Agents, Epistemic Justification, and Defeasibility

Donald Nute

Department of Philosophy and
Artificial Intelligence Center
The University of Georgia
Athens, GA 30605, U.S.A.
dnute@uga.edu

Abstract. As software agents become more intelligent they will gather information from a wide range of sources. Inevitably, they will encounter conflicts in the information they receive. These conflicts will represent error since at least one of any set of incompatible propositions must be false. To be truly intelligent, these agents will need to diagnose and correct these errors. They will need an epistemology: principles and methods for determining when they are justified in accepting a proposition. The epistemological problems faced by software agents are similar to those that humans face. I will explore the possibility of building an adequate theory of epistemic justification based on recent work on nonmonotonic or defeasible reasoning. The theory to be sketched is intended to apply to both human and software agents. The theory owes much to the work of Robert Audi and John Pollock, but I consider both foundationalist and coherentist approaches where Audi and Pollock favor a foundationalist theory of defeasible epistemic justification.

1 Introduction

Error is the mother of epistemology. Attempts to develop an account of epistemic justification are most often motivated by the recognition that we sometimes believe something that is false. The importance of avoiding this kind of error often plays such a central role in epistemology that a theory of justification is expected to guarantee complete freedom from error. But while such a Cartesian ideal may not be impossible, it so impoverishes our epistemology that we find ourselves able to justifiedly believe very little. Most of human 'knowledge', including all of empirical science, must be abandoned if we demand epistemic certainty. It is, I would claim, not rational to cling to such a draconian account of justification. We must form a workable model of the world around us to effectively pursue our goals including our very survival. The epistemic maxim to minimize error cannot exceed in importance the maxim to maximize true beliefs. The price we must pay to have a stock of beliefs adequate for achieving our goals and adapting to a changing environment is that we must be willing to tolerate a reasonable risk of error.

Traditionally the quest for epistemic certainty has spawned foundational structures for knowledge. Some set of foundational beliefs are held to be self-justifying, certain, incorrigible. All other justified beliefs must be inferred from

these by deductively valid methods. But few beliefs can qualify for the position of foundations on this view, and deductively valid arguments severely restrict possible growth from these impoverished beginnings. Other more liberal accounts of the structure of epistemic justification appear to be needed. Coherence theories have been proposed to replace foundationalism. The coherentist holds that no belief is completely self-justifying, certain, or incorrigible. (Or more modestly, not beliefs about the sensible meet these qualifications.) Instead, our beliefs are arranged in a structure of justification where each is supported by the others. The beliefs 'cohere' with one another. This coherence justifies our acceptance of the beliefs as a set, and the coherence of the set of beliefs also justifies our acceptance of individual beliefs. While coherence theories would seem to avoid the error of giving error too much importance in our epistemologies, these theories run into problems of their own. One problem is to explain what the coherence of a set of beliefs consists of. Another is to show how an epistemic agent can establish the coherence of his beliefs and thus be in a position to claim that any of his beliefs are justified. A third problem is that coherence seems like it might be totally internal to a set of beliefs and independent of the way the sensible world actually is. There appears to be a danger that one could have a coherent set of beliefs that has little or no correspondence to the world.

Of course, foundationalism and coherentism do not exhaust the variety of epistemologies that have been proposed. But these two approaches to developing a theory of the structure of epistemic justification will be enough to fuel a discussion of the role of epistemology in artificial intelligence research. John McCarthy has claimed that AI has taken epistemology away from the philosophers. The thrust of this claim seems to be that we are ready to abandon the empty theorizing of the philosophers and get on with something practical and concrete. By building intelligent systems that possess knowledge, we will finally be able to point to our artifacts and say, "This is what knowledge is; this is how it works." John Pollock on the other hand cautions us that we will never build truly intelligent systems until our systems have an epistemology. I think the point Pollock is trying to make is something like this: We can build so-called knowledge based systems by representing what we take to be the knowledge of human experts. The question of how that knowledge is justified either is not addressed at all or it is settled before the knowledge is incorporated into the system. We may use confidence factors or assign probabilities of errors to bits of knowledge as we construct these knowledge bases and we can build systems that compute with these measures of confidence or error. But eventually we want to build systems that extract information and 'knowledge' directly from the world through a variety of means. Inevitably the 'knowledge' these systems derive from the world will involve inconsistencies, and sooner or later we will build systems that can detect these inconsistencies. At some point if our computing systems are to be truly intelligent they must discover and evaluate error themselves. When this happens, they will have to 'decide' what to believe and what to reject. They will need to have built into them a theory of justification, and epistemology.

Humans have evolved to receive information about the world through a particular set of sensory organs. Our memories and even our inferential mechanisms work in certain ways. Unlike other species of terrestrial animals, we are able to study our own bodies and our own behavior. We can develop descriptions of how we come to have the beliefs we do. We can also examine these processes critically and prescribe ways to modify them that promise both to reduce the false beliefs we have and to increase the true beliefs we discover. We can engage in both psychology and epistemology. When we build our machines, we have options that we don't have when it comes to our own bodies. We can design the senses our machines will have, construct huge redundant memories that work quite differently from our own, and build in whatever inferencing mechanisms we think will be most effective. We can also build into our machines an understanding of these very aspects that will allow the machines to recognize discrepancies in the information they receive and the conclusions they draw. And we have the opportunity to build in principles and mechanisms for resolving some or all of these discrepancies, for identifying and *correcting* error. Of course there will be both computational and theoretical limits on the abilities of our machines to examine all their beliefs and to detect and correct all their errors. And this is surely the most important goal of epistemology, *not to avoid error but to detect and correct it.*

The kinds of epistemologies we construct for our machines should be radically different from those we construct for ourselves. Of course the details will and should be different simply because our perceptual apparatuses and our memories work differently. But I see no reason why the *structure* of justification should be different. This structure should play two roles. First it should explain how beliefs are generated and how error is recognized and corrected. Second we should be able to articulate this structure at least partially and at least on some occasions well enough to justify some particular belief *explicitly.*

The kind of view I will examine here has its roots, most appropriately, in both the philosophical and AI communities. I will concentrate on foundationalism and coherentism as candidate theories for the structure of justification, but I will augment these accounts with recent work on nonmonotonic reasoning. I will take as my starting points Robert Audi's 'moderate foundationalism' and John Pollock's 'direct realism', two accounts of how we might develop a foundationalism with *defeasible* foundations. But I will also consider how a nonmonotonic reasoning system might be used to explain the notion of coherence in a coherentist theory of justification. With this initial review of the problems and issues, I will next present a particular account of nonmonotonic reasoning that I call defeasible logic. With this in hand, I will reexamine foundationalism and coherentism with particular attention given to the work of Audi and Pollock. I don't expect to build a conclusive case for a particular 'defeasible' epistemology. In fact, I will point out some features of such an account that some might consider unintuitive or even fatal flaws for such an account. Nevertheless, we can outline a theory of justification that answers some of the questions raised by traditional

versions of foundationalism and coherentism even though we will leave a great deal of work yet to be done.

2 Defeasible Logic and its Language

Suppose that $\vdash$ is the consequence relation of some formal system Σ. Then Σ is monotonic just in case for any two sets S and T of formulas and any formula ϕ of the language of Σ, if $S \vdash \phi$, then $S \cup T \vdash \phi$. We noted earlier that any reasoning system that preserves truth must be monotonic, but a reasoning system that preserves *justification* will *not* be monotonic. That means that a belief that ϕ might be justified based on our belief in some set of propositions S, but there could be a set of propositions T such that if we came to believe all the propositions in $S \cup T$ we would no longer be justified in believing ϕ.

The nonmonotonic formalism I will describe here first appeared as the foundation for a system of defeasible deontic logic in ([17]) and was later presented without the deontic operators in ([19]). Defeasible systems use rules whose consequents may not be detachable even when their antecedents are derivable ([10, 9, 6, 7, 24, 23, 16]). Detachment of the consequent of one of these *defeasible* rules may be *defeated* by a fact or another rule. The conflicting rule may either *rebut* the first rule by supporting a conflicting consequent, or it may simply *undercut* the first rule by identifying a situation in which the rule does not apply ([24]). Defeasible logic uses strict rules, defeasible rules, and undercutting defeaters.

