From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Philosophy

2017

New Work on the Old World Order

Nichols, Cory

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp019g54xm31s

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

NEW WORK ON THE OLD WORLD ORDER

Cory Musgrave Nichols

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE DEPARTMENT OF

PHILOSOPHY

Advisor: Boris Kment

Secondary Advisor: Gideon Rosen

November 2017

© Copyright by Cory Musgrave Nichols, 2017.

All rights reserved.

Abstract

This dissertation is a collection of three independent but related papers on conditionals and modal language in general.

The first paper, “Rethinking Similarity”, discusses some problem cases (some new, some known) for the standard style of semantics for counterfactual conditionals, and develops a positive theory designed to account for these cases. According to what I call the Similarity First Paradigm, which includes the influential theories of David Lewis, Robert Stalnaker, and Angelika Kratzer, a counterfactual of the form *If A were the case, then C would be the case* is true if and only if (roughly) the *closest* A-worlds are C-worlds, where the closeness of one world to another is a matter of how *similar* they are in the right kinds of ways. According to my view, context determines a range of relevant ways that A might be true, which I model as subsets of the set of A-worlds, and the counterfactual is true iff the closest worlds in each of those sets are C-worlds.

The second paper, “Strict Conditional Accounts of Counterfactuals”, introduces several novel problem cases for the dynamic strict conditional analysis of counterfactuals, due to recent work by Kai von Fintel and Thony Gillies, and works through some possible amendments of the view that seem to fail in response to these cases. I also consider the leading competing account of the conditional data motivating the strict conditional view, and show that it appears to have no special trouble accounting for my cases as well.

The third paper, “New Puzzles for Kratzer’s Modals”, introduces three novel problem cases for Angelika Kratzer’s extremely influential semantic framework for modal language (e.g. *must* and *can*), which is intimately related to her analysis of conditionals. In the first two cases I

propose amendments to the view that I think may handle the data, in one case appealing to Kratzer's own analysis of conditionals. In the third case I can see no such possibility, but I do offer a diagnosis of the problem that is instructive about the analysis of modality in language in general.

Dedicated to the amazing women in my life:

To my moms, Margie and Nancy, without whom I would not exist.

To my sisters Alejandra, Diana, and Briana, for whom I would cease to exist.

To my beloved sister Jesse, who has ceased to exist, but for whose existence
I would happily trade my own, if I only could.

For Cortney, whose existence constantly brightens mine:
If ninjas could dance, they would dance the mambo!

And for Blythe, whose existence I am partly responsible for:
may you seek and find both bewilderment and understanding.

Acknowledgements

This section was surprisingly difficult and time-consuming to write (sorry Jo!!!). I guess I was consumed by the threat of leaving out someone crucial, or even of remembering someone but not saying quite the right thing. Like everyone else in this position, I'm sure I've forgotten many people deserving of mention, most of whom I will surely remember just as soon as it's too late to do anything about it. Let it be implied in each subsection below that I apologize to anyone whose name should have been included, and that the exclusion thereof implies nothing about their significance to my philosophical career.

Additionally, the organization of this section is a mess. I have grouped people mainly by academic institution, with a handful of close personal friends, including the primarily non-academic ones, reserved for the end. The lines between all the important categories are blurry. (That's life, huh?)

Amongst the professors I have had the pleasure of working with during my time as a graduate student, the ones who stand out as having contributed the most directly to my dissertation are Boris Kment, Alan Hájek, Gideon Rosen, and Karen Lewis. If we used full dissertation committees in our department rather than individual advisors, I would have liked these four individuals to have comprised mine.

Boris Kment was about the best advisor I could have hoped for. One hears horror stories of advisors who are uninterested, absent, hostile, or unreasonable. Boris was the antithesis of these qualities. He was endlessly generous with his time and attention, and sometimes we'd end up talking for several straight hours in his office after losing track of time. (Once I missed dinner plans!) He would also happily provide meticulous comments on several drafts of the same paper, usually within a few days of receiving the draft. And he surpassed the standard requirements of a dissertation advisor in other ways, e.g. offering comparably detailed feedback on every mundane component of my job market materials. If anything, I was often the derelict one. Occasionally after long stretches of not hearing from me, Boris would check in to see how I was doing, which was often just the jostle I needed to get back on track. Most of all, Boris

was just the right balance between supportive and critical. He'd be direct when he thought an idea didn't work, or a passage needed rewriting, and he'd make incredibly helpful, constructive suggestions when he thought something was promising, even if he didn't agree. There's no doubt in my mind I will continue to seek his advice throughout the rest of my academic career.

Alan Hájek, as anyone who has met him knows, is a *mensch*: warm, kind, generous, brilliant, and yet still somehow entirely unassuming. I first met Alan in the summer of 2013 at a conference on conditionals in Barcelona, though his work had been important to me since my undergraduate days. (I wrote my writing sample for grad school applications on his counterfactual skepticism.) He invited me to visit the Philosophy Department at the Australian National University in Canberra, and I took him up on it. The couple of months I spent there were amongst the most intellectually engaging and productive of my life, and produced the first, rudimentary version of my Kratzer paper (a much better version of which appears in this dissertation). Since then Alan has continuously provided extraordinary support, advice, and feedback. And remarkably, when it comes to conditionals, Alan and I seem to agree on almost everything. (Morgenbesser's coin is the only serious exception I'm aware of.) I couldn't ask for a better coconspirator. (Here's hoping we turn out to be right!)

It's no exaggeration when people say that Gideon Rosen is one of the treasures of the Princeton philosophy department: brilliant, charismatic, has an astonishing scope of knowledge, always finds just the right words to capture a thought. As a graduate student I had the privilege to work closely with Gideon on a couple of papers, solicit comments on a few others, have him on my general exam committee, and receive feedback at a couple of informal talks. Most importantly, I wrote the later, more substantial drafts of my Kratzer paper (again, found within this dissertation) with Gideon, which is the main reason he is listed as my secondary advisor. My major regret about graduate school, in general, is not having taken more frequent advantage of the opportunity to meet with the stellar faculty; this applies to Gideon perhaps more than anyone else.

Karen Lewis has been a paragon of supererogation, meeting with me many times to discuss my work despite having no professional obligations to me whatsoever. Every time I asked her to meet it would have been easy, and perfectly reasonable, for Karen to say she didn't have the time, or was already overburdened by her commitments to her own students and department. But she always found time to meet, and always came with thoughtful comments. Indeed, each of the papers in my dissertation owes a great deal to Karen's insights. Additionally, our positive views are very similar, and discussing her account often informed and guided my approach to my own. I look forward to further such guidance in the future; I suspect we may discover we are even more like-minded than we realize.

Other people from the Princeton Philosophy Department who must be mentioned include:

Michael Smith, for being an all-around fantastic human, an outstanding department chair, a masterful singalong-leader, and also for demonstrating that moral philosophy can be approached with the same rigor, clarity, precision, and careful analysis that as an undergrad I had mistakenly associated primarily with logic and M&E areas. I might not have developed a secondary interest in moral philosophy without exposure to Michael.

Delia Graff Fara, who sadly passed away shortly before these acknowledgments were written. I had the pleasure and honor of serving as Delia's research assistant for a few semesters on a book on vagueness that she planned to write but, unfortunately, never will. On a more personal level, I always looked forward to running into her around the department and at official functions. Her warm nonchalance and down-to-earth demeanor were reliably refreshing in the often-stuffy academic world. The loss of Delia is devastating to me, Princeton, and the philosophy community in general. She'll be dearly, dearly missed.

Tom Kelly, for encouragement during my first couple of years as a grad student, a period which severely tests one's self-confidence. Tom gave me very positive feedback in our first-year seminar and in subsequent unit work on epistemology that I did with him individually and in his and Sarah McGrath's

moral epistemology seminar. This early support and feedback helped me feel at home in the department and overcome much of the "imposter syndrome" that is common amongst graduate students.

And I would be remiss not to also mention the following members (some former) of the Princeton community: John Burgess, Shamik Dasgupta, Adam Elga, Anna Faiola, Christiane Fellbaum (Linguistics), Dan Garber, Frank Jackson, Jo Kelly, Sarah-Jane Leslie, Sarah McGrath, Ben Morison, and Peter Singer.

And of course I could fill an ocean liner with the Princeton grad students (or visitors) who in some way significantly affected my time there as a graduate student – and here I'm especially sure I'm leaving out some important names – including: Ashley Atkins, Dan Berntson, John Colin Bradley, Kezia Chuaqui, Ellie Cohen, Adam Crager, Rachel Cristy, Anthony Cross, Simon Cullen, Jordan Delange, Robin Dembroff, Josh Gillon, Mark Harris, Sukaina Hirji, Robbie Hirsch, Eric Hubble, Rishi Joshi, Brett Karlan, Maddie Kilbride, Laura Kotevska, Kyle Landrum, Yena Lee, Adam Lerner, Eden Lin, Errol Lord, John Mackay, Barry Maguire, Jimmy Martin, Brennan McDavid, Michaela McSweeney, Angela Mendelovici, Carla Merino, Patrick Miller, Sam Preston, Joe Rachiele, Domenica Romagni, Vanessa Schouten, David Schroeren, Whitney Schwab, Mor Segev, Gabe Shapiro, Derek Shiller, Simon Shogry, Jack Spencer, Nat Tabris, Gerard Vong, and Dan Wodak.

Of course there are also many people who greatly impacted my undergraduate career at Rutgers:

I take the following inference pattern to be truth-preserving: P is a claim about physics or the philosophy thereof; Tim Maudlin says P; therefore, P. The amount of support and kindness Tim and Vishnya Maudlin showed me during my time at Rutgers was above and beyond: they gave me great feedback; they let me go on too long so I could work through my own thoughts; they had coffee with me after class to talk about my ideas (or whatever else came up); and they invited me into their home for dinner on multiple occasions. They also instilled in me a healthy amount of skepticism towards the style of

unapologetic armchair metaphysics that was quite popular at Rutgers at the time (though I confess to a lingering attraction). And their initial support helped give me the confidence to pursue philosophy in the first place, rather than be intimidated by it.

Bob Matthews' upper-level philosophy of language course introduced me to the subfield I now call my own. He then allowed me to enroll in his graduate seminar on his then-new book *The Measure of Mind* (highly recommended!); it was just me and a handful of Rutgers grad students, so, naturally, I was terrified. Bob gave me brutally honest feedback. I wrote a paper for that seminar that I thought was great, and he tore it apart (figuratively). When I went to his office to defend it, I found myself spewing a string of buzz-words I'd recently learned from *SEP* articles. Bob stopped me and asked me to state my view without any philosophy lingo. I couldn't do it. I then realized I didn't have a clear grasp of what I was arguing for. That moment taught me to be leery of relying too heavily on industry jargon and picture thinking, which still guides my approach to philosophy today.

One of John Hawthorne's best qualities is a willingness to take up someone else's perspective, treat it seriously, and help them think it through, even when he disagrees. I reckon John and I differ quite a bit when it comes to first-order views (e.g. on conditionals!). But a typical conversation with him feels like we agree on everything. He's never dismissive, and he'll help you hunt through logical space to find the best version of your view. That sort of encouragement is important for a fledgling philosophy student, full of self-doubt and reticence. And John also showed me that philosophers can be fun: after Jack Woods, he's my second favorite pool opponent (extremely high praise in my book).

Jeff King came to Rutgers just a few semesters before I graduated, but still managed to make an impact on my experience there. I took his upper-level undergraduate metaphysics course, and then pestered him until he found the time for an independent study with me the following year. (He'd just been named chair, so he had very little free time.) Jeff met with me every week or two at French Roast on the Upper West Side to comb through Lewis' *On the Plurality of Worlds* and Ted Sider's *Four-Dimensionalism*. This led

to my honors thesis on Sider's stage theory of persistence, and was one of my first careful studies of metaphysically informed semantics, which turned out to be my main interest.

I did an independent study on counterfactuals with Brian Weatherson in my last semester at Rutgers. At the time I was groping around for the ideas that would later become my positive account of counterfactuals, defended below in this very dissertation. Brian helped me hone those nascent ideas into something vaguely coherent and sensible; I then revisited them a few years later, and they became the basis for *Rethinking Similarity*. Brian's feedback was invaluable for this first stage of what I expect will be a huge part of my life's work.

Martin Bunzl taught the massive intro course that I took just after switching my major to philosophy. I was one of those undergrads who wanted to talk constantly in class, and Martin kept me appropriately in check. But he also recognized my genuine enthusiasm and managed to foster it. He told me to come to his office hours and advised me as to which courses to take next if I was serious about pursuing philosophy. This early guidance was invaluable, and I might never have ended up in a PhD program without it.

My main undergraduate comrades were Zee Perry and Sophie Ban. We took lots of courses together and spent hours and hours outside of class talking about both philosophy and non-philosophy, chain-smoking cigarettes (well, just Sophie and me), and other sorts of shenanigans. There are the folks who major in philosophy because of a passing interest, or because they think it will be easy, or because they don't know what else to do, and there are the folks who major in philosophy because it keeps them up at night. It was clear from the start which group we three were in.

Other members of the Rutgers community at the time who must be mentioned include Frank Arntzenius, Mercedes Diaz, Trip McCrossin, Al Nigrin (Cinema Studies), Ted Sider, Holly Smith, and Dean Zimmerman.

My time as a visiting graduate student at the ANU, as I've already mentioned, stands out in my mind as among the best time I've spent anywhere. While almost everyone I met there was absolutely lovely, there are a handful of other faculty members aside from Alan who stand out as having contributed greatly to making my time there so memorable, particularly via feedback in meetings that helped along my dissertation papers in their earlier stages. I'm thinking especially of R. A. Briggs, Daniel Nolan, Paolo Santorio, and Wolfgang Schwartz.

Just as important to my overwhelmingly positive experience in Canberra were the other awesome folks who made up the philosophy community at the time. There are too many to name, but a handful jump to mind: Arif Ahmed, Alma Barner, Adrian Currie, Jessica May Isserow, Colin Klein, Christopher Hunter Lean, Erick Llamas, Nathan Pensler, Gabe Rabin, and Alex Sandgren.

There are a few handfuls of others who either do not fit neatly into one of the previous groups, or for some other reason I feel should be mentioned separately:

Jack Woods, a.k.a. Jacques Bois, a.k.a. J-Wo, a.k.a. my partner in crime, as in pool. There are friends who will show up to help you move, and then there are friends who will show up to help you move at 9 a.m., after a late night, on three hours' sleep, in central New Jersey. Moreover, the greatest indicators of how well I get along with a philosopher are: (1) how much time can I stand talking about philosophy with you? and (2) how much do we have to talk about that *isn't* philosophy? Jack scores highest marks in both areas. If I were to be stuck on a desert island with just one other person, I would be hard-pressed to think of someone better than Comrade Woods.

Matt Moss is another obvious choice in the desert island scenario. Together with Jack, I think of the three of us as a sort of triumvirate, though of what exactly I cannot say. Matt is one of those friends I would be there for under any conceivable circumstances, as I know he would for me. If I were to abandon academia tomorrow, I am certain our friendship would not suffer in the slightest, and we would continue to spend

late nights talking nonsense in long, syntactically overwrought sentences, sharing “friendship vodka”, and dancing to strange techno in even stranger dark rooms in Brooklyn, or wherever we happened to find ourselves. Few in this world are nearer and dearer to my heart than Matthew Moss.

Noel Swanson is a kind of Renaissance man: it’s rare to find someone equally well-versed in opera, bourbon, baseball, physics, philosophy, and Star Wars trivia. We started grad school together, then lived for a summer (with Alberto Barros) in the East Village. Many of my best memories from those early years involve Noel. Moreover, I started writing this dissertation while visiting him and his family in the Bay Area. I’d been through a bad breakup and was terribly unproductive. Noel sweetly invited me back home with him over Spring Break, and that change of pace and scenery was just what I needed. I wrote the first draft of “Rethinking Similarity” there, so Noel’s generosity helped this dissertation happen.

Alberto Barros – what a glorious human! After knowing him for only a semester or two, he was already one of my dearest friends; living together for one of the funnest summers of my life helped to crystallize this. Rarely has anyone else made such an impression in such a short time. Alberto is one of those friends who lives thousands of miles away, and who I might only get in touch with a few times a year, but each time we talk it feels like little has changed since the last time. He’ll continue to be a source of joy and brightness as long as we’re both alive. (And I owe you a visit to Brazil!)

Another major regret from my grad school days will be that I did not spend more time with Raphael Krut-Landau while we were in such close proximity. I took a liking to Raffi immediately – we were introduced under the 1879 Hall archway, I believe, and he wore some sort of leather jacket, yet was charmingly soft-spoken and unassuming (classic Raffi). I suppose the trouble is that we became closer after we lived far away: I moved from Princeton back to New York, and then he lived in Paris for some time (fun memories there!), and then Philly. But if there’s an ounce of justice in this awful world, there will come a time when we both live in the same city. And may the gods have mercy on that city.

Catharine Diehl is one of my favorite replicants. (She will know what this means.) She's been so kind and hospitable every time I've visited Berlin (there is a permanent impression of me in her couch), even before we were good friends. We've danced all night, we've commiserated over the Trump coup and contemplated communism, and we've co-worked for countless hours in more Berlin cafes than I can shake a *weissweinschorle* at. (And, speaking of work, no, Catharine, "Work from Home" by Fifth Harmony is not, in *any* sense of "about", a song about telecommuting!)

A very important but overlooked statistic, when measuring up any friend, is the fun-had-to-time-spent ratio. A comparably underappreciated statistic, especially when measuring up an *academic* friend, is the things-learned-to-time-spent ratio. Nate Charlow stands a good chance of taking the gold in each category. We've been close only a couple of years, and have not spent much time outside conferences or other professional settings, yet we've had a string of absolute blasts and some of the highest quality shop-talks I can think of. This friend's a keeper.

A handful of other comrades from the academic world who cannot be forgotten include: Vladan Djordjevic, Dan Harris, Eliana Horn, Nathan Howard, Alex Skiles, Brian Leahy, Rachel McKinney, Lauren Ross-Feldman, Adam Sennett, Una Stojnic, Andreas Walker, Lee Walters, and Zsofia Zvolenszky.

Now come the (mostly) non-academics – that is, the folks with whom I have a primarily non-academic relationship:

To my mothers, Nancy Musgrave and Margie Nichols. You are the absolute best, best, best.

Nancy likes to tell people that if she'd been able to custom-order a child from a catalogue she would've ordered one exactly like me. Flattering (and a counterfactual!), but surely a bit hyperbolic. But the reverse claim, about her and Margie, is basically true.

I cannot imagine better parents. I've met so many people whose parents didn't accept them, didn't encourage them, weren't there, didn't make them feel loved. Or even just parents whose values didn't match their own, or who didn't understand who they were, or who they weren't especially proud of or especially excited to see when they came home. My moms are the opposite in every way. They make me beam with pride on a regular basis, whether I'm thinking of their work with battered women in the 70s, with AIDS victims in the 80s, or with transgender kids more recently. Most of all, they've always been an endless fountain of love and support, through the very, very toughest of times. Some people dread family gatherings, complain at the holidays, and move far away to get away from their families. I always feel bad for those folks, and grateful for my own situation. I can't wait for holidays. I love my moms more than anything.

My other mom, Margie (the original Dr. Nichols in the family), used to have a pin that read: "I may have a PhD, but I'm not stupid!" At one point (so I'm told), in a desperate expression of youthful rebellion, I yelled at her that I would never be like her, and I would *never* get a Ph.D.! Well, I'm about to have a Ph.D., so I guess I *am* stupid.

And to Tom, my father – or uncle? – or something! – and Pat, who have always loved and supported me. Our relationship is atypical, but no less special for it. I love you dearly, and get much of what made me a philosopher from you.

Cortney Norris: My partner, my companion, my love. You've given me so much support, even at times when I must have been maximally unpleasant to be around (e.g. first time on the job market, ugh!). When I met you, at a party, we had an argument (about the merits of the Harry Potter franchise), and you made me look foolish in front of a bunch of people. Well, that did it, you won me over from that very moment (though it took a bit longer for me to do the same). I absolutely adore, cherish, treasure the life we've built together over the subsequent few years: our apartment, our yard, our conversations, our habits, our

vacations, our meals, our combined wine collection, the way we complement each other, the way we fit together just right on the couch on a lazy Sunday, the way you balance me out, the way you appreciate me, the way you actually empty the dishwasher instead of just taking out forks as needed (like I do), the way I'm certain you just got that 30 Rock reference, and last but certainly not least, our weird private language that nobody else would understand (see: our use of the definite article). We will never run out of things to say to each other, and I can't wait to never do so.

Christopher Englese (a terrific filmmaker, Google him if you're reading this): you magnificent, sexy beast. You have been my rock for most of my existence on this stupid rock. Other than my moms, you have been by my side through more hell and highwater than anyone else in my life. You remember high school. You remember Jesse. You remember my darkest days, and my brightest. You're the only friend I could have the god-awfullest of fights with and then always find the heart to forgive, and to be forgiven by, no matter what. You're beautiful, you're brilliant, you're one in a billion, you're a blessing on my life, if I believed in that kind of thing. I grew up with two moms and four sisters, and I always wanted a freakin' brother. Well I had one, it turns out, in you. Nothing will ever tear us apart.

To the rest of my Jersey City crew (though you'll probably never read this): Pete, Farah, Amaechi, C-Lo, my fellow Hudson School alums, and surely some others I'm forgetting about. My roots are with you. I love you all.

To Andrew Ritchie – one day we'll solve the mystery of transubstantiation in Catholic coffee shops! – and to Natasha Umer – you can't be a communist and do all your shopping at AllSaints! – and to the rest of my Brooklyn crew (you know who you are (and you'll never read this either)).

And lastly, but not leastly, to New York, the city I call home. I love everything about you, even the things I hate. I will die here, (even) if it's the last thing I do.

Oh, and to Lisa Simpson. You're my fucking hero.

Cory Musgrave Nichols

Brooklyn

2017

Table of Contents

Abstract	iii
Dedications	v
Acknowledgments	vi-xvii
Introduction	1
Rethinking Similarity	4
Strict Conditional Accounts of Counterfactuals	47
New Puzzles for Kratzer's Modals	89

NEW WORK ON THE OLD WORLD ORDER

CORY NICHOLS

PRINCETON UNIVERSITY

INTRODUCTION

The unifying theme of the three papers in this dissertation, and the reason for its title, emerged after the papers were more or less complete. I did not set out with the topic of world orderings in mind, either as an object of criticism or even as a topic of interest *per se*. Rather, I came to realize after the fact that the central problem underlying most of the problem cases I discuss is an overreliance on orderings of possible worlds. As a result, the practice of appealing to orderings to do heavy theoretical lifting in general is not even discussed explicitly within the papers. But the theme is there nevertheless: world orderings are getting us into all kinds of trouble.

This is clearest in the case of "Rethinking Similarity", which I consider the central paper of the dissertation (though this designation has no official significance). A direct appeal to a similarity-based ordering on possible worlds (*closeness* is the theoretical term for the relation that induces the ordering) in the semantics for counterfactuals is the trademark feature of what I call the Similarity First Paradigm. This includes the standard Lewis-Stalnaker semantics, which has been the benchmark account within philosophy for decades, but it also includes other major views that are designed to deliver more or less the same truth conditions in most ordinary cases, e.g. (most notably): Angelika Kratzer's so-called restrictor analysis (which is the benchmark account within linguistics); Jonathan Bennett's variant of the Lewisian approach; and Kai von Fintel's dynamic strict conditional account. In this paper I discuss several classes of problem cases that arise quite straightforwardly from the emphasis on *the closest* antecedent-worlds. To

avoid these problems I offer a positive theory of counterfactuals that appeals to the similarity/closeness ordering in a less central way. Specifically, *relevance* replaces similarity: context selects subsets of the total set of antecedent-worlds according to their relevance, and it is the closest worlds in each of these sets, rather than the closest antecedent-worlds overall, that matter.

The problem with overreliance on world orderings is also fairly clear in "New Puzzles for Kratzer's Modals". Of the three main problem cases I discuss, two of them could be described similarly to the problem for the Similarity First Paradigm. Specifically, in the cases entitled *Last Call* and *Death or Taxes* the trouble is, broadly speaking, that we are not always required, as the view tells us we are, to bring about the "best" world we can, or choose the "best" option available to us, in the sense of "best" that is appealed to by the theory (namely, via an ordering induced on the worlds in the modal base, i.e. the set of worlds quantified over by modals like *must*, given the particular circumstances of the context). The amendments to the view that I suggest on behalf of the Kratzerian (though I do not personally endorse any version of her view) do not do away with the world ordering, but they do introduce additional structure into the semantics that modulate the role the ordering plays so that other factors contribute significantly as well. (In one case it is a relativization of the modals and the ordering to agents, along with a more sophisticated process for determining the ordering; in the other case it is the postulation of implicit conditional or conditional-like content as part of the meaning of the modal in addition to the universal quantification over the ordered worlds.)

In "Strict Conditional Accounts of Counterfactuals" (forthcoming in *Linguistics and Philosophy*) the problematicity of world orderings is slightly more subtle: the modal domain expansions appealed to by Kai von Fintel's semantics for counterfactuals are structured by the

same similarity/closeness relation appealed to in general by views falling under the Similarity First Paradigm. The first major problem for the view discussed in the paper arises because domain expansions of this kind turn out to be too coarse-grained; some other systematic basis for the expansions would be needed to solve this particular problem. And, perhaps more importantly, a plausible general diagnosis of my problem cases in this paper (a diagnosis which I discuss towards the end) is that von Fintel's view is designed to replicate the standard truth conditions only when counterfactual discourse occurs in an order that is outwardly progressing along the dimension of closeness, and to deliver its alternative truth conditions only when counterfactual discourse occurs in the reverse order; but it seems unlikely that this order of progression even correlates with the relevant data, let alone is central to explaining it.

So in each case, sometimes more blatantly than others, some or all of the problems for the view(s) under discussion arise as a result of overreliance on orderings of possible worlds. Is there something deeper to explore here? Are these symptoms of a broader problem? Should we be leery of world orderings in general, and go looking for similar problems for other views that appeal to them heavily as well? The pattern is certainly suggestive, but more investigation would be needed to say anything with great certainty. I hope to have a chance to do such investigation in my future research. For the moment the discovery of the pattern will have to do.

Cory Nichols

Princeton University

RETHINKING SIMILARITY

CORY NICHOLS

PRINCETON UNIVERSITY

1. THE OLD PARADIGM

Suppose I'm holding a match and a matchbook, and I say:

(MATCH): If I struck this match right now, it would light.

According to the predominant paradigm within philosophy and linguistics, a *counterfactual conditional* like this is true if and only if the match lights at the *most similar* possible worlds where I strike it, i.e. the worlds that are the most like the actual world in the right kinds of ways. In this case they are the worlds that depart from actuality enough so that I do strike the match, but no more – perhaps a few neurons fire differently in my brain, sending a nerve impulse down my arm, leading to a flick of the wrist, dragging the matchhead across the striker… And if every world like this unfolds in a way that leads to the match lighting, then MATCH is true.

David Lewis puts the central idea nicely: "Roughly, a counterfactual is true if every world that makes the antecedent [the *if*-clause] true *without gratuitous departure from actuality* is a world that also makes the consequent [the *then*-clause] true" (1973, p. 41, emphasis added). So, according to the standard conception of similarity in the relevant sense, if a world departs from actuality more than is required to make the antecedent true, it is not amongst the most similar antecedent-worlds. The standard approach thus gives similarity the starring role, and so might be called the *Similarity*

First Paradigm (SFP).[1] Given a *closeness* ordering that ranks worlds according to their overall similarity, the following is a good approximation:

> **SFP:** a counterfactual of the form *If A were the case then C would be the case* is true at a world *w* iff C is true at all the closest worlds to *w* where A is true.

The SFP must be replaced. It may yield correct verdicts on simple cases like MATCH, which are about a particular event at a specific place and time. But many counterfactuals are about a *variety* of ways in which A could be true, some more like actuality than others, and what would happen in each of these scenarios. For these conditionals, as we'll see, the SFP's characteristic emphasis on similarity yields incorrect predictions by privileging the closest A-worlds and ignoring other, less similar A-worlds that are nonetheless also relevant to the truth conditions.

To describe the problem as *pandemic* would not be extravagant: it arises for several apparently disparate species of counterfactuals, not just an isolated few. In the next section (2) I discuss a slew of such cases, some new and some known, organized into different classes, and offer a common diagnosis. In the following section (3) I propose an alternative paradigm that gives *relevance* semantic priority over similarity: a counterfactual is true iff the *relevant* A-worlds are C-worlds. On my preferred way of spelling this out: there are various *ways in which A could be true*, which I call *A-scenarios*, modeled in the theory as subsets of the set of A-worlds; context determines which of these are *relevant*; each of the relevant A-scenarios is associated with a set of worlds where it is witnessed, i.e. where A is true *in that way*; and the union of all of these sets is the set of relevant A-worlds. Similarity still has a role to play, but it is a less central one. In the

[1] See, e.g., Stalnaker (1968, 1981), Lewis (1973, 1986b), Kratzer (1986, 2012), and Bennett (2003). Von Fintel (2001)'s view agrees with these views in most ordinary cases, so my arguments apply to him; Gillies (2007) refines von Fintel's view but stops short of endorsing a similarity-based ordering; if he did so, his view would fall under the scope of my arguments too.

following section (4) I revisit my problem cases to see how my view handles them. In the following section (5) I explore the implications of my view for a series of perennial topics from the counterfactuals literature. And in section (6) I consider and reject the two most initially plausible SFP-friendly solutions, thus bolstering the motivation for my proposal. I end with a brief conclusion (7).

