editor, *Belief Revision*, pages 29–51. Cambridge University Press, New York, 1992.

J. Eckroth and J. R. Josephson. Anomaly-driven belief revision by abductive metareasoning. *Advances in Cognitive Systems*, page to appear, 2014.

T. Eiter and G. Gottlob. The complexity of logic-based abduction. *Journal of the ACM*, 42 (1):3–42, 1995.

R. Elio and F. J. Pelletier. Belief change as propositional update. *Cognitive Science*, 21(4): 419–460, 1997.

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

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

T. Finin and G. Morris. Abductive reasoning in multiple fault diagnosis. *Artificial Intelligence Review*, 3(2):129–158, 1989.

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

M. Flores, J. Gámez, and S. Moral. Abductive inference in bayesian networks: finding a partition of the explanation space. *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 470–470, 2005.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The annals of statistics*, 28(2):337–407, 2000.

T. Fung and R. Kowalski. The IFF proof procedure for abductive logic programming. *The Journal of logic programming*, 33(2):151–165, 1997.

M. Gales and S. Young. The application of hidden markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304, 2008.

A. Garcez, D. Gabbay, O. Ray, and J. Woods. Abductive reasoning in neural-symbolic systems. *Topoi*, 26(1):37–49, 2007.

P. Gärdenfors. An epistemic approach to conditionals. *American philosophical quarterly*, 18(3):203–211, 1981. ISSN 0003-0481.

P. Gärdenfors. *Knowledge in flux: Modeling the dynamics of epistemic states*. The MIT press, 1988.

M. R. Genesereth. An overview of meta-level architecture. In *Proceedings of the AAAI Conference*, pages 119–124, 1983.

D. H. Glass. Coherence measures and inference to the best explanation. *Synthese*, 157(3): 275–296, 2007.

D. H. Glass. Inference to the best explanation: A comparison of approaches. In *Proceedings of the AISB*, 2009.

A. Goel and J. Ramanujam. A neural architecture for a class of abduction problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 26(6):854–860, 1996.

R. Goldman, C. Geib, and C. Miller. A new model of plan recognition. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 245–254. Morgan Kaufmann Publishers Inc., 1999.

S. Hansson. A survey of non-prioritized belief revision. *Erkenntnis*, 50(2):413–427, 1999. ISSN 0165-0106.

C. Heinze, S. Goss, and A. Pearce. Plan recognition in military simulation: Incorporating machine learning with intelligent agents. In *Proceedings of IJCAI-99 Workshop on Team Behaviour and Plan Recognition*, pages 53–64. Citeseer, 1999.

J. Hintikka. Two papers on symbolic logic: Form and quantification theory and reductions in the theory of types. *Acta Philosophica Fennica, Fasc VIII*, 1955.

V. Iranzo. Bayesianism and inference to the best explanation. *Theoria, An International Journal for Theory, History and Foundations of Science*, 23(1), 2008.

Y. Jin and M. Thielscher. Iterated belief revision, revised. *Artificial Intelligence*, 171(1): 1–18, 2007. ISSN 0004-3702.

F. Johnson and S. Shapiro. Dependency-directed reconsideration belief base optimization for truth maintenance systems. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 313, 2005.

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

B. Juang and L. Rabiner. Hidden markov models for speech recognition. *Technometrics*, pages 251–272, 1991.

A. C. Kakas, A. Michael, and C. Mourlas. ACLP: Abductive constraint logic programming. *The Journal of Logic Programming*, 44(1-3):129–177, 2000.

A. Kakas, R. Kowalski, and F. Toni. Abductive logic programming. *Journal of logic and computation*, 2(6):719–796, 1992.

P. D. Karp. *Hypothesis Formation and Qualitative Reasoning in Molecular Biology*. PhD thesis, Stanford University, 1989.

S. Khemlani and P. Johnson-Laird. The need to explain. *The Quarterly Journal of Experimental Psychology*, 64(11):2276–2288, 2011.