We define atomic formulas in the usual way. A literal is any atomic formula or its negation. All and only literals are formulas of our language. Where ϕ is an atomic formula, we say ϕ and $\sim\phi$ are the complements of each other. $\neg\phi$ denotes the complement of any formula ϕ, positive or negative.

Rules are a class of expressions distinct from formulas. Rules are constructed using three primitive symbols: $\rightarrow$, $\Rightarrow$, and $\rightsquigarrow$. Where $A \cup \{\phi\}$ is a set of formulas, $A \rightarrow \phi$ is a *strict rule*, $A \Rightarrow \phi$ is a *defeasible rule*, and $A \rightsquigarrow \phi$ is an *undercutting defeater*. In each case, we call A the *antecedent* of the rule and we call ϕ the *consequent* of the rule. Where $A = \{\psi\}$, we denote $A \rightarrow \phi$ as $\psi \rightarrow \phi$, and similarly for defeasible rules and defeaters. Antecedents for strict rules and defeaters must be non-empty; antecedents for defeasible rules may be empty. We will call a rule of the form $\emptyset \Rightarrow \phi$ a *presumption* and represent it more simply as $\Rightarrow \phi$. While we allow variables in rules, we will treat such rules as schemata for all their instantiations.

Strict rules can never be defeated. They not only do not have exceptions, but could not have exceptions. We may think of them as expressing a necessary connection between antecedent and consequent. Examples of the kind of claim we would represent by a strict rule are "Bachelors are not married" and "Penguins are birds." Defeasible rules represent weaker connections which can be defeated. Examples are "Penguins live in Antarctica" and "Birds fly." An example of a presumption is "Presumably, there is no life on the moon." Undercutting defeaters are too weak to support an inference. Their role is to call into question

an inference we might otherwise be inclined to make. We represent such caveats using "might" as in "A damp match might not burn."

A defeasible theory includes an initial set of facts and a set of rules, but it must include more. Two rules with consequents ϕ and $\neg\phi$ conflict with each other and might defeat each other. Schurtz [27] suggests that this principle should be extended to any rules with incompatible consequents. Makinson suggests that we should detach the consequents of norms (that is, normative defeasible rules) iteratively "while controlling for 'consistency with the condition' in a piecemeal way" ([11], page 20.) But recognizing exactly when the consequents of a set of rules will lead to inconsistency is a serious problem, particularly for first-order logic where the question is not decidable. We will address this problem by explicitly incorporating the notion of a conflict set in our defeasible theories. Each conflict set will represent a minimal set of incompatible formulas, and a set of rules conflict if their consequents comprise a conflict set. We want the set of conflict sets in a theory to reflect the necessary relations embodied in our strict rules. If $\{\phi, \psi\}\rightarrow\chi$ is in our theory, then $\{\phi, \psi, \neg\chi\}$ should be one of our conflict sets. This idea will be extended below so that our conflict sets are "closed" under the strict rules of a theory.

Another component of a defeasible theory will be a precedence relation. This relation provides a way of adjudicating conflicts between conflicting rules. When we want to apply a defeasible rule $A\Rightarrow\phi$, we must look at all conflict sets to which ϕ belongs. In each such set, there must be one member ψ different from ϕ such that for every rule with consequent ψ either the antecedent of the rule fails or $A\Rightarrow\phi$ takes precedence over the rule. More will be said about the precedence relation after the definition of a defeasible proof has been presented.

The theories we will consider, then, each consist of a set of literals (representing initial facts about the world,) a set of rules, a set of conflict sets (closed under the strict rules in the theory,) and a precedence relation.

Definition 1. *A* **closed defeasible theory** *is a quadruple* $\langle F, R, C, \prec\rangle$ *such that*

1. *F is a set of formulas,*
2. *R is a set of rules,*
3. *C is a set of finite sets of formulas such that for every formula ϕ,*
   (a) $\{\phi, \neg\phi\} \in C$, and
   (b) for every $S \in C$ and $A\rightarrow\phi$ in R, if $\phi \in S$, then $A \cup (S - \{\phi\}) \in C$, and
4. *$\prec$ is an acyclic binary relation on the non-strict rules in R.*

3 Extension Semantics for Defeasible Logic

The most common approach to developing a semantics for nonmonotonic systems considers supersets of a theory that satisfy certain constraints. These supersets are often called extensions of the theory. Typically, every default rule in the system is either failed (the antecedent is not contained in the extension,) defeated,

or applied (its consequent is in the extension.) An extension is in some sense a *smallest* set that satisfies this requirement. The sense of 'smallest' used here is not always simply the set theoretic notion. A single theory may have multiple extensions. When a theory has multiple extensions, one option is to take the consequences of the theory to be the intersection of all the extensions of the theory, the so-called "skeptical" approach. In general, either there are no algorithms for generating the extensions of a theory or generating the extensions is computationally expensive. (However, see [4] for some work on testing to see if a formula is a member of an "admissible set" without having to generate the entire set.) Furthermore, a particular theory may not have an extension. Examples of the kind of semantics I have in mind include default logic ([26]), autoepistemic logic ([14, 8]), or in the deontic arena, allowed entailments ([15]). In at least some of these *extension* semantics, the extensions of a theory are the fix-points of a function from sets of propositions to sets of propositions, where the function used is defined in terms of the underlying nonmonotonic theory. Reiter's semantics for default logic is a prime example of a fix-point semantics.

The extension semantics presented here for our defeasible language is due to Donnelly ([3].) First, we want a set of literals that includes all the initial literals in a theory and that also *complies with* all the rules in that theory.

Definition 2. *Let T be a closed defeasible theory. A set K of literals is* **T-compliant** *if and only if*

1. $F_T \subseteq K$,
2. $\phi \in K$ *if there is* $A \rightarrow \phi \in R_T$ *such that* $A \subseteq K$, *and*
3. $\phi \in K$ *if there is* $A \Rightarrow \phi \in R_T$ *such that* $A \subseteq K$, *and for all* $S \in C_T$, *if* $\phi \in S$ *then there is* $\psi \in S - (F_T \cup \{\phi\})$ *such that*
   (a) *for all* $B \rightarrow \psi \in R_T$, $B \not\subseteq K$,
   (b) *for all* $B \Rightarrow \psi \in R_T$, *either* $B \not\subseteq K$ *or* $B \Rightarrow \psi \prec_T A \Rightarrow \phi$, *and*
   (c) *for all* $B \rightsquigarrow \psi \in R_T$, *either* $B \not\subseteq K$ *or* $B \rightsquigarrow \psi \prec_T A \Rightarrow \phi$.

Compliance with the rules in a theory is not enough. If it were, we could take our extensions to be those T-compliant sets having no proper subsets that are T-compliant. The problem with this approach is that we can still have gratuitous beliefs represented in such a set. Take for example a theory T for which $F_T = \{\phi\}$, $R_T = \{\phi \Rightarrow \psi, \chi \Rightarrow \theta, \theta \Rightarrow \chi, \{\chi, \theta\} \Rightarrow \sim\psi\}$, $\prec_T = \emptyset$, and C_T just contains all pairs of atomic formulas and their negations. Then intuitively we only want ϕ and ψ in our extensions of T. But the set $S = \{\phi, \sim\psi, \chi, \theta\}$ is also a smallest T-compliant set. We would say that from the point of view of T, belief in χ and θ is gratuitous. But each supports the other once they are accepted, and together they defeat the inference to ψ. However, if we remove both χ and θ from S, nothing in the remaining set $\{\phi\}$ together with the rules in R_T requires us to put either of χ and θ back into the set. That is, we can throw these two literals away and what is left does not force us to put them back. This should not happen with an extension of a defeasible theory.

