Defeat and Preservation in Epistemology, Metaphysics and Ethics

Joshua Edward Pearson
Research Statement

Learning new information can often "defeat" your justification for a particular belief you held. For instance, I may now be justified in believing I'll go for a walk later, but learning that it will rain may defeat that justification. Defeat is at the heart of my doctoral research. Understanding it has crucial implications for the nature belief, the interaction between belief and rational action, and the problem of skepticism.

For example, consider the following influential principle constraining defeat:

> **Preservation:** If you're justified in believing C, and B is a live possibility (you're not justified in believing not-B), then learning B will not defeat your justification for C.

Preservation seems intuitive. Very plausibly, I'm justified in believing Carlos Alcaraz won't win the next forty Grand Slams. At the same time, it's a live possibility that he wins the next one. If so, Preservation tells us that learning he has won the next one will not defeat my justification for believing he won't win the next forty. That sounds right.

Yet Preservation leads to skepticism. Consider Alcaraz again. It's of course a live possibility that he wins the next Grand Slam. However, if I were to learn he has indeed won it, without learning anything further (say that he's been injured, or retired), I would not be justified in believing he won't also win the Grand Slam after that. So, per Preservation, it must *now* be a live possibility that he wins the next two. This reasoning repeats. For learning he's won the next two would similarly be incompatible with justifiably believing he won't win the next three, and so on. A skeptical conclusion eventually follows: I'm not, after all, justified in believing Alcaraz won't win the next forty Grand Slams. Similar arguments can be constructed for a vast number of our purportedly justified beliefs.

It gets worse. In my thesis, I show that giving up Preservation is not nearly enough if one wants to avoid skeptical conclusions of this kind. I isolate an important but under-theorized weakening of Preservation, which I call "Anticipation", and argue that any anti-skeptics compelled to give up Preservation should give up Anticipation, too. This generates a significant challenge: no existing theory of justified belief can accommodate these counterexamples to Anticipation. I tackle this challenge head-on and develop a novel theory of justified belief, plausible in its own right, but that can further predict the required counterexamples to Anticipation. (See my "Belief Revision Revised", *PPR*, 2025.)

My research on defeat is not over. I plan to generalize and develop my doctoral research and further investigate its consequences for rational inquiry—such as its consequences regarding Kripke's "dogmatism paradox". However, over the past year I have also begun developing my doctoral research in another direction, turning to a project on counterfactuals. My guiding methodology is to pursue systematic analogies between problems concerning counterfactuals and those about defeat. I've seen more and more how exploring these analogies can make significant progress on longstanding problems concerning counterfactuals, such as their semantics, the question of counterfactual skepticism, and the meaning of "might"-counterfactuals.

For instance, consider counterfactual skepticism—the thesis that we know almost none of the counterfactuals we assert in ordinary life. The famous arguments for this view use premises that are now widely disputed. However, by exploiting analogies to defeat we can construct new arguments for counterfactual skepticism. Consider:

> **Counterfactual Preservation.** If *Had A, would C* and *Had A, might B* are both true, then so is *Had A&B, would C.*

While Counterfactual Preservation seems plausible, and is entailed by standard theories like David Lewis's, it—analogously to Preservation—leads to counterfactual skepticism. Consider a counterfactual we'd ordinarily take ourselves to know is true, say — *Had I left early, I would have avoided traffic*. Now, had you left early, it might have been that someone else in your neighborhood left early, too. Counterfactual Preservation therefore yields: *Had I left early and one other in my neighborhood had too, I would have avoided traffic.* But if that had happened, a second neighbor might have left early as well, so similarly we derive: *Had I left early and so had two others in my neighborhood, I would have avoided traffic.* Iterating leads to the absurd conclusion: *Had I left early and so had everyone in my neighborhood, I would have avoided traffic.* The skeptic concludes the initial counterfactual was never known to begin with.

Responding to this challenge requires developing a new theory that denies Counterfactual Preservation. Here, again, exploring structural analogies to epistemology is fruitful. Here's one way to do this (I'm exploring multiple). On a simplified version of Lewis's semantics, *Had A, would C* is true when all the closest *A*-worlds are *C*-worlds. Consider the following subtle modification. Call a world "*A*-eclipsed" when there's an *A*-world significantly closer to actuality than it. We'll then say that *Had A, would C* is true when *C* is true at all the *A*-worlds that aren't *A*-eclipsed. It turns out that this change to Lewis's view, which mimics anti-Preservation theories of belief, allows Counterfactual Preservation to fail. We thus arrive at a novel, anti-skeptical theory of counterfactuals, that needs to be taken seriously.

There are also analogies to ethics I plan to explore. Consider:

> **Permissibility Preservation.** If you ought to ϕ, and it's permissible to ψ, then you still ought to ϕ conditional on ψ-ing.

There's a lot to like this ethical analogue to Preservation. Suppose it's permissible to bring your own food to a party, and conditional on doing so, you're not obligated to eat the host's food. Then, per Permissibility Preservation, it must be that you weren't initially obligated to eat the host's food—bringing your own was always an option. Further, the restriction that ϕ is permissible means it sidesteps "gentle murder" problems. For example, suppose you ought not steal, either from the poor or from the rich. Even so, conditional on stealing, it seems you ought to steal from the rich. This is consistent with Permissibility Preservation, as the act that "defeats" your obligation not to steal from the rich is an act which is itself impermissible—stealing. Finally, this principle doesn't appear susceptible to the kinds of skeptical arguments outlined above.

Yet Permissibility Preservation may still face counterexamples. Suppose you can have duties to yourself, bestowed upon you by your rights. And suppose further that you can permissibly waive those rights. Then there's some permissible action—waiving your right—such that, conditional on performing it, an obligation is "defeated"—the duty to yourself. So far, those who have defended duties to oneself have made sense of them by claiming they are cases in which obligations fail to iterate—you're obliged to ϕ but this fact is not itself obligatory. Understanding such duties instead as counterexamples to Permissibility Preservation looks like an attractive alternative.

My future research will thus pursue key issues in epistemology, metaphysics and ethics in a united manner. I expect this work to yield numerous articles and the foundations for a book-length project.