J. Kruger and D. Dunning. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999.

H. Levesque. A knowledge-level account of abduction. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1061–1067, 1989.

I. Levi. Subjunctives, dispositions and chances. *Synthese*, 34(4):423–455, 1977. ISSN 0039-7857.

P. Lipton. *Inference to the best explanation*. Psychology Press, 2004.

T. Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10 (10):464–470, 2006.

M. C. Mayer and F. Pirri. Abduction is not deduction-in-reverse. *Logic Journal of IGPL*, 4(1):95–108, 1996.

R. A. Miller, J. Pople, Harry E., and J. D. Myers. INTERNIST-I: An experimental computer-based diagnostic consultant for general internal medicine. In J. A. Reggia and S. Tuhrim, editors, *Computer-Assisted Medical Decision Making*, Computers and Medicine, pages 139–158. Springer, 1985. ISBN 978-1-4612-9567-9.

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

M. Pagnucco. *The role of abductive reasoning within the process of belief revision*. PhD thesis, University of Sydney, 02/1996 1996.

G. Paul. Approaches to abductive reasoning: An overview. *Artificial Intelligence Review*, 7(2):109–152, 1993.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988. ISBN 1558604790.

C. S. Peirce. *Collected Papers of Charles Sanders Peirce*, volume 5. Harvard University Press, 1958.

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

G. D. Plotkin. A note on inductive generalization. *Machine intelligence*, 5(1):153–163, 1970.

I. Pomeranz and S. Reddy. On undetectable faults and fault diagnosis. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 29(11):1832–1837, 2010. ISSN 0278-0070.

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

H. E. Pople, J. Myers, and R. Miller. Dialog: A model of diagnostic logic for internal medicine. In *IJCAI*, volume 4, pages 848–855. Citeseer, 1975.

H. Pople. The formation of composite hypotheses in diagnostic problem solving: an exercise in synthetic reasoning. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, volume 2, 1977.

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

S. Psillos. Inference to the best explanation and bayesianism. *Induction and deduction in the sciences*, pages 83–91, 2004.

A. Ram. A theory of questions and question asking. *Journal of the Learning Sciences*, 1 (3-4):273–318, 1991.

J. Reggia. Virtual lateral inhibition in parallel activation models of associative memory. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pages 244–248. Morgan Kaufmann Publishers Inc., 1985.

J. A. Reggia, D. S. Nau, and P. Y. Wang. A formal model of diagnostic inference. i. problem formulation and decomposition. *Information Sciences*, 37(1):227–256, 1985a.

J. A. Reggia, D. S. Nau, P. Y. Wang, and Y. Peng. A formal model of diagnostic inference, ii. algorithmic solution and application. *Information Sciences*, 37(1):257–285, 1985b.

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

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

G. Schurz. Patterns of abduction. *Synthese*, 164(2):201–234, 2008.

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

R. Stalnaker. Iterated belief revision. *Erkenntnis*, 70(2):189–209, 2009.

N. Tennant. On the degeneracy of the full AGM-theory of theory-revision. *Journal of Symbolic Logic*, 71(2):661–676, 2006. ISSN 0022-4812.

N. Tennant. *Changes of Mind: An Essay on Rational Belief Revision*. Oxford Univ Press, 2012.

P. Thagard. Explanatory coherence. *Behavioral and Brain Sciences*, 12:435–502, 1989.

J. Weisberg. Locating ibe in the bayesian framework. *Synthese*, 167(1):125–143, 2009.

G. Welch and G. Bishop. An introduction to the kalman filter. Technical Report TR 95-041, University of North Carolina at Chapel Hill, 1995.

M. Williams. Iterated theory base change: A computational model. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 1541–1549, 1995.

M. Williams. Anytime belief revision. In *International Joint Conference on Artificial Intelligence*, volume 15, pages 74–81, 1997.

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

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

Z. Zhuang, M. Pagnucco, and T. Meyer. Implementing iterated belief change via prime implicates. *AI 2007: Advances in Artificial Intelligence*, pages 507–518, 2007.