Definition 3. *Let T be a closed defeasible theory. A set of literals E is a* **T-extension** *if and only if*

1. *E is T-compliant, and*
2. *there is no* $K \subseteq E$ *with* $K \neq \emptyset$ *such that*
   (a) $F_T \subseteq E - K$,
   (b) $\phi \notin K$ *if there is* $A \rightarrow \phi \in R_T$ *such that* $A \subseteq E - K$,
   (c) $\phi \notin K$ *if there is* $A \Rightarrow \phi \in R_T$ *such that* $A \subseteq E - K$, *and for all* $S \in C_T$, *if* $\phi \in S$ *then there is* $\psi \in (S - (F_T \cup \{\phi\})$ *such that*
      i. *for all* $B \rightarrow \psi \in R_T$, $B \not\subseteq E - K$,
      ii. *for all* $B \Rightarrow \psi \in R_T$, *either* $B \not\subseteq E - K$ *or* $B \Rightarrow \psi \prec_T A \Rightarrow \phi$, *and*
      iii. *for all* $B \rightsquigarrow \psi \in R_T$, *either* $B \not\subseteq E - K$ *or* $B \rightsquigarrow \psi \prec_T A \Rightarrow \phi$.

We will say that a literal ϕ is *defeasibly entailed* by a defeasible theory T just in case ϕ is in every T-extension.

Definition 4. *Where T is a closed defeasible theory and ϕ is a literal, $T \approx \phi$ if and only if for every T-extension E, $\phi \in E$.*

4 Proof Theory for Defeasible Logic

As was mentioned before, several well-known approaches to nonmonotonic reasoning develop an extension semantics but provide no proof theory to go with it. I will show in the next section that in at least some cases, no constructive proof theory is possible. In the case of defeasible logic, the proof theory came first and the semantics came much later. While a semantics without a proof theory might be intellectually satisfying in some respects, practical application is difficult or impossible. And it certainly isn't an attractive model for the nonmonotonic reasoning of ordinary people.[1] After we present the proof theory we will investigate the relationship between the proof theory and the semantics.

The defeasible logic presented here was presented in [19] and is a refinement of the system presented in [17]. Our proof theory will provide a constructive way to establish that a particular formula is derivable from a theory without having to generate an extension for the theory. Every theory will have a unique closure in the logic.

To apply a defeasible rule, it will sometimes be necessary to show that a conflicting rule is not satisfied, that is, that its antecedent conditions are not derivable. Thus, a proof will include both positive and negative defeasible assertions.

Definition 5. *σ is a* **positive defeasible assertion** *iff there is a defeasible theory T and a formula ϕ such that $\sigma = T \vdash\!\sim \phi$. σ is a* **negative defeasible assertion** *iff there is a defeasible theory T and a formula ϕ such that $\sigma = T \sim\!\dashv \phi$. σ is a* **defeasible assertion** *iff σ is either a positive defeasible assertion or a negative defeasible assertion.*

[1] Of course, we should want our formal systems to *resemble* the reasoning system of ordinary people, but we generally don't want a formal system that mirrors *exactly* what ordinary people do. After all, ordinary people often reason very badly. What we want to do is to discover those patterns where their reasoning is at its best and then try to extend those patterns to correct other cases where their reasoning goes astray.

A negative assertion $T \sim\!\!| \ \phi$ is intended to make a stronger statement than $T \not\sim \phi$. $T \sim\!\!| \ \phi$ indicates that there is a demonstration that ϕ does not follow from T. Before we can detach the consequent of a defeasible rule, we will need to establish that it is not defeated by some conflicting rule. To do this, we will often need to show that the antecedent of a conflicting rule cannot be satisfied. What we need, then, are complementary notions of derivation and refutation. We will sometimes need to show that some formula is refutable in order to show that another formula is derivable, and we will also sometimes need to show that some formula is derivable in order to show that another formula is refutable. Of course, by "refutable" we do not mean that a formula can be shown to be false. Instead, we mean only that we can show that it is not derivable. I think both of these uses of the terms "refute" and "refutation" occur in ordinary usage, but it is important to keep in mind which is intended here.

Our defeasible proofs will have a tree-structure rather than the usual linear structure. Those familiar with logic programming will know that a query can succeed, fail finitely, or fail infinitely. A query fails infinitely when an attempt to prove it can proceed indefinitely without succeeding and without failing in the usual sense. One way a query can fail infinitely is when there is circularity in a theory and the same new query to be proved occurs repeatedly. For example, if our theory contains no formulas and only the single rule $\phi \rightarrow \phi$, then the use of the rule $\phi \rightarrow \phi$ in an attempt to find a proof can lead us to try to prove ϕ by proving ϕ *ad infinitum*. This is what an automated theorem prover will do if it has no way to check for loops; but of course a proof theory is not committed to any particular method for generating proofs. All that is necessary for ϕ to follow from T is that there is one successful proof of ϕ from T. However, when showing that T *refutes* ϕ, we will need to show that *all* efforts to prove ϕ from T must fail. One way an attempt can fail is by looping. Giving our proofs a tree structure makes it possible to recognize and use such infinite failures within the proof theory. Alternatively, we might try a labeled logic *a lá* Gabbay [5] where the labels carry the same information as the descendants of a node in a proof tree.

Definition 6. *$\mathcal{T}$ is a* **defeasible argument tree** *iff $\mathcal{T}$ is a finite tree and there is a defeasible theory $th(\mathcal{T})$ such that for every node n in $\mathcal{T}$ there is a formula ϕ such that n is labeled either $th(\mathcal{T}) \ \mid\!\sim \phi$ or $th(\mathcal{T}) \ \sim\!\!| \ \phi$.*

Definition 7. *Let $\mathcal{T}$ be a defeasible argument tree and let n be a node in $\mathcal{T}$. The* **depth of** *n* **in** *$\mathcal{T}$* **is** *k ($dp(n, \sigma) = k$) iff n has $n-1$ ancestors in $\mathcal{T}$. The* **depth of** *$\mathcal{T}$* **is** *k ($dp(\mathcal{T}) = \|$) iff $k = max\{j$: there is a node $m \in \sigma$ such that $dp(m, \mathcal{T}) = |\}$.*

So far, we have used the notion of the antecedent of a rule succeeding or failing informally. Before presenting our basic defeasible proof theory, we will say precisely what it means for sets of formulas to succeed or fail at a node in a defeasible argument tree.

Definition 8. *Let $\mathcal{T}$ be a defeasible argument tree, $th(\mathcal{T}) = T$, n be a node in $\mathcal{T}$, and A be a set of formulas.*

1. *A* **succeeds at** *n iff for all* $\phi \in A$, *n has a child labeled* $T \mathrel{|\!\sim} \phi$.
2. *A* **fails at** *n iff there is* $\phi \in A$ *such that n has a child labeled* $T \sim\!\!|\ \phi$.

Definition 9. $\mathcal{T}$ *is a* **defeasible proof** *(***d-proof***) iff* $\mathcal{T}$ *is a defeasible argument tree,* $th(\mathcal{T}) = T$, *and one of the following conditions holds for every node n in* $\mathcal{T}$.