Readers who are interested primarily in the positive proposal might skip or skim over the following portions to save time: from the first full paragraph on p. 6, beginning with "So far…", up to the second full paragraph on p. 8, beginning with "Here is my diagnosis…"; the second half of section 4, beginning at the bottom of p. 15 with "In the case of RAIN…"; and all of section 6.

2. The Problem

Counterfactuals with disjunctive antecedents demonstrate the problem most clearly:

(USA): #If I were in Miami or Havana I'd be in the USA.[2]

This is false. If I were in Miami I'd be in the USA, but of course if I were in Havana (Cuba) I wouldn't. But, on the aforementioned standard conception of similarity, the most similar worlds where I'm in Miami or Havana are likely ones where I'm in Miami. Suppose, for instance, I'm just half an hour outside Miami, in Fort Lauderdale. Then all it takes to get me to Miami, so to speak, is a quick decision to hop in my car and go for a drive. This requires very little *departure from actuality*, to borrow Lewis's expression. But getting me to Havana takes much more: Cuba is farther; it's unreachable by car; and most US citizens still can't go there without a lot of red tape.[3] Since greater departure from actuality makes a world more distant from the actual world,

[2] We precede a sample sentence with # to indicate that the sentence is intuitively false/infelicitous.
[3] This example becomes less forceful as US-Cuba relations warm. Imagine it uttered in 1990.

the closest worlds where I'm in Havana must be farther than the closest worlds where I'm in Miami. So the closest worlds where I'm in Miami or Havana are ones where I'm in Miami, and thus in the USA. And according to the SFP no other worlds matter, so USA is true. Privileging the closest antecedent-worlds neglects the fact that there are two relevant ways that the antecedent could be true – by me being in Miami, and by me being in Havana – and they both ought to have an effect on the truth conditions.[4]

One might suspect the problem lies in some idiosyncratic feature of disjunction.[5] But we find the same problem with determiners like *any*, *some*, or *a*:

(USA-any): #If I were in any city with over a million Cubans I'd be in the USA.

(USA-some): #If I were in some city with over a million Cubans I'd be in the USA.

(USA-a): #If I were in a city with over a million Cubans I'd be in the USA.

These fail for the same reason. The closest worlds where I'm in any city, or some city, or a city with over a million Cubans (Miami and Havana happen to be the only ones) are worlds where I'm in Miami, and thus in the USA. Disjunction is not the problem.

Comparatives and interval-denoting terms can do the trick as well:

(ATLST): #If I were at least 6 feet tall, I'd be exactly 6 feet tall.[6]

(BTWN): #If I were between 6 and 7 feet tall I'd be exactly 6 feet tall.

[4] It has been pointed out before that it may be possible to force a true (or at least better) reading of cases like USA, e.g.: "If I were in Miami or Havana, it would be Miami, not Havana; and then I'd be in the USA; so if I were in Miami or Havana, I'd be in the USA (rather than Cuba)". I find such speeches quite awkward, but perhaps they do force true readings. But we can force the more natural false reading as well, e.g. by stressing *or*, and/or by adding *either*: "#If I were in *either* Miami *or* Havana, I would be in the USA". I find it impossible to hear this as true. And anyway, all I need for my point is that there is some false reading not accounted for by the standard approach.

[5] Loewer (1976) and McKay & van Inwagen (1977), e.g., offer alternative interpretations of the logical form of counterfactuals with disjunctive antecedents. But even if these interpretations are plausible, this will not affect my other cases, including the non-disjunctive variants of USA.

[6] I borrow this case from Alan Hájek (ms); the following one is obviously derivative of it.

These sound false. But given that I'm a bit under 6 feet tall, the most similar worlds where I'm at least 6 feet tall, or between 6 and 7 feet tall, are likely ones where I'm exactly six feet tall. After all, other things being equal, worlds where I grow to six feet and keep growing depart further from actuality than ones where I stop at six feet.

Counterfactuals concerning spans of time can also generate the problem, such as the following now-famous case adapted from Nute (1982). I left my coat at the bar last night, and there were two shady characters who might've stolen it if they'd noticed: Earl earlier in the night, and Nate later in the night. Absent some special reason to think it would not have been Earl, the following seems false:

(COAT): #If my coat had been stolen last night it would've been stolen by Nate.

But the SFP seems to predict COAT to be true. Worlds where my coat is stolen must depart from the actual course of events somewhere. The later this departure occurs in a given world the longer that world preserves the actual history, so the less it departs from our world in total. So worlds with later thefts must generally be closer than worlds with earlier thefts.[7] So it would've been Nate.

So far each class of cases has involved a particular kind of locution: *or*, *any*, *some*, *at least*, *between*, and references to times. So one might guess that all the trouble is due to some idiosyncratic syntactic/semantic features of these particular locutions and the way these features happen to interact with counterfactuals. But there are plenty of problem cases without such locutions. First, the earlier cases can be rephrased to avoid using them. Suppose it had occurred to

[7] The later Lewis (1986b) further spelled out his account in an attempt to yield the result that similarity at or after the antecedent time – the time the antecedent is about – matters less than similarity before it, so one might think COAT has been adequately responded to, since the antecedent of COAT is about a particular time (last night). But it is doubtful he was successful (see Bennett 2003 and Elga 2001), and the result seems undesirable anyway for reasons unforeseen at the time (see Kment 2006). Anyway, it is easy to come up with temporal cases where the intuitively relevant antecedent-worlds differ *before* the antecedent time, e.g. THIEF on the next page.

members of the Cuban diaspora in Florida to coin a term for the mereological sum of Miami and Havana: *Miavana*. Then roughly the same content as USA would be expressed by:

(MIAVANA): #If I were in Miavana I'd be in the USA.

And similar thoughts could be expressed without lexical fiction:

(CULTURE): #If I were surrounded by Cuban culture I'd be in the USA.

(ACCESS): #If I had access to authentic Cuban food I'd be in the USA.

These all sound about as bad as the original USA, but the SFP will count them as true for the same reason. Exactly analogous remarks apply to the variants of USA involving *any*, *some*, and *a*, and, with a little imagination, ATLST and BTWN. And COAT has a counterpart with no mention of time. While leaving the bar, I say:

(THIEF): #If a thief had my coat right now, it would be Nate.

This seems just as bad as COAT. If virtually every problem case with an idiosyncratic locution has an equally problematic counterpart without that locution, then a story about those locutions won't explain much at all.

Furthermore, there are other cases lacking such idiosyncratic expressions or paraphrases thereof that generate the problem simply by being sufficiently general that a variety of scenarios where the antecedent is true are relevant. Suppose the team from Phoenix (where it almost never rains) is beating the team from Seattle (which is famous for rain) by a small margin in fair weather. The following seems false:

(RAIN): #If it were raining Seattle still wouldn't be winning.

Given their experience, Seattle would likely fare *better* if it were raining. But on the standard conception of similarity the most similar worlds where it rains are likely ones where it *just barely* rains, which presumably has a negligible impact on the game. When looking for the closest rain-worlds, so to speak, once we've adjusted one or more weather conditions such as humidity, temperature, or air pressure just enough so that it rains ever so slightly, any further adjustments constitute additional, unnecessary departure from actuality, making the corresponding worlds too dissimilar to be amongst the closest rain-worlds. But this is a wrong-headed way to evaluate RAIN: when thinking about *what would happen if it were raining*, we are interested in a broader variety of scenarios where it rains. The SFP favors the minimal-rain scenario merely because it is a bit more like what actually happens.

So far we've seen a series of counterfactuals each of which is about a variety of related possibilities falling under one description, and the Similarity First Paradigm makes the wrong predictions because it privileges the most similar of those possibilities over the others. But the problem may arise even for simpler cases like the first one we discussed, MATCH. Alan Hájek (ms) worries about what he calls the *implausible specificity problem*.[8] Here is my version of it. Recall MATCH: *If I struck this match right now, it would light*. We supposed earlier that the closest worlds where I strike the match are ones where a cluster of neurons in my brain fire differently, sending the right sort of impulse down my arm. Well, there will likely be many different possible neuron-clusters whose firing would've led to the striking of the match. Suppose one of them – call it *xyz* – would've required less departure from actuality than the others, say by involving fewer

[8] This is not to be confused with another, more famous problem due primarily to Hájek (2007), and discussed also by Hawthorne (2005) and K. Lewis (2016). This problem is often called the problem of *counterfactual skepticism*, and involves the claim that most ordinary counterfactuals are false.

neurons; then *xyz* is the cluster that *would've* fired. This is already counterintuitive, because it means:

(MATCH-xyz): #If I struck this match right now, it would be a result of neuron cluster *xyz* firing, and not any other.

But it gets worse: anything causally necessitated by the firing of *xyz* will obtain at the closest A-worlds too; and given how fine-grained this event is, this will lead to all sorts of other fine-grained counterfactual "facts", such as, for some *m*, *j*, and *a*:

(MATCH-m): #If I struck this match right now, it would take exactly *m* milliseconds to light.

(MATCH-j): #If I struck this match right now, it would emit exactly *j* joules of energy.

(MATCH-a): #If I struck this match right now, I would hold it precisely at angle *a*.

But these sound, as Hájek puts it, *implausibly specific*. Asking precisely what angle I would hold the match at if I struck it feels a bit like asking precisely how many hairs would be on Sherlock Holmes's head if he were real.

It is an open question whether proponents of the similarity-based approach can adequately respond to Hajek's implausible specificity problem, e.g. by finding a workable notion of similarity that ignores the fine-grained differences between *xyz* firing and other clusters firing. But this still would not solve all of the problem cases above, since no reasonable conception of similarity can ignore the much greater difference between, e.g., me being in Miami and me being in Havana.[9]

[9] I defend this claim more extensively in section 6 as part of a discussion of two SFP-friendly alternative solutions that I reject, one of which appeals to the putative context-sensitivity of the closeness ordering, and one of which relaxes the closeness constraint to *the sufficiently close* A-worlds. I postpone this discussion so I can introduce my positive view.

Here is my diagnosis of what has gone wrong in all these cases. For many counterfactuals, there exists *a variety of ways the antecedent could be true*, each of which is relevant: disjunctive antecedents mention these ways within their disjuncts; terms like *any* quantify over a generalized description of these ways; comparative and interval terms denote a spectrum of these ways by highlighting a relevant dimension of possible departure from actuality; temporal terms do something analogous by highlighting a relevant period of time of possible departure; and many otherwise unremarkable antecedents are sufficiently general that they could be true in a variety of different ways, multiple of which are relevant. When multiple ways of being true are relevant for a given antecedent, the counterfactual is about what would happen at *all* of them, not just whichever most resemble actuality. Relevance trumps similarity.

It is worth pausing to emphasize the significance of my proposal. First, the Similarity First Paradigm has been, by far, the predominant paradigm within the counterfactuals literature, in both philosophy and linguistics, for around 50 years. Its status as orthodoxy is so firmly established that it is often appealed to without argument by authors working on related subjects, e.g. van Inwagen (1997), Manley & Wasserman (2008),[10] Hare (2011), Strevens (2012), and Wedgwood (2013). Second, counterfactuals have often been employed to do heavy-duty theoretical work, even quite recently, in analyses of many other philosophically important concepts, such as: dispositions (Lewis (1997), Manley & Wasserman (2008)); causation (Lewis (1986a), Hitchcock (2001)); knowledge (Nozick (1981)); laws of nature (Lange (2009)); decision theory (Gibbard & Harper

[10] Interestingly, Manley & Wasserman (pp. 69-70) discuss a case analogous to my comparative and interval cases, and assume as a premise in one of their arguments that the standard (Lewis-Stalnaker) account's prediction about it is correct. They imagine that a concrete block built to withstand drops from up to 20 meters (but no more) is in fact suspended from more than 20 meters in the air. Observing that the closest worlds where the block is dropped are presumably ones where it is dropped from its current height, at which it breaks, they claim the counterfactual *If the block were dropped from higher than one half meter, it would break* is true, since its current height is higher than one half meter. This strikes me as the wrong judgment; rather, we should think this counterfactual false, and add this case to the list of counterexamples to the Similarity First Paradigm.

(1981)); and others. So if I am correct, and the standard view is not just false but *fundamentally flawed*, this constitutes a radical revision to the status quo and could have serious repercussions throughout philosophy.

3. The New Paradigm

My diagnosis naturally suggests an alternative: *relevance* takes priority over *similarity*. Call this the *Relevance First Paradigm* (RFP). How should we fill in the details? We'll want to substitute relevance for similarity, and of course relevance will depend upon context. So, as a first pass, we can start by restating the truth conditions in those terms (I'll begin using 'A>C' to represent *If A were the case then C would be the case*, as is standard):

RFP-1: A>C is true at a context *c* iff *C* is true at all the relevant worlds at *c* where *A* is true.

But so far this says nothing about what relevance is, and different conceptions of it will yield different finer-grained versions of the view.[11] I will now spell out the proposal a little more carefully, at two levels of further specificity: first, I will give *ways A could be true* a formal representation within the semantics, so that they have an influence on the set of relevant worlds; second, I will suggest how to model this in more detail using only familiar theoretical resources, at least as a working view. This latter step will provide a clearer statement of the view and allow us to test its predictions; if it needs to be amended later, it can be.

Let's work backwards from the problem. The main observation motivating the new approach is that many counterfactuals are about a variety of ways the antecedent could be true,

[11] Strictly speaking the RFP, as stated, is compatible with the SFP, since in principle it could turn out that the *relevant* worlds are always the *most similar* ones. But this would be surprising, since the pragmatic factors that determine *relevance*, in the appropriate sense, are distinct from the factors generally taken to determine similarity.

and the old approach ignored that. So we want, first of all, to include in the semantics some representation of these different ways of being true, and to do so in a way that ensures that each of them has an appropriate effect on the truth conditions. Given a possible worlds framework,[12] this suggests a two-part semantics: there are the *ways of being true*, and there are the corresponding *possible worlds* that will ultimately deliver the truth conditions we need, given the ways of being true. So we would like an account of both: what is a way of being true, and how can we represent its contribution to the truth conditions via the resources provided by possible worlds?

If the counterfactual is about a particular set of ways A could be true, and we would like those ways to influence the set of worlds that provides the truth conditions, the most natural way to achieve this is for each of those ways for A to be true to contribute a set of worlds where A is true *in that particular way* to the total set of relevant worlds. This means the set of relevant worlds would be comprised of a variety of A-worlds, each of which makes A true in one of the relevant ways. To simplify the statement of the theory, let's call a way of being true a *scenario*, so that an A-scenario is a way A could be true, in the sense we have in mind.[13] And let's extend our notion of relevance to scenarios as well, so that the relevant worlds are determined by the relevant scenarios. And, since scenarios are meant to capture the different ways a given A could be true, let's say that a world *witnesses an A-scenario* iff A is true at that world in the way captured by that scenario. Lastly, let's introduce an *assignment function f* that maps each relevant scenario onto the set of worlds it contributes to the truth conditions; the union of all these sets is the set of relevant worlds. (I will say more about both scenarios and the assignment function momentarily.)

---

[12] Plenty of philosophers and linguists are skeptical towards possible-worlds-based analyses for various reasons. I am sympathetic, but for present purposes, given the familiarity of possible worlds and the precision and clarity they provide, I appeal to them in my analysis. I am not opposed to abandoning them if some other sort of analysis turns out to be better.

[13] I intend the term *scenario* to be theory-neutral – unlike the terms *situation*, *state of affairs*, or *possible world*, there are no widely established theoretical associations with the term *scenario* that I am aware of. I also expect the reader will find it intuitive; I have already used it several times.

How does this work? Let's say the counterfactual USA, e.g., is intuitively about two ways the antecedent could be true, *me being in Miami* and *me being in Havana*, and it says (falsely) that I would be in the USA in either case. We've called these *scenarios* in the theory, so there are two scenarios relevant to the evaluation of USA. The assignment function *f* associates each with a set of worlds, and the set of relevant worlds is the union of these. Since each scenario is mapped by the function to a set of worlds *witnessing* that scenario, i.e. where that scenario obtains, the set of relevant worlds will include some worlds witnessing the Miami-scenario, i.e. where I am in Miami, and some witnessing the Havana-scenario, i.e. where I am in Havana. In these latter worlds I am in Cuba, not the USA, so USA is false, as it should be.

The proposal has now reached an intermediate level of generality that is worth pausing to describe. The first-pass formulation of the RFP merely replaced similarity/closeness with relevance: A>C is true iff the *relevant* A-worlds are C-worlds, rather than the *closest* ones. Now, as a second pass, we've honed this a bit: the relevant A-worlds are the ones that witness the relevant scenarios, given some assignment function that maps scenarios to sets of worlds. Let's tidy this up with a more formal statement of the view. Let S_c stand for the set of A-scenarios s_i that are relevant at context *c*. Let *f* be the assignment function that maps each of the s_i to a set W_i of worlds that witness it. The union of the W_i is the set of relevant worlds, and A>C is true iff C is true at all of them, i.e. iff $\cup W_i \subseteq C$.[14] Equivalently, A>C is true iff *f* maps every scenario *s* in S_c onto a subset of *C*:

RFP-2: A>C is true at a context *c* iff $\forall s \in S_c(f(s) \subseteq C)$.

[14] We represent a proposition set theoretically as the set of worlds where it is true, so the union set of the W_i is a subset of *C* iff *C* is true at every member of that set.

The first-pass version of the Relevance First Paradigm replaced closeness/similarity with relevance; now the second-pass version has sketched a framework for spelling this out in terms of relevant scenarios and a mapping from scenarios to worlds. But this still leaves open two major questions: what are scenarios, and what is the assignment function that maps them to worlds?

Here my personal convictions begin to weaken. First, plenty of familiar theoretical entities are reasonable candidates to play the role of scenarios, e.g. sets of worlds, states of affairs, situations, or various conceptions of propositions. A proper comparison of the options is needed, but would lead us too far astray for present purposes. So, as a working view, I will simply construe scenarios as sets of worlds, so that an A-scenario is identified with the set of worlds witnessing it, i.e. the worlds where A is true in the characteristic *way* with which that scenario was associated. I confess some reservations about this identification, which will be explained later. But it makes no ontological commitments beyond the standard appeal to possible worlds, and it allows for an attractively clear and simple formalization of the theory. If this formalization must be complicated further down the road, it may.

Second, we must decide which worlds are associated with each scenario, i.e. what characterizes the assignment function *f*. There are again, in principle, many options: it could be that *all* worlds realizing a given scenario are relevant; or just the closest worlds, on either the standard or some other conception of closeness; or the *sufficiently close* worlds; or something else. Again, the theoretical considerations raised by this question again outrun the scope of this paper. So I will again adopt, with some reservation, a simple working view: *f* maps each scenario *s* to the set of *closest* worlds witnessing *s*, on the standard conception of closeness. Since I have construed scenarios as the sets of worlds witnessing them, *f* effectively selects a subset of each of these sets,

namely the closest members of each. The union of all these subsets, then, is the set of relevant worlds, and A>C is true iff C is true at all of them.

We are finally in a position to provide our third and most fine-grained formal statement of the view. Let $\leqslant$ be an ordering relation such that $w \leqslant w'$ iff w is at least as close to the actual world as w', and let $\leqslant W$ represent the subset of a set of worlds W containing its closest members. Since we've construed scenarios as sets of worlds, and we've construed f as a function from sets of worlds to their subsets containing their closest members, we can now say for a given scenario s that $f(s) = \leqslant s$. And A>C is true iff C is true at the closest worlds witnessing each of the relevant A-scenarios, so:

RFP-3: A>C is true at a context c iff $\forall s \in S_c(\leqslant s \subseteq C)$.

I have now offered my proposal at three levels of generality. First, counterfactuals involve quantification over the *relevant* A-worlds, rather than the *closest* A-worlds, where the relevant worlds may vary with regard to closeness. Second, context determines a set of relevant *ways A could be true*, which I have called scenarios, and the set of relevant worlds is composed of worlds witnessing each scenario, as provided by an assignment function *f*. Lastly, scenarios are construed as sets of worlds, and the assignment function is construed as a function from these sets to their closest members. I expressed some reservations about these last choices; my reservations will be easier to explain below, once we have seen the view in action. In the meantime, it is worth clarifying that I am personally most strongly committed to the second, intermediate characterization of the view. I am convinced the correct analysis of counterfactuals must involve the individuation of different ways in which A could be true, and the selection by context of a privileged subset of these as the relevant ones. I am not so invested in the third, finer-grained proposal, but I cautiously adopt it as a working view for present purposes. If some other conception

of scenarios, or some other assignment function, or even some theoretical entities other than possible worlds, turn out to better represent the truth conditions of counterfactuals, I am happy to revise the corresponding details of the formal semantics.

Before we move on to evaluate the performance of the theory, two more questions are worth briefly addressing. First: Do we really need the division of labor between scenarios and worlds? Isn't it enough to replace closeness with relevance, and rely on an account of the relevant worlds to do the rest of the work? And second: What determines the set of relevant scenarios?

The first answer is that I believe any truth-conditionally adequate procedure for selecting the relevant A-worlds will ultimately have to appeal to scenarios, or ways of making A true, or something equivalent, in the kinds of cases under discussion. Whatever the selection procedure is, in many cases it must be capable of selecting a fairly eclectic set. Suppose the antecedent is *If I were somewhere with a large Cuban population*… Presumably the relevant set of worlds must include some where I'm in Miami, some where I'm in Havana, and some where I'm in other places, e.g. perhaps Tampa, Los Angeles, New York, and so on, even though none of these possibilities has explicitly been mentioned. And it must include *the right* worlds of each kind, for worlds where I'm in Miami and also a billionaire are not relevant. This seems to require a sophisticated process for carving up the set of all A-worlds that involves first grouping them together into subsets according to some unifying features shared by the members of each group, and then choosing the privileged worlds from amongst each group. I don't see how the needed results could be achieved without the structure provided by this grouping step. So I say the unifying feature is that every world in a given group makes A true in the same way, and I call the groups scenarios.

It's important to stress, however, that because the fine-grained version of the proposal (RFP-3) identifies scenarios with sets of A-worlds, it requires no ontological commitments beyond

those to possible worlds and set theory that we have already taken for granted. We could actually state the view without using the term *scenario*, if we like: given a counterfactual A>C at a context *c*, various factors at *c* coordinate to select a privileged set of sets of A-worlds – think of this as a function *g* from A and *c* onto a set *S* of subsets of the set of A-worlds[15] – and the assignment function *f* maps every member of *S* onto a subset of *it*, viz. the set containing all and only its closest members; the union of all *these* subsets is the set of relevant worlds, and A>C is true iff C is true at every world in this set. (In plainer words: context carves the set of A-worlds up in a certain way into subsets; then we take the closest members from each subset; if all these are C-worlds, the counterfactual is true.)

So nothing but a little theoretical simplicity is sacrificed by incorporating scenarios into the semantics, and this additional sophistication may provide the structure needed to handle the data. Admittedly, my argument for this latter claim has been a bit hand-wavy so far. A stronger, clearer case for the appeal to scenarios can be made later on once we have explored some of the implications of my view and some unsuccessful alternatives.[16]

The second answer is more complicated. Presumably the selection of a privileged set of relevant scenarios involves the harmonization of many contextual parameters. This is bound to be true of any theoretical appeal to relevance of the kind I have in mind. Suppose the best account of proper nominal reference is one according to which a proper name like *Smith* always refers to *the relevant Smith*, as selected by context from a set of candidate Smiths. This selection process will

[15] One might wonder whether a *partition* on the set of all A-worlds might do the trick. It won't. The cells of a partitions are by definition *non-overlapping*, since a partition of a set assigns each member of the set to exactly one of the subsets induced by the partition. But we want the sets of worlds representing the relevant scenarios to be able to overlap. For the antecedent *If Sophie and Tina came to the party*…, presumably the relevant A-scenarios are *Sophie coming to the party* and *Tina coming to the party*. But suppose Sophie and Tina have recently started dating, and are joined at the hip lately. Then the closest worlds where Sophie comes to the party are likely ones where Tina does too, and vice versa. So these worlds had better be in both the set of Sophie-worlds and the set of Tina-worlds.
[16] In particular, see the sections below on *might* counterfactuals and the *reordering approach*.

likely turn out to depend upon an intricate complex of factors, e.g. who has been mentioned recently, who the speakers are acquainted with, who is present, what the topic of conversation is, etc. Thoroughly cataloguing these factors and identifying the relationships between them will be a demanding task requiring a great deal of additional research. But this *per se* is no reason to reject the view that *Smith* refers to the relevant Smith, and the analogous complexity of relevance in my theory is no reason to reject it.[17]

A proper investigation of relevance will have to wait. In the meantime, a performance evaluation of the new view is in order, which will also provide some insight into which scenarios are relevant along the way. In the next section I revisit the earlier problem cases to see how my view handles them.

4. SOLUTIONS

Recall the first case, USA. We have already noted that the intuitively relevant scenarios are *me being in Miami* and *me being in Havana*. Since the closest worlds witnessing the scenario where I'm in Havana will be ones where I'm in Cuba, not in the United States, USA will be false.

In the case of USA-any, whose antecedent is *If I were in any city with over a million Cubans*, which scenarios are relevant is largely determined by which cities have over a million Cubans. As it turns out, the only such cities are Miami and Havana, so the relevant scenarios are the same as those for USA, so USA-any is false too. But if New York happened to have over a million Cubans, *me being in New York* would be a relevant scenario too. (Analogous remarks apply

[17] It is worth noting that, Lewis himself appeals in a similar way to a nebulous complex of contextual factors, which putatively determine the similarity facts: "the truth conditions for counterfactuals are fixed only within rough limits; …they are a highly volatile matter, varying with every shift of context and interest" (1973, p. 92). (See section 6ii below, on the "reordering solution", for more on this.) He never spells out in much more detail what these contextual factors are, but this per se is not a reason to reject his semantic framework.

to USA-some and USA-a.) This is one way in which the actual facts can affect which scenarios are relevant.

In the case of ATLST, whose antecedent is *If I were at least 6 feet tall*, intuitively a range of scenarios where I am at least 6 feet tall is relevant, including *me being exactly 6 feet tall*, *me being 6-foot-1*, *me being 6-foot-2*, etc. Of course in any world realizing one of these last two scenarios I am not exactly 6 feet tall, so ATLST is false. (Analogous remarks apply to BTWN.) But are all scenarios where I am at least 6 feet tall relevant? Apparently not, because *me being 9 feet tall* is typically not relevant, since it seems true that if I were at least 6 feet tall I still wouldn't be 9 feet tall. (I'd be the tallest person in recorded history!) What explains this?

One plausible answer is that similarity plays a role here as well, in that scenarios where I am 9 feet tall are too far-fetched or dissimilar, given the context, to be relevant. This suggests a *minimal closeness* constraint on relevance, to the effect that a scenario must be sufficiently close to actuality to be relevant.[18] If this kind of constraint applied in general, the intuitively irrelevant scenarios in cases like ATLST might be eliminated from the relevant set simply in virtue of failing to be witnessed by any sufficiently close world. This would be another significant factor in the selection of the relevant A-scenarios.

However, one can imagine contexts in which the scenario where I am 9 feet tall might be relevant. Suppose we are in a medical school classroom discussing the hormonal imbalance known as *gigantism*, which leads humans to grow far taller than average, and we've been discussing how the physiological makeups of various individuals in the room would be affected if they were to

[18] There may be independent motivation for such a constraint. We routinely make stand-alone *would* claims, like *you would love Rome* or *I wouldn't lie to you*, which seem to be about a fairly wide variety of possibilities. One plausible account of this phenomenon is that the sorts of modal quantifications involved are typically implicitly restricted to something like the *sufficiently close worlds*, subject of course to contextual variation, so that any world too far-fetched is ignored. And Furthermore, Kai Von Fintel (2001) appeals to a similar notion, which he calls the *modal horizon*. I discuss this in "Strict Conditional Accounts of Counterfactuals" (forthcoming).

grow to various unusual heights. It might no longer sound true that I would not be 9 feet tall if I were at least 6 feet tall, and ordinarily far-fetched scenarios might suddenly be entertainable. This is another example of how contextual factors – in this case presumably the topic and purpose of the conversation, at least – can affect which scenarios are relevant.

In the case of COAT, whose antecedent is *If my coat had been stolen last night*, intuitively a range of different times at which my coat could've been stolen are relevant. The fact that worlds where my coat is stolen late in the night (by Nate) preserve similarity to the actual world for a longer period is neither here nor there; if my coat could just as easily have been stolen earlier in the night (by Earl) then some scenarios where Earl steals it are relevant as well. (And, of course, in these scenarios my coat is not stolen by Nate, so COAT is false.)

It's not easy to state exactly which theft-scenarios are relevant, though. We wouldn't want to say, for example, that for every time *t* last night the scenario *my coat is stolen at t* is relevant, because there may have been many times at which my coat wouldn't have been stolen. Suppose between 8 and 8:30 the bar was packed full of police officers, and no one would have dared to attempt a theft. But of course there are *some* possible worlds, albeit far-fetched ones, where some fool does. If we were required to consider for every *t* the closest worlds where my coat is stolen at *t* then these far-fetched worlds would be relevant. Here again a minimal threshold of closeness might suffice to rule out the irrelevant possibilities.

In the case of RAIN, whose antecedent is *If it were raining*, intuitively a range of different ways in which it could rain are relevant. The fact that worlds where it just barely rains are more similar to the actual course of events than ones where it moderately rains is neither here nor there. When we wonder what would happen if it were raining, it is typically appropriate to consider scenarios where it rains moderately as well. In this case, since Seattle has more experience playing

in the rain, presumably some of the worlds witnessing these scenarios will be ones where Seattle performs better, making RAIN false.

