Defeasibility in Epistemology

Dissertation Summary

My dissertation is an exercise in what we might call *nontraditional formal epistemology*. It's both common and natural to define formal epistemology in opposition to its more traditional cousin: Where traditional epistemology approaches (normative) questions relating to belief, knowledge, and reasoning relying on the classical method of conceptual analysis, formal epistemology approaches these same questions drawing on tools from math and logic. In principle, various formal tools could be used, but, in practice, formal epistemologists usually draw on either the Bayesian framework—which combines probability theory and inductive logic—or epistemic logic. The use of these standard tools to answer epistemological questions can, then, be called *traditional formal epistemology*, and what's going on in my dissertation can be thought in opposition to it. At any rate, my dissertation relies on a completely different formal framework, namely, that of logics for defeasible reasoning.

This framework originated in the field of artificial intelligence in response to the challenge to represent the information that would let a machine exhibit intelligent behavior. Efforts to meet this challenge quickly made it clear that ordinary logic is utterly inadequate for this task, since much of this information takes the form of defeasible generalizations. Thus, the statements "Birds fly" and "Things that look red are red" appear to express sensible principles of reasoning—principles we'd seem to constantly rely on in our everyday life—even though they allow for exceptions. Defeasible logics then are, roughly, logics of such defeasible generalizations, and the thesis of my dissertation is that they can be of great help in answering important normative questions in epistemology. The thesis is supported by developing three independent and equally important applications of defeasible logics. Accordingly, the dissertation is divided into three parts.

Part 1 is concerned with simple epistemic rules, such as "If you perceives that X, then you ought to believe that X" and "If you have outstanding testimony that X, then you ought to believe that X." The problem is that it's almost too easy to imagine cases where rules like these come into conflict—such as the one where you perceive a red object and are told that it is blue. One popular response to the problem suggests that these rules have implicit hedges, or unless clauses specifying the conditions under which they fail to apply. Another response suggests that these rules are contributory, or that they do not, in fact, specify what beliefs one ought to have, but only what counts in favor or against having them. Drawing on the defeasible logics framework, I devise a model for each of these seemingly very different views on rules and establish a type of equivalence result between them, which suggests that the views themselves are much closer than standardly thought. This result also has far-reaching ramifications for various claims about rules and views on rules advanced in the literature.

Part 2 shifts the focus from rules to epistemic requirements, such as the intuitive and widely accepted "If X is supported by your (total) evidence, then you ought to believe that X" and "You ought to believe that X if you believe that your evidence supports X." It's naturally seen as doing two things. First off, I use a defeasible logic to work out a new solution to an important puzzle about epistemic rationality: In case one's (total) evidence can be misleading about what it itself supports—as many epistemologists think—then the above two requirements can come into conflict, suggesting that there are dilemmas of rationality. My defeasible logic-based solution has a number of attractive features when compared to the other solutions from the literature, even though it does comes with an unorthodox perspective on epistemic requirements, a perspective on which they are defeasible. I also show—and this is the second major idea of this part of the dissertation—that defeasible epistemic requirements can be naturally thought of as epistemic ideals, and that the defeasible logic used to solve the puzzle can be naturally seen as the formal backbone of the *conflicting-ideals view* that David Christensen has been advocating for in his recent work. In effect, I'm proposing to understand this view as a move away from the default metaepistemological position according to which epistemic requirements are strict and governed by a strong, but never explicitly stated logic, toward the more unconventional view, according to which requirements are defeasible and governed by a comparatively weak logic. This illuminates the view and helps counter some common objections to it.

Finally, Part 3 applies logics for defeasible reasoning in the context of the burgeoning debate about the epistemic significance of disagreement. The general aim here is to get a better grip on the intuitively appealing *conciliatory views*—which say, roughly, that you're to become less confident of your belief in the face of a disagreement with an epistemic peer—and, in particular, their behavior in scenarios involving higher-order disagreements, such as disagreements over conciliatory views themselves and disagreements over epistemic peerhood. It turns out that the core idea motivating conciliatory views can be naturally expressed in a certain defeasible logic, with conciliationism emerging as a well-behaved defeasible reasoning policy. My defeasible logic-based model of conciliationism also turns out to be very useful: Among other things, it lets us address a well-known challenge for conciliatory views, namely, that they would seem to self-defeat and issue inconsistent directives in scenarios involving disagreements over their own truth. (Also, in the course of pursuing the philosophical goals of Part 3, I devise an intuitive formal argumentation theory framework and show that it extends *default logic*, or the particular defeasible logic used to model conciliationism. This result is an instance of a known general theorem, but it is of independent interest due to the intuitive character of the framework and its prospective applications.)