1. *n is labeled* $T \mathrel{|\!\sim} \phi$ *and either*
   (a) $\phi \in F_T$,
   (b) [**Strict Detachment**] *there is* $A{\rightarrow}\phi \in R_T$ *such that A succeeds at n, or*
   (c) [**Defeasible Detachment**] *there is* $A{\Rightarrow}\phi \in R_T$ *such that A succeeds at n and for all* $S \in C_T$, *if* $\phi \in S$ *then there is* $\psi \in S - (F_T \cup \{\phi\})$ *such that*
      i. *for all* $B{\rightarrow}\psi \in R_T$, *B fails at n,*
      ii. *for all* $B{\Rightarrow}\psi \in R_T$, *either B fails at n or* $B{\Rightarrow}\psi \prec_T A{\Rightarrow}\phi$, *and*
      iii. *for all* $B{\rightsquigarrow}\psi \in R_T$, *either B fails at n or* $B{\rightsquigarrow}\psi \prec_T A{\Rightarrow}\phi$.
2. *n is labeled* $T \sim\!\!|\ \phi$ *and*
   (a) $\phi \notin F_T$,
   (b) [**Failure of Strict Detachment**] *for all* $A{\rightarrow}\phi \in R_T$, *A fails at n, and*
   (c) [**Failure of Defeasible Detachment**] *for all* $A{\Rightarrow}\phi \in R_T$, *either*
      i. *A fails at n, or*
      ii. *there is* $S \in C_T$ *such that* $\phi \in S$ *and for all* $\psi \in S - (F_T \cup \{\phi\})$, *either*
         A. *there is* $B{\rightarrow}\psi \in R_T$ *such that B succeeds at n,*
         B. *there is* $B{\Rightarrow}\psi \in R_T$ *such that B succeeds at n and* $B{\Rightarrow}\psi \not\prec_T A{\Rightarrow}\phi$, *or*
         C. *there is* $B{\rightsquigarrow}\psi \in R_T$ *such that B succeeds at n and* $B{\rightsquigarrow}\psi \not\prec_T A{\Rightarrow}\phi$.
3. [**Failure by Looping**] *n is labeled* $T \sim\!\!|\ \phi$, *n has an ancestor m in* $\mathcal{T}$ *such that m is labeled* $T \sim\!\!|\ \phi$, *and every node in* $\mathcal{T}$ *between n and m is labeled with a negative defeasible assertion.*

The Failure by Looping condition in Definition 9 requires justification. To apply a defeasible rule, we sometimes need to show that a conflicting rule is not satisfied. We do this by showing that every attempt to derive the antecedent of the conflicting rule fails. Suppose the antecedent contains ϕ. Then we must explore every way that we might derive ϕ. If we discover one way of deriving ϕ that requires us to derive ϕ, then we can be sure that if ϕ is to be derived at all then there must be some way of doing it that is not circular. Thus, we can reject any circular attempts to establish ϕ. This in no way allows us to avoid examining all the non-circular ways that ϕ might be established. What Failure by Looping amounts to is the recognition that if there is no non-circular way to establish a formula, then there is *no* way to establish it.

Definition 10. ϕ **is defeasibly derivable from** T *(*$T \mathrel{|\!\sim_D} \phi$*) iff there is a d-proof* $\mathcal{T}_\phi$ *such that the top node in* $\mathcal{T}_\phi$ *is labeled* $T \mathrel{|\!\sim} \phi$.

Definition 11. ϕ **is defeasibly refutable in** T *($T \sim\!|_D\ \phi$) iff there is a d-proof $\mathcal{T}_\phi$ such that the top node in $\mathcal{T}_\phi$ is labeled $T \sim\!| \phi$.*

For $\mathcal{T}_\phi$ as described in Definition 10 or 11, we will say that $\mathcal{T}_\phi$ *establishes* $T \mid\!\sim_D \phi$ or $T \sim\!|_D \phi$ respectively.

Definition 12. *A set of formulas A* **is defeasibly derivable from** T *($T \mid\!\sim_D A$) iff for all $\phi \in A$, $T \mid\!\sim_D \phi$.*

Definition 13. *A set of formulas A* **is defeasibly refutable in** T *($T \sim\!|_D A$) iff there is $\phi \in A$ such that $T \sim\!|_D \phi$.*

Given our informal notion of refutation, it should not be possible to both derive and refute the same formula from a defeasible theory. After all, a refutation is supposed to be a demonstration that the refuted formula *cannot* be derived. The following theorem establishes this essential property for our basic defeasible logic.

Theorem 1. **[Coherence]** *If $T \sim\!|_D \phi$, then $T \not\mid\!\sim_D \phi$.*

Another important property of our proof theory is that defeasible rules cannot produce any new contradictions. Any contradictions derivable in the theory depend only on the strict rules and the initial set of literals in the theory.

Definition 14. *$T \vdash_D \phi$ iff $\langle F_T, \{A \rightarrow \psi : A \rightarrow \psi \in R_T\}, C_T, \prec_T \rangle \mid\!\sim_D \phi$.*

Theorem 2. **[Consistency]** *If $S \in C_T$ and for all $\phi \in S$, $T \mid\!\sim_D \phi$, then for all $\phi \in S$, $T \vdash_D \phi$.*

Now we return to the relationship between our proof theory and our semantics. Donnelly [3] proves the following results.

Theorem 3. **[Soundness]** *If $T \mid\!\sim_D \phi$, then $T \models\!\!\approx \phi$.*

Theorem 4. *If $T \sim\!|_D \phi$, then ϕ is in no T-extension.*

Donnelly also shows that our proof theory is *not* complete with respect to our semantics. Makinson and Schlecta [12] introduce the notion of a "floating conclusion". This is a formula that belongs to the intersection of all extensions of a theory even though no rule supporting the formula is satisfied in every extension. Floating conclusions are possible in our semantics. But since our proof theory is sound, we will be unable to derive any floating conclusions. So long as a theory has multiple extensions, there will be the possibility of floating conclusions. This implies a more general conclusion, not just the conclusion that the defeasible logic presented here is not complete. Given *any* extension semantics for nonmonotonic reasoning and *any* proof theory for that semantics, if floating conclusions are possible in that semantics then the proof theory cannot be both sound and complete. I will present an argument for this in the next section.

5 Soundness and Incompleteness

Audi and Pollock both propose varieties of foundationalism where foundational beliefs are defeasible. Does a theory of defeasible reasoning fit better with a foundationalist than with a coherentist epistemology? Defeasible reasoning has a distinctly coherentist flavor. Pollock and Cruz propose that S's belief at t that ϕ is justified just in case S instantiates at t an undefeated argument that has ϕ as its conclusion. At first glance, this sounds like a foundationalist principle. The arguments S could instantiate are finite and presumably their ultimate premises must be foundational beliefs for S. But the requirement that these arguments must be *undefeated* by any of S's other beliefs suggests a holistic account, a coherentist account.

The defeasible logic presented in this paper includes a constructive proof theory and an extension semantics. Donnelly ([3]) showed that this defeasible logic is sound with respect to the semantics. However, Donnelly also showed that the proof-theory is *not* complete with respect to the semantics. Reiter ([26]), McDermott and Doyle ([13]), Moore ([14]), and others have presented nonmonotonic formalisms based on an extension semantics without providing a proof theory for their semantics. We can show that there cannot be a sound and complete constructive proof theory for most of these semantics.

First, let's clarify what we mean by a constructive proof theory. In the languages of most nonmonotonic formalisms there are expressions that correspond to the defeasible rules in defeasible logic. These expressions include components clearly identifiable as the antecedent and consequent of the expression. For example, in autoepistemic logic we have expressions of the form $\phi \wedge L\psi \supset \chi$ where L is an epistemic operator which we read as 'for all I know'. $\{\phi\}$ is the antecedent of the conditional and χ is the consequent, at least in the sense I will use these terms here. We will give the expressions that play this role the generic name *defaults*. Suppose a nonmonotonic formalism Σ has a proof theory that defines a derivability relation $\mid\!\sim_{\Sigma}$. We will call a sequence σ of Σ-formulas a K-argument in Σ iff for every member of σ_i of σ, either

1. σ_i is a member of K, or
2. σ_i is derivable from $\{\sigma_1, \ldots, \sigma_{i-1}\}$ using only the first order fragment of Σ, or
3. for some $j \leq i$, there is a default in K with antecedent A and consequent σ_i such that $A \subseteq \{\sigma_1, \ldots, \sigma_{i-1}\}$.

We will say that the proof theory for Σ is *constructive* if and only if for every Σ-theory K and every Σ-formula ϕ, $K \mid\!\sim_{\Sigma} \phi$ *only if* there is a K-argument σ in Σ such that $\sigma_{\ell(\sigma)} = \phi$ and for every $i \leq \ell(\sigma)$, $K \mid\!\sim_{\Sigma} \sigma_i$. Of course, this is not in general a sufficient condition for $K \mid\!\sim_{\Sigma} \phi$.

We will say that an extension semantics for Σ is *grounded* if and only if for every Σ-extension E of a Σ-theory K and every Σ-formula ϕ, if $\phi \in E$ then there is a K-argument σ in Σ such that $\sigma_{\ell(\sigma)} = \phi$ and $\{\sigma_1, \ldots, \sigma_{\ell(\sigma)-1}\} \subset E$. Of course, this is not a sufficient condition for $\phi \in E$, but for most semantics that have been proposed for nonmonotonic systems it is a necessary condition.

However, groundedness is not generally a condition satisfied by the *intersection* of the set of extensions of a theory. Makinson and Schlecta ([12]) introduce the notion of a floating conclusion for an extension semantics. Adapting this notion to our discussion, we will define a Σ-formula ϕ to be a *floating conclusion* of a Σ-theory K just in case ϕ is a member of every Σ-extension for K, but there is no K-argument σ in Σ such that $\sigma_{\ell(\sigma)} = \phi$ and $\{\sigma_1, \ldots, \sigma_{\ell(\sigma)-1}\}$ is contained in every Σ-extension for K.

Our semantics for defeasible logic admits floating conclusions as does pretty much every extension semantics that has been proposed for an nonmonotonic formalism. And that is why there cannot be a constructive nonmonotonic proof theory that is both sound and complete with respect to any of these semantics. For suppose there were and suppose ϕ were a floating conclusion for the theory K. Then if our proof theory were complete, we would have $K \mathrel{|\!\sim}_{\Sigma} \phi$. And since our proof theory is constructive, there must be a Σ-argument σ such that $\sigma_{\ell(\sigma)} = \phi$ and for every $i \leq \ell(\sigma)$, $K \mathrel{|\!\sim}_{\Sigma} \sigma_i$. Since ϕ is a floating conclusion for K, there must be some $i \leq \ell(\sigma)$ and some Σ-extension E of K such that $\sigma_i \notin E$. However, since our proof theory for Σ is sound and since $K \mathrel{|\!\sim}_{\Sigma} \sigma_i$, $\sigma_i \in E$, a contradiction.

6 Priorities, Specificity, and Normative Reasoning

One method for determining priorities of rules in defeasible theories is to use *specificity*. A rule with antecedent A is said to be more specific than a rule with antecedent B, relative to a theory T, if we can derive all of B from A using only the rules in T, but not *vice versa*. The precedence relation in a theory might be based on such a notion of specificity. A more interesting case, though, is where we already have some explicit relation on the rules of a theory that we want to use as the *core* of our precedence relation. We might then extend this core precedence relation by using specificity in all those cases where the core precedence relation does not settle the matter. To state this condition precisely, we will let $R^\circ = \{r : r \in R \text{ and the antecedent of } r \text{ is not empty }\}$.

Definition 15. *Let $\Gamma \subseteq \prec_T$. Then T is* **Γ-specific** *iff for all non-strict $r_1, r_2 \in R_T$ such that $(r_1, r_2) \notin \Gamma$, A is the antecedent of r_1, and B is the antecedent of r_2,*

1. *if $\langle A, R_T^\circ, C_T, \prec_T\rangle \vdash_D B$ and $\langle B, R_T^\circ, C_T, \prec_T\rangle \sim\!|_D A$, then $r_2 \prec_T r_1$, and*
2. *if $\langle A, R_T^\circ, C_T, \prec_T\rangle \sim\!|_D B$ or $\langle B, R_T^\circ, C_T, \prec_T\rangle \vdash_D A$, then $r_2 \not\prec_T r_1$.*

A special case will be where we use specificity as our sole criterion for adjudicating conflicts between non-strict rules. This amounts to taking the empty set as our core precedence relation.

Definition 16. *A theory T* **preserves specificity** *iff T is $\emptyset$-specific.*

An advantage of using Γ-specific or specificity preserving theories is that we can now *compute* whether one rule takes precedence over another. This becomes particularly important when we add a deontic operator O (for 'ought') to our

logic and represent norms (rules with deontic consequents.) The proof theory for deontic defeasible logic requires additional principles to handle the interaction between norms and other rules and to resolve some of the paradoxes in standard deontic logics. Such a deontic extension of defeasible logic is described in [17, 18].

Let's call a theory T that contains norms a *primary* theory, and let's call a theory $T^{\prec}$ that contains rules about which rules in T take precedence over which other rules in T a *precedence* theory for T. By doing proofs or refutations in $T^{\prec}$, we can determine in at least some cases whether $A{\Rightarrow}\phi \prec_T B{\Rightarrow}\psi$. Specificity is one way to assign priorities for rules, but there are others. The principles of *lex superior* and *lex posterior* are familiar examples. Suppose we have a theory T containing rules representing both federal laws of the United States and laws of some particular state within the United States. In the language of a second theory $T^{\prec}$, we have names for all the rules in T and predicates **federal**, **state** and $\sqsubset$. Then $T^{\prec}$ can contain all instances of the form

$$\{\mathbf{federal}(r_1), \mathbf{state}(r_2)\}{\Rightarrow} r_2 \sqsubset r_1.$$

This would be a special instance of the principle of *lex superior*. Let

$$\Gamma = \{\langle r_1, r_2\rangle : T^{\prec} \mid\!\sim_D r_1 \sqsubset r_2\}$$

and suppose T is Γ-specific. Now we can use a combination of *lex superior* and specificity to determine precedence among the laws in our theory T of normative use. And we can use proofs and refutations in $T^{\prec}$ to determine the core of the precedence relation for T. Our complete defeasible theory now has two components, a primary theory and a theory of precedence for the primary theory. Notice that *lex superior* takes priority (at a higher level) over specificity since the definition of Γ-specificity guarantees that specificity is only applied when the core precedence relation Γ does not determine priority.

Clearly, matters become more complex when we try to add *lex posterior* to our precedence theory and when we try to include the laws of all fifty states in the U.S. It may even turn out to be impossible to do this without violating the requirement that $\prec_T$ must be non-cyclical. It is an open question whether for any non-cyclic relation Γ on the non-strict rules in a set R of rules, any set F of facts, and any conflicts set C for the language of F and R, it is possible to construct a theory $T = \langle F, R, \prec, C\rangle$ that is Γ-specific.

Notice in the example of *lex superior*, the rule used to express this principle is defeasible. But why shouldn't this be a strict rule? Indeed, why should we need defeasibility at all in a precedence theory? The doctrine of states' rights found in the U.S. Constitution illustrates the need for defeasible rules in precedence theories. Once again, *lex superior* should tell us that Constitutional law should take precedence over either federal or state law. So we get a nice, neat three-level normative system. But the principle of states' rights is that the Constitution prohibits the federal government from enacting laws limiting the powers of the state governments in certain areas. If this principle of states' rights were to prevail, then the principle of *lex superior* would have to be defeasible. It would

have exceptions in any case where a federal law contradicted a state law and the subjects of these two conflicting laws fell under the umbrella of the Constitutional guarantee of states' rights.

7 Defeasible Epistemic Justification

The idea that I want to explore is that defeasible logic or something like it might provide the underlying structure for a system of epistemic justification. The basic notion is that a belief would be justified either if it were defeasibly derivable from some defeasible theory or if it were included in some extension(s) of some defeasible theory. The immediate question, of course, is the source of the initial theory.

We might begin building our defeasible theory with facts and strict rules. These would, presumably, represent analytic propositions or relations that we can know directly. Examples might include 'Squares have four sides.' or 'Bachelors do not have wives.' But we don't have to start with facts and strict rules. Indeed, we don't have to have any facts or strict rules in the defeasible theory at all. We can derive propositions or build extensions from defeasible theories that consist entirely of defeasible rules. In place of facts, we would have *presumptions*, that is, defeasible rules with empty antecedents.