Finally, what about the MATCH cases? Here things get tricky. In the original, simple case of MATCH – *If I were to strike this match right now, it would light* – the most natural thing to say is that there is one relevant scenario: *me striking this match right now*. If so then, on my working view, the worlds associated by the assignment function with this scenario are just the closest worlds where I strike the match right now. But for anyone worried about Hájek's implausible specificity problem, this is cause for concern. (I am such a person, hence some of my reservations about this version of the view.) For, recall, if the closest worlds where I strike the match are all ones where it is a result of neuron cluster *xyz* firing, then not only will it have to have been *xyz* that led to the striking, but also anything causally necessitated by the firing of *xyz* would have to have occurred. So MATCH-xyz and the related cases would still be true, counterintuitively.

There are a couple of ways one might try to avoid this consequence. For one, perhaps our choice of assignment function needs to be reconsidered. Suppose it is not just the *closest* worlds witnessing each scenario that are relevant, but some other, more inclusive subset. Then the relevant worlds where I strike the match might include some worlds where a neuron cluster other than *xyz* is responsible for the striking, and thus where the match takes a different length of time to light, emits a different number of joules, and so forth. If there are enough of these worlds in the set, the implausible specificity that Hájek is worried about might be avoided. Alternatively, the individuation of scenarios might be more fine-grained than we have been supposing, at least in some cases. Perhaps the set of relevant scenarios for MATCH includes not just a single scenario, *me striking this match right now*, but a multitude of finer-grained scenarios, e.g. *me striking this match in way* w_1, *me striking this match in way* w_2, *me striking this match in way* w_3, etc. If these

scenarios were sufficiently varied then the set of relevant worlds would be accordingly varied, and we might avoid the implausible specificity problem this way as well.[19] Whether some plausible alternative assignment function or method of individuating scenarios can achieve this result is unclear; for now I set it aside as a subject for future research.

So far the Relevance First Paradigm – particularly the version of it spelled out in RFP-3 above, according to which the relevant worlds are the closest worlds witnessing each scenario – seems better equipped to handle the problem cases than the traditional Similarity First Paradigm. But of course it is important to see how well the view performs on other issues, and what the competition is like. In the next section I discuss the implications of my view for some perennial topics from the counterfactuals literature; in the following section I discuss two potential alternative solutions compatible with the SFP, and argue that neither is promising.

5. IMPLICATIONS

Let's consider the implications of my proposal for some perennial topics from the counterfactuals literature. There are far too many to investigate here, but several of the more well-known issues are worth discussing.

First let's consider a triad of logical principles that famously fail for counterfactuals: *Antecedent Strengthening* (AS); *Transitivity* (TR); and *Contraposition* (CP):

(AS): A>C

[19] Though it's not obvious we will entirely. There are uncountably many moments at which I could've struck the match, uncountably many angles at which I could've held it, etc. Unless there are uncountably many relevant worlds, some of these combinations of moments and angles will be ruled out as counterfactual possibilities. Whether this yields acceptable results will depend on the details, and may vary from case to case. But if it requires uncountably many relevant scenarios, the appeal to scenarios in addition to worlds may seem less plausible.

∴ (A & B)>C

If I went to the movies (A) I'd have a good time (C); but this doesn't entail that if I went to the movies *and got mugged* (A&B) I'd have a good time. AS fails.

(TR): A>B, B>C

∴ A>C

If I quit my job tomorrow (B) I'd have trouble paying my rent next month (C); if I won the lottery today (A) I'd quit my job tomorrow; but this doesn't entail that if I won the lottery today I'd have trouble paying my rent next month. TR fails.

(CP): A>C

∴ ¬C > ¬A

Meryl Streep didn't show up to our party, but the paparazzi did. If Meryl had showed up (A) the paparazzi would've been there (C); but this doesn't entail that if the paparazzi hadn't been there (¬C) Meryl (still) wouldn't have showed up (¬A). (Maybe she only skipped it to avoid the paparazzi.) CP fails.

Similarity-based views get these results quite easily. But so do I. In the case of AS, the closest worlds witnessing the scenario *I go to the movies* might be ones where I have a good time (because, among other things, I don't get mugged); but the closest worlds witnessing the quite different scenario *I go to the movies and get mugged* are presumably ones where I don't have a good time. For TR, the closest worlds witnessing the scenario *I quit my job tomorrow* might be ones where I can't pay my rent, and the closest worlds witnessing the scenario *I win the lottery today* might be ones where I quit my job tomorrow (for a quite different reason); but those worlds

needn't be ones where I also can't pay my rent. And for CP, the closest worlds witnessing the scenario *Meryl Streep shows up at the party* might be ones where the paparazzi are there (since they actually were there, and surely wouldn't have been deterred by her presence); but the closest worlds witnessing the scenario *the paparazzi aren't there* needn't be ones where Meryl doesn't come (since maybe she would've come if the paparazzi hadn't). My account delivers these standard results just as easily as similarity-based views.

Next let's consider another pair of related principles. First, the *Uniqueness Assumption* (UA)[20] says that for any A there is a unique closest A-world. Lewis and Stalnaker famously disagree here: Lewis allows for ties in the closeness ordering so that there may be multiple closest A-worlds, so he rejects UA; Stalnaker, on the other hand, accepts UA, so for him A>C is true iff *the unique closest A-world* is a C-world. For Similarity First views UA goes hand-in-hand with the logical principle *Conditional Excluded Middle* (CEM):

(CEM): For all A, C: $(A>C) \vee (A>\neg C)$

CEM says that, for every A and C, either *If A were case then C would be the case* is true or *If A were the case then C would not be the case* is true. Suppose there is a fair coin that never gets flipped. To accept CEM is to accept that, necessarily, either *If the coin had been flipped it would've landed heads* is true or *If the coin had been flipped it would've landed tails* is true. To deny CEM

[20] UA is often discussed alongside another principle called the *Limit Assumption* (LA), which is the assumption that there is always at least one closest A-world. Stalnaker accepts LA, but Lewis rejects it. He worries in some cases there may turn out to be an infinite series of closer and closer worlds, e.g. if there is a line that is actually one inch long and a series of worlds where it is longer than one inch but closer and closer to its actual length (Lewis 1973, pp.19-21). Lewis complicates his semantics to accommodate such cases, so that A>C is true iff some (A&C)-world is closer than any (A&¬C)-world, i.e. iff, when examining closer and closer A-worlds, so to speak, we reach a point at which all the remaining A-worlds in the ordering are C-worlds. I could, if I wanted, easily amend my view analogously, so that A>C is true iff some (A&C)-world witnessing each scenario is closer than any (A&¬C)-world witnessing that scenario.

is to allow that possibly neither of these is true (which is not to deny that *If the coin had been flipped it would've either landed heads or tails* is true).

Why do UA and CEM go hand-in-hand for the Similarity Paradigm? If there is always a single closest A-world, then that world will by definition decide the truth or falsity of every C, thus making true either (A>C) or (A>¬C). If there can be multiple closest A-worlds, however, then since by definition any two worlds must differ for some C, that C will be neither true at every closest A-world nor false at every closest A-world; this will entail the falsity of both (A>C) and (A>¬C).[21] Lewis rejects UA and CEM along with it; Stalnaker accepts both.

What does my view say? Strictly speaking, I can take or leave UA, since on my view the truth conditions of counterfactuals are not always determined by the facts at the closest A-world(s). If UA is true then there will be a unique closest world witnessing each scenario; but since there can be multiple relevant scenarios, there may still be multiple relevant A-worlds, in which case CEM would fail anyway. So for my view the direct connection between UA and CEM is severed.

But it is likely that Stalnaker's primary motivation for accepting UA was not some metaphysical intuition about the similarity of possible worlds, but rather its consequences for the logic of counterfactuals, especially CEM. The fact that these principles are divorced on a non-similarity-based view like mine demonstrates that it is not UA *per se* that guarantees CEM, but rather the existence of a unique world privileged by the semantics. This suggests a variant of UA – call it *UA** – according to which there is always a unique *relevant* world, drawn (perhaps arbitrarily) from amongst the worlds associated with the relevant scenarios by the assignment function. The resulting view might be preferred by a Stalnakerian who nonetheless wanted to

[21] Strictly speaking there is a third option, which is that there may be multiple closest worlds (i.e. UA is false) but a single world from amongst the closest is chosen as the "privileged" world, so that A>C is true iff the privileged world is a C-world (in which case CEM would hold). This is structurally similar to one of the Stalnaker-inspired variants of my view discussed below.

embrace the multiplicity of A-scenarios. Then counterfactuals would always quantify over a single A-world, so CEM would be validated, but that world might witness any of the relevant scenarios, so it needn't be the closest A-world. So in the case of USA, for example, if both the scenarios *me being in Miami* and *me being in Havana* were relevant, USA would sometimes be true, sometimes false, depending on whether the unique relevant world happened to be a Miami-world or a Havana-world. Furthermore, if we had no way of knowing which it was, then even if USA were strictly speaking true, it would be *unassertible*, which might explain why it sounds bad. (Analogous remarks apply to the other cases.) I don't find this explanation satisfying, but someone with the same motivations and intuitions as Stalnaker might.

I want the set of relevant worlds to be representative of the full range of relevant A-scenarios, so I reject UA*. I'm inclined to reject the original UA as well, but I needn't do so. Whether I do or not, my view will invalidate CEM, and is thus more like Lewis's on this matter. I agree with Lewis here anyway, so I am happy with this result.

There is perhaps more at stake, however, in the related dispute over so-called *might*-counterfactuals, i.e. those of the form *If A were the case, then C might be the case*. (I'll call them *mights*, and their 'would' counterparts *woulds*, and occasionally abbreviate these to *If A, might C* and *If A, would C*.) This is a major topic of discussion in the relevant literature, and, as promised earlier, it also provides motivation for my appeal to scenarios in addition to worlds. For whether we prefer a Lewisian or Stalnakerian treatment of mights, it will be difficult to get their truth conditions right without introducing scenarios into the semantics. Suppose Klaus lives in Miami, and we run in the same social circles, so if I were there I might run into him. But suppose he never goes to Havana, and never would. The following seems false:

(KLAUS): #If I were in Miami or Havana I might run into Klaus.

I might run into him in Miami, but we know I wouldn't in Havana. Like the USA conditionals, KLAUS is about both the Miami-scenario and the Havana-scenario. The natural way to represent this in my theory is, again, to ensure that some worlds witnessing each scenario be relevant to the conditional. But I don't think this could be achieved without appealing to scenarios or something equivalent.

Here is why. There are two major approaches to might-counterfactuals, due primarily to Lewis and to Stalnaker, respectively. On Lewis's view *If A, might C* is true iff *some* closest A-worlds are C-worlds. The most closely analogous version of my view is one according to which *If A, might C* is true iff *some* relevant A-worlds are C-worlds. But this would make KLAUS true. Since I might run into Klaus if I were in Miami, some of the closest Miami-worlds must be Klaus-worlds, so some of the relevant A-worlds will be Klaus-worlds, so KLAUS will be true. In fact, once one relevant scenario makes the consequent true it won't even matter what additional scenarios are relevant. Consider:

(KLAUS+): #If were in Miami or Havana or Antarctica or a thousand miles from wherever Klaus was I might run into Klaus.

Surely this must be false. But so long as the Miami-scenario is relevant and some of the Miami-worlds are Klaus-worlds, the set of relevant worlds will contain Klaus-worlds no matter which other scenarios are relevant. But this seems wrong; the truth of KLAUS requires that *each* scenario be witnessed by some Klaus-worlds.

A better marriage of my view and the Lewisian approach to mights is a view according to which, at least sometimes, *If A, might C* is true iff some worlds associated with each scenario are C-worlds. But this requires grouping the A-worlds into those where I'm in Miami and those where I'm in Havana, and quantifying over each group separately. After all, for Lewis the *might* in might-

counterfactuals is an existential quantifier. If this quantifier takes scope over the entire set of relevant worlds, it will be sufficient for the truth of the counterfactual that *any* of the relevant worlds be C-worlds, regardless of which scenario(s) they are associated with. But the truth conditions of KLAUS require that there be a particular *distribution* of C-worlds, namely some associated with each scenario, which requires quantification individually over each of the relevant scenarios (which, recall, are sets of worlds). I can't see how we could achieve this result without the additional structural complexity achieved by introducing scenarios into the framework.

Stalnaker's approach doesn't fare much better. Stalnaker interprets the *might* in might-counterfactuals as epistemic and wide-scoping, so that *If A, might C* just means *It (epistemically) might be that if A, would C*. On the traditional conception of the epistemic *might*, this is roughly equivalent to *For all I know, if A were the case, C would be the case*. But recall that Stalnaker also endorses the Uniqueness Assumption, i.e. the claim that there is always a single closest A-world. Since on my view it is not the closest A-worlds but the relevant ones that matter, the closest analogue of Stalnaker's view is one according to which there is always a single relevant world, chosen arbitrarily from amongst the relevant scenarios. (I called this UA* earlier.) But if this is the case then KLAUS will be true, since for all I know the relevant world was selected from the Miami-scenario, and some of the Miami-worlds are Klaus-worlds.

Of course one needn't make the change from UA to UA*. The combination of UA with my scenarios-based account is a view where each A-scenario is associated with a single closest world witnessing it, and A>C would be true iff the closest world witnessing each A-scenario was a C-world. Again, this view is unlikely to be embraced by Stalnaker or Stalnakerians, since it invalidates CEM. But combined with Stalnaker's epistemic interpretation of might-counterfactuals, it correctly yields the result that KLAUS is false. For on this analysis KLAUS

would be true iff, for all I know, the closest Miami-world *and* the closest Havana-world are both Klaus-worlds. Since I know Klaus wouldn't be in Havana, I know the embedded conditional is false, so KLAUS is false. But this result again requires a version of the view that appeals not just to a set of relevant worlds, but which separates the A-worlds into groups and ensures that the relevant worlds be drawn from each group. It's hard to see how to deliver the correct truth conditions without this additional step.[22] More needs to be said to establish this conclusion decisively. But scenarios make the job easy.

Finally, another interesting logical consequence of my view is that it severs the connection between what is known as *Strong Centering* (SC) and what I call the *True Antecedent Principle* (TAP). These are often conflated in the relevant literature, but they are quite different. SC is a claim about the closeness ordering to the effect that every world is closer to itself than any other world. (*Weak Centering* says merely that every world is *at least as close* to itself as any other world.) TAP, on the other hand, is a logical principle relating the truth conditions of counterfactuals to the truth values of their antecedents and consequents at the actual world:

(TAP): If (actually) A, then A>C iff (actually) C.

In other words, if A is actually true, then A>C is true if C is true and false is C is false. Often this seems natural. Suppose I say:

(MIJON): If Mijon were at this party he'd be wearing a bowtie.

[22] To be clear, this is not to suggest that no semantics formulated in terms of a set of relevant worlds can be adequate. Indeed, my initial formulation of the Relevance First Paradigm was stated this way. And Karen Lewis (2016, ms) also defends a relevance-based view, according to which closeness is understood as a function of both similarity and relevance. On her view relevance can trump similarity, so a world can be among the "closest" even if it is not amongst the most similar. As far as I can see, nothing about her conception of relevance rules out spelling it out in terms of scenarios.

And suppose we then find out Mijon is, in fact, at the party. Our inclination would be to think that what I said was true if Mijon is wearing a bowtie but false if he isn't. In cases like this, the counterfactual stands or falls with the consequent when the antecedent is actually true.

For Similarity First views, SC directly validates TAP.[23] If every world *w* is closer to itself than any other, then whenever A is true at *w* the closest A-world to *w* is *w* itself. And whenever *w* itself is the closest A-world, any view that says that A>C is true iff the closest A-worlds are C-worlds will say that A>C is true iff *w* is a C-world, i.e. iff C is actually true. Hence TAP.

My view, however, divorces SC from TAP. Suppose SC is true. Then whenever the actual world realizes one of the relevant A-scenarios, the closest (and only) world that *f* associates with that scenario will be the actual world. But whenever other A-scenarios not witnessed by the actual world are relevant (if that is possible) then some non-actual worlds will be associated with them, so TAP will fail for some values of C. But I think this is a desirable result, because I think there are many cases much like the ones discussed earlier – to my knowledge ignored until now – where TAP fails because A and C are both true but A>C is false. For example, even if I were in Miami, and thus in the USA, it would still seem false that if I were in Miami *or* Havana I'd be in the USA. Even worse, suppose I'm sitting on the steps of the Sacré Coeur in Paris, admiring the Eiffel Tower, and I say:

(EIFFEL): If I were here *or anywhere else in the world* I'd be able to see the Eiffel Tower.

This, of course, is ludicrous. But since I am, in fact, "here" or anywhere else, and I can see the Eiffel Tower, TAP tells us EIFFEL is true.

[23] Von Fintel's view is an exception, strictly speaking, because in certain cases counterfactuals quantify over a domain including not only the closest A-worlds but others as well. But in ordinary cases his view predicts the same truth conditions as other similarity-based views, so in these cases the entailment would hold. (And in ordinary settings my problem cases would apply to him as well.)

My view can get the result that MIJON stands or falls with its consequent when its antecedent is true, but that EIFFEL is false even though its antecedent and consequent are true. In the case of MIJON, it seems plausible that only one scenario is relevant, namely *Mijon being at the party*. If Mijon is actually at the party, and SC is true, then *f* will map this scenario onto the actual world, which will be the only relevant world. Then MIJON will be true iff the actual world is a C-world, i.e. iff Mijon is actually wearing a bowtie. In the case of EIFFEL, however, there are many relevant scenarios (though it will presumably be vague which they are). One of them, of course, will be *my being here* (i.e. Paris), but many others will involve my being in places very far away, e.g. Australia. There is no remotely close world where I can see the Eiffel Tower from Australia, so the consequent of EIFFEL will be false at any worlds witnessing any Australia-scenarios, so the counterfactual will be false. I take it to be a point in favor of my view that it can account for cases like EIFFEL where TAP intuitively fails – a virtue not shared by the Similarity First Paradigm.[24]

6. Alternative Solutions

In this section I will consider two potential alternative solutions compatible with the Similarity First Paradigm. Experience has shown that many who consider my data find these options plausible at first glance, so it is important for the motivation of my positive proposal to demonstrate that their prospects are dimmer than they seem. The first might be called the *close enough solution*, and involves a modification of the truth conditions of the Similarity First Paradigm so that *If A*,

[24] We could still preserve TAP, if we wanted, by embracing a principle about scenarios analogous to SC, to the effect that whenever an antecedent-scenario is witnessed by the actual world no non-actual scenarios are relevant. Call this *Strong Scenario Centering* (SSC). The combination of SSC and SC would entail TAP: by SSC, whenever some scenario *s* is actual, no non-actual scenarios are relevant; by SC, the actual world @ is the unique closest world where *s* obtains; by my semantics, *f* will map *s* onto @, and @ will be the only relevant A-world; in this case A>C will be true iff C is true at @.

would C is true iff the *sufficiently close* A-worlds are C-worlds. The second might be called the *reordering solution*, and appeals to the putative context-sensitivity of the closeness ordering.

I. The Close Enough Solution

Some readers will have wondered whether the Similarity First Paradigm's only mistake was in its use of the superlative *closest*, and whether we might solve the problem by simply relaxing this constraint to *close enough*. In other words, perhaps we should amend the original semantics like so:

> **SFP+:** A>C is true at a world *w* iff *C* is true at all the *sufficiently close* worlds to *w* where *A* is true.

Rather than privileging the *very closest* A-worlds and ignoring the rest, this version of the semantics sets a threshold of closeness, presumably dependent upon context, and the truth of the conditional requires that C be true at every A-world that falls within that threshold. And perhaps this threshold is often rather strict, so that the "close enough" worlds are more or less the closest ones, but it can be expanded in contexts where a broader range of scenarios are relevant. In the case of USA, for example, if the threshold of closeness were expanded enough to include some worlds where I am in Havana (perhaps just the closest ones), then USA would be false.

This solution turns out to be too crude, but to see why we will have to consider a new case. So far most of the cases discussed have been intuitively false counterfactuals that the SFP predicts to be true. This is because, in each case, the SFP predicts that the consequent will need to be true at fewer antecedent-scenarios than it intuitively does (just the closest ones, rather than a wider range) in order for the whole conditional to be true. So in effect the SFP makes it easier than it ought to be for counterfactuals to be true; hence the slew of false cases wrongly predicted to be

true. But there are also intuitively true cases where it makes the right predictions, but for the wrong reasons. Suppose again I'm not far from Miami. I've just watched a documentary on Cuban food and culture, and at the end of it I'm hungry and craving some Cuban food. Knowing that Miami and Havana would be amongst the best places to get some, I say:

(FOOD): If I were in Miami or Havana I'd get some Cuban food.

This could very well be true. And the original SFP semantics, given how I've described the case, will predict it to be true, but for the wrong reasons. Intuitively what's required for FOOD to be true is that I would get Cuban food in either Miami or Havana. But the SFP will predict the truth of FOOD so long as I get Cuban food in Miami, since, for familiar reasons, Miami-worlds are ceteris paribus closer than Havana-worlds. Change the case a bit so that I *wouldn't* get Cuban food in Havana – perhaps I'm bizarrely fussy about Cuban food, and I'll only eat it in Miami – and intuitively FOOD is false. But the SFP will still predict it to be true, so long as I get Cuban food at the closest Miami-worlds, since the Havana-worlds are farther.

But doesn't this seem like another case where the modified SFP+ semantics could deliver the right results? If the threshold of closeness expands to include the closest Havana-worlds as well as the closest Miami-worlds, then won't it turn out that FOOD is true if I get Cuban food at those worlds too? Surprisingly, no. Suppose again I would get Cuban food in either Miami or Havana, so FOOD is true. Expanding the threshold of closeness to include the closest Havana-worlds will thereby also include any other A-worlds at least as close. But among these will likely be worlds where I'm in Miami but I don't get Cuban food, which would make FOOD false. Here's why. The closest worlds where I'm in Miami, at which I get some Cuban food, don't require much departure from actuality, since I'm close to Miami, hungry, and craving Cuban food; but it wouldn't require much additional departure from actuality to change my mind so that, say, I'm in

the mood for sushi instead, so I don't get Cuban food; but these are still worlds where I'm in Miami, so the antecedent of FOOD is true but its consequent is false. And if any worlds like this, where I'm in Miami but I fail for some reason to get Cuban food, are closer than or as close as the closest worlds where I'm in Havana, then they will be counted amongst the sufficiently close A-worlds. So FOOD would be false according to SFP+, even though it is intuitively true.

To summarize the problem a bit more succinctly: the truth of FOOD requires that I get Cuban food at the closest worlds where I'm in Miami and the closest worlds where I'm in Havana; but since the latter worlds are farther than the former worlds, there are likely worlds in between them where I'm in Miami but don't get Cuban food for some reason; so if we expand the threshold of closeness to include Havana-worlds, we'll include these worlds too, which will falsify (the intuitively true) FOOD.

Of course it will vary from one case to another whether there are any such worlds within the threshold of closeness falsifying an intuitively true conditional. The claim is not that there always will be. The claim is that there will be whenever the farthest antecedent-worlds within the threshold require a degree of departure from actuality that is greater than the amount required to find a world that makes the antecedent true and the consequent false. And if there is a way of making the antecedent true that requires little departure (e.g. me being in Miami), and a way of making the consequent false that requires little additional departure beyond that (e.g. I suddenly crave sushi), there will likely be such worlds.

The "close enough" solution is too coarse-grained. Simply expanding the threshold of closeness wide enough to include all the intuitively relevant worlds will often include other, intuitively irrelevant worlds along with them. A more nuanced selection process is needed.

II. The Reordering Solution

Another possibility that will have occurred to some readers involves an appeal to a putative context-sensitivity in the closeness ordering. Perhaps counterfactuals are indeed just about the *closest* antecedent-worlds, as the SFP suggests, but the closeness ordering on worlds can vary across contexts in a way that counts all and only the intuitively relevant worlds as closer than any others. After all, some readers may be thinking, doesn't Lewis say something like this anyway?

This solution turns out to be less promising than it sounds. For one, Lewis does say something *like this*, but nothing that would directly lend much independent motivation to this solution. He grants that the closeness ordering may vary across contexts in certain kinds of cases. But the cases motivating Lewis's discussion here are unlike mine, and his remarks about them are not immediately helpful. The most notable example is the *Caesar in Korea* case (1973, pp. 66-67). In different settings either of the following may seem true:

(CAESAR-1): If Caesar were in command in Korea he would've used catapults.

(CAESAR-2): If Caesar were in command in Korea he would've used nuclear weapons.

Lewis suggests that different conversational settings may affect how different aspects of similarity are weighed against one another, yielding different counterfactual judgments. When emphasizing Caesar's actual knowledge of military technology, say, we may prefer worlds with an ancient Caesar somehow transplanted into the 20th Century, using catapults in the Korean War; when emphasizing his ruthlessness in battle, say, we may prefer worlds with a modernized Caesar using atomic bombs. But this case seems fairly anomalous, for it involves atypical conversational settings and an extremely far-fetched counterfactual scenario. My cases are nothing like this: we needn't imagine an atypical conversational setting to evaluate USA and FOOD, and there is nothing particularly far-fetched about my being in Miami or Havana. So even if atypical settings

like those imagined in the Caesar cases can induce atypical variations in the similarity ordering, this does not motivate the suggestion that such variations are commonplace.

Furthermore, it's not even clear there is a viable way of applying the reordering solution. The natural way of spelling out this idea is that different aspects of similarity are assigned values according to the context and weighed against each other (call this a *similarity weighting*) to generate an ordering relative to that context. Here is Lewis: "Overall similarity consists of innumerable similarities and differences in innumerable respects of comparison, balanced against each other according to the relative importances we attach to those respects of comparison" (1973, p. 91). He then draws a useful analogy to the comparison of cities: when I ask whether New York is more like Philadelphia or Los Angeles, say, it depends on what respects of comparison are important. If we're focused on architecture, history, and geography, Philadelphia is the clear winner; if we're focused on size, commerce, and cultural influence, it must be Los Angeles.

Lewis suggests this might explain the two possible judgments about the Caesar scenario: emphasizing similarity regarding Caesar's actually military knowledge yields an ordering where catapult-worlds are closest, so CAESAR-1 is true; emphasizing similarity regarding his ruthlessness yields an ordering where nuclear-worlds are closest, so CAESAR-2 is true. But my problem cases, such as FOOD, aren't much like this. The Caesar case involves two different contexts with two different similarity weightings, and each weighting will prefer a different set of worlds based on the features that characterize the worlds in the set (Caesar's military knowledge, or his ruthlessness). But FOOD involves a single context with a single similarity weighting that, on this proposal, would need to yield an ordering that counts two otherwise comparatively distant sets of worlds as equally close. The Caesar case is analogous to choosing some standards of comparison that count Philadelphia and New York as more similar in one conversation, and

choosing standards that count Los Angeles and New York as more similar in a different conversation. The current proposal, applied to cases like food, is more like choosing a *single* set of standards that counts Philadelphia and Los Angeles as equally similar to New York.[25]

The disanalogy between the Caesar case and my cases undermines the claim that the similarity ordering already exhibits the kind of context-sensitivity we would need to appeal to on the current proposal. But, more importantly, it's not at all clear that there exists a plausible single weighting that will count the Miami-worlds and Havana-worlds as equally close without generating other problematic results. Here's why. One way the closest Miami-worlds differ from the closest Havana-worlds, presumably, is that I have less money in the Havana-worlds. After all, since I'm fairly close to Miami it only costs me, say, a few dollars' worth of gas to get there. But getting to Havana costs a lot more – let's say about a thousand dollars for a plane ticket and other travel expenses. So in the closest Havana-worlds I have about a thousand dollars less to my name, while in the closest Miami-worlds I have about the same as I do now. So if the relevant similarity weighting counts these worlds as equally close, then the difference of a thousand dollars more or less must be irrelevant to it; i.e. the amount of money I have in a world, at least within a thousand dollars or so, must not make a difference per se to how close that world is. But if this is the case then worlds where I'm in *Miami* with a thousand dollars more or less should be counted as equally close as well. But if I were in Miami with a thousand dollars more I might spring for the expensive new sushi place everyone's been talking about, and if I were in Miami with a thousand dollars less I might be pinching my pennies and not eat out at all.[26]

---

[25] This is not to be confused with *indeterminacy* regarding which is more similar. It's easy to imagine, in the absence of some specific set of criteria, denying that there is a determinate fact of the matter whether Philadelphia or Los Angeles is more similar to New York. It's harder to imagine spelling out some specific, non-arbitrary criteria for comparison and finding that the two cities are exactly tied.

[26] The case is even stronger if we add on more scenarios. Suppose I say "If I were in Miami or Havana or Tampa or New York City or anywhere else where I could find good Cuban food, I would get some Cuban food". The more scenarios are relevant, the more worlds we have to treat as equally close that we normally would not, and the more

The claim is not that that every case will run into a problem of this sort; it will depend on the details. The point of this example is simply to demonstrate that even if there is generally a way of adjusting the similarity weighting to yield an atypical ordering that counts two given sets of disparate worlds as equally close, there is no guarantee that these adjustments will propagate throughout the rest of the ordering in ways that yield acceptable verdicts about counterfactuals.

Of course in principle we could always gerrymander any ordering we like, since for any set of worlds there trivially exists *some* ordering that ranks all and only those worlds as highest in the ordering. So some ordering or other ranks all and only the intuitively relevant A-worlds as "closest". But if this ordering cannot be shown to be the product of some set of similarity-based judgments or criteria, then the grounding of counterfactuals in similarity that is constitutive of the SFP is lost. And it is ultimately the reduction of counterfactual judgments to judgments of similarity that gives the account its explanatory virtue. Lewis, again, puts it nicely:

> Somehow, we *do* have a familiar notion of comparative overall similarity, even of comparative similarity of big, complicated, variegated things like whole people, whole cities, or even – I think – whole possible worlds. However mysterious that notion may be, if we can analyze counterfactuals by means of it we will be left with one mystery in place of two. (1973, p. 92)

The Reordering Solution may be tempting, but it is questionable whether it can provide sufficiently nuanced results to make the correct predictions about our cases while remaining sufficiently grounded in similarity to retain SFP's explanatory virtues.