Supposing for a moment that the defeasible theory upon which our justificational structure rests consists of defeasible rules only, where would we get these defeasible rules? Then answer is that they would come from the usual sources: perception, memory, testimony, etc. Robert Audi has proposed what he calls a *modest foundationalism* that incorporates a feature like this. Audi characterizes foundationalism generically in two ways, one regarding knowledge and the other regarding justification. It is his account of foundationalism regarding justification that concerns us here:

> II. For any S and any t, the structure of S's body of justified beliefs is, at t, foundational, and therefore any inferentially (hence non-foundationally) justified beliefs S has depend on non-inferentially (thus in a sense foundationally) justified beliefs of S's. ([2], p. 179)

Audi expands this generic account to describe *fallibilist foundationalism*.

> III. For any S and any t, (a) the structure of S's body of justified beliefs is, at t, foundational in the sense indicated by thesis II; (b) the justification of S's foundational beliefs is at least typically defeasible; (c) the inferential transmission of justification need not be deductive; and (d) non-foundationally justified beliefs need not derive *all* their justification from foundational ones, but only enough so that they would remain justified if (other things remaining equal) any other justification they have (say, from coherence) were eliminated. ([2], p. 179)

What exactly does Audi mean when he says that some of our beliefs may be both foundational and defeasible? Let's consider what he says on another occasion

about perception. Audi considers two principles for perception. One is called the *vision principle*:

> ...when a visual belief arises in such a way that one believes something in virtue of either seeing *that* it is so or seeing it *to be* so, normally the belief is justified and is always prima facie justified.) ([1], p. 25.)

The second, weaker principle Audi calls the *visual experience principle*:

> ...when on the basis of an apparently normal visual experience (such as the sort we have in seeing a bird nearby), one believes something of the kind the experience seems to show (for instance that the bird is blue), normally this belief is justified. ([1], p. 25.)

Audi then proposes similar principles for the other senses, and later for memory and testimony:

> ...normally, if one has a clear and confident memory belief that one experienced a given thing, then the belief is justified. Similarly, we might call such beliefs prima facie justified. ([1], p. 68.)

> ...we might say that at least normally, a belief based on testimony is thereby justified (that is, justified on the basis of the testimony) provided the believer has overall justification for taking the attester to be credible regarding the proposition in question. ([1], p. 138.)

Audi gives many examples of how defeasible or prima facie justification might be defeated, but he does not attempt to develop a general theory of defeasible reasoning or to develop a formal system of defeasible reasoning. How can we incorporate Audi's account into a theory of epistemic justification based on defeasible logic? First, I suggest that being in a perceptual situation, having a memory experience, or receiving testimony can cause us to believe a proposition based on the perception, memory, or testimony, and it can also cause us to accept a presumption concerning that same proposition. Remember that a presumption is a defeasible rule with an empty antecedent (represented symbolically by $\Rightarrow\phi$,) and rules do not represent formulas in defeasible logic. Believing that ϕ and accepting $\Rightarrow\phi$ are two quite different things. This difference is all the greater if we hold that only formulas may express propositions. What I am suggesting, then, is that the perceptual, memorial, or testimonial experience can generate both belief in a proposition and acceptance of a presumption (a defeasible rule.) Ultimately, the belief is only justified if it is derivable from or belongs to some extension(s) of a defeasible theory containing the corresponding presumption. We might say that the belief is *situationally justified* if it is indeed supported by the believer's defeasible theory in the appropriate way, but that the believer *justifiedly believes* something only if she believes it in part because she notices that it is supported by her defeasible theory. Of course, a believer might also believe a proposition on the basis of a perceptual experience even though she has accepted other defeasible rules (including other presumptions arising from

perception, memory, or testimony) that defeat the presumption corresponding to the present perceptual, memorial, or testimonial belief. In that case, of course, the believer's belief is *not* justified on the account we are developing.

There is some difficulty in squaring this account with Audi's proposals. For Audi, foundational beliefs are non-inferential. But on the view I am developing, at least in the foundational version, perceptual and other beliefs that are usually candidates for foundational beliefs are only justified if they are derivable from the believer's defeasible theory. But notice that the foundational belief does not depend for its justification on an inference from some independent reason for holding the belief. Rather the purpose of the derivation is to show that the belief is not *defeated* by some other parts of the defeasible theory that the believer has accepted. If nothing else in the believer's defeasible theory competes with the presumption that corresponds to the belief, then the belief is an immediate consequence of the assumption. Since the presumption is a defeasible rule with an empty antecedent condition, the belief in the proposition expressed in the head of the presumption does not depend on inference from any other belief.

John Pollock also supports a form of foundationalism based upon defeasible reasoning. But unlike Audi, Pollock has extensively explored the nature of defeasible reasoning and attempted to capture it in a formal system (for example in [20–22, 24, 23].) Not surprisingly, Pollock (and his co-author Joseph Cruz) are fully aware of the inferential nature of justification even in the case of foundational beliefs.

> We can then understand the foundationalist as asserting that belief in P is justified for a person S if and only if S instantiates an undefeated argument supporting P. ([25], p. 38)

I believe Pollock intends this principle to apply to foundational as well as to what Audi would call "inferential" beliefs. Certainly Pollock's account of defeasible reasoning is different from the account offered here in many details. However, I believe the nondoxastic foundationalism that Pollock calls *direct realism* is largely compatible with the account of the role that perception, memory, and testimony play in generating elements of an initial defeasible theory that plays a key role in a correct theory of justification.

Leaving aside any facts and strict rules that might be included in a believer's defeasible theory, there are at least two other important sets of defeasible rules to consider besides presumptions. These are ordinary defeasible rules such as those expressed in English as 'Birds fly' or 'Matches burn when you strike them'. A better reading of these as *rules* would be 'Take something's being a bird as evidence that it flies' or 'Take something's being a match as evidence that it will burn if you strike it.' These do not represent propositions and they are not the content of beliefs. Instead they are imperatives - rules - and they can be accepted. A believer might add them to her defeasible theory as a result of testimony or through reflection on other beliefs (through induction, for example.) We can also have a second tier of defeasible rules that refer to other defeasible rules in the lower tier. These are rules that determine the precedence relation for the lower tier in the theory. How are these rules generated?

Pollock and Cruz discuss *epistemic norms* which both govern the generation and the correction of our beliefs.

> The concept of epistemic justification can therefore be explained by explaining the nature and origin of the epistemic norms that govern our reasoning. We have been calling this "the procedural concept of epistemic justification". ([25], pp. 123-124.)

> The internalization of norms results in our having "automatic" procedural knowledge that enables us to do something without having to think about how to do it. It is this process that we are calling "being guided by the norm without having to think about the norm". This may be a slightly misleading way of talking, because it suggests that somewhere in our heads there is a mental representation of the norm and that mental representation is doing the guiding. Perhaps it would be less misleading to say that our behavior is being guided by our procedural knowledge and the way in which it is being guided is described by the norm. What is important is that this is a particular way of being guided. It involves non-intellectual psychological mechanisms that both guide and correct (or fine tune) our behavior. ([25], p. 128.)

I propose that these psychological mechanisms cause us both to form beliefs and to accept the defeasible rules in our theories. Our defeasible reasoning faculty is then a major component of the corrective mechanism that allows us to "fine tune" our beliefs. We *belief* a proposition or *accept* a defeasible rule because of the action of these mechanisms, whether or not the belief is justified. We can be *situationally justified* in believing a proposition because that proposition is supported (proof-theoretically or semantically) by our defeasible theory whether or not we believe that proposition. The perceptual, inductive, reflective, and other mechanisms that produce a believer's acceptance of a particular defeasible rule does not need to be completely reliable since the defeat mechanisms built into our reasoning allows us to identify many of the defeasible rules that support incorrect propositions.

Notice that we can believe a proposition without accepting the presumption that would support it, and we can accept a presumption without believing the proposition it supports. A proposition ϕ might be the consequent of a defeasible rule we accept that has a non-empty antecedent. Because that rule and others are included in our defeasible theory, our defeasible theory might support ϕ and we might believe that ϕ for that reason. On the other hand, we might include $\Rightarrow\phi$ in our theory but not believe that ϕ because the presumption is defeated. For example, I might think I see my best friend in a store, but then remember that he is at a conference in another town. We could interpret this as a case where the presumption was accepted but the belief was defeated. If I later learn that my friend cancelled his trip, the conclusion from the presumption can be reinstated and I may again believe that I saw my friend at the store.

I have said that it is not as critical that the processes or mechanisms that lead us to accept defeasible rules into our theories be completely reliable because

defeasible reasoning allows us to correct for some of our errors. But we still might ask whether a particular defeasible rule is epistemically justifiable. If we do this, then we are extending the notion of epistemic justifiability to include not only beliefs but also the policies we accept concerning what we will and will not take as evidence for particular propositions. I think Pollock may have something like this in mind as well when he talks about epistemic norms. If we accept a defeasible rule and then find that each opportunity where we might apply it is one where it is defeated, we might reject the rule. But memory might also play a role in such cases. If we accept a defeasible rule and then either have no opportunity to apply it or find that each opportunity where we might apply it is one where it is defeated, we may forget the rule in time. However this may be, I will not explore any further the issues involved in providing justification for the defeasible rules we accept in this paper.

The picture I am proposing, then, is that perception, memory, reasoning, reflection, and other processes produce presumptions, defeasible rules about connections between concepts, and defeasible rules about the precedence of other defeasible rules. Along with Pollock, I tip my hat to the reliabilists that these defeasible rules should play a role in the justification of our beliefs because the processes that produce them are reliable (but still fallible.) Actual beliefs are then justified because they are supported, proof-theoretically or semantically, by the defeasible theory whose contents are the results of these processes.

8 Coherence and Defeasibility

Let's call perception, memory, generalization and other reliable processes that produce the presumptions, defeasible rules, and other components of a defeasible theory that an epistemic agent accepts *belief initiators.* Let's call the set of presumptions, defeasible rules, etc., produced by an agents belief initiators his *initial theory.* Then we could define an agent's set of beliefs as *coherent* just in case the set of propositions he believes is a nonmonotonic extension of the agent's initial theory. In this section we will consider some of the consequences of adopting a coherentist theory of justification that explains coherence in terms of nonmonotonic extensions of an agent's initial theory.

This proposal is not intended to rely too heavily on the defeasible logic presented earlier in this paper. Defeasible logic is essentially propositional and may be too weak for our purposes. Some other nonmonotonic formalism may serve our purposes better. But whatever formalism is used will be assumed to include a semantics that involves multiple extensions.

Using nonmonotonic extensions to define coherence explains how a coherent set of beliefs can depend on the way the world is without requiring that any perceptual, introspective, or other beliefs are entirely self-justifying, certain, or incorrigible. An agent can accept a presupposition generated by perception without believing the proposition that presupposition supports. This may happen because the presupposition is defeated by other presuppositions or defeasible rules that the agent accepts. Perception, for example, generates presuppositions

and thereby initiates the processes that result in what we call perceptual beliefs, but perception is also self-correcting. Not only may perception be corrected by perception, but it may also be corrected by memory or other belief initiators that generate presuppositions or defeasible rules that defeat a particular perceptual presupposition. Only those presumptions that are not defeated produce belief.

Since a single initial theory may have multiple extensions regardless of which nonmonotonic formalism we adopt, there is the (logical) possibility that two agents could have the same initial theory but hold different and incompatible sets of beliefs each of which is coherent for the agent that holds it. Or to put the matter a bit differently, even allowing that only complete extensions of an agent's initial theory should count as the content of a coherent belief state for an agent, there will be several incompatible sets of beliefs that an agent could coherently hold. What can we say to this? It seems that the only thing we can say is that if an agent's belief set is a nonmonotonic extension of his initial theory, then the agent is justified in believing everything that he believes. That means that one and the same agent could be justified in believing ϕ and could also, without any change in his experience (that is, in his initial theory) be justified in believing $\sim\phi$, although of course he would not be justified in believing $\phi \wedge \sim\phi$. This may seem counterintuitive, but it is a consequence of the position we are sketching. The suggestion of some possibility like this has often been offered as a criticism of coherentist theories of justification. Notice, though, that additional perceptual or other experience could generate new presuppositions and defeasible rules, extending our agent's initial theory. The extended theory will in general not have the same extensions as the earlier theory. In fact, since our extensions are nonmonotonic, it will generally not even be the case that any extension of the earlier theory will be contained in any extension of the later, extended theory. An agent who at time t_1 could coherently hold a set of beliefs that includes a belief that ϕ or coherently hold a set of beliefs that includes a belief that $\sim\phi$ might not at a later time t_2 be able to coherently hold a set of beliefs that includes a belief that ϕ (or that $\sim\phi$.) Additional experience can even correct beliefs that we *justifiedly* held at an earlier time.

The position that an agent is justified in believing all the propositions in some nonmonotonic extension of his initial theory is the kind of position that is called *credulous* in the literature on nonmonotonic reasoning. A corresponding *skeptical* position would be that an agent is justified in believing a proposition only if it belongs to *every* nonmonotonic extension of his initial theory. But the intersection of the nonmonotonic extensions of a theory will normally not itself be an extension of the theory. So we would have to change our account of what it is for a set of beliefs to be coherent. There are computational problems in determining whether a set of propositions is a nonmonotonic extension of a defeasible theory. And there may be no effective method for constructing all extensions, or even one extension, of a defeasible theory. At best we may have to guess an extension which we may then be able to verify. This is not to say that the problem of generating an extension is NP-hard; the problem may be much worse than that. For the problem is not that constructing a nonmonotonic extension of

a defeasible theory is intractable; it may be impossible. The skeptical approach would require us to construct *all* extensions and then find their intersection.

Perhaps some connectionist mechanism can help us with at least the credulous approach to a defeasible coherentism. Perhaps an initial theory can be represented in some sort of connectionist architecture. Then by inputting activation into the nodes corresponding to the presuppositions in the theory and updating activation of nodes in an asynchronous way, the network will eventually stabilize with patterns of activated nodes corresponding to the propositions in some nonmonotonic extension of the initial theory. If the activation levels of the nodes in the network had been updated in a different order, then the final pattern of activation levels might correspond to a different extension of the theory. A particular example that seems relevant is a Hopfield net designed to find the front side of a Neckar cube, the line drawing that seems to flip back and forth between two different orientations. Updating nodes in a network that represents relationships between the vertices of a Neckar cube asynchronously can produce a stable state that brings either of the competing faces to the front depending on the actual order that the nodes are updated. I do not know how to build a connectionist system that will work in this way even for a very simple defeasible theory, but work that has been done using connectionist systems to solve constraint satisfaction problems suggests that we might explore this possibility further.

The very sketchy account I have offered for developing a coherentist theory of justification based on an extension semantics for nonmonotonic reasoning tells us when an epistemic agent justifiedly believes some set of propositions. But this account certainly allows for the possibility that an agent might justifiedly believe something without being able to provide that justification, without being able to provide good reasons for his belief. In fact, if demonstrating that a large set of beliefs is a nonmonotonic extension of a theory is as difficult as I suspect it is, it's hard to see how an agent could *ever* explain or even *know* that *any* of his beliefs were justified. So if our account were to end here, we might conclude that an agent could be justified in believing ϕ, but he might never be justified in believing that he is justified in believing ϕ. I believe Pollock's account of epistemic justification allows this possibility. Remember that Pollock and Cruz say that an agent is justified in believing ϕ just in case he "instantiates an undefeated argument supporting" ϕ. ([25], p. 38) And Pollock and Cruz say that they are developing "the procedural concept of epistemic justification" but note that "there may be other concepts that can reasonably be labeled 'epistemic justification'". ([25], p. 124.) Indeed, there are other concepts of epistemic justification and many of them are very concerned with the ability of an agent to provide good reasons for his beliefs.

9 Foundationalism and Defeasibility

While a procedural concept of epistemic justification might be founded on an extension semantics (although this is certainly *not* what Pollock and Cruz pro-

pose,) it is less likely that an adequate concept of epistemic justification that hopes to provide an epistemic agent with the means to provide good reasons, to himself and others, for his beliefs can be build on the kind of extension semantics that have been proposed for nonmonotonic formalisms. Even if it were computationally possible to determine that a large set of beliefs corresponded to a nonmonotonic extension of an agent's initial theory or, even more difficult, to the intersection of *all* nonmonotonic extensions of an agent's initial theory, this is hardly something that an agent could do regularly and as his initial theory changes with new experience.

The form of such an account should at this point be obvious: S is justified at t in believing that ϕ if and only if $T \mid\sim \phi$ where T is S's initial theory at t. Since defeasible proofs are finite and constructive, at least for finite defeasible theories, it is in principle possible for S to produce a defeasible proof for any of his justified beliefs. The concept of epistemic justification captured by this principle is a form of defeasible foundationalism.

Pollock's direct realism is a nondoxastic foundationalism because beliefs may have justifications that are not themselves beliefs. For example, perceptual beliefs may be justified by the perceptual state an agent is in. The defeasible foundationalism described here is also nondoxastic: beliefs are justified by the presumptions, defeasible rules, and other components of the initial defeasible theory accepted by an agent. The agent's perceptual state may generate the presumption that supports a perceptual belief; indeed, there may be no difference between being in a particular perceptual state and accepting the corresponding presumption. But an important difference is that the foundations in this epistemology are not beliefs but presumptions, defeasible rules, and the other components of an agent's initial theory. There really are no foundational beliefs (at least no foundational perceptual or memorial beliefs) since any belief justified at t_1 could be defeated at a later time t_2 by further experience.

Where the derivability relation used in this analysis of epistemic justification is based on the proof theory for defeasible logic, and the semantics for a coherence theory of epistemic justification is based on the semantics for defeasible logic presented here, it is clear that we get two accounts of epistemic justification that are not extensionally equivalent. The proof theory for defeasible logic has been shown to be sound but not complete for the semantics. So there can be beliefs that are justified on the coherentist account that are not justified on the foundationalist account even when we fix the agent and the agent's initial theory. For other nonmonotonic formalisms, formalisms for which an extension semantics but no proof theory is known, a foundationalist account is not even possible. If we look at any nonmonotonic formalism for which both a semantics and a proof theory are known, then we have shown that if the proof theory is constructive and the semantics admits floating conclusions (as we might expect any extension semantics that allows multiple extensions to do,) then the proof theory will not be complete with respect to the semantics. For any such nonmonotonic formalism, including but not limited to defeasible logic, the foun-

dationalist and the coherentist accounts of epistemic justification will not be extensionally equivalent.

The foundationalist and coherentist versions of epistemic justification presented here are not incompatible so long as we recognize that they are intended to play different roles. The foundationalist concept of epistemic justification is an account of what may count as good reasons for the beliefs an agent has; the coherentist concept offers a procedural account of epistemic justification that the agent may not be able to marshal in defense of his beliefs. Since in general nonmonotonic proof theories will not be complete with respect to their corresponding semantics, agents will sometimes be (procedurally) justified in believing ϕ when they cannot provide good reasons for believing ϕ. However, since we would want to insist that any nonmonotonic proof theory is sound with respect to the corresponding semantics, an agent will be (procedurally) justified in believing ϕ whenever he can provide good reasons for believing ϕ.

10 Conclusions

How would we build an epistemology for an artificial agent?

References

1. Robert Audi. *Epistemology: A Contemporary Introduction to the Theory of Knowledge*. Routledge, London and New York, 1998.
2. Robert Audi. Contemporary modest foundationalism. In Louis P. Pojman, editor, *The Theory of Knowledge: Classical and Contemporary Readings*, pages 174–182. Wadsworth/Thomson Learning, Belmont, California, third edition, 2003.
3. S. Donnelly. Semantics, soundness, and incompleteness for a defeasible logic. Master's thesis, Artificial Intelligence Center, The University of Georgia, 1999.
4. P. M. Dung, R. A. Kowalski, and F. Toni. Synthesis of proof procedures for default reasoning. In *Proceedings of the international workshop on logic programming synthesis and transformation*, pages 313–324. Springer Lecture Notes on Computer Science 1207, 1996.
5. D. Gabbay. *Labelled Deductive Systems*, volume 1. Oxford University Press, Oxford, 1996.
6. H. Geffner. *Default Reasoning: Causal and Conditional Theories*. PhD thesis, UCLA, 1989. Research Report 137, Cognitive Systems Laboratory, Department of Computer Science.
7. H. Geffner and J. Pearl. A framework for reasoning with defaults. In H. Kyburg, R. Loui, and G. Carlson, editors, *Knowledge Representation and Defeasible Reasoning*, Studies in Cognitive Systems, pages 69–88. Kluwer Academic Publishers, Boston, 1989.
8. K. Konolige. On the relation between default theories and autoepistemic logic. *Artificial Intelligence*, 35:343–382, 1988.
9. R. Loui. Defeat among arguments: A system of defeasible inference. *Computational Intelligence*, 3:100–106, 1987.
10. R. Loui. *Theory and Computation of Uncertain Inference and Decision*. PhD thesis, The University of Rochester, 1987. Technical Report 228, Department of Computer Science.

11. D. Makinson. On a fundamental problem of deontic logic. In H. Prakken and P. McNamara, editors, *ΔEON'98: 4th International Workshop on Deontic Logic in Computer Science*. Università degli Studi di Bologna, 1998.
12. D. Makinson and K. Schechta. Floating conclusions and zombie paths: two deep difficulties in the 'directly skeptical' approach to inheritance nets. *Artificial Intelligence*, 48:199–209, 1991.
13. D. McDermott and J. Doyle. Non-monotonic logic I. *Artificial Intelligence*, 13:41–72, 1980.
14. R. Moore. Semantical considerations on non-monotonic logic. *Artificial Intelligence*, 25:75–94, 1985.
15. M. Morreau. Reasons to think and act. In D. Nute, editor, *Defeasible Deontic Logic*, Synthese Library, pages 139–158. Kluwer Academic Publishers, Dordrecht, Netherlands, 1997.
16. D. Nute. Defeasible logic. In D. Gabbay and C. Hogger, editors, *Handbook of Logic for Artificial Intelligence and Logic Programming*, volume III. Oxford University Press, Oxford, 1994.
17. D. Nute. Apparent obligation. In D. Nute, editor, *Defeasible Deontic Logic*, Synthese Library, pages 287–315. Kluwer Academic Publishers, Dordrecht, Netherlands, 1997.
18. D. Nute. Norms, priorities, and defeasibility. In P. McNamara and H. Prakken, editors, *Norms, Logics and Information Systems*, pages 201–218. IOS Press, Amsterdam, 1999.
19. D. Nute. Defeasible logic: theory, implementation, and applications. In O. Bartenstein, U. Geske, M. Hannebauer, and O. Yoshie, editors, *Web Knowledge Management and Decision Support: Proceedings of INAP 2001, 14th International Conference on Applications of Prolog (Revised Papers)*, pages 151–169. Springer-Verlag, Berlin Heidelberg New York, 2003.
20. J. Pollock. Defeasible reasoning. *Cognitive Science*, 11:481–518, 1987.
21. J. Pollock. Oscar: A general theory of rationality. *Journal of Experimental and Theoretical Artificial Intelligence*, 1:209–226., 1989.
22. J. Pollock. *How to Build a Person*. Bradford/MIT, Cambridge, MA, 1990.
23. J. Pollock. Self-defeating argument. *Minds and Machines*, 1:367–392, 1991.
24. J. Pollock. A theory of defeasible reasoning. *International Journal of Intelligent Systems*, 6:33–54, 1991.
25. J. Pollock and J. Cruz. *Contemporary Theories of Knowledge*. Rowman and Littlefield, Totowa, NJ, second edition, 1999.
26. R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.
27. G. Schurtz. Defeasible reasoning based on constructive and cumulative rules. In R. Casati, B. Smith, and G. White, editors, *Philosophy and Cognitive Sciences*, pages 297–310. Hölder-Pichler-Tempsky, 1994.