In fact, this example may provide some additional motivation for the appeal in my account to scenarios in addition to worlds. In some cases, different ways of making an antecedent true may

---

aspects of similarity we have to assign low or zero weight, so the more likely we are to count other, intuitively irrelevant worlds as equally close too.

require different standards for closeness. After all, the closest worlds where I am in Miami involve changes to the actual facts that are somewhat different from the corresponding changes in the closest worlds where I am in Havana. It is important to keep my financial situation roughly the same in the Miami-worlds, but not so (or less so) in the Havana-worlds. It isn't perfectly clear how to represent this difference theoretically, but one very natural idea is that these different scenarios require different similarity weightings. Recall, from the very beginning of this paper, Lewis's intuitive gloss on the closest worlds as those "that make the antecedent true without gratuitous departure from actuality" (1973, p. 41). But departure that counts as gratuitous with respect to one antecedent may be necessary for another. Likewise, when there are multiple, disparate relevant ways of making a single antecedent true, departure that is necessary for making A true in one way may be gratuitous when making A true another way. And the Miami-Havana case shows us that what matters is not just *degree* of departure but also *kind* of departure: the closest Havana-worlds are not only more far-fetched than the closest Miami-worlds, but they also depart from actuality in different ways. So to properly understand the theoretical idiom *the closest A-worlds* in terms of the departure from actuality required to make A true, we may have to build into the semantics the flexibility to apply different standards of similarity to different ways of making A, true, even for a single A. On my proposal, this amounts to relativizing the closeness ordering to individual scenarios.[27] If I am correct, this provides another motivation for the additional semantic structure of grouping A-worlds into different scenarios; a semantics stated simply in terms of world relevance with no such grouping, as far as I can see, would be unable to deliver these results.

[27] More formally: suppose there are three relevant scenarios, s_1, s_2, and s_3, and a different similarity ordering appropriate for each one, $\leqslant_1$, $\leqslant_2$, and $\leqslant_3$, respectively. Then f would take as its input not just scenarios but scenario-ordering pairs, and map each scenario onto the set of worlds that are closest according to that scenario's associated ordering, i.e. s_1 onto the $\leqslant_1$-closest worlds, s_2 onto the $\leqslant_2$-closest worlds, and s_3 onto the $\leqslant_3$-closest worlds.

7. CONCLUSION

I have introduced several problem cases for the Similarity First Paradigm that I do not think it can solve. The SFP is undoubtedly the predominant paradigm in both philosophy and linguistics, so if I am correct a major change to the status quo is in order. I suggested an alternative paradigm based on relevance, which I think is better suited to handle my problem cases. On my view: counterfactuals are not always about a single possibility, denoted by the antecedent, and whatever is true at the closest worlds realizing that possibility; rather, they are often about a variety of related possibilities, described by a single antecedent, and what is common of all of them.

I also chose a more specific working view falling under that paradigm as a toy theory that I used to explore the consequences of my proposal. According to this view: the various antecedent-possibilities are *scenarios*, modeled as a set of sets of A-worlds; context privileges a subset of them as the *relevant scenarios*; an assignment function selects the closest worlds in each set as the *relevant worlds*; and A>C is true iff C is true at all the relevant worlds. Ultimately, I am not especially attached to my toy theory; in particular, I am leery of the appeal to the closest worlds witnessing each scenario, because I am worried about Hájek's implausible specificity problem. If some other version of my proposal turns out to be superior, so be it.

It should be emphasized that on my view similarity still has a role to play. We observed that similarity seemed important when selecting the relevant A-scenarios, since excessively far-fetched ones are generally irrelevant. And on the more fine-grained version of the view *f* assigns the closest worlds realizing each scenario to the set of relevant worlds. But on my view similarity is not the *primary* factor; relevance comes first.

I have also explored the implications of my view for some perennial topics from the counterfactuals literature, and found no cause for concern. In fact, I found it to be a virtue of my

view that it breaks from orthodoxy in rejecting the connection between *Strong Centering* and what I dubbed the *True Antecedent Principle*. I also found that the consideration of so-called *might* counterfactuals provided further support for my appeal to scenarios.

Finally, I considered two common SFP-friendly alternative solutions and argued that they were inadequate to handle the cases I have discussed. Moreover, a problem with one of them revealed another potential motivation for my view.

There is of course much more to say about all of this. This paper is intended to lay the initial groundwork for an alternative style of counterfactual semantics that can be improved upon in a number of ways in the future. I expect to develop the view further in subsequent research.

WORKS CITED

Bennett, Jonathan (2003). *A Philosophical Guide to Conditionals*. Oxford University Press.

Elga, Adam (2001). Statistical Mechanics and the Asymmetry of Counterfactual Dependence. *Philosophy of Science* 68 (S1).

von Fintel, Kai (2001). Counterfactuals in a Dynamic Context. *Current Studies in Linguistics* series 36, 123-152.

Gibbard, Alan and Harper, William L. Counterfactuals and Two Kinds of Expected Utility. In *Ifs: Conditionals, Belief, Decision, Chance, and Time*. Harper, William L., Stalnaker, Robert, and Pearce, Glenn (eds.). D. Reidel.

Gillies, Anthony S. (2007). Counterfactual Scorekeeping. *Linguistics and Philosophy* 30 (3): 329-360.

Hájek, Alan (2007, unpublished). Most Counterfactuals Are False.

Hájek, Alan (manuscript in progress). Begging to Differ with Similarity Accounts of Counterfactuals.

Hare, Caspar (2011). Obligation and Regret When There is No Fact of the Matter About What Would Have Happened if You Had not Done What You Did. *Noûs 45 (1): 190-206.*

Hawthorne, John (2005). Chance and counterfactuals. *Philosophy and Phenomenological Research* 70 (2):396–405.

Hitchcock, Christopher (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *The Journal of Philosophy* 98 (6): 273-299.

van Inwagen, Peter (1997). Against Middle Knowledge. *Midwest Studies in Philosophy* 21: 225-236.

Kment, Boris (2006). Counterfactuals and explanation. *Mind* 115 (458):261-310.

Kratzer, Angelika (1986). Conditionals. *Chicago Linguistics Society* 22 (2):1–15.

Kratzer, Angelika (2012). *Modals and Conditionals*. Oxford University Press.

Lange, Marc (2009). *Laws & Lawmakers: Science, Metaphysics, and the Laws of Nature*. Oxford University Press.

Lewis, David (1973). *Counterfactuals*. Blackwell.

Lewis, David (1986a). Causation. Reprinted with postscripts in *Philosophical Papers, Volume II*. Oxford University Press.

Lewis, David (1986b). Counterfactual Dependence and Time's Arrow. Reprinted in *Philosophical Papers, Volume II*. Oxford University Press.

Lewis, David (1997). Finkish Dispositions. *The Philosophical Quarterly* 47 (187): 143-158.

Lewis, Karen (2016). Elusive Counterfactuals. *Noûs 50: 286-313.*

Lewis, Karen (manuscript in progress). Counterfactual Discourse in Context.

Loewer, Barry (1976). Counterfactuals with Disjunctive Antecedents. *The Journal of Philosophy* 73: 531-537.

Manley, David and Wasserman, Ryan (2008). On Linking Dispositions and Conditionals. *Mind* 117 (465): 59-84.

McKay, Thomas and van Inwagen, Peter (1977). Counterfactuals with Disjunctive Antecedents. *Philosophical Studies* 31: 353-356.

Moss, Sarah (2012). On the Pragmatics of Counterfactuals. *Noûs* 46 (3): 561-86.

Nichols, Cory (forthcoming). Strict Conditional Accounts of Counterfactuals. *Linguistics and Philosophy*.

Nozick, Robert (1981). *Philosophical Explanations*. Harvard University Press.

Nute, Donald (1982). Topics in Conditional Logic. *Mind* 91 (361):136-138.

Stalnaker, Robert C. (1968). A Theory of Conditionals. In Nicholas Rescher (ed.), *Studies in Logical Theory*. Blackwell.

Stalnaker, Robert C. (1981). A Defense of Conditional Excluded Middle. In William Harper, Robert C. Stalnaker & Glenn Pearce (eds.), *Ifs*. Reidel.

Strevens, Michael (2012). The Explanatory Role of Irreducible Properties. *Noûs* 46 (4): 754-780.

Wedgwood, Ralph (2013). Rational 'Ought' Implies 'Can'. *Philosophical Issues* 23: 70-92.

Strict Conditional Accounts of Counterfactuals

Cory Nichols[28]

This paper is forthcoming in Linguistics and Philosophy. *All citations should refer to the official version.*

1. A Peculiar Asymmetry

Until fairly recently it seems to have gone unnoticed that counterfactuals – conditionals of the form *If A were the case, then C would be the case* – exhibit a surprising asymmetry: so-called *Sobel sequences* like those below are fine in one order but infelicitous in reverse. Suppose Jeff and Lars are delightful party guests alone, but together they always fight. Consider:

(PARTY): (a) If Jeff had come to the party, it would've been fun. (b) If Jeff and Lars had come to the party, it wouldn't have been fun.

(PARTY-R): (a) If Jeff and Lars had come to the party, it wouldn't have been fun. (b) #If Jeff had come to the party, it would've been fun.

The typical reactions are that PARTY sounds fine throughout but PARTY-R(b) is infelicitous. Here is another case: the Yankees have just won the World Series, and baseball legend Derek Jeter is sure to make an appearance at the victory parade. Consider:

[28] Special thanks to Boris Kment, Karen Lewis, Alan Hájek, Gideon Rosen, Paolo Santorio, Jack Woods, and Matt Moss for invaluable discussion and/or comments on earlier versions of this paper; to audiences at New York Philosophy of Language Workshop, Princeton Philosophical Society, the Eighth Barcelona Workshop on Issues in the Theory of Reference, Boğaziçi University, and the Second Belgrade Conference on Conditionals for extremely useful questions and feedback; and to anonymous reviewers in this journal for helpful suggestions.

(SOPHIE): (a) If Sophie went to the parade, she would see Jeter. (b) If Sophie went to the parade and got stuck in the back of the crowd, she wouldn't see Jeter.

(SOPHIE-R): (a) If Sophie went to the parade and got stuck in the back of the crowd, she wouldn't see Jeter. (b) #If Sophie went to the parade, she would see Jeter.

It isn't clear what accounts for this asymmetry. The standard semantics for counterfactuals is known as the variably strict account (VSA), due primarily to David Lewis (1973, 1986) and Robert Stalnaker (1968, 1981). On this view worlds are ranked according to their *closeness*, where the closeness of a world is a function of how *similar* it is in certain ways to the actual world. A counterfactual A>C is true, then, iff (roughly) all the closest A-worlds are C-worlds.[29] Lewis puts the central thought nicely: "Roughly, a counterfactual is true if every world that makes the antecedent true *without gratuitous departure from actuality* is a world that also makes the consequent true" (1973, p.41, emphasis added). But if facts about similarity comparisons between worlds underwrite the truth values of counterfactuals, why should their order of utterance make any difference? The similarity facts are not generally taken to be so deeply context-sensitive. Some further explanation is needed.

2. Strict Conditionals

Philosophers had long dismissed the possibility of a strict conditional account of counterfactuals, according to which A>C simply means *necessarily, if A then C*:

[29] Lewis and Stalnaker famously disagree about several further details of the variably strict account, and there are many controversial issues surrounding the notions of similarity and closeness. For a good overview see Bennett (2003). I can ignore these controversies for my purposes.

(SCA): A>C is true iff □(A⊃C)

The necessity operator seems too demanding: it might be true that if I went to the zoo (A) I would have fun (C); but surely there is *some* possible world where I go to the zoo and don't have fun (A&¬C), e.g. if I get mauled by a lion; so A>C is true but □(A⊃C) is false. So the idea of treating counterfactuals as strict conditionals has historically not been taken seriously.

More recently, however, Kai von Fintel (2001) and Anthony Gillies (2007) have given us reason to take the idea seriously.[30] In natural language we routinely quantify over restricted domains, and these restrictions can rapidly change throughout a conversation. Looking in the fridge, I say: "All the wine is gone; luckily, there's more in the cellar". Lest I be interpreted as contradicting myself, the first part of my utterance must be understood as quantifying over a different domain – something like *the immediately accessible wine* – from that of the second part of my utterance – something like *the wine at our disposal*. Von Fintel's view begins with the observation that strict conditionals systematically restricted in a similar way might preserve the standard truth conditions in ordinary cases while predicting non-standard truth conditions in idiosyncratic cases like reverse Sobel sequences. Lewis (1973), writing before these data were known, dismissed this style of analysis as ad hoc and defeatist. But von Fintel's account seems neither, and purports to explain the peculiar asymmetry mentioned above.

The view has two major parts. First, there is the strict conditional analysis: A>C just means □(A⊃C). Second, there is a *dynamic modal domain*: the set of worlds quantified over evolves throughout a conversation as speakers discuss new possibilities. More specifically, the

[30] Von Fintel and Gillies differ on some of the details. Most important for present purposes is that Gillies ultimately hedges on whether to appeal to a closeness-based ordering as von Fintel does. As a result his view is more difficult to evaluate, so I will focus on von Fintel's view. Even abandoning closeness, however, will only avoid the first of my four classes of problem cases discussed in the next section.

operative modal domain D_c at any context point *c* is demarcated by the *modal horizon*, i.e. the "outer" limit that determines which worlds are included in the domain, and the modal horizon is expanded as necessary to accommodate new possibilities under discussion. Why? A quantified modal claim of the form *All φ-worlds are ψ-worlds* interpreted at a domain including no φ-worlds is vacuously true. So in order to give speakers a chance of saying something non-vacuous, conversational participants typically broaden the modal horizon until it reaches some A-worlds (A being the antecedent of the conditional), which demarcates a new domain D_{c*} at a new context c*. But when the domain already includes A-worlds no such change is needed, so $D_{c*} = D_c$.

A few further details are needed for this account to predict the asymmetry:[31]

(i) Expansion is *closeness-based*: worlds are ordered by closeness in the traditional way, and the modal horizon is understood as a location in the ordering. In other words, when the modal horizon is broadened, this amounts to admitting more distant worlds into *D*.

(ii) Expansion is *conservative*: when *D* expands to include some A-worlds, it expands only as much as is necessary to do so. Combined with (i), this means expansions will only include the *closest* A-worlds for the A inducing the expansion.

(iii) Expansion is *coarse-grained*: when *D* expands to include the closest A-worlds, it also includes all other worlds at least as close, whether or not they are A-worlds.[32]

---

[31] These clauses and terminology are features of my preferred presentation of the view, not von Fintel's own.

[32] For the interested reader, we could spell out the view more formally as follows. Let ≤ be an ordering relation such that $w \leq w'$ iff *w* is at least as close as *w′*, let $\boldsymbol{\varphi}$ represent the set of worlds where φ is true, and let the abbreviation $\forall w \in \boldsymbol{\varphi}$ represent the restricted quantification *all worlds* w *in* $\boldsymbol{\varphi}$. Then:

> (SCA): A>C uttered at a context *c* including domain D_c is interpreted as □(A⊃C) at a new context *c** quantifying over domain D_{c*}, such that:
> (i) if $D_c \cap \mathbf{A} \neq \emptyset$ then $D_{c*} = D_c$
> (ii) if $D_c \cap \mathbf{A} = \emptyset$ then $D_{c*} = \{w: \forall w' \in \mathbf{A}(\forall w'' \in \mathbf{A}(w' \leq w'') \supset w \leq w')\}$

Now we are in a position to see how the semantics works. Suppose there are no worlds in D_c where I go to the zoo. Then when I say at *c* that if I went to the zoo I would have fun, D_c expands until it includes some worlds where I go to the zoo. Because of (i) these will be the *closest* zoo-worlds; because of (ii) the expansion will stop there. The counterfactual is then evaluated as a strict conditional – $\Box$(*I go to the zoo* $\supset$ *I have fun*) – at the new context c* quantifying over the new domain D_{c*}, and thus is true iff every world in D_{c*} makes the antecedent false or the consequent true. The only worlds in D_{c*} where the antecedent is not false are (the closest) worlds where I go to the zoo; so if all those worlds make the consequent true, i.e. I have fun, the strict conditional is true. So A>C is true so long as the closest A-worlds are C-worlds. So whenever expansion occurs the dynamic strict conditional account predicts the same truth conditions as the standard variably strict account.

But something different happens with reverse Sobel sequences. For convenience, let's say informally that φ is a *closer possibility than* ψ when the closest φ-worlds are closer than the closest ψ-worlds. In the forward versions of Sobel sequences, then, the first conditional's antecedent denotes a closer possibility than the one denoted by the second. In PARTY, for example, the possibility of Jeff being at the party without Lars must be closer than the possibility of both of them being there. After all, if it's true that it would've been fun if Jeff had come, and also true that it *wouldn't* have been fun if Jeff *and Lars* had come, then the closest worlds where Jeff comes must be ones where Lars doesn't. Let's make this clearer with a diagram, where @ =

In something closer to ordinary English: if the intersection of D_c and **A** is non-empty, i.e. there is an A-world in the initial domain, then the "new" domain D_{c*} is just D_c; but if the intersection is empty, i.e. the initial domain includes no A-worlds, then D_c is replaced by an expanded domain D_{c*} consisting of the set of worlds *w* such that: for any world *w*′ in **A**, if *w*′ is at least as close as any world *w*′′ in **A** (i.e. if w′ is a closest A-world), then *w* is at least as close as *w*′. (This last clunky bit says that D_{c*} is the set of *w*s at least as close as any closest A-world. This is necessary because expansion is coarse-grained, i.e. not every *w* included in an expansion need be an A-world.)

the actual world, J = *Jeff comes to the party*, L = *Lars comes to the party*, F = *the party is fun*, and position from left to right represents distance from the actual world:

Diagram 1

==========>

@	...	*w1*	...	*w2*	…
~J, ~L		J, ~L, F		J, L, ~F	

At the actual world, @, Jeff and Lars didn't come. The closest worlds where Jeff does come, labeled *w1*, are closer than the closest worlds where both Jeff and Lars come, labeled *w2*. Supposing these worlds are not in the initial domain, SCA predicts that the domain for the forward version of the sequence will evolve in the normal way: PARTY(a) induces an expansion to include *w1*-worlds, where Jeff is at the party, but not *w2*-worlds, where Lars is there too; then (b) induces a second expansion to include the *w2*-worlds as well. But in the reverse case the larger (*w2*) expansion occurs first, so the smaller (*w1*) one is unnecessary. That is, an utterance of PARTY-R(a) (which is the same conditional as PARTY(b)) expands the domain to include *w2*-worlds, which of course also includes the closer *w1*-worlds; but then no domain change is needed to accommodate the subsequent utterance of PARTY-R(b), whose antecedent is *if Jeff had come to the party*, for the *w2*-worlds are themselves such worlds (as are the *w1*-worlds). Since no analogous method of domain contraction exists[33], the counterfactual is evaluated at this

[33] Von Fintel observes that in some cases we can eliminate the infelicity of a reverse Sobel sequence by explicitly signaling that the possibility just mentioned is to be ruled out, e.g.: "If Jeff and Lars had come to the party, it wouldn't have been fun. *But Lars wasn't at the party*. So if Jeff had come to the party, it would've been fun." In these cases, he says, the effect of the intermediate claim is to induce a domain contraction (to where is unclear). But in ordinary cases such contractions do not occur.

domain, and is thus true iff every world in it makes the material conditional *Jeff comes to the party* ⊃ *it is fun* true. But *w2*-worlds are worlds where Jeff comes and it isn't fun, because Lars is there too. So the counterfactual is false in this context. (SOPHIE-R is exactly analogous.)

So a strict conditional semantics supplemented with the technology of a dynamic modal domain seems competent to get basic cases right, and appears to have an elegant explanation of the asymmetry exhibited by Sobel sequences.

Before moving on, it is worth noting that there is another major motivation for von Fintel's framework. Conditional antecedents are one of a handful of locutions that grammatically license the appearance of *negative polarity items* (NPIs), so-called because they characteristically occur in linguistically "negative" environments. Two paradigm NPIs are *any* and *ever*, which are permitted under the scope of negation, as in the sentences "We don't have any wine" or "I don't think she ever drinks", but not in their "positive" counterparts "We have any wine" and "I think she ever drinks". But NPIs turn out to occur in a handful of other, non-negative environments as well (the term is something of a misnomer), including the antecedents of conditionals, e.g. "If we'd drunk any wine, it would've been red" or "If she ever drinks, she drinks wine".

The further details of the relevant data and literature exceed the scope of this paper (and the expertise of its author). In short, it appears the unifying feature that explains the linguistic distribution of NPIs must be related to the entailment patterns validated by the environments they prefer,[34] so we would like a theory of conditionals whose logic agrees with those patterns. Von Fintel's dynamic strict conditional account does so, but Lewis's variably strict account does not, so this is a point in favor of von Fintel. All of this is orthogonal to the data I will discuss here,

[34] Particularly patterns involving *downward entailment/monotonicity*. Von Fintel attributes this observation to Ladusaw (1979); see von Fintel (2001, p. 132-3) for more on this, and von Fintel (1999) for more on NPIs in general.

but the fact is worth mentioning, if only as a dialectical signpost, that there remains a second motivation for von Fintel's view that is completely unaddressed by this paper. So at best I will be able to claim by the end to have undermined one of the two.

3. Trouble for the Strict Conditional Account

The discussion in the relevant literature focuses primarily on pairs of counterfactuals in which one antecedent is a logically strengthened version of the other, typically involving an additional conjunct. For example, *Jeff and Lars come to the party* is a strengthened version of *Jeff comes to the party*. But this constitutes a fairly narrow class of cases, and exploration of further data reveals several classes of cases about which the view, as it stands, makes bad predictions.

I. Intermediate Worlds

Recall that domain expansion, according to SCA, is coarse-grained: when the domain expands to include the closest A-worlds for some counterfactual, it includes all other worlds at least as close. This turns out to be too permissive, as the following cases demonstrate.

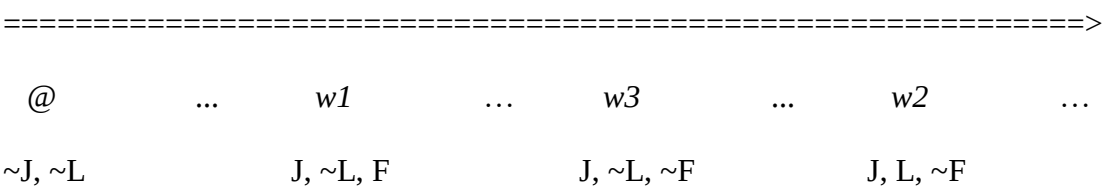
First, given the earlier description of the party scenario, this seems true:

(NOTLARS): (a) If Jeff and Lars had come to the party, it wouldn't have been fun. (b) If Jeff but not Lars had come to the party, it would've been fun.

We know from a moment ago that the closest worlds where Jeff comes to the party are ones where Lars doesn't come, and it's fun (*w1*-worlds). And we know that the closest worlds where they both come, at which it's not fun (*w2*-worlds), are farther than these. (All this was represented by Diagram 1.) Now suppose further that the *w2*-worlds are *much* farther than the

w1-worlds, i.e. involve a significantly greater degree of departure from actuality.[35] For example, suppose Jeff lives down the street and nearly came to the party, but Lars was officiating a wedding in California, 3,000 miles from the party in New York. Then *w1*-worlds wouldn't require much departure from actuality – say, Jeff changes his mind on a whim and decides to come after all – but *w2*-worlds would require a fair amount of departure from actuality – say, Lars is willing to disappoint his friends, skip their wedding, cancel his trip to the West Coast, and come to a party where Jeff will be. If the *w1*- and *w2*-worlds are this far apart then there are likely *intermediate worlds* in between them where Jeff but not Lars comes to the party, *but it still isn't fun*. For example, suppose Jeff is characteristically gregarious, but had his day gone a bit differently he would've been uncharacteristically ornery, which wouldn't have been fun. Or suppose the party almost ran out of beer, and an additional large group almost showed up who would've drunk up the last of it, which also wouldn't have been fun. If any worlds like these, where Jeff comes to the party without Lars but it isn't fun (*w3-worlds* in Diagram 2), require less departure from actuality than *w2*-worlds, then the ordering will be:

Diagram 2:



[35] It's tempting to think and speak as though there is a metric on similarity/closeness, such that one could make good sense of a claim like "world *w* is three times as far from @ as world *w*′ is". Lewis is skeptical of this idea (1973, 50-52), but does not reject the possibility. But we needn't assume such a metric to make good sense of weaker claims like "world *w* is *very* far from @", just as I needn't assume that my preferences are determinate and precise (though they could be!) to believe that I *greatly* prefer chocolate to vanilla. In fact, it is sufficient for my cases here merely that we can make good sense of intuitions of the form: "world *w* involves at least as much departure from @ (of the relevant kind) as world *w*′ does". And indeed, these seem to be the very sorts of judgments that are presumed by proponents of similarity-based accounts to underlie our capacity to evaluate counterfactuals.

And if this is the case NOTLARS(b) must be false. For (a) would expand the domain to include *w2*-worlds, thereby also including both *w1*- and *w3*-worlds; and since *w3*-worlds make the antecedent of (b) true but its consequent false – Jeff but not Lars comes to the party, but it isn't fun – (b) would be false. But intuitively, of course, it's true.

It is important to appreciate that it is the structure of the case that matters, not the particular details. This intermediate worlds problem can arise whenever more departure from actuality is required to make the antecedent of an earlier counterfactual true than to make a subsequent counterfactual's antecedent true but its consequent false. And of course there are no constraints on how far apart the various antecedent-possibilities in a counterfactual sequence may be. So once the recipe is clear, cases like the previous one are easy to cook up. Here is an even simpler one. Suppose Tina almost came to the party too, and it would've been fun. Then:

(TINA): (a) If Lars had come to the party, it would've been fun. (b) If Tina had come to the party, it would've been fun.

But since Tina almost came to the party, but Lars was never going to come, worlds where Tina comes are likely to be much closer than worlds where Lars comes. Then the same sorts of intermediate worlds, e.g. where Tina comes but we run out of beer, may be included in the domain expansion induced by (a), thus falsifying (b).

Analogous variants of SOPHIE are easy to come by too: suppose Sophie lives in Australia, and never seriously considered flying to New York for the parade, but had planned to watch it on TV. The following may well be true:

(ONTV): (a) If Sophie had gone to the parade, she would've seen Jeter. (b) If she had watched it on TV, she would've seen him too.

But if worlds where she watches on TV are much closer than ones where she goes in person, and it doesn't require much additional departure from actuality to find worlds where she watches on TV and somehow misses Jeter – e.g. if her TV reception cuts out – then these intermediate worlds will be included in the expansion induced by (a), thus falsifying (b). So intuitively such worlds ought not to be included in expansions; coarse-grained expansion is too permissive.

II. Falsifying Antecedents

At this point the following line of thought is natural: the intermediate worlds problem was a result of coarse-grained expansion, which includes not just the worlds inducing the expansion (the closest A-worlds), but also any other world at least as close, some of which falsify subsequent conditionals in unexpected ways. So if we could screen off these problematic intermediate worlds in a principled way, perhaps we could avoid this problem.

Whether or not such worlds can in fact be successfully screened off will be discussed below when considering possible responses on behalf of SCA. But even if they can be, this solution will do nothing to avoid the remaining cases. In particular, this next class of cases involves sequences in which even the first conditionals' antecedent-worlds falsify the second conditionals, and these are the very worlds the relevant expansions are meant to include. Suppose Lars wanted to go to the beach instead of the wedding. Then this might well be true:

(BEACH): (a) If Lars had come to the party, it would've been fun. (b) If he hadn't been at a wedding on the West Coast that day, he would've gone to the beach.

Since Lars is great fun at parties, (a) is presumably true. And we can easily imagine that (b) is true. But the closest worlds where the antecedent of (a) is true, i.e. where Lars is at the party in New York, are worlds that make (b) false, i.e. where it's true that he isn't at the wedding on the West Coast, but it's false that he goes to the beach (assuming he can only do one of the three). So even an expansion including *only* the closest antecedent-worlds for (a) would make the antecedent of (b) true but its consequent false.

An analogous version of the SOPHIE/ONTV case is easy to imagine:

(COUCH): (a) If Sophie had gone to the parade, she would've seen Jeter. (b) If she hadn't fallen asleep on her couch just before it began, she would've watched it on TV.

Given that Sophie lives in Australia, worlds where she goes to the parade in New York are worlds where it's true that she doesn't fall asleep on her couch just before it begins, but where it's false that she watches it on TV.

In the previous section we observed that coarse-grained expansion would lead to problems by admitting falsifying intermediate worlds into the domain. But even the most fine-grained expansion possible, adding to the domain only the antecedent-worlds for the conditional inducing the expansion, would lead to problems as well.

III. Complex Evolution

At this point another line of thought is natural: perhaps the original version of SCA, according to which domain expansions by default endure, is too simple. Perhaps what these cases show is that

domains are often reset or contracted. So perhaps if we could explain when expansions endure and when they do not, we could avoid this problem.

Whether or not there is an adequate solution to the previous cases in this vicinity will also be discussed in the response section below. But even if a mechanism for domain contraction or resetting is built into the formal framework, with a corresponding reliable method for predicting when expansions endure, the semantics will not yet be sophisticated enough to deal with the following cases. Suppose the problematic cases above, such as BEACH and COUCH, really did induce some sort of domain contraction or resetting. Then we would expect subsequent domain expansions to continue to occur in the usual way: when there are no A-worlds in the domain, the modal horizon is broadened just enough to include the closest ones, and the counterfactual in this context should have the ordinary truth conditions. But this cannot be right. This would mean that after a putative contraction we would generally *not* find the same sorts of infelicities observed in the original reverse Sobel sequences. For what causes these infelicities, according to SCA, is when an enduring earlier expansion causes a later conditional to be evaluated at the previously expanded domain rather than the domain at which it would ordinarily be evaluated. So in cases where such an expansion is reversed, we should generally not expect it to be able to generate the same sort of infelicity in later conditionals.

But in fact we find the opposite. Consider a variant of BEACH:

(BEACH2): (a) If Jeff and Lars had come to the party, it wouldn't have been fun. (b) If Lars hadn't been at a wedding on the West Coast that day, he would've gone to the beach. (c) #If Lars had come to the party, it would've been fun.

The tension between (a) and (c) of this sequence is exactly analogous to the one between (a) and (b) in the original PARTY-R[36], so the explanation ought to be essentially the same. This means the expansion induced by BEACH2(a), after being reversed for the evaluation of (b) (as is the current proposal), would then have to be *reinstated* for the evaluation of (c), rather than (c) inducing its own expansion in the usual way. For if (c) induced its own ordinary expansion, the domain at that point would include the closest worlds where Lars comes to the party, at which Jeff does not come and it is fun, but not the worlds where they both come and fight. So we would evaluate (c) in the usual way, at the domain including only the closest A-worlds, and should judge it to be true. But it is in fact infelicitous. So something additional must be added to the semantics to reinstate the earlier domain in this case, even though it is not necessary for the accommodation of the new antecedent.

An exactly analogous variant of the Sophie case is just as easy to construct:

(COUCH2): (a) If Sophie had gone to the parade and gotten stuck in the back, she wouldn't have seen Jeter. (b) If she hadn't fallen asleep on her couch just before it began, she would've watched it on TV. (c) #If she had gone to the parade, she would've seen Jeter.

Since worlds where Sophie is at the parade in New York are *a fortiori* worlds where she does not fall asleep on her couch in Australia just before it begins, but where she also doesn't watch it on TV, these worlds would falsify (b). So the domain expansion that includes them, induced by (a),

[36] The meticulous reader will have noticed that BEACH2(c) is about Lars coming to the party, rather than Jeff. The example is simpler this way. We can suppose that if Lars had come Jeff would still not have come, so that the closest worlds where Lars comes are ones where the party is fun. This is compatible with the claim that *if* they both had come, it wouldn't have been fun.

must be reversed for the evaluation of (b). But then if (c) induced another expansion in the normal way the domain would expand just enough to include the closest worlds where Sophie goes to the parade, at which she sees Jeter, so (c) should be true. But it is in fact infelicitous. So, again, the earlier domain must be reinstated for some reason, even though this is not necessary for the accommodation of the new antecedent.

These cases show that even if some mechanism for domain contraction is built into the account that can distinguish between the cases that would require such contractions and the cases that would require, rather, a persistence of the previous expansion, SCA will still fail to predict these later-occurring infelicities. Still more complexity would need to be added to the view.

IV. Independent Antecedents

Finally, even if all the previous cases could be accounted for via additional complexity in the semantics/dynamics – and that is a big *if*! – there are cases that apparently generate the same infelicity, but which seem to elude any analysis of the kind SCA has to offer:

(SASHA): (a) If Sophie's twin sister Sasha, who is just like her in almost every way, went to the parade and got stuck in the back, she wouldn't see Jeter. (b) #If Sophie went to the parade she would see Jeter.

(GROUP): (a) If Jeff went to the parade and got stuck in the back, he wouldn't see Jeter. (b) If Lars went to the parade and got stuck in the back, he wouldn't see Jeter. (c) If Tina went to the parade and got stuck in the back, she wouldn't see Jeter. (d) If you or I went to the parade and got stuck in the back, we wouldn't see Jeter. (e) #If Sophie went to the parade she would see Jeter.

In these cases, unlike the other cases generating the infelicity, there is no direct logical relationship between the infelicitous sequence-final conditionals and the earlier conditionals in the sequences. In SASHA, the proposition that *Sasha* does not see Jeter is perfectly compatible with the proposition that *Sophie* sees Jeter, so the inclusion in the domain of worlds where the former is true has no truth-conditional bearing on conditionals about the latter. (Similar remarks apply to the second example, of course.)

Furthermore, note that it is immaterial which worlds are closer, the Sasha-worlds or the Sophie-worlds – we don't even need to know anything about who was more likely to go to the parade in order to recognize the infelicity of SASHA(b). And suppose again that Sophie lives in Australia and never seriously considered going to the parade, and suppose as well that Sasha lives a few blocks from the parade and almost went. Then the possibility of Sasha going is presumably much closer than the possibility of Sophie going. So an utterance of (a) should only expand the domain to include the closest Sasha-worlds, which would not include any Sophie-worlds. So a subsequent utterance of (b) should expand the domain again to include the closest Sophie-worlds, which are by hypothesis all ones where she sees Jeter. Whatever the explanation of the infelicity of SASHA(b) is, it seems like it *can't* be the same as SCA's explanation of the infelicity of SOPHIE(b). But intuitively their explanation should be essentially the same. This suggests the original explanation was wrong.

Readers familiar with the relevant literature may have noticed the similarity of this case to one mentioned by Sarah Moss (2010, p. 21). She observes that the same infelicity found in reverse Sobel sequences can be generated by certain non-modal claims, like the following (adapted from her case):

(OFTEN): (a) Often when people go to parades they get stuck in the back of the crowd and can't see the thing they came to see. (b) #If Sophie went to the parade, she would see Jeter.

Since a construction like (a) is "not a counterfactual, or even a modal sentence, it does not prompt any expansion of the domain over which counterfactuals quantify" (ibid), and thus does not fall under the purview of the SCA explanation of infelicities of this kind. Moss claims this datum is a point in favor of her account, since she appeals to a more general pragmatic/epistemic principle that plausibly applies to cases like OFTEN as well as sequences of conditionals. (See section 5 below for a more detailed discussion of her view.)

But a reasonable response on behalf of von Fintel might go something like this: "It is a nice *bonus feature* of your account that it also explains analogous infelicities generated by non-counterfactual constructions. But this virtue of your account is not a *defect* in mine, for all I have offered is an account of counterfactuals. For all I have said, analogous mechanisms exist elsewhere that have similar effects on modal domains to generate analogous infelicities. But these data are not my responsibility." However, SASHA and GROUP *are* sequences of counterfactuals, and whatever putative effects on the modal domains are responsible for the infelicity of the sequence-final conditionals, they are effects generated by the earlier counterfactuals. So apparently these data *do* fall under the purview of von Fintel's account, and an explanation of them *is* his responsibility, not just a nice bonus. In principle it is possible that there is an additional factor, outside the scope of an analysis of counterfactuals, interacting with the ordinary mechanisms at play in counterfactual discourse that replicates the same or a similar effect. But there is more pressure on the strict conditional account to be able to explain data of the kind it purports to be *an account of* than there is to explain data of another kind that exhibits

similar behavior. And given the structure of the framework, it isn't clear at all how such a story could go.

Recap

Let's summarize the findings of the current section before moving on to possible solutions. The *intermediate worlds cases* showed that coarse-grained expansion would in many cases allow worlds into the domain that would falsify subsequent conditionals that were intuitively felicitous. The *falsifying antecedents cases* showed that even the most fine-grained method of expansion, including only the antecedent-worlds of the conditional invoking the expansion, will deliver similar incorrect predictions about other sequences that are intuitively felicitous. The *complex evolution cases* showed that even if domain expansions were reversed for some reason in the previous cases, the account would fail to predict later infelicities generated by conditionals occurring after these putative reversals. And the *independent antecedents cases* showed that even if all the previous cases could be accounted for, there are other counterfactual sequences apparently generating the same infelicity that seem to elude explanation of the kind offered by the strict conditional theory.

4. Responses

In this section I consider possible responses to my problem cases on behalf of the strict conditional account. I find none of them to be fully satisfying.

i. Antecedent-World-Only Expansion

One option that can be quickly ruled out is the idea of trading in coarse-grained expansion for *maximally fine-grained* expansion, i.e. expansion that includes *only* the closest antecedent-

worlds for the conditional inducing the expansion. In the case of PARTY-R, for example, an A-world-only expansion induced by (a) would include only the closest worlds where Jeff and Lars both come to the party. The new domain would thus exclude the problematic intermediate worlds that threatened to falsify (b), where Jeff but not Lars comes to the party, but, say, Jeff is unusually ornery so it isn't fun after all. This wouldn't address the other problem cases, but it would avoid the intermediate worlds cases, which would be a good first step.

This option is a non-starter, however, for it validates a disastrous logic of counterfactuals, at least in contexts with sufficiently small initial domains. Here are three examples:

Antecedent-to-Consequent Movement (ACM):

$D \cap \mathbf{A} = \emptyset$

(A&B)>C

∴ A>B

Antecedent Conjunct Elimination (ACE):

$D \cap \mathbf{A} = \emptyset$

(A&B)>C

∴ A>C

Antecedent Entailment Substitution (AES):

$D \cap \mathbf{B} = \emptyset$

A>C

A ⊨ B

∴ B>C

ACM and ACE go hand-in-hand. Whenever the domain *D* contains no A-worlds (i.e. $D \cap \mathbf{A} = \emptyset$), an utterance of (A&B)>C will expand *D* to include the closest A&B-worlds. These will now be the *only* A-worlds in *D*, since there were previously none. And of course they are all B-worlds, since they are all A&B-worlds. They must also all be C-worlds, since (A&B)>C is true. So all A-worlds in *D* will be both B-worlds and C-worlds. So A>B will be true (ACM), and A>C will be true (ACE).

Apply this to the case of PARTY-R: suppose there are no worlds in *D* where Jeff comes to the party. An utterance of (a) will expand *D* to add the closest worlds where Jeff and Lars come to the party. Then these will be the only worlds in *D* where Jeff comes to the party, so it will be true that if Jeff came to the party then Lars would come to the party. And since all these worlds are worlds where the party isn't fun, it will be true that if Jeff came to the party it wouldn't be fun. But if the closest worlds where Jeff comes to the party are ones where Lars doesn't come, and it's fun, these should both be false.

An even more extreme case will underscore the implausibility of ACM and ACE. Suppose there are no worlds in *D* where I go to the zoo. Then an utterance of "If I went to the zoo and got mauled by a lion, I would be traumatized" will add to *D* the closest worlds where I go to the zoo and get mauled by a lion. Then these will be the only worlds in *D* where I go to the zoo, so an utterance of "If I went to the zoo, I would get mauled by a lion" will be true. And if all these worlds are worlds where I am traumatized, an utterance of "If I went to the zoo, I would be traumatized" will be true as well. But these (one hopes) are both false.

AES is similar. Suppose A entails B. An utterance of A>C will expand *D* to include the closest A-worlds, which will also be B-worlds due to the entailment. If there were previously no

B-worlds in *D*, these will now be the only B-worlds in *D*. If A>C is true, these will all be C-worlds, so B>C will be true.

Apply this to a variant of the previous case: supposing *I go to the zoo* entails *I leave the house*, if there are no worlds in *D* where I leave the house, then an utterance of "If I went to the zoo, I would see an elephant" will expand *D* to include the closest worlds where I go to the zoo, which will be the only worlds in *D* where I leave the house; if I see an elephant at these worlds, then I see an elephant at every world in *D* where I leave the house; so an utterance of "If I left the house, I would see an elephant" will be true. But this is obviously false.

These logical consequences are devastating. In addition, this strategy would only avoid the intermediate worlds cases. So this option can be ruled out conclusively.

ii. Medium-Grained Expansion

Coarse-grained expansion made SCA vulnerable to the intermediate worlds problem, and maximally fine-grained expansion validated unacceptable logical principles. But perhaps something in between would get things right: perhaps if domains expanded by adding more than just the closest A-worlds, but less than every other world at least as close, maybe they would add enough worlds to avoid validating the unacceptable logical principles, but not enough to include the problematic intermediate worlds.

There are too many possible versions of this approach to rule it out with the same confidence as the previous one, but there is reason for pessimism. Let's work backwards from the problem to the potential solution. In the intermediate worlds cases, the problematic worlds were ones that falsified the second conditional in a given sequence by departing from actuality enough to make its antecedent true *and then some*, making the consequent false via some additional departure. Recall that in NOTLARS the problematic intermediate worlds falsifying (b)

(*If Jeff but not Lars had come to the party, it would've been fun*) were ones where Jeff comes to the party and Lars doesn't, but something else makes the party unfun, e.g. Jeff is uncharacteristically ornery or another group shows up and drinks all the beer. It is natural to feel that these worlds ought to be irrelevant, since the extra features of them that falsify the conditional – Jeff being ornery, the extra group showing up – were not crucial to bringing about the truth of the antecedent (*Jeff but not Lars comes to the party*). They seem to depart in irrelevant ways, reminding one of what Lewis insisted the closeness ordering ought to avoid: *gratuitous departure from actuality*. A sensible thought, then, is: What if domain expansions systematically excluded worlds that differ in ways not relevant to the antecedent?[37]

The trouble is, the problematic intermediate worlds are irrelevant to the evaluation of (b), but the domain expansion that admits them is induced by (a), the previous conditional, *with a different antecedent*. So for this solution to work, they would have to be excluded from the expansion on the basis of irrelevance to (a). But of course there is no guarantee that the same things will be relevant to the antecedents of both (a) and (b). Consider the intermediate worlds where Jeff is unusually ornery so the party isn't fun. On the current proposal these worlds would be excluded from the expansion induced by (a) on grounds of irrelevance. But suppose the closest worlds where the antecedent of (a) – *Jeff and Lars come to the party* – is true *are* ones where Jeff is unusually ornery. Perhaps he never would've showed up to the same party as Lars in the first place unless he was looking for a fight. Then the departure from actuality involved in making Jeff ornery *is* required to make the antecedent of (a) true, so it *is* relevant after all. In this

[37] A nearly equivalent variant of this view, perhaps closer in spirit to Lewis's account, would be to count worlds that differ in ways not relevant to A as farther than the closest A-worlds. On this view, at least formally, coarse-grained expansion is preserved, but the problematic intermediate worlds are counted as too far in the relevant ordering to fall within the modal horizon. Ignoring some technical differences, this would deliver more or less the same results as the proposal currently under consideration, since the same worlds would be included in or excluded from the domain, albeit in virtue of being too distant, rather than being simply irrelevant. As a result my objections to the current proposal would apply, *mutatis mutandis*, to this variant as well.

case the intermediate worlds falsifying (b) would not depart in ways irrelevant to (a), so they would not be excluded from an expansion induced by (a). And in that case they would be in *D* to falsify (b) after all.

Here is a second problem. Suppose as we just did that the closest worlds where Jeff and Lars both come are ones where Jeff is ornery. And suppose that the closest worlds where Jeff alone comes are ones where he is characteristically gregarious. And suppose finally that in the actual world Jeff was in a lukewarm mood, which is why he skipped the party. In this case, relative to the evaluation of (a), worlds where Jeff is gregarious seem to depart from actuality gratuitously. After all, Jeff's actual mood was lukewarm, and in the closest antecedent-worlds for (a) he is ornery, so any worlds where he is gregarious depart in a way that is irrelevant to (a). So on these grounds an expansion induced by (a) should exclude any worlds where Jeff is gregarious, among which are the closest worlds where Jeff (but not Lars) comes to the party. After such an expansion the only worlds in *D* where Jeff and not Lars come will be ones where Jeff is not his usual, gregarious self. Supposing this would not be fun, NOTLARS(b) would then be false, and it would be true instead that if Jeff but not Lars came, then Jeff would not be his usual, gregarious self, and it would not be fun.

Finally, note that the closest A-worlds for a conditional like (b) may be the falsifying intermediate worlds for some other true conditional (b′). So, for the reasons just mentioned, we *would* want them to be included by a preceding expansion for the purposes of evaluating (b); but, to avoid the intermediate worlds problem, we *would not* want them to be included in such an expansion for the purposes of evaluating (b′). For example, suppose Jeff was invited to two parties, ours and Sasha's, and if he had been in a better mood he would've gone to a party, but it would've been Sasha's and not ours. Then the following sequence seems felicitous:

(MOOD): (a) If Jeff and Lars had come to the party, it wouldn't have been fun. (b) If Jeff had been in a better mood that day, he would've gone to Sasha's party.

Note that MOOD(a) is the same conditional as NOTLARS(a), so (assuming a relevantly similar context) they should induce the same domain expansion. And for NOTLARS(b) to be true, this expansion ought to include the closest worlds where Jeff (but not Lars) is at our party, at which he is in a gregarious mood and the party is fun.[38] But these very worlds are ones in which he is in a better mood (than his actual lukewarm mood), but he is at our party instead of Sasha's. So for MOOD(b) to be true, the expansion ought *not* to include these worlds. But of course the expansion must either include or exclude the worlds where a gregarious Jeff comes to our party; it cannot do both.

It may be premature to rule out the possibility that some other version of the current proposal will be successful, but given these initial difficulties I cannot imagine what it would be.

iii. Sophisticated Domains

My first two classes of problem cases – the intermediate worlds cases and the falsifying antecedents cases – involve sequences of counterfactuals in which (according to SCA) the first conditional would induce an expansion that (according to me) would inadvertently include worlds that would falsify the second conditional when evaluated at this expanded domain. Alternate methods of expansion that would exclude some of these worlds led to undesirable

[38] We would also want such worlds, where Jeff but not Lars comes to our party, to be included in order to avoid the entailment from "If Jeff and Lars had come to the party, it wouldn't have been fun" to "#If Jeff had come to our party, Lars would've come as well" (an instance of the undesirable ACM inference from the previous section).

results. So perhaps the proper solution is that expansions turn out to be more fragile, i.e. easily reversed in some way, than the simple version of SCA allows, and so the second conditionals in these cases are not in fact evaluated at the expanded domains.

A few questions are immediately raised by this proposal, two pertaining to the technical details, and a third concerning a related explanatory burden. The first technical question is: When an expansion is reversed, what is the resulting state of the modal domain? It might return to its most recent previous state, or some other earlier state, or a null state, or the singleton set containing the actual world, or something else. Different answers to this question will make different predictions, but I will set this issue aside entirely here.

Second, how are the complex evolution cases to be accounted for within the semantic framework? A reversed expansion cannot simply be "erased", because the same infelicity can be generated after a conditional that purportedly induces such a reversal. Recall:

(COUCH2): (a) If Sophie had gone to the parade and gotten stuck in the back, she wouldn't have seen Jeter. (b) If she hadn't fallen asleep on her couch just before it began, she would've watched it on TV. (c) #If she had gone to the parade, she would've seen Jeter.

According to SCA, first (a) induces an expansion to include the closest worlds where Sophie goes to the parade and gets stuck in the back. Call this domain D_1. Next, on the current proposal, (b) induces a domain contraction of some kind to exclude these worlds and any intermediate worlds that falsify (b). Either this contraction first reduces the domain to some minimal state (i.e. just the actual world, or no worlds at all), after which another expansion occurs to include the closest worlds where Sophie doesn't fall asleep on her couch before the parade, or the domain

simply contracts directly to this point. Call this domain D_2. Since the closest worlds where Sophie doesn't fall asleep on the couch are ones where she watches the parade on TV, and these are closer than any worlds where she goes to the parade in person, D_2 will include no worlds where she goes in person. So (c) would then induce another expansion, this time to include the closest worlds where Sophie goes in person, at which she *doesn't* get stuck in the back, and does see Jeter. Call this domain D_3. So (c) should be true, since D_3 includes none of the worlds where she goes to the parade and misses Jeter. But it is in fact infelicitous.

So the mechanics of domain expansion must be more complex. The explanation of the infelicity of (c) ought to be essentially the same as that of the original SOPHIE-R(b), since COUCH2 minus its (b) conditional is effectively identical to SOPHIE-R.[39] So COUCH2(c) must expand the domain not to D_3, as predicted by the simple mechanics, but rather to D_1, the domain including worlds where Sophie goes to the parade and gets stuck in the back, which was previously associated with the (a) conditional. But why should it do that? The antecedent of (c) is *Sophie goes to the parade*, not *Sophie goes to the parade and gets stuck in the back*. Recall that, according to SCA, expansions occur simply to accommodate new possibilities being introduced, so that speakers have a chance of saying something non-trivial. This was a purely mechanistic explanation in terms of an independently plausible conversational process. But in the present case the normal method of expansion must be circumvented and replaced by something more complex. This cries out for explanation.

So the second task for the current proposal is to provide a more sophisticated mechanics for domain expansion to account for cases like COUCH2, which require reinstating domains

[39] The only difference between them is that COUCH2 is a past tense counterfactual while SOPHIE-R is present tense – "if Sophie *had gone* to the parade, she *would've* seen Jeter", as opposed to "if Sophie *went* to the parade, she *would* see Jeter". This was done to maintain tense agreement between COUCH2 and COUCH, which made more sense as a past tense counterfactual with an assumption that Sophie in actuality did fall asleep on the couch.

from previous points in the discourse. I will not fuss over the formal details here, but I suspect the best option is to build into the account some sort of running record of previous domains and the conditionals associated with them, and a relation *R* between conditionals (or antecedents, perhaps) that holds whenever one conditional is to be evaluated at the domain associated with another earlier conditional. (The *R*-relation, of course, is a black box in the theory that will have to be replaced by a more detailed account of what prompts the reinstatement of an earlier domain. But it will help in the meantime to have a sort of placeholder for the missing part of the theory.) There is nothing terribly wrong with this additional complexity per se, though the account will have suffered some loss of theoretical simplicity and elegance vis-à-vis the original, simpler SCA.

More important than the technical apparatus, however – and this brings us to the third and most pressing question for this proposal – will be the choice of what plays the role of the *R*-relation in the theory. Of course it is not enough simply to say that some or other relation exists that unifies the cases generating the infelicity. The account of *R* will have to not only make the right predictions about which particular cases call for reinstatements, but also provide some non-ad hoc explanation of why such reinstatement does or does not occur. This predictive and explanatory power will be the measure of success for this proposal.

It is important to appreciate the significance of this explanatory challenge. Von Fintel's original view provides an account of how modal domains behave throughout a discourse, which makes certain desired predictions about a particular class of data. The mechanics of domain evolution alone provided an explanation of such behavior. These predictions and the corresponding explanation provided the primary motivation for the theory. But my examples show that the theory overgenerates and makes undesirable predictions about a variety of other,

similar data. So the scope of data actually *accounted for* by the account, as it stands, turns out to be narrower than expected. To be complete the account will have to be amended to cover the new data as well; to do this it will have to build domain contraction into the mechanics. But now with both expansion and contraction in the mechanics, the picture is more complicated: we had a story about when expansions occur, but that story turned out to be insufficient; we now need a story about when contractions occur and why, in order to restore the theory's scope of explanation to full generality.

What might such a story look like? It will help to take inventory of the data in need of explanation. Let's revisit a sample of the earlier cases for comparison, reprinted here for convenience:

(SOPHIE-R): (a) If Sophie went to the parade and got stuck in the back of the crowd, she wouldn't see Jeter. (b) #If Sophie went to the parade, she would see Jeter.

(ONTV): (a) If Sophie had gone to the parade, she would've seen Jeter. (b) If she had watched it on TV, she would've seen him too.

(COUCH): (a) If Sophie had gone to the parade, she would've seen Jeter. (b) If she hadn't fallen asleep on her couch just before it began, she would've watched it on TV.

(COUCH2): (a) If Sophie had gone to the parade and gotten stuck in the back, she wouldn't have seen Jeter. (b) If she hadn't fallen asleep on her couch just before it began, she would've watched it on TV. (c) #If she had gone to the parade, she would've seen Jeter.

(SASHA): (a) If Sophie's twin sister Sasha, who is just like her in almost every way, went to the parade and got stuck in the back, she wouldn't see Jeter. (b) #If Sophie went to the parade she would see Jeter.

In the classic reverse Sobel sequences like SOPHIE-R, the first conditional expands *D* to include the closest A-worlds, and these worlds falsify the second conditional when it is evaluated at *D*. So the aforementioned *R*-relation between conditionals or antecedents of course holds in these cases. In the intermediate worlds and falsifying antecedents cases like ONTV and COUCH, the same sort of expansion occurs, so according to the simple version of SCA the second conditional should be falsified, either by the problematic intermediate worlds or again by the first conditional's A-worlds, respectively. But in these cases the second conditional is intuitively true, so the putative expansion must be reversed, so *R* must *not* hold. In the complex evolution cases like COUCH2, the third conditional is infelicitous, and it is the first conditional that is responsible for its infelicity, so *R* must hold between these. But the second conditional is perfectly felicitous, so here as well the putative expansion induced by the first conditional must be reversed, so *R* must *not* hold between the second conditional and the first. Since the first conditional must be able to produce the infelicity of the third even after the reversal occurs, the *R*-relation must hold across domain contractions, so to speak, and in these cases an earlier state of *D* must be reinstated. And in the independent antecedents cases like SASHA, the closest A-worlds for the first conditional do not falsify the second conditional, nor necessarily do any of the other worlds included in the expansion, yet the second conditional is infelicitous in a way that is analogous to the other cases. And in this case there is no previous state of *D*, e.g. one including worlds where Sophie goes to the parade and does not see Jeter, that could be reinstated to falsify the second conditional. So here the infelicity is completely mysterious.

So what special relationship obtains between the relevant conditionals or antecedents in SOPHIE-R, COUCH2, and SASHA, but not in ONTV or COUCH? Isolate the first two sequences reprinted above (SOPHIE-R and ONTV), and one might think the relation was

entailment-based: *Sophie goes to the parade and gets stuck in the back of the crowd* entails *Sophie goes to the parade*, but *Sophie goes to the parade* does not entail *Sophie watches the parade on TV*. But the remaining sequences defeat this hypothesis: *Sophie goes to the parade* and *Sophie goes to the parade and gets stuck in the back* do entail *Sophie doesn't fall asleep on her couch just before the parade begins* – not logically, of course, but in some significant sense of entailment – but these pairs are not infelicitous. And *Sophie's twin sister Sasha… goes to the parade* does not in any sense entail *Sophie goes to the parade*, but this sequence is infelicitous.

Likewise, at a glance it might appear to be some syntactic/structural relationship between the conditionals in the sequences that is relevant. In the infelicitous SOPHIE-R and (the infelicitous portion of) COUCH2, *Sophie goes to the parade and gets stuck in the back of the crowd* contains as a syntactic constituent *Sophie goes to the parade*, but the same cannot be said of the conditionals in the felicitous ONTV, COUCH, or (the felicitous portion of) COUCH2. But neither can the same be said of the infelicitous SASHA. Moreover, SOPHIE-R could easily be rephrased to eliminate this structural relationship between the antecedents:

(SOPHIE-R'): (a) If Sophie went to the parade and got stuck in the back of the crowd, she wouldn't see Jeter. (b) #If the group of parade attendees included Sophie, she would see Jeter.

(SOPHIE-R"): (a) If Sophie went to the parade and got stuck in the back of the crowd, she wouldn't see Jeter. (b) #If the parade were attended by Sophie, she would see Jeter.

The infelicity survives. Clearly it is not explained by syntactic structure.

Neither a logical nor syntactic relationship is the *R* we are looking for. What more subtle feature might explain the relevant patterns? Admittedly there are too many possibilities in logical space to take exhaustive inventory of the options. But one more possibility worth considering – partly because it has arisen during informal discussion of my data – is the notion of *topic*. Suppose the above cases where I have claimed SCA incorrectly predicts infelicity in fact involved mid-sequence topic changes, and suppose this were the sort of thing that could somehow change the operative modal domain. Then even when the first conditional in a sequence expanded the domain to include some problematic worlds, if the second conditional induced a topic change then it would not be evaluated at this earlier domain, and the infelicity would not be predicted. Domains could not be simply "reset", of course, as we learned from the complex evolution cases. But perhaps if each topic were associated with its own dynamic domain evolving in the manner initially proposed by von Fintel, then *unification under a single topic* could play the role of the *R*-relation.

But it is doubtful any plausible notion of topic will be up to the task. First we can rule out the *commonsense* notion of topic. Most of my problem cases intuitively involve no change of topic in the ordinary sense, though verification of this claim is best left to the reader. It is corroborated, however, by the bizarreness (signaled by **?** below), if not quite infelicity, of inserting an explicit signal of topic change mid-sequence (as one often does to smooth over an abrupt change of topic):

(ONTV?): (a) If Sophie had gone to the parade, she would've seen Jeter. (b) **?**On a different topic, if she had watched it on TV, she would've seen him too.

(COUCH?): (a) If Sophie had gone to the parade, she would've seen Jeter. (b) **?**On a different topic, if she hadn't fallen asleep on her couch just before it began, she would've watched it on TV.

Additionally, there are a few theoretical notions of topic established within the relevant linguistics literature,[40] but they are unlikely to do the job either. *Grammatical* notions of topic are largely grounded in sentence structure (at least in English) and anaphora resolution (see, e.g., Roberts (2011), Cornish (2006)). But both conditionals in ONTV, for instance, have the same general structure with the same grammatical subject (viz. Sophie, modulo the substitution of "she" for "Sophie" in (b)). And it was demonstrated a moment ago that the grammatical structure of the infelicitous cases could be rearranged without change to the infelicity of the sequence. Moreover, all the anaphora in ONTV(b) are bound by constituents of (a). Together these facts strongly suggest that the transition from (a) to (b) does not even constitute a change of *sentence* topic, let alone a change of *discourse* topic (which would be the more appropriate candidate for the *R*-relation).

The other notable linguistic conception of topic identifies or associates topics with *Questions Under Discussion* (QUDs) – implicit background questions that guide discourse along various paths of inquiry, thus organizing conversational contributions correspondingly (see, e.g., Roberts (op. cit.), (2012)). For example, consider the sequence:

(DRINKS): Jack had the beer; Una had the wine; I had the whiskey.

[40] Thanks to an anonymous referee for this journal for pointing this out.

According to QUD theory, these statements form a coherent narrative because they answer a cluster of related questions, such as: *What did Jack have?*; *What did Una have?*; and *What did [the speaker] have?*, or *Who had the beer?*; *Who had the wine?*; and *Who had the whiskey?*; as well as the more general question: *Who had what?* But it is unlikely the concept of QUDs tracks the patterns of in/felicity observed in our data. For example, the most natural candidates for QUDs unifying the conditionals in the original PARTY under a single topic would be: *What if Jeff came to the party?* or *If Jeff came to the party, what would it be like?* But it's hard to imagine why these QUDs should not also unite the conditionals in NOTLARS in precisely the same way.

Finally, and more generally, it is unlikely that any notion of topic *whatsoever* will be adequate to completely avoid intermediate worlds cases, given their structure. Recall that if the first conditional's A-worlds are significantly farther from actuality than the second conditional's A-worlds, there will likely be some worlds in between that falsify the second conditional. All that is needed for this situation to arise is that some scenario making the second conditional's antecedent true but its consequent false should require less departure from actuality than the antecedent of the first conditional. But this relationship is a matter of comparative similarity, which has nothing inherently to do with topic, and it is extremely unlikely that these two disparate notions will turn out to be systematically coordinated in a way that could provide the needed theoretical resources. More specifically, it is unlikely that any notion of topic worthy of the name will guarantee, for any two possibilities falling under the same topic, that the closest worlds realizing each will always fall within a narrow enough threshold of closeness to each other so that no problematic intermediate worlds exist between them.

So whatever special relation holds between the pairs of conditionals or antecedents that generate the infelicity, it does not seem to be logical, structural, or topic-based. I believe that no

other candidate exists that will prove to be suitable, though I cannot claim to have decisively established this here, only that these few initially plausible ones are not; further work would be needed to establish the more general negative claim. We are therefore in the suboptimal position of having gone as far as identifying the work that would need to be done to save the theory – identifying a satisfactory *R*-relation – without being able to go so far as to confidently conclude that the work cannot be done. Should we throw up our hands and call it a draw? I think not. Pending a satisfactory answer to this challenge for the strict conditional view, an abductive case could be made against it via a compelling case for a competing account of what unites the infelicitous sequences. So it will be helpful to end my discussion with a brief exploration of the major alternative solution on offer, due primarily to Sarah Moss. I will stop short of actually endorsing Moss's view, but I think it is more promising than the strict conditional view, and thus worth mentioning given the state of the dialectic.

5. Moss's Alternative Solution

There is an alternative style of explanation of the asymmetry of Sobel sequences, due originally to Sarah Moss (2012), which is worth discussing for comparison. According to Moss, the infelicitous reverse sequences can be explained without abandoning the traditional variably strict semantics, by appeal to an independently plausible pragmatic/epistemic principle, viz. (roughly):

(PEP): For possibilities P and Q and speaker S, it is infelicitous for S to assert Q if:

(i) P is salient;

(ii) P and Q are incompatible; and

(iii) S is unable to rule out P.

PEP seems independently plausible due to its applicability to non-conditional cases. Suppose you and I are waiting for a New Jersey Transit train, which is due momentarily. Compare the following sequences:

(TRAIN): Me: Our train will be here any minute.

You: New Jersey Transit trains often run very late.

(TRAIN-R): You: New Jersey Transit trains often run very late.

Me: #Our train will be here any minute.

In the first case, when the possibility of the train being very late has not yet been mentioned, it may be perfectly felicitous for me to outright assert that our train will arrive any minute. But in the second case, once the possibility of lateness is salient, the very same assertion is infelicitous. Moss's principle, together with some reasonable assumptions about salience, seems to explain this: when our attention is focused on our train, mentioning the fact that NJ Transit trains are frequently very late naturally raises to salience the possibility that *our* NJ Transit train is very late. (Let this be P.) Our train being here any minute (let this be Q) is incompatible with its being very late. Unless I have some way of ruling out P – say, if I can see our train down the track – it is infelicitous to assert Q. But when P is not salient, Q may be perfectly assertible.[41]

What does this have to do with counterfactuals? According to Moss, analogous explanations apply to the infelicitous reverse Sobel sequences. In SOPHIE-R, for example, (a) makes salient the possibility that if Sophie went to the parade she might get stuck in the back and

[41] Note that this makes no commitments about *truth* or *falsity*. If we have not considered the possibility that our train is late, but it is, it may be perfectly *assertible* that the train will arrive any minute, even if it is false.

not see Jeter, which is incompatible with – or at least clashes with[42] – the claim that if she went to the parade she *would* see him.[43]

Moss's account needs to be developed further, but it is worth pointing out that the cases discussed in this paper do not appear to be problematic for her style of explanation. First, recall the intermediate worlds cases: in these cases the expansions induced by the first conditionals unexpectedly included worlds that falsified the second conditionals. But Moss's explanation appeals to no such expansions, so these intermediate worlds would not be quantified over by the second conditionals. And the corresponding possibilities are not mentioned, either, so they would not be salient. So this problem never arises at all for a Moss-style explanation.

Neither do the falsifying antecedents cases: here the closest antecedent-worlds for the first conditionals themselves falsify the second conditionals, so including these worlds in the domain incorrectly predicts infelicity. In BEACH, e.g., worlds where Jeff comes to the party are *a fortiori* worlds where he isn't at a wedding thousands of miles away, but he also isn't at the beach, so (b) should be false. But, again, Moss does not appeal to domain expansions, so these antecedent-worlds needn't be relevant to subsequent conditionals. And even if BEACH(a) makes salient some possibilities about what would or might have happened if Jeff had come to the party, this has no immediate bearing on our judgments about what he would've done if he weren't at the wedding.

---

[42] It is controversial whether so-called *might* counterfactuals – "If A were the case, C *might* be the case" – are *incompatible* with their consequent-negated ordinary (*would*) counterparts – "If A were the case, C *would not* be the case" – or whether they merely clash in some other way. (See especially Stalnaker (1981).) In either case, we would expect infelicity to result from co-assertion.

[43] The positive accounts of counterfactuals defended by Jonathan Ichikawa (2011), Karen Lewis (2015, forthcoming), and Nichols (ms) offer solutions to the problem of reverse Sobel sequences that bear similarities to the Moss-style explanation. An extensive comparison of the solutions on the table is a topic for future research, but it is worth mentioning that there are several proposed variations of the style of explanation discussed in this section.

So Moss's view outperforms von Fintel's view on the first two classes of cases simply by not predicting infelicity where there is none. More impressive, however, is that her view correctly predicts infelicity in the complex evolution cases. Salience can endure throughout an extended stretch of conversation, even if the salient item is not constantly attended to. Consider the following variation of TRAIN-R:

AMTRAK: You: New Jersey Transit trains often run very late. Amtrak trains, however, are extremely punctual, and very clean. Metro-North trains are quite clean too.

Me: #Our (NJ Transit) train will be here any minute.

Clearly the salience of the possibility of lateness is able to survive a brief interlude, since my assertion is still infelicitous. The same observation applies to complex evolution cases like COUCH2. If (a) makes salient the possibility that if Sophie went to the parade she might get stuck in the back, there is no reason to expect this salience to vanish as soon as a slightly different possibility is mentioned. So (c) should be expected to clash with (a) even after the brief digression created by (b).

Moss's view even seems likely to predict the infelicity of independent antecedents cases like SASHA and GROUP. In the train case, making salient the general possibility of NJ Transit trains being late thereby made salient the particular possibility of our NJ Transit train being late. Similarly, making one possibility P salient often thereby makes some similar possibility P′ salient as well:

(HANDS): You: For all I know I'm a brain in a vat! I may not be a real person! I may not even have a body!

Me: #Thank god *I* have a body.

Since your speech makes salient the possibility that you are a bodiless brain in a vat, it automatically makes salient the possibility that I am one as well, even though you have said nothing about me. Likewise, in SASHA, since Sasha and Sophie are so similar, the salience of the possibility that Sasha might get stuck in the back if she went to the parade plausibly brings about the salience of the corresponding possibility involving Sophie. And in GROUP, the salience of several different people possibly getting stuck in the back of the parade plausibly makes salient the general possibility of one getting stuck in the back of the parade, which naturally brings about the salience of the same possibility involving Sophie. Since this is a possibility in which she wouldn't see Jeter, it is infelicitous in these cases to assert that if she went to the parade she *would* see him. Moss's view might have to be developed more to accurately predict these sorts of salience relations, but there is clear independent motivation for their existence, and, with any luck, an account of them could be largely inherited from an adequate account of salience.

Finally, it is important to fully appreciate the difference between the two styles of explanation on offer. According to Moss, counterfactual assertions (and other types) can raise to salience certain related possibilities that interfere with the assertibility of subsequent counterfactual claims that would otherwise be assertible. These relationships between assertions, possibilities, and salience are largely independent of any ordering on worlds. But von Fintel's explanation is quite different: counterfactual assertions can expand the domain of modal quantification outward (in the direction of greater world-distance); and this may include worlds that interfere with the interpretation of subsequent counterfactuals about nearer possibilities, since these may be evaluated at the expanded domains including not just their *closest* A-worlds

(i.e. the only ones relevant on the standard truth conditions), but others as well. The infelicity under discussion, by design, thus occurs only at the moments when we turn our attention from farther away possibilities towards closer ones. And, as a corollary, the standard truth conditions for counterfactuals are replicated only when sequences of counterfactual discourse are arranged in an outwardly progressing order. But why should they be so arranged? In ordinary discourse we are not generally concerned with ordering the series of possibilities under discussion according to their distance from actuality. The in/felicity patterns displayed by sequences of counterfactuals seem to be tracking something else.

So Moss's view seems fairly well equipped to handle my problem cases, insofar as they arise for her view at all. This ought not to be taken as an all-things-considered verdict in favor of her account over von Fintel's: my considerations could conceivably be outweighed by other problems for her view and/or virtues of his that I have not discussed. And there are other views in the vicinity of Moss's that may fare better in the long run. But with respect to the cases discussed within this paper, Moss's explanation is at a clear advantage.

6. Conclusion

One of the two primary motivations for the dynamic strict conditional account of counterfactuals is the asymmetry exhibited by reverse Sobel sequences. I have presented four classes of sequences about which this account, as it stands, makes incorrect predictions. I have also considered a few possible extensions of the account that one might have expected to offer some improvement, but which look unpromising upon closer inspection. Though moments of this discussion were inconclusive, we now have at the very least a clearer view of the work that would lie ahead for the strict conditional view. Finally, I have argued that the major competing style of explanation on the table seems to handle these cases fairly easily or avoids them

altogether. Further research must be done to more decisively adjudicate between these options. And it bears repeating that the second major motivation for the strict conditional view, concerning so-called negative polarity items, remains completely unaddressed by this paper. But I take myself to have raised a series of objections that at least will need to be addressed by proponents of the strict conditional view, and at most have undermined one of the two arguments for it.

WORKS CITED:

Bennett, Jonathan (2003). *A Philosophical Guide to Conditionals*. Oxford University Press.

Cornish, Francis (2006). Discourse Anaphora. In Keith Brown (ed.), *Encyclopedia of Language and Linguistics (2nd edition)*, Elsevier Ltd. (Oxford, RU), pp. 631-638.

von Fintel, Kai (1999). NPI Licensing, Strawson Entailment, and Context Dependency. *Journal of Semantics* 16 (2): 97-148.

von Fintel, Kai (2001). Counterfactuals in a Dynamic Context. *Current Studies in Linguistics* series 36, 123-152.

Gillies, Anthony S. (2007). Counterfactual Scorekeeping. Linguistics and Philosophy 30 (3): 329-360.

Ichikawa, Jonathan (2011). Quantifiers, Knowledge, and Counterfactuals. *Philosophy and Phenomenological Research* 82 (2):287-313.

Ladusaw, William A. (1979). *Polarity Sensitivity as Inherent Scope Relations*. Ph.D. Dissertation, University of Texas, Austin.

Lewis, David (1973). *Counterfactuals*. Blackwell.

Lewis, David (1979). Scorekeeping in a Language Game. *Journal of Philosophical Logic* 8 (1): 339-59.

Lewis, David (1986). Counterfactual Dependence and Time's Arrow. Reprinted in *Philosophical Papers, Volume II*. Oxford University Press. 32-52.

Lewis, Karen (2015). Elusive Counterfactuals. Noûs 49 (2).

Lewis, Karen (forthcoming). Counterfactual Discourse in Context. Noûs.

Nichols, Cory (ms). Rethinking Similarity.

Moss, Sarah (2012). On the Pragmatics of Counterfactuals. Noûs 46 (3): 561-86.

Roberts, Craige (2011) Topics. In Claudia Maienborn, Klaus von Heusinger & Paul Portner (eds.) *Semantics: An International Handbook of Natural Language Meaning*. Mouton de Gruyter. 1908-1934.

Roberts, Craige. 2012. Information structure: Towards an integrated formal theory of pragmatics. Semantics and Pragmatics 5.

Stalnaker, Robert C. (1968). A Theory of Conditionals. In Nicholas Rescher (ed.), *Studies in Logical Theory*. Blackwell.

Stalnaker, Robert C. (1981). A Defense of Conditional Excluded Middle. In William Harper, Robert C. Stalnaker & Glenn Pearce (eds.), *Ifs*. Reidel.

NEW PUZZLES FOR KRATZER'S MODALS

Cory Nichols

Princeton University

1. Introduction

Modal language is remarkably versatile. Everyday locutions such as *must* and *can*, and others like them, are able to express a seemingly limitless variety of different types of permissions and obligations. For example, consider what might be expressed by an utterance of the sentence "you must be quiet" in the following scenarios: we are in a legally designated noise-pollution-free zone; we are in an avalanche zone with a town beneath us; we are at a funeral during the delivery of the eulogy; we are playing the party game of charades and it is your turn; we are soldiers attempting to sneak through enemy territory undetected. In each of these cases the particular sort of obligation expressed by *must* depends on at least two major contextual factors: the relevant *circumstances*, and the *source* of the obligation. In the first scenario, for example, the circumstances include our location (the noise-free zone), while the source of the obligation is the law. In the second scenario, the circumstances include the precariousness of the environment, the vulnerable placement of the town, and perhaps facts about avalanches and acoustics, while the source of the obligation is presumably morality (and perhaps prudence). The remaining cases are similar, involving details about the funeral, the party, and the war, while expressing the demands of social etiquette, the rules of the game, and the facts of self-preservation.

Modal language is also quite *pervasive*: not only *must* and *can*, but a long list of other terms also express the same or similar content. For example, *ought*, *should*, *have to*, *need to*,

obligated to, and *required to* express something similar to *must*, while *might*, *may*, *could*, *possibly*, *permitted to*, and *allowed to* express something similar to *can*. There are subtle shades of difference between the terms in each set,[44] but on the whole they seem intimately related, as attested by the fact that they are often more or less interchangeable: "You must pay a fine", "You have to pay a fine", and "You are required to pay a fine", for example, all seem to say the same thing. The fact that a large repertoire of related terms has arisen in natural language[45] for the purpose of expressing similar modal relations is a phenomenon of interest both because it is linguistically distinctive and because modality is a philosophically central subject.[46] So modals like those listed above constitute an important class of related locutions bearing something like a family resemblance, and we would like an account of them.

These observations create theoretical pressure to provide a semantic analysis that is unified along two dimensions. One, we would like to capture what these distinct but similar terms (*must*, *ought*, *obligated*, etc.) share in common, while respecting their differences. Two, we would like to explain what different occurrences of each individual term share in common, for example how a single word like *must* manages to express an impressive variety of different sorts of obligations and permissions in different contexts. Despite the variation, some features must be invariant across usages of *must*; these features constitute the *semantic core* of the term. The first

---

[44] For example, is not incoherent to say: "You don't *have to*, but you *should*"; see Lassiter (forthcoming).
[45] This phenomenon is not limited to English, of course. See, for example, Kratzer 1991, p. 650, for a survey of analogous expressions in German.
[46] Two things are worth noting about the term *modal*. First, philosophers sometimes use the term to mean something like "analyzable in terms of possible worlds" or "possible-worlds based", in which case, e.g., an account of "must" that didn't appeal to possible worlds at all would not be considered modal. Linguists, however, use the term to designate a linguistic class of terms like *must*, *can*, *might*, *should*, *may*, etc., with no assumption or commitment as to whether the proper analysis of these terms involves possible worlds. This is how I will use the term here. Second, a semantic account of modal terms is not the same as a philosophical account of *modality*. A philosophical account of modality as a metaphysical phenomenon might seek to explain the nature of possibility, broadly construed, and what grounds modal facts. It also might take certain sorts of modal facts to be primitive, and/or might explain certain kinds of modal facts in terms of others. A semantic account of the meanings of modal terms in natural language makes little or no commitment about any of that; rather it models the meanings of these terms using the theoretical resource of possible worlds.

dimension of analysis is a topic for another time. The second dimension, the shared semantic content of terms like *must*, is the one that interests us here.[47]

How might we approach such a project? The noted versatility of *must* suggests it should be interpreted relative to (at least) two major parameters, which, following Angelika Kratzer (1991), we might call the *circumstantial background* and the *normative background*.[48] The circumstantial background consists in certain details of one's situation that are relevant to determining one's permissions and obligations; the normative background consists in certain details of the given system of conventions, requirements, or other constraints from which some particular set of permissions and obligations issue, e.g. the law, morality, etiquette, etc.[49] (Let's call such a system a *flavor of modality*, as Kratzer does.) For example, if I am legally forbidden from parking in zone *z* on weekdays, the circumstantial background might include (but is not necessarily limited to) the facts that: today is Tuesday; today is not a holiday; I lack the special permit required to park in zone *z* on weekdays; etc.; and the normative background might include the relevant dictates of the law, e.g.: that zone *z* is a parking zone of such-and-such type; that a certain permit is required to park in zones of that type; that holidays and weekends are exceptions; etc. Formally, we can construe each of these backgrounds as a set of propositions, one supplied by the details of my situation, the other supplied by the operative flavor of

---

[47] As is done by Kratzer herself and others in the associated literature, I will focus primarily on the term *must* here, but inevitably slip into speaking in terms of other modals when they seem more or less interchangeable. It is clear, after all, from Kratzer 1991, esp. section 9, "The Semantic Field of Modal Expressions", that she wants the basic framework to apply, mutatis mutandis, to all sorts of other modal terms (even modal suffixes like -ible/-able).

[48] Kratzer's terminology is actually slightly more complicated than this. A *conversational* background is, roughly, any contextual factor that supplies the content of one of her two major semantic parameters (discussed below). An *epistemic* background is a conversational background that issues from epistemic sources; a *circumstantial* background is a non-epistemic one that issues from the relevant circumstances; and a *normative* background is one that issues from a normative source. See Kratzer (1991, pp. 645-6) for more on this. For my purposes I can focus on these latter two categories and ignore the other two.

[49] Readers familiar with the relevant literature may find the omission of epistemic modality throughout this paper rather glaring. It is extremely controversial whether epistemic modals belong in the same category as non-epistemic ones (so-called *root* modals). (See, e.g., Yalcin (2007).) So I have decided to avoid discussion of them entirely here. Nothing important hangs on this.

modality. The first question to ask, then, is how these sets of propositions interact with each other to contribute to the meaning of *must*-claims. The most obvious answer would be that MUST(P) is true iff all the worlds compatible with both sets of propositions are P-worlds. As we'll see, this analysis appears to be too simple. Angelika Kratzer (1977, 1981a, 1991) has proposed a more sophisticated analysis, which has since become the benchmark account of modal language that serves as the starting point for most contemporary discussion of the topic. In this paper we'll explore some problematic consequences of this view that have apparently gone unnoticed for several decades.

Here is a road map of the rest of this article. In the next section I will outline the simple account, which I call the Classical View, and discuss two of its shortcomings that motivate Kratzer's alternative semantics. In the following section I will outline Kratzer's semantics, and explain how it is meant to accommodate those shortcomings. In the following three sections I will introduce three novel problem cases for Kratzer's account; after each I will provide some commentary, considering and responding to some possible objections, and then drawing some conclusions about what can be learned from the given case. For the first two cases I have a positive proposal to offer for how best to amend the Kratzer-style analysis to solve the problem. (I do not endorse the view, I merely think it is the best option for this kind of analysis.) In the third case, I cannot see how the view could be adequately amended, though I do have a diagnosis to offer. At the very end I will briefly summarize my findings.

2. The Classical View[50]

[50] Kratzer (1991) refers to this as "the standard anaylsis", but her account has so long been the standard analysis that it now seems quaint to describe it that way.

Here is a seductively simples account of the phenomena under discussion. Claims of the form MUST(P) are assessed relative to a particular modal flavor *m* (e.g. the law) and a set of relevant background circumstances *b* (e.g. the legally relevant facts about the context). Formally, *m* and *b* each may be construed as a set of propositions, M and B, which encode, respectively, the dictates of the modal flavor and the details of the circumstances. And MUST(P), relative to M and B, is true iff P is true at all the worlds compatible with M and B. In the case of the law, for example, this analysis tells us that legally I must ϕ, given my circumstances, iff I ϕ in all the worlds compatible with my circumstances and the laws.[51] The intuitiveness of this proposal should be evident: if my legal obligations are determined by the content of the laws and the (legally relevant) circumstances I find myself in, then the set of worlds where those laws are obeyed and those circumstances obtain is a natural candidate to capture the meaning of the legal *must* within our theory.

Let's formalize this a bit more. If we follow standard practice and take the denotation of a set of propositions to be the set of worlds in which all of its members are true, then for formal purposes we may treat M and B as equivalent to the sets of worlds in which their members are true, which I will represent symbolically in boldface as **M** and **B**. Then the set of worlds compatible with both M and B will be the intersection of these sets, i.e. $\mathbf{M} \cap \mathbf{B}$. By this convention the proposition P itself also denotes a set of worlds, **P**, and P is true at all the worlds in a set *W* iff *W* is a subset of **P**. Put all of this together and we can say that P is true at all the

[51] Some modals plausibly express a blend of modal flavors, especially when the demands of various similar normative systems agree. (See, e.g., Szabo & Knobe (forthcoming).) For example "You mustn't steal" could express some combination of moral, legal, and social (etiquette) obligation, especially given that each of these flavors of modality is largely concerned with governing our interactions with others. I doubt that this complication will be relevant to my discussion here.

worlds compatible with M and B iff $\mathbf{M} \cap \mathbf{B} \subseteq \mathbf{P}$. We can now state the formalization of the Classical View's analysis of *must*:

CV-must: $\text{MUST}_{M,B}(P) =_{df} \mathbf{M} \cap \mathbf{B} \subseteq \mathbf{P}$

And treating *can* as the dual of *must* in the usual way, such that CAN(P) is true iff *some* of the worlds compatible with M and B are P-worlds, i.e. iff the intersection of **M**, **B**, and **P** is not empty, we can state the corresponding analysis of *can*:

CV-can: $\text{CAN}_{M,B}(P) =_{df} \mathbf{M} \cap \mathbf{B} \cap \mathbf{P} \neq \varnothing$

At first pass, this appears to capture the intuitively correct truth conditions for *must* and *can*: when the law requires some action in a given situation, I perform that action in that situation in *every* world where the law is not violated; when the law permits an action in a given situation, I perform it in that situation in *some* world where the law is not violated. Likewise, the obligations and permissions of other modal flavors, such as morality, rationality, or etiquette, are reflected in the events of the corresponding worlds where they are complied with. Unfortunately, however, this style of analysis appears to be too simple.

3. Puzzles for the Classical View

Angelika Kratzer (op. cit.) has described several problems for the Classical View. Here I will focus on two of them.

3.1 Violation-based Requirements

For some flavors of modality, the kinds of requirements they generate can (in the broadest sense of "can") be violated; for others they cannot. The laws of physics, metaphysics, and logic, on the one hand, are what we might call *unbreakable* – they are never violated at a world in which they

apply. (What would it be for a world to violate its own physical laws?) But the dictates of morality, rationality, judicial law, social etiquette, games, and so forth are *breakable* – they can be violated and, at least in our world, frequently are. So these sources of normativity can apply at worlds where they are not perfectly obeyed. Furthermore, breakable flavors of modality can and often do provide for what to do in case of a violation. Etiquette may dictate that you promptly offer your guest a drink upon arrival, but it might also dictate that *if you fail to do so*, you apologize profusely. Likewise, morality may often require that we make amends for wrongs done, judicial law imposes penalties for various infractions, games may punish cheating, and so on.

Why is this a problem for the Classical View? Consider the legal case.[52] Suppose the law says you mustn't commit murder, and it also says that if you *do* commit a murder you must turn yourself in (i.e. failure to do so constitutes an additional violation). You commit a murder. You must now turn yourself in. But if MUST(P) means P is true in all worlds compatible with the laws and the circumstances, then "You must turn yourself in if you commit a murder" means that you turn yourself in in all worlds compatible with the laws where you murder. But of course there are no such worlds, because murdering itself is forbidden, so in none of the worlds compatible with the laws does anyone murder in the first place. But the fact that you have committed murder is part of the circumstances generating the obligation to turn yourself in (since non-murderers are not so obligated). So there are no worlds compatible with both the circumstances and the laws. Since the laws and the circumstances are incompatible, the

[52] I have chosen legal cases for most of my discussion throughout this paper because intuitions about them tend to be relatively clear and uncontroversial, compared to, say, ethical cases. Nothing crucial should hang on this, since Kratzer's account is meant to apply to every flavor of modality. Nonetheless, occasionally I will consider non-legal cases as well, lest the reader worry that my observations here really expose some idiosyncratic feature(s) of the law in particular, rather than of modals in general.

intersection of the sets of worlds they denote will be empty. In this case every requirement will turn out to be trivially true, because it is trivially true for any φ that you φ at every world in the empty set. Correspondingly, every permission will be trivially false, because it is trivially false for any φ that I φ at some world in the empty set.

In short, the first problem for the Classical View is as follows. Many normative systems provide for what to do in cases of violations, thus generating (as it were) secondary requirements that arise only in cases of violations of other requirements. But, since any violation of a set of requirements is by definition incompatible with that set of requirements, and thus occurs at no world compatible with it, there is no way to capture the content of these secondary requirements if obligations and permissions are analyzed solely in terms of what occurs at worlds compatible with the total set of requirements.

3.2 Conflicting Requirements

Some modal flavors can be violated in worlds where they apply; others can even be internally inconsistent, yet still (partially) apply. That is, subsets of a system of laws or rules can be inconsistent or demand inconsistent things without automatically undermining the rest of the system of which they are a part. Some pair of rules of a game might be inconsistent without rendering it unplayable; one judge's ruling might conflict with another's without undermining every other part of the legal system; some claim morality can demand incompatible things in the case of genuine moral dilemmas; and so on.

Focus again on a legal case. Local legislature passes a law requiring everyone over the age of 18 to vote. But there is another, forgotten law on the books prohibiting anyone under the age of 21 from voting. (And suppose the legal system in question is arranged so that neither law

takes precedence.) What are 20-year-olds (for example) to do on election day? It seems to be indeterminate what the law requires or permits of them regarding voting. But intuitively this inconsistency in the law does not invalidate all the other laws, e.g. prohibitions against murder, theft, and so forth. But the set of worlds compatible with the laws is empty, because there is no world where 20-year-olds comply with both the requirement that they vote and the requirement that they not vote. So, as in the previous case, all statements of requirement will be trivially true and all statements of permission trivially false.

In short, the second problem for the Classical View is as follows. Minor inconsistencies in a set of requirements need not undermine the legitimacy of the remainder of those requirements. But if the set is inconsistent then no possible world is compatible with it, and thus no world satisfies the totality of requirements. So there is no way to capture the content of the other, non-conflicting requirements if obligations and permissions are analyzed in terms of what occurs at worlds compatible with the total set of requirements.

4 Kratzer's Alternative

Angelika Kratzer (op. cit.) has proposed a more sophisticated analysis to accommodate these problem cases. Her account retains the two basic parameters of the Classical View: the modal flavor *m*, and the circumstantial background *b*. And *m* and *b* each still contribute to the semantics a set of propositions M and B, respectively, but they interact with each other differently. B is still associated formally with the set of worlds **B** where all the members of B are true, just as before. But rather than likewise associating M with another set of worlds **M**, and taking the intersection of **B** and **M** as the set of worlds putatively quantified over by *must* and *can*, Kratzer takes **B** itself to comprise this set of worlds, which she dubs the *modal base*. M then induces a partial ordering

on the worlds in the modal base, as determined by some procedure that "ranks" worlds according to how close they come to making all of the propositions in M true. This part is key: it may be that not all worlds in **B** make all members of M true, and if M is internally inconsistent, or incompatible with something in B, then *none* of the worlds in **B** will make all members of M true. But worlds may make *more or less* of the propositions in M true, and in doing so will do better or worse at approximating the "ideal" described by M. The worlds in **B** that come closest to making all of M's members true are the best-ranked worlds according to M; and if all *these* worlds are P-worlds, according to Kratzer, then MUST(P) is true (relative to *b* and *m*).

This raises some immediate questions about the ranking procedure, but before addressing them it will be helpful to spell out the semantics more carefully. Let $\leq_M$ be an M-induced ranking on worlds (the structural details of which are to be discussed shortly), such that $w \leq_M w'$ (read as: *w is at least as good as w′ according to M*) iff w is at least as highly ranked as w′ is according to the ordering procedure supplied by M. And let the expression $\leq_M W$ stand as shorthand notation for the best-ranked members of a set of worlds *W* according to M.[53] Kratzer's analysis says that MUST(P) is true, given M and B, iff all of the M-best worlds in **B** are P-worlds; we can now formally state this as follows:

K-must: $\text{MUST}_{M,B}(P) =_{df} \leq_M \mathbf{B} \subseteq \mathbf{P}$

And, again, treating *can* as the dual of *must* in the usual way, such that CAN(P) is true iff *some* of the M-best worlds in **B** are P-worlds, we can state the corresponding analysis of *can*:

K-can: $\text{CAN}_{M,B}(P) =_{df} \leq_M \mathbf{B} \cap \mathbf{P} \neq \varnothing$

The modal operators *must* and *can* now have the capacity to scope over a more inclusive set of

[53] We could easily define this set more rigorously, if we wanted, as the set of worlds w in *W* such that for any other world w′ in *W*, $w \leq_M w'$:

$\leq_M W =_{df} \{w: w \in W \,\&\, \forall w'(w' \in W \supset w \leq_M w')\}$

worlds than just the worlds compatible with the circumstantial background B and the dictates of the modal flavor M: namely, the worlds compatible with B, ranked by M, even when they are all incompatible with something in M. This will provide the flexibility to handle the problem cases that the Classical View seemed unable to accommodate. But before we see how, we need to address one more detail of the view.

It was noted a moment ago that some immediate questions are raised by the characterization of $\leq_M$ as tracking which worlds *make more of the propositions in M true*. Most saliently: what does it mean for one world to make more or less of the propositions in a set true than another world does, especially if that set can in principle be infinite?[54] Kratzer defines a limited notion of this comparative relation, for the special case where the propositions in M made true by one world are a subset of those made true by another world. (This notion would need to be replaced or generalized in some way to account for other cases, as we'll see shortly.) This yields a weak, partial ordering on worlds. It is weak (rather than strict) because it allows for ties (i.e. $w \leq w'$ and $w' \leq w$), since two worlds may make exactly the same propositions in M true. And it is partial (rather than total) because it allows for incomparability (i.e. neither $w \leq w'$ nor $w' \leq w$), since two worlds may make very different propositions in M true. To see why more clearly, let's consider an extremely simplified case. Suppose there are only three propositions in M: P_1 = *nobody murders*; P_2 = *nobody steals*; and P_3 = *nobody tax-evades*. Then $w \leq w'$ iff w makes at least many of $\{P_1, P_2, P_3\}$ true as w' does. So, e.g., if w_1 makes all of P_i true, both w_2 and w_3 make P_1 and P_2 true but P_3 false, and w_4 makes P_1 true but P_2 and P_3 false, then $\leq$ yields the ordering: $w_1 \leq \{w_2, w_3\} \leq w_4$.

[54] It seems implausible that the law provides an infinite set of propositions as its dictates, but it's conceivable that some other modal flavor could, or at least this is not ruled out by anything in the formal details.

	Worlds	Ordering
w_1:	P_1, P_2, P_3	w_1
w_2:	$P_1, P_2, \neg P_3$	$\leq$
w_3:	$P_1, P_2, \neg P_3$	w_2, w_3
w_4:	$P_1, \neg P_2, \neg P_3$	$\leq$
		w_4

One might call these worlds "nested", in that the set of relevant propositions made true by each world stands in some sort of subset relation to those of every other one. This yields a subset of the set of all worlds that is totally (but weakly) ordered throughout. But add another world w_5, which makes P_1 false but P_2 and P_3 true, and it is impossible to fit it into the ordering in such a way that it is related by $\leq$ to every other world. $w_1 \leq w_5$, because w_1 makes none of the P_i false, and w_5 makes one of them false. But w_5 is incomparable to each of the remaining worlds $w_2 \ldots w_4$, since each of those worlds makes P_1 true but P_3 false, while w_5 makes P_1 false but P_3 true. The ordering on this larger set is merely partial.

Why does this matter? Suppose the circumstantial background ruled out w_1, but not $w_2 \ldots w_5$. Which would be the best-ranked world(s) in the set? w_2, w_3, and w_5 each make 2 of the 3 propositions in M true; w_2 and w_3 are symmetrically related by $\leq$ (each is as good as the other), so neither is better than the other; and w_5 is related by $\leq$ to neither w_2 nor w_3, so it is better than neither, nor vice versa. Since the ordering is partial with respect to the candidates for best world(s), it does not yield a clear verdict in this case. And if there is no determinate set of best-ranked worlds, the result of the ordering procedure will be null or indeterminate, and so in these cases we will be no better off than we were before.

Kratzer ultimately complicates her semantics so that it yields a determinate verdict even in cases like this, where a total ordering is not possible (see Kratzer (1991), and Swanson (2008) for further discussion). But the ordering procedure, as it stands, is unable to rank non-nested sets of worlds, i.e. sets that fall outside the special case scenario described above, so whenever there are non-nested subsets of the modal base, it will yield a mere partial ordering. This may be acceptable for certain flavors of modality, but it is doubtful it will be acceptable for all. Suppose, for instance, I have gotten myself in an unfortunate situation in which my only options are to break a promise to pick a friend up at the airport or to steal an innocent person's car in order to do so. Barring extremely unusual circumstances, morality demands breaking the promise rather than committing the theft. It may still be that breaking the promise is *wrong*, and so the worlds where I break the promise contain a moral transgression not found in the worlds where I steal the car in order to fulfill the promise. And of course the worlds where I steal some poor bystander's car in order to pick up my friend contain a moral transgression not contained by the worlds where I merely break my promise instead. So, on the simple ordering procedure, neither class of world's violations are a subset of the other's, so the worlds in each set will be incomparable to those in the other, and neither will be considered morally "better". But in this case morality makes a clear and determinate recommendation (barring some sort of emergency): you must choose to break the promise rather than steal the car. Either choice would involve a moral transgression; but *given* the choice, morality has a preference for which moral transgression you ought to commit.[55]

[55] It might appear that the law works the same way, but this is actually far from clear. Given a world with one minor crimes and no other crimes, and another world with one major crime and no other crimes, there is some ordinary sense in which it seems right to call the former world "legally better". But it isn't clear this ordinary sense is the one relevant to the semantic evaluation of modal claims, since it isn't clear there are any statements of the form (legally) MUST(P) that follow from this observation. That is, it isn't clear there is any sense in which the law, over and above forbidding each crime, also requires you to choose the one crime over the other (though it punishes the major crime

There is a second, even more straightforward problem with the simple ordering procedure, which is that it has no sensitivity for the way in which, or degree to which, a world violates some particular modal flavor. If the dictates of the law the constitute M in the legal case are propositions like *nobody murders*, and the ordering is determined by how many of these propositions a given world makes true or false, then there will be no preference between worlds with one murder and worlds with a billion murders. After all, worlds with only one murder falsify the proposition that nobody murders just as much as worlds with multiple murders. So if there has already been a murder, all worlds compatible with the circumstances will falsify the proposition that nobody murders, whether they have no more murders than the one already committed or a billion more, so other things being equal they will be ranked as equally good. In this case it would be false that we must not murder, given that someone already has. This is obviously absurd.

There might be various ways around this second problem, e.g. by opting for some more fine-grained type of propositions in M so that new murders falsify new murder-related propositions. But this seems ad hoc and, fortunately, is probably unnecessary. It seems both of these problems can be solved by appealing to an ordering procedure that tallies up the individual violations at each world, rather than the propositions falsified by each world. If the ordering procedure "knows" what constitutes a murder, it can keep track of how many instances of murder there are (given some adequate way of individuating and enumerating distinct events and/or actions), even if each murder falsifies the same relevant proposition in M (*nobody murders*). This will allow it to rank worlds with one murder as ceteris paribus better than worlds with multiple murders, and so on. And if there is some way of weighting different violations

---

more severely).

against each other, at least for flavors of modality for which this is appropriate, then the procedure should be able to rank worlds with minor infractions like promise-breakings as ceteris paribus better than worlds with major ones like auto thefts. Describing such a procedure will admittedly require a bit more explanation than the simple subset relation between sets of propositions that we started with. But I see no reason to think this can't be done adequately. For the remainder of this paper we will "spot" the Kratzerian such an explanation, and take it to be unproblematic to appeal to an ordering procedure that counts and weights violations at worlds, rather than merely keeping track of which propositions are made true or false by them. If this type of ordering procedure does turn out to be problematic for reasons unforeseen at the moment, this will only make my case stronger, since I am granting for the sake of argument that the Kratzerian has a readily available solution to the worries just mentioned. At any rate, I have designed my major problem cases so that, as far as I can see, these complications should not turn out to be relevant anyway.

So how does all of this help to avoid the problems that were raised for the Classical View? In the first case, the problem of Violation-based Requirements, the trouble was that some requirements are conditional upon previous violations, but on the Classical View such requirements would be evaluated against an empty set of worlds, since no worlds are compatible with both the laws and circumstances when the circumstances include a violation. Kratzer's account avoids this by allowing quantification over imperfect worlds. Given that a murder has occurred, none of the worlds compatible with the circumstances are compatible with the laws. But of the worlds where murders do occur, worlds where the murderers turn themselves in come closer to complying with all the laws than the worlds where the murderers not only murder but also fail to turn themselves in (which, recall, is an additional crime in our imagined scenario.) So

the former worlds are legally "better", and thus preferred by the ordering. So we can achieve the result that turning oneself in is required, given that one has committed murder.

In the second case, the problem of Conflicting Requirements, the trouble was that some sets of requirements can be partially internally inconsistent without being completely undermined, but on the Classical View any requirements issuing from such a system would also be evaluated against an empty set of worlds, since no worlds whatsoever are compatible with an inconsistent set of laws. Kratzer's account avoids this by, again, allowing quantification over imperfect worlds. Given that the laws are inconsistent, no world satisfies all of them. But of the "imperfect" worlds, ones where the rest of the laws are obeyed come closer to perfect compliance than ones where additional violations occur as well. So the former worlds are legally "better", and thus preferred by the ordering. So the requirements issuing from the consistent subset of the totality of requirements are not undermined by the inconsistent subset.

Kratzer's alternative account looks promising as a solution to these problems.[56] But it has shortcomings of its own, some of which have been noticed[57] and others of which, evidently, have not. In the remainder of this paper I will raise three novel puzzle cases for Kratzer's account, each of which I believe exposes a different shortcoming. In the first two cases I have a constructive proposal for how to improve the view; in the third case I do not.

5 New Puzzles

Here are three new puzzles for Kratzer's semantic framework; each case is followed by some

[56] There is another major motivation for Kratzer's account, which I will not address at all here, which is the analysis of *graded* modals. Terms like "extremely probable", "barely possible", or "more likely than" seem difficult to capture on the Classical View, but the additional structure offered by the introduction of a world ordering may suffice. I take it this subject is completely orthogonal to the cases I discuss below.

[57] See, e.g., Zvolenskzy (2002), Swanson (2008).

commentary, including possible objections and responses to them. In the first two cases, I offer a constructive proposal for how to amend or supplement the view in response. In the third case, I have no such proposal to offer, but I do have a diagnosis of the problem.

5.1 Case 1: Last Call

Dee is a bartender. Towards the end of the night she has one last customer, Buzz, who is visibly drunk. The law explicitly permits bartenders to eject drunk customers, but it also permits letting them stay the night in the bar to sober up. The law also forbids public drunkenness. If Dee ejects Buzz, Buzz will be drunk in public, which will constitute an additional legal violation. If Dee lets Buzz sleep it off in the bar, no such violation will occur. So worlds where she lets Buzz stay are legally "better" than worlds where she kicks Buzz out, and are therefore preferred by a Kratzer-style ordering. So, according to Kratzer's semantics, Dee must let Buzz stay. But this clashes with the most basic datum of the case, which I take to be non-negotiable: Dee is explicitly permitted to eject Buzz.

5.2.1 Objection to Case 1

One natural thought is that perhaps the fact that Buzz is drunk is not in the set of propositions that constitute the circumstantial background B. In this case there would be worlds in the modal base **B** where Buzz is not drunk, and in these worlds ejecting him leads to no further legal violation. Then these worlds would be just as legally good, according to the ordering, as worlds where Dee lets Buzz stay and sober up, so the latter would not be preferred by the ordering, and thus Dee would be both permitted to eject Buzz and permitted to let him stay.

There are several things to say in response. First, the circumstantial background represents the legally relevant facts (in the legal case, of course), and, intuitively, Buzz's drunkenness *is* legally relevant, both in general and in these particular circumstances. In general, if there are laws restricting the situations in which someone may or may not be intoxicated, then whether one is intoxicated seems generally to be a legally relevant fact. And in this scenario specifically, given that the law explicitly states that bartenders may eject *drunken* customers, Buzz's drunkenness seems straightforwardly legally relevant.

Furthermore, we can alter the case slightly to *force* the fact that Buzz is drunk into the circumstantial background. Imagine the law says that *only* drunk customers may be ejected from bars. (Perhaps not a likely law, but surely a possible one.[58]) In this case Dee's legal right to eject Buzz is contingent upon Buzz's drunkenness; she may lawfully eject Buzz *if and only if* Buzz is drunk. Moreover, any world where Dee ejects a sober Buzz contains an additional violation not contained by the other worlds: Dee ejecting a sober customer. So these worlds would rank no better on the legal ordering than worlds where Dee ejects a drunk Buzz. In one set of worlds, Dee is in violation for ejecting a sober customer; in the other set of worlds, the customer is in violation for public drunkenness. But worlds where Dee lets Buzz stay avoid both of these violations without incurring any others (other things being equal), and so they are still the preferred worlds on the ordering. So Dee must let Buzz stay.

Next, let's suppose anyway that Buzz's drunkenness were not part of the circumstantial background, so the modal base included worlds where Buzz was sober. Then Kratzer's account would no longer predict the truth of *Dee must let Buzz stay*, but it would predict the truth of *it*

[58] We could make the case more plausible if we changed the law to state that only customers who were drunk *or breaking some law* may be ejected. I leave it to the reader to verify that the rest of the argument would still go through.

must be that Buzz is sober or Dee lets Buzz stay.[59] The only worlds in the modal base where Dee ejects Buzz that are as legally good as worlds where she lets him stay are ones where Buzz is sober. So in every best world in the modal base the disjunction *Buzz is sober or Dee lets Buzz stay* is true. This disjunctive *must*-claim strongly implies that Dee's legal freedom to eject Buzz is contingent upon Buzz's being sober, which we have stipulated of the case to be false.

Relatedly, consider the conditional "If Buzz is drunk, then Dee must let him stay". This seems false. But according to Kratzer's account of conditionals (1981b, 1986, 1989, 2005), their antecedents (*if*-clauses) act as *restrictors* on the modal base, which exclude any worlds where the antecedent is false. The effect here would be to add the proposition that Buzz is drunk to the modal base, which would ensure that Buzz were drunk at every world in **B**, and then to evaluate the consequent (the *then*-clause), *Dee must let [Buzz] stay*, at the updated modal base. Since the legally best worlds where Buzz is drunk are ones where Dee lets him stay, the above conditional would come out true.[60] And for analogous reasons, the intuitively true conditional "If Buzz is drunk, then Dee may eject him" would come out false. Now, within the context of her broader semantic program, Kratzer's account of conditionals is intimately related to her modal framework. But it is theoretically separable from it, so one might embrace the latter without embracing the former. Other leading accounts of conditionals (e.g. Stalnaker (1968, 1981), Lewis (1973, 1986), Bennett (2003), von Fintel (2001), Edgington (1995, 2008)) are likely to

[59] This statement is admittedly a bit awkward, at least to my ear, but I still find it interpretable. Informal polling and personal linguistic judgements suggest that the unnaturalness may be due to the fact that, at least in American English, (non-epistemic) *must* seems to be used more commonly as an auxiliary to a verb of action, i.e. as found in sentences of the form *[agent] must [action]*, rather than as a sentential complement verb, i.e. as found in sentences of the form *it must be that P*. Sentences of the latter form are nonetheless interpretable, as in "according to the constitution, it must be that Supreme Court nominees are confirmed by the Senate", though their counterparts with the auxiliary must are preferred, as in "according to the constitution, Supreme Court nominees must be confirmed by the Senate". When *must* appears as a sentential complement verb it seems more natural as an epistemic, as in "It must be that the train is running late".

[60] There is another series of problems for Kratzer involving conditionals with embedded modals discussed by Zvolenszky (2002). This is a different problem.

evaluate the consequent of the conditionals in this paragraph (*Dee must let [Buzz] stay*, and *Dee may eject [Buzz]*) from the perspective of a world or information state according to which the antecedent (*Buzz is drunk*) is true. It outruns the scope of this paper to explore whether these accounts of conditionals, combined with a Kratzer-style analysis of the consequent, will yield the same counterintuitive truth conditions, but I suspect they will. At any rate, Kratzer's holistic program delivers the wrong results here, and the culprit seems to be the account of *must*.

5.2.2 Assessment of Case 1

I don't see a tenable way to avoid the first puzzle, but our observations do suggest a way in which the account might be amended to solve it. And it turns out to be an amendment we may want to adopt for independent reasons anyway. The thought is simple: Buzz's public drunkenness, given the situation, seems irrelevant to determining what Dee's legal duties are, precisely because it is not a violation committed by Dee. In this case – though not in all cases, as we'll see momentarily – the law provides, fairly explicitly, that one is permitted to perform a certain action despite the fact that it will lead to someone other individual being in violation of the law. This suggests that the fact that Buzz will be in violation – though not the fact that he is drunk, as I argued above – can be ignored for the purposes of evaluating Dee's legal options. But of course it can't be ignored for the purposes of evaluating *Buzz*'s legal options, since legally he must not be drunk in public.

How might we represent this in the kind of semantic framework outlined by Kratzer? The most natural option is to relativize the semantics to particular agents, so that different obligations and permissions may apply to different individuals. Even though the legally best worlds *overall* may be ones where Dee lets Buzz stay and sober up, not all violations are relevant to evaluating

Dee's duties, so the legally best worlds relative to her behavior might ignore violations committed by others, even in some cases where they are a result of her actions. The way to represent this formally would be to co-index the modal statement, on the left side of the analysis, and the world ordering, on the right side of the analysis, to the same agent (in this case Dee). Using α as a variable representing the agent of evaluation, we can add this simple parameter into the Kratzer semantics like so:

K-must: $\text{MUST}_{M,B,\alpha}(P) =_{df} \leq_{M,\alpha}\mathbf{B} \subseteq \mathbf{P}$

K-can: $\text{CAN}_{M,B,\alpha}(P) =_{df} \leq_{M,\alpha}\mathbf{B} \cap \mathbf{P} \neq \varnothing$

Now statements of permission and obligation are intuitively relativized to a particular agent, and assessed relevant to an ordering of worlds customized for that agent.

But this minor formal adjustment is the easy part. The real challenge is to provide an account of just *how* the ordering procedure yields customized orderings for different agents. This is not as simple as it may seem. It cannot be, for example, that Dee's ordering simply ignores any violations committed by others, for these are often crucial for determining whether our own actions count as violations themselves. Certain kinds of actions, such as violence in self-defense, are often considered permissible when done in response to a violation committed by another, but not usually otherwise. Punching a stranger on the street with no provocation is generally considered a crime, but punching them after they have assaulted you is often not. Other kinds of actions, such as aiding and abetting, accomplice, or incitement, may only constitute violations in connection to another's violation(s). Aiding someone in their ordinary grocery shopping, for example, is not usually a crime; but if they steal the groceries, it may be. And certain kinds of violations, such as conspiracy, may require the coordination of multiple agents, since one cannot commit conspiracy alone. In all these cases the ordering procedure, as part of the process of

determining the obligations and permissions of the agent α, would need to "consider" the actions of individuals other than α and whether they constituted violations, and then to be able to "decide" which of these were relevant and which were to be ignored. In the case of Dee and Buzz, Buzz's public drunkenness constitutes a violation, but is to be ignored when evaluating Dee's obligations. (Though perhaps not entirely – it might be illegal for Dee to let Buzz drive home, since he is drunk.) But if Buzz were to assault Dee, this violation *would* be relevant to the assessment of her obligations, since the normal obligation not to use violence against him might be suspended in this situation.

Can an adequate and plausible account of how the ordering process works be provided for the agent-relative version of Kratzer's semantics? This remains to be seen. The process would be significantly more complicated and context-dependent than the simple one Kratzer describes in her work. It might also have to take a different form for each flavor of modality, since some flavors might conceivably not "care" at all about others' actions, or might not distinguish between different agents, or might otherwise differ in some fundamental way that would affect the process. I see no reason to rule out the possibility that such an account can be provided, but I will not attempt to provide it here. (I should also note that I do not *endorse* this or any other variation of Kratzer's semantics, I merely propose this as the best way for the Kratzerian to solve the present puzzle.)

Next, would the agent-relativization amendment be ad hoc? I think not. It is intuitive enough that different agents have different sets of permissions and obligations, so there is nothing unnatural or particularly surprising about the suggestion. Furthermore, there may be independent reasons to want this bit of additional complexity in the semantics. For one, there seem to be cases where our individual obligations and permissions can conflict with each other

in ways that make agent-specific modal statements seem more appropriate than agent-neutral ones. For instance, if by some awful misfortune we were in a situation where only one of our children could be saved, morality might demand of each of us that we save our own child, but not demand in general either that it be the case that my child be the one who is saved or that your child be the one who is saved.[61] Or if an evil warlord has gravely wronged both of our families, the prevailing code of honor in the relevant society may require of each of us individually that we be the one to kill him, but not more generally either that it ought to be me who kills him or that it ought to be you. So one might be more comfortable characterizing these situations in terms of agent-specific modal statements – statements about what is required *of a particular individual* – than in terms of agent-neutral statements about what ought to be the case, full stop. More generally, it often seems less natural to ask questions like "What ought to be the case?", and more natural to ask agent-specific questions like "What ought S to do?". To be clear, the claim here is not we obviously would want to make such distinctions, just that we very well might, depending on our reactions to various sorts of cases. Those who do find such distinctions desirable will likely also find the proposal of agent-relativization to be a welcome improvement on the view, rather than an ad hoc epicycle.

What about cases not involving individual agents? Obligations can be shared; some laws apply to corporations and other organizations; there are laws about how the government ought to be organized, or even governing other laws (e.g. the constitution); and some flavors of modality may simply deal in agent-neutral permissions and obligations, such as physical/nomological or

[61] You might think morality can never make inconsistent demands like this, e.g. if you are a traditional consequentialist. But even if that is the case, we wouldn't want it to be the semantics of *must* to tell us that, rather than ethical considerations. And there might be other flavors of modality that can make these sorts of demands, e.g. ones grounded in individuals' preferences, desires, plans, or self-interest.

logical. Formally, this is no problem. In such cases the agent parameter may simply be null, or it may be occupied by other types of "agents", such as corporations or governments.

One last point is making before moving on. There is a worry that analogous problem cases may exist in which a permitted action by an agent α leads to a violation *by α*, rather than some other agent. For example, suppose US law made it extremely easy to renounce one's citizenship, so that anyone may do so, at any time and in any circumstances, simply by saying "I hereby renounce my citizenship" in the presence of witnesses. (Again, not a very likely law, but surely a possible one.) And suppose further that doing so without first leaving the country or making other prior legal arrangements would automatically place one in violation of immigration law, as an undocumented "alien". If I am on US soil and have made no such legal arrangements, then renouncing my citizenship would immediately place me in violation; this world is legally worse, ceteris paribus, than worlds where I refrain from renouncing. So, legally, must I refrain from renouncing my citizenship? It appears not. Were I to do so I would be in violation, of course, but renouncing my citizenship would not be the violation. If this is the case, then simply ignoring certain violations of others will not do. The ordering procedure would also need to ignore certain possible violations of the agent α being evaluated. It can't just ignore all of α's future actions, however, since plotting or intending to commit some future action, for example, may constitute a violation only if the future action is itself a violation. So the ordering procedure would have to be sophisticated enough to "know" which of α's other actions to ignore and which to consider. Formally, I don't see why this would be a problem, especially once we have already committed to a similarly sophisticated procedure for evaluating the actions of others. But it does indicate another bit of explanatory work the proponent of the agent-relativist variation of Kratzer's framework would have to do.

The Last Call case shows us that we are sometimes explicitly permitted to bring about a suboptimal state of affairs, at least provided it is worse due to violations of others for which we are not responsible. The natural solution I proposed was to relativize modals to agents by adding an additional parameter to the semantics. More specifically, I proposed that the simple ordering procedure described by Kratzer (which we had already complicated somewhat in the earlier sections of this paper) be replaced with a more sophisticated one that is equipped to discriminate between relevant violations and irrelevant ones. This would incur an additional explanatory burden, but it looks promising as a solution to the first problem.

6.1 Case 2: Death or Taxes

It's Tax Day, and Dorothy has forgotten all about it. There are only 15 minutes left until the deadline, and there is no way she can possibly file on time, given the complexity of the tax law and the filing process. It seems inevitable that she will miss the deadline. But wait – there is another option: she can always kill herself. (Suppose there is no prohibition on suicide.) Worlds where Dorothy kills herself before the deadline are worlds where she avoids missing it and thus violating the corresponding legal requirement. So these worlds contain one less legal violation, ceteris paribus, than worlds where she stays alive 15 minutes longer to miss the deadline, and are thus legally better according to the ordering procedure. So legally, according to Kratzer's account, Dorothy must kill herself. But of course it's *not* the case that Dorothy must kill herself.[62]

[62] The choice of the name "Dorothy" is an homage to the poet Dorothy Parker, who published a book of poetry entitled *Death and Taxes*, and who *did not* kill herself. In fact, in the poem "Resumé" (from a different collection, entitled *Enough Rope*), she famously (and cleverly) advises *against* suicide.

6.2 Objections to Case 2

6.2.1 Bite the Bullet (Dorothy Must Die)

First-hand experience shows that some will at least entertain the counterintuitive response that Dorothy, in fact, must kill herself. It seems absurd at first glance, they might say, but given the choices available to her, the only legally permissible option is suicide. So, legally, she must kill herself. Of course we would ordinarily never describe her situation this way, perhaps for pragmatic reasons, but, strictly speaking, it is true.

It goes without saying that this conclusion is extremely counterintuitive. But, as the saying goes, one's *modus tollens* is another's *modus ponens*. The most we can do in response to the person willing to bite this bullet is to point out additional counterintuitive consequences that come along with it. (It will be instructive to see what they are, anyway.)

First, note that any legal expert would almost certainly disagree. Try and imagine consulting an attorney, a judge, or a legal scholar for advice, and being told (even "strictly speaking") that one must kill oneself. One might be advised that one *could* kill oneself, if one really wanted to, and that would avoid violating the tax code, but not that one legally *must*.

Second, suppose in the scenario described there were additionally some sort of second-order "metalaw" in the relevant legal system stating that it can never require its citizens to kill themselves.[63] It would be bizarre to characterize this situation as a kind of legal conflict. But that is essentially what it would be, if our theory of *must* tells us that the tax laws (along with the lack of prohibition on suicide) give rise to a legal obligation for Dorothy to kill herself, but another law in the legal system tell us this cannot be the case. To be clear, it seems obvious that, as a

[63] This could be entirely non-trivial. At a certain point in the history of Japanese law, the tradition of ritual suicide by self-disembowelment, known as *seppuku* (or more colloquially in English, via a related term, as *hara-kiri*), was used as a form of capital punishment for severe crimes. A law forbidding legally mandated suicide in our imagined scenario could have been introduced specifically as a response to such traditions.

matter of fact, the metalaw would take precedence in this case. But it's unclear how Kratzer's theory can accommodate this: since suicide is perfectly legal, worlds where Dorothy chooses to commit suicide to avoid missing the deadline are legally "better" than worlds where she chooses to stay alive and misses it. Including a metalaw prohibiting government-mandated suicide changes little, since one of Dorothy's options is voluntary suicide. But one expects that even the bullet-biter will deny here, even if not in the original case, that Dorothy must kill herself.

Note that this is not simply another case of inconsistency in the laws themselves. There is no law that explicitly states that taxpayers unable to make the filing deadline must kill themselves, with another law contradicting it. If that were the case, the metalaw would take precedence and the first-order law would be overridden, and the theoretical representation of this would be that the set M of legal dictates would not reflect the content of the overridden law. In our case, rather, there is a set of laws and circumstances such that the only completely legally permissible course of action available to the agent is suicide, and our theory of *must* makes the surprising prediction that the agent thereby *must* choose suicide.[64] This appears not to be the case, especially in this variation of the scenario including the metalaw. The first problem is that it's unclear how the semantics could account for this. The second problem is that even if the semantics can account for it, the bullet-biter must characterize this new scenario as one in which a *pro tanto* legal obligation was overridden by a dominant law. That seems to be the wrong way to characterize the case. (Compare it to a case with an actual first-order law mandating that anyone unable to file on time must commit suicide, and a metalaw overriding *that*.)

[64] Stated abstractly, the following general principle might have seemed highly intuitive: When there is only one legally permissible course of action available to an agent, the agent legally must choose that course of action. The current case, I submit, is a counterexample.

Third, biting this bullet invites a barrage of others. If it is true that *I must* ϕ whenever ϕ-ing is the only (legally permissible) way to avoid some violation, then a host of other counterintuitive requirements apparently hold in structurally analogous cases. Suppose voting is mandatory for all citizens, but it's too late for me to vote before the deadline. If I can renounce my citizenship to avoid missing the deadline, I must do so. Suppose I must renew my vehicle registration every year, but for some reason I can't this year. If I can somehow get rid of my vehicle before the registration lapses, I must do so. And in any other situation where the only way to avoid a violation is to kill myself, I must do so there as well. Though the predicted requirements in these first two examples are not as severe as suicide, I take it they are nonetheless about as counterintuitive as the verdict that Dorothy must kill herself to avoid missing the tax deadline. They also all have something distinctive in common, which we shall return to in a moment: in each case, the putatively required action would be a way of *avoiding* some legal requirement, rather than *meeting* it. (Avoiding filing taxes rather than filing, avoiding voting rather than voting, avoiding registering the vehicle rather than registering it.) There may be good independent reasons to want our theory to be able to make this distinction, and the most natural way to do so given the formal details may help accommodate the Dorothy case. More on this below.

6.2.2 Smaller Backgrounds

It is crucial to this case that the best worlds in the modal base are ones where Dorothy kills herself. This requires that the rest of the worlds in the base be ones where she fails to file her taxes on time. And *this* requires that the circumstantial background B include the facts about her situation that preclude her from filing on time (e.g. that it is 15 minutes before the deadline and

that she hasn't started the filing process yet). But what reason do we have to think that these facts must be in B? These may be "relevant facts about her circumstances" in some intuitive sense, but that doesn't mean that they must be part of the set of background circumstances appealed to in a technical sense by the theory. If these facts were excluded from B, there would be worlds in the modal base **B** where she does file on time, and these worlds would be as good as the worlds where she kills herself, so it would be false that she must kill herself.

The first thing to say in response is that while it's true that the technical sense of background circumstances appealed to by the theory needn't exactly match our intuitive sense of "relevant circumstances", it ideally should not stray *too* far from it. For it is this intuitive notion that lends initial plausibility to both the Classical View (which appealed to it too) and the Kratzer semantics. Here is Kratzer offering an intuitive gloss on the main thought motivating her view:

> Circumstances create possibilities: the set of possible worlds compatible with them. These worlds may be closer or further away from: what the law provides; what is good for you; what is moral; what we aim at; what we hope; what is rational; what is normal; what you recommended; what we want… To all of those ideals correspond normative conversational backgrounds. Those conversational backgrounds can function as ordering sources for a circumstantial modal base. (1991, p. 646)

Here and elsewhere, Kratzer appeals to the notion of circumstances (sometimes implicitly) in an apparently non-technical sense.[65] The intuitiveness of this appeal to circumstances gives her

[65] See, e.g., also: "Certain conclusions may be drawn from what is known about the circumstances of the crime" (643); "The different occurrences of *can*… mean 'possible given the relevant circumstances'" (640).

view initial plausibility. Divorcing the technical notion of the circumstantial background too far from the intuitive notion appealed to thus sacrifices some of that initial plausibility.

The next thing to say in response is that we may be able to modify the case to force the relevant facts (preventing Dorothy from filing on time) into the circumstantial background. (We made a similar move when discussing the previous case involving the bartender.) Suppose the tax filing process has been standardized into an online process that takes a minimum of three hours, without exception. And suppose there is a special tax discount, designed to encourage timely filing, that reduces one's tax burden proportionately to when one starts the filing process, but only up to one hour prior to the deadline. After that point one is ineligible to request the discount, and doing so is considered tax fraud. Dorothy has not even started, and it is 15 minutes before the deadline. She is ineligible to apply for the discount, so she must file her taxes without the discount. But, of course, she still has the suicide option, and in those worlds she doesn't miss the deadline, so they are better than worlds where she files (late) without the discount. So Kratzer's semantics predicts that she must kill herself.

How is this case different from the original one? In this case, the fact that Dorothy had not started the filing process one hour before the deadline must be included in the circumstantial background B, since it is in virtue of this fact that she must file *without* the discount. But then there are no worlds in **B** where she starts earlier than that, and since the details of the tax-filing system preclude anyone from filing in less than three hours from start to finish, there are no worlds in **B** where she files on time. In this case the legally best worlds in **B** must be the suicide-worlds, so Dorothy must commit suicide. But that claim is no more plausible after the inclusion of this additional detail of the tax code than it was before.

Even if there were something incoherent about this variant of the case, or something wrong about my assessment of it (though I can't see what it would be), the current proposal would still face trouble. For one, even if we could exclude from B the facts about Dorothy's circumstances that render her unable to file on time, thus including in **B** worlds where she does file on time, these would presumably all be worlds where she does something differently in the past. After all (returning to the original version of the case now), it is 15 minutes until the deadline, and there is no way she can file before the deadline in that much time, so any world where she does (barring miracles or feats of superhuman powers) must be a world where she started the process earlier. This means that it is true of Dorothy, right now at 15 minutes before the deadline, that she legally must have done things differently in the past.[66] It seems strange to think that amongst our legal requirements at a given time may be some that require us to do something differently at some earlier time. We do sometimes make claims about what should have occurred in the past – e.g. "Knowing what I know now, I should've trusted my instincts", or "The vote ought to have been declared invalid" – but they seem typically to express a retrospective judgment that at some point in the past *it was* the case that something ought to be, rather than that right now it ought to be that something in the past is different. This may be one of the easier bullets to bite for this line of response, but it is another counterintuitive consequence nonetheless. Presumably we would like to know more about the role of time/tense in the semantics in order to make good sense of it.

Now suppose this consequence is acceptable, and we are happy admitting worlds into the modal base **B** where Dorothy does file on time, even if all of those worlds are ones where she does something differently in the past. The next thing to note is that the worlds where Dorothy

[66] Until now we have not spoken of requirements being relativized to particular times. Presumably the theory will eventually require such relativization, since our requirements can change over time.

files on time are *as good as* worlds where she kills herself before the deadline, but not *better than them*. So this strategy would avoid the undesirable verdict that she must kill herself, but it would fail to yield the intuitive verdict that she must file her taxes by the deadline. Instead, what would be true of Dorothy would be some sort of disjunctive modal, such as *Dorothy must file her taxes or kill herself*. To say the least, this is not how we typically interpret the law.

This result generalizes similarly to the final problem raised for the previous response. Take any situation where, intuitively, one must ϕ, but there is also some permissible course of action ψ available to them that would allow them to avoid ϕ-ing. Earlier we imagined scenarios where I must vote on voting day, or I must renew my vehicle registration, but for some reason am unable to do so, so according to the theory I must take the extreme measures available to me in order to avoid violating (renouncing my citizenship or giving up my vehicle). Well, suppose the details of my circumstances that prevent me from fulfilling my intuitive obligation are not in the circumstantial background B (analogously to the current strategy in the Dorothy case). Then there will be worlds in **B** where I do vote today, or where I do renew my registration. But there will also be worlds where I take the extreme measures instead, and these will be ranked just as highly in the ordering. So rather than it being true that I must vote, or that I must renew my registration, it will be true that I must vote *or* renounce my citizenship, or that I must renew my registration *or* give up my vehicle. Analogous results apply to any other case where I can avoid some intuitive requirement by killing myself or taking similar evasive measures. And note that this will be true even of cases where I have no trouble fulfilling the obligation: even if I could easily file my taxes on time, or vote today, or renew my registration, if *any other* legally

permissible course of action is available to me that would allow me to avoid doing so, my requirements will have the disjunctive form of those just mentioned.[67]

It is important to appreciate that to accept these results would be to embrace either: (i) a radical error theory to the effect that a huge swath of the claims of the form *must* ϕ that we make and assent to every day are false, and rather some weaker claims of the form *must* ϕ *or* ψ are true instead (and often with ψ being something rather extreme); or (ii) a theory of paraphrase, according to which a huge swath of the claims we make that seem to be of the form *must* ϕ are in fact tacitly disjunctive, of the form *must* ϕ *or* ψ (again, with ψ often being something extreme). This is not necessarily fatal to the theory, but it is a serious theoretical cost in either case.

Two further things are worth noting about the "disjunctive option". One, there can also be *explicitly* disjunctive requirements, rather than the implicitly or "surprisingly" disjunctive requirements just mentioned, and it is unclear how the theory could distinguish between them. Sometimes the law really does issue commands explicitly in the form of disjunctions: the sentences for some crimes may have such a form: *you must pay a fine or do community service*; the tax code may provide multiple filing options: *you may file jointly (with your spouse) or separately*; and contracts may stipulate terms in disjunctive form: *a broken heating unit must be fixed or replaced, at the owner's discretion*. These sorts of explicitly disjunctive modals, on the face of it, seem indistinguishable on Kratzer's account from the surprisingly disjunctive ones above, since the disjunctivity of each arises in the same way: in some of the legally best worlds one disjunct is satisfied, and in the rest of the legally best worlds the other disjunct is satisfied.

[67] In fact, one wonders whether this extends even to everyday, "constant" requirements, like the requirements that one not murder, steal, assault, etc. Is killing myself at any given moment a way of "avoiding" those requirements? I don't know. (Killing yourself isn't a way of not not murdering, as selling your car is a way of not renewing the registration.) But if so, then perhaps virtually all requirements will turn out to have the form ϕ *or die* (as it were). This would create even more pressure to amend the theory in the kind of way I propose momentarily.

But one might want one's theory of modals to make some principled distinction between them at the semantic level.

This brings us to the second further point about the disjunctive option, which is that all the putative implicit disjuncts in these cases have something distinctive in common: they are all ways of *avoiding* a requirement, rather than *meeting* it. (This is a distinction which seemed relevant earlier as well.) Killing oneself to avoid missing the tax deadline, renouncing one's citizenship to avoid a mandatory vote, and giving up one's car to avoid letting one's registration lapse, are not ways of *meeting* the corresponding requirements, but rather of escaping them in some way. In fact, in light of this one could characterize the disjunctive option as the claim that whenever there is some permissible way of avoiding an intuitive requirement ϕ, what one in fact must do is not ϕ but *ϕ or (permissibly) avoid ϕ-ing*. (By contrast, none of the explicitly disjunctive cases discussed in the previous paragraph have this form.) The distinction between ϕ-ing and avoiding ϕ-ing seems like a substantive one, and the fact that all the surprisingly disjunctive cases have this feature in common seems suspicious. One would prefer a way of handling these cases that made this distinction, and which avoided the surprising disjunctive analysis.

6.3 Assessment of Case 2

There may be a better way to handle these cases. Let's work backwards from the desired solution. If we want to be able to deny that Dorothy must kill herself, *and* to be able to say that Dorothy must file her taxes rather than merely that she must file her taxes *or* kill herself, we want to somehow exclude worlds where she kills herself from consideration when interpreting her tax-related obligations. But we don't want to exclude suicide-worlds in general, because

there is no prohibition on suicide in our scenario, and ruling out such worlds from the modal base would mean Dorothy must *not* commit suicide. So it sounds like what we need (at least) is a temporary elimination of suicide-worlds from the modal base for the purposes of evaluating Dorothy's tax-related requirements, but an inclusion of these worlds when evaluating her suicide-related requirements. So we want the modal base to be able to vary depending on which of these modals is being evaluated. Now, of course it's already the case that the modal base varies from one modal to the next: when the relevant circumstances change, this changes the modal base, and when the modal flavor changes, this changes which circumstances are relevant, which changes the modal base. But in this case no variation of either the modal flavor or the circumstances are necessary. We can make the claim that Dorothy (legally) must pay her taxes or the claim that Dorothy (legally) may commit suicide at precisely the same context, with none of the circumstances having changed. So the modal base must change for some other reason, either because a different *subset* of the circumstances is considered relevant (i.e. the circumstantial background does not always contain the totality of the relevant circumstances) or because some other restriction on it is imposed.

Consider the first option: perhaps the modal base is determined not by the totality of legally relevant circumstances, but by some privileged subset of them. Well, what circumstances would have to be different to exclude the suicide-worlds in the one case, but not the other? If we want to exclude worlds where Dorothy kills herself from the modal base **B** for the evaluation of the claim that she must file her taxes, and we want to do so by adjusting the circumstantial background B, then we would have to add to B something like the proposition that Dorothy does not kill herself before the deadline, or that she is alive at the time of the deadline. But why would a proposition like this be in the set of relevant circumstances 15 minutes before the deadline?

These facts may not even be settled yet. It would be bizarre to consider them part of her circumstances at the time of evaluation. So perhaps the second option is better: perhaps some other restriction on the modal base is imposed, for some reason or another.

What could this reason be? Let's think about what we're doing when we assert that Dorothy must pay her taxes by the deadline. It's not plausible that we are assuming it is already part of her circumstances that she will survive until the deadline, but we do seem to be doing something like *supposing* it. And supposing that P for the purposes of some assertion is very much like asserting a *conditional*, i.e. a statement of the form *if… then…*, with P as the *if*-clause.[68] (From here on I'll follow standard practice in referring to the *if*-clause as the *antecedent* and the *then*-clause as the consequent, and abbreviating the entire conditional as $A \rightarrow C$, read as "if A, then C".) In this case, the supposition would be something like *Dorothy stays alive until the deadline*, and the assertion made would be *Dorothy must file her taxes,* so the corresponding conditional would be *if Dorothy stays alive at the deadline, she must file her taxes*. And this seems more or less to capture what we want, namely that Dorothy's legal obligation to file her taxes is contingent upon her staying alive until the deadline. The assertion that she may commit suicide before the deadline, however, certainly does *not* suppose that she will be alive at the deadline, so this modal would not be evaluated within the same conditional setting.

The proposal at this point is to treat at least some modals of the form *must* ϕ as implicitly expressing some sort of conditional or conditional-like content, with the content of the antecedent provided by some condition(s) under which the modal would apply. How would this fit into the semantic picture we've already drawn for *must* and other modals? Quite nicely, actually. Angelika Kratzer, in other work (e.g. 1981b, 1986, 1989, 2005), already defends a

[68] In fact, Dorothy Edgington (1995, 2008) argues that this is *precisely* what a conditional is. To assert *if A then C*, she claims, is to assert C on the supposition that A.

theory of conditionals that, by design, coordinates nicely with her broader modal framework. According to her, the antecedent of a conditional acts as a *restrictor* on the modal base by adding its content (the proposition A, in A→C) to the set of propositions determining the base. The effect is to temporarily shrink the base by excluding from it any worlds incompatible with A. Outside the scope of the conditional, this restriction is removed,[69] and the modal base (typically) returns to its previous state.[70] And this is precisely what we want: the supposition that Dorothy stays alive until the deadline is temporarily added to the modal base, so that worlds where she commits suicide before then are ruled out; then the assertion that she must file her taxes is evaluated at this modal base, and assuming there are worlds in it where she manages to file on time,[71] they will be the legally best worlds in the base, so it will be true that she must file. (And, importantly, it will be false that she must kill herself before the deadline, since we will have excluded worlds where she does.) Then the supposition that Dorothy stays alive until the deadline may be dropped later on for the purpose of evaluating other modals for which it would be inappropriate, such as the claim that she may commit suicide if she chooses to.

This proposal has two surprising implications for modals of the kind we have been studying: first, they are quite context-sensitive (in a new way not previously predicted); second, at least sometimes, they tacitly have a very different form than that which they present on the

---

[69] Though the suppositional force of a conditional can be extended throughout a discourse beyond the initial *if… then…* statement. This phenomenon is called *modal subordination*, and will be discussed shortly.

[70] It might not return to exactly its previous state, if other, longer-lasting restrictions are made that survive the exit from the supposition of the antecedent. For example, suppose I'm considering buying a car, and I say: "if it were a Subaru, I could take it off-roading up in the mountains." This introduces a temporary restriction of the base to worlds where I buy a Subaru. Then I may continue: "…But I'm afraid of heights, so I would never take any car up into the mountains." This introduces a restriction of the base to exclude worlds where I take any car up into the mountains. Then I may exit the supposition that I buy a Subaru: "…But what if it were a Volkswagen? They make some nice models…" And the assumption that I would not take a car into the mountains may still hold: "…Of course Volkswagens don't all have all-wheel-drive, like Subarus do. But that wouldn't matter so much, since I'm not taking it into the mountains."

[71] Recall the earlier problem that all the worlds in the base where she files on time would be ones where she does something differently in the past.

surface level. Is there independent evidence for these implications? Yes, in both cases. First, notice that the modal claims *Dorothy must file her taxes by the deadline* and *Dorothy may kill herself before the deadline (without filing taxes)* each sound fine independently, but seem to clash when uttered together as part of a single speech: *#Dorothy must file her taxes by the deadline, and she may kill herself before the deadline (without filing taxes)*. A natural explanation of these data is that each statement requires evaluation relative to a different set of parameters when evaluated separately, and that conjoining them into a single statement forces evaluation relative to a single set of parameters, which causes the clash. We might even make these parameters explicit in order to clear up confusion caused by asserting the statements in close proximity. Consider the following discourse segment:

Speaker 1: Dorothy must file her taxes by midnight. But, of course, she could kill herself before then instead.

Speaker 2: Wait, what do you mean? Does she have to file her taxes or not?

Speaker 1: Well, *if she stays alive that long*, then yes, she does. But she could commit suicide instead, and then she wouldn't have to.

That fact that it is natural to tease apart the two modal statements by using a conditional that makes explicit the conditions under which the one obligation would apply suggests that there is something right to our proposal. (In fact, it offers evidence not only for the context-sensitivity component but also for the conditional component.)

All of this patterns with another, well studied phenomenon that is closely structurally analogous: implicit quantifier domain restrictions. We often quantify over implicitly restricted domains, as in when we say "There's no more beer", meaning not, of course, that there is no more beer *anywhere whatsoever*, but rather that there is no more beer in some contextually

relevant location, e.g. the fridge, the party, the house, etc. In the very same context, finding that the fridge is empty, I could say "There's no more beer", meaning there's no more beer in the fridge, or "There's more beer", meaning there's more somewhere else in the house. But of course saying them together causes a clash: "#There's no more beer and there's more beer". And if we were to cause some confusion with such a statement, it would be natural to clarify by making the restrictions explicit, e.g.: "There's no more beer *in the fridge*, but there's more *in the garage*." Since modals and conditionals are treated semantically as expressions that quantify over *modal* domains, these data are closely analogous.

There is also independent evidence for the existence of modals that implicitly express conditionals or content equivalent to conditionals, namely the phenomenon of *modal* subordination (e.g. Craige Roberts 1989, 1997). The classic examples are ones of the form:

Speaker 1: A wolf might come in.

Speaker 2: It would eat us.

Speaker 1 introduces a possibility, *a wolf coming in*, and Speaker 2 makes an assertion about it. It is widely noted that Speaker 2's assertion is equivalent to a conditional: "*If a wolf came in*, it would eat us". And these sorts of discourses can go on indefinitely, with the same initial possibility serving as the implicit antecedent to what are in effect a series of implicit conditional statements:

Speaker 1: A wolf might come in.

Speaker 2: It would eat us.

Speaker 3: That would be terrible.

Speaker 4: I might get away; I'm very fast.

Speaker 5: I wouldn't; my leg is injured.

…

It has also been noted that extralinguistic context can supply content of the implicit antecedent. Suppose, looking at a high-powered stereo system in an electronics store, I assert (discourse-initially): "My neighbors would kill me!" (Stone 1997) It is obvious that the possibility I have in mind is one where I buy the stereo for my home, and this modal expresses the same content as the corresponding conditional: "*If I bought that stereo*, my neighbors would kill me!" So there is plenty of precedent for the idea that many modals express content that is implicitly conditional in form, with the implicit content determined by the context.

Now to deliver on another promise: how can the conditional view account for the intuitive distinction between *meeting* a requirement and *avoiding* it? Let's review the cases that made this distinction seem desirable. In Dorothy's case, the disjunctive view yielded the result not that Dorothy must file her taxes, but rather that she must file her taxes *or* kill herself. In the analogous cases of the voter who could renounce their citizenship to avoid voting, and the vehicle-owner who could give up their vehicle to avoid renewing the registration, similar disjunctive requirements were predicted. And not only was it counterintuitive that these requirements should take this disjunctive form, but it also seemed conspicuous that the putative implicit disjunct in each case happened to be a way of avoiding the requirement, rather than meeting it. Now let's compare the conditional view's analyses of these cases, keeping this distinction in mind. In Dorothy's case, the implicitly conditional content of the modal *Dorothy must file her taxes* takes the form of a suppressed antecedent clause: *If Dorothy stays alive until the deadline*… In the voter's case, the implicitly conditional content of the modal *I must vote* would presumably be: *If I remain a citizen*… And in the vehicle owner's case, the implicitly conditional content of the modal *I must renew my registration* would presumably be: *If I keep my*

vehicle… What do all these conditional contents have in common? They express the *conditions of applicability* of the given requirement. Dorothy's staying alive until the tax-filing deadline is the *condition* under which the requirement to file her taxes would *apply* to her; if she dies before then, it does not apply, and she is not considered in violation for not having filed prior to her death.[72] Similarly, my retaining my citizenship is the condition under which the voting requirement applies to me, and my keeping my vehicle is the condition under which the requirement to renew my registration applies. The reason the conditional view yields more intuitively acceptable results in these cases is because the conditional interpretations of these modals rule out the option of *avoiding* one's obligation precisely by supposing that the applicability conditions for that obligation obtain. The cases of requirement-avoiding are cases of changing one's situation so that the requirement no longer applies to them, rather than performing the action that the requirement commands. If we assume that one's situation remains such that the obligation does apply to them, then the only way to avoid violating the requirement is to *meet* it.[73]

Now that the proposal is fully on the table, there remains the task of describing it formally. There are a number of possible ways to do this, depending on various theoretical choices, and I will not take the time to explore them thoroughly here. For instance, one might

[72] Note that this is non-trivial: the law could state that suicide is permissible, but one must file taxes for the previous tax year before committing suicide. In this case Dorothy's requirement to file taxes would *not* be conditional upon her being alive past the tax deadline, and worlds where she kills herself (without filing) to avoid missing the tax deadline would *not* be legally better than worlds where she misses the deadline.

[73] This distinction is already established in the literature on practical rationality, and there is a promising analysis of it that harmonizes with our discussion here. Errol Lord (2011) prefers the terminology *exiting* and *complying*, where to exit a requirement is to somehow make it the case that the requirement no longer holds, and to comply with a requirement, by contrast, is to satisfy the demands of the requirement, i.e. to do what is asked. His arguments outrun the scope of our discussion, but he argues that there is good reason to think of certain kinds of rational coherence requirements as having a form much like the implicit conditionals under discussion here, e.g.: *If I believe I ought to ϕ, then I must intend to ϕ*. And part of the reason for interpreting them this way is to capture the distinction between exiting and complying. Were there only space here, a more thorough comparison of his discussion to ours would be valuable.

treat the kinds of modals with apparent implicit conditional content as special cases, and attribute to other modals the simpler, non-conditional form that they exhibit on the surface. Or one might treat all simple modals as implicitly conditional, with some cases having null or trivial antecedents. I have nothing personally invested in this decision. (To clarify, like the previous amended version of the semantics, I do not *endorse* this conditional view, I am merely proposing it as the best way for a Kratzerian to respond to my puzzle case.) For simplicity, let's choose the second option. Modals of the surface form *It must be that P* express (typically) implicit conditionals of the form *If A, then it must be that P*, where A encodes the applicability conditions for the requirement in the consequent. (Nevermind what determines the applicability conditions – presumably it is something in the modal flavor M.) Recall that on Kratzer's view of conditionals, the antecedent A simply acts as a temporary restrictor on the modal base, i.e. eliminates from the base any ¬A-worlds. In our case this is equivalent to intersecting the set of worlds **B** with the set of A-worlds **A**, and taking $\mathbf{B} \cap \mathbf{A}$ as our new modal base for the purposes of evaluating the modal in the consequent.[74] We can then plug this into the previous formal machinery (keeping the agent-relativization feature we added earlier), with the addition that *must* be relativized yet again to a new parameter, the applicability conditions A, and the result is:[75]

K-must: $\text{MUST}_{M,B,\alpha,A}(P) =_{df} \leq_{M,\alpha}(\mathbf{B} \cap \mathbf{A}) \subseteq \mathbf{P}$

[74] Why not just think that the applicability conditions A are part of the content of B? As noted above, it seems implausible in Dorothy's case that her circumstances include the applicability conditions for the requirement that she file her taxes by the deadline. After all, how can it be part of her circumstances at *t* that she be alive at *t* + 15 minutes, when the truth or falsity of that proposition may not even be settled yet at *t*? (Similar remarks apply in the analogous cases of the voter and the vehicle owner.) I suppose one might be inclined to describe the situation as one involving a temporary, hypothetical addition of this proposition to B. But this is essentially equivalent to the conditional interpretation. On this construal of the proposal the central innovation would be that B sometimes contains temporary, hypothetical additions that would not normally be thought of as part of the background circumstances.

[75] Note that this "amendment" to the Kratzer semantics, unlike the previous amendment to include agent-relativization, needn't necessarily be construed as a *modification* of the view, so much as a supplement to it. It could be that Kratzer's semantics is correct for the purely modal components of these claims, but that sometimes natural language sentences of the form *It must be the P* also have a conditional component, which is not captured by her original semantics, and this more complicated semantics simply incorporates both components.

K-can: $\text{CAN}_{M,B,\alpha,A}(P) =_{df} \leq_{M,\alpha}(\mathbf{B} \cap \mathbf{A}) \cap \mathbf{P} \neq \varnothing$

Now let's take this for a test run to make sure we've got what we want. In Dorothy's case, in the context where it's intuitively true that she must pay her taxes before the deadline, the implicit applicability conditions are that she remain alive until the deadline, so this will be our A. The circumstantial background B includes facts about her situation; assuming B doesn't include the facts that render her unable to file before the deadline, e.g. that she did not start filing early enough to complete the process on time, then there will be some worlds in **B** where she does begin early enough and thus meets the deadline. And prior to the restriction induced by A there will also be equally legally good worlds where she kills herself before the deadline instead, thus *avoiding* a violation of the tax code. But *after* the restriction induced by A, these latter worlds will no longer be in the modal base, since these are worlds where the applicability conditions do not hold, i.e. where she fails to remain alive until the deadline. Now the legally best worlds in the temporary new modal base (now $\mathbf{B} \cap \mathbf{A}$) will be ones where Dorothy files on time, so it will be true that she must do so. However, mentioning the possibility of her committing suicide instead can force a new context in which the supposition that Dorothy remains alive until the deadline is dropped, so the restriction induced by **A** is undone. Now we are back to our previous modal base **B**, among whose best-ranked worlds are both ones where Dorothy files on time and ones where she kills herself, and so it is not the case that she *must* kill herself, but it is the case that she *may*. We seem to have gotten just about everything we wanted.

One standing problem, as noted earlier, is that all the worlds in the modal base where she does file on time will be ones where she does something differently in the past, relative to the time of evaluation for the modal. We noted there was something a bit bizarre about thinking that Dorothy could have at time *t* (15 minutes before the deadline) a requirement to do something

differently at some previous time *t*- (perhaps several hours before *t*). If we are uncomfortable with this result we might be tempted to insist that the facts about Dorothy's timeline are in fact included in B, at least for the evaluation of the restricted modal (*Dorothy must file on time*). But this would yield the result that there are *no* worlds in the base where she files on time, so that modal would be false. There would also be no worlds in the base where she kills herself, due to the restriction induced by A, so it would also be false that she must kill herself. But now what would be the best worlds in **B**? Presumably ones where she files late, but pays whatever fines or other penalties she incurs as a result. This doesn't sound so bad at first: she cannot file on time, so she is required to pay the penalties. But think about the timeline again: the time of evaluation is prior to the deadline, but she is already required to pay a fine that is imposed as a punishment for missing the deadline. This seems wrong too: before the deadline her obligation is to file regularly, and then once the deadline has past, and she has not filed, only then does she incur the new obligation to pay a fine. If she were to die suddenly just before the deadline, she would not die with an unfulfilled obligation to pay the fine. It's true that she *will* have to pay a fine, but not yet that she *does* have to pay it.[76]

At any rate, it isn't clear which counterintuitive consequence the Kratzerian should pick here: it is either the case that Dorothy is required at time *t* to have done things differently at some prior time, or that she is required at *t* to do something that intuitively should not become a requirement until some later time. But either bullet is easier to bite than the result that Dorothy must choose death before taxes.

[76] Magnify the time spans here, and the counterintuitiveness increases. Imagine an analogue of the case where it is settled days in advance that Dorothy will not be able to file on time. Then she has already incurred a legal obligation to pay a fine for missing the deadline days before the deadline has arrived. This seems implausible.

7.1 Case 3: The Joker

The National Art Gallery is a government-owned museum filled with billions of dollars' worth of fine art. The Joker, notorious criminal and modern art skeptic, walks in with a can of spray paint and destroys several Rothko paintings within a matter of minutes. The damage is assessed at 100 million dollars. Legally, the Joker must repay the government for the damage he has done to the Rothkos. So the following seems true: *the Joker must pay 100 million dollars to the government*. But the Joker is destitute. His net worth is virtually nothing. Given his financial circumstances, there is no way he can pay anyone that much money. If his financial circumstances are legally relevant – and surely they often are – then there are no worlds compatible with the relevant circumstances where he pays the government what he must (or anything close to it). So *a fortiori* none of the *legally best* worlds compatible with the relevant circumstances are ones where he pays what he must. So the Kratzer semantics predicts it is false that he must pay 100 million dollars. (In fact, it predicts that he *must not* pay 100 million dollars!) But this seems to be his legal requirement. Furthermore, aside from the unusually high dollar amount, this sort of situation seems entirely unremarkable: we are generally legally liable for the damages we cause to the property of others, and we can easily do far more damage than we are able to pay for.

7.2 Responses

7.2.1 Smaller Backgrounds

The first question to ask is whether the Joker's financial circumstances are in fact relevant in this case, in the sense we have in mind. Of course our financial circumstances are legally relevant in many situations, e.g. when determining how much we owe in taxes, whether we are eligible to

file for bankruptcy or certain kinds of financial aid, etc. But this doesn't necessarily mean they are legally relevant in every case, at least not in the sense appealed to by the theory. So it could be that the Joker's financial circumstances are not part of the circumstantial background, so there are worlds in the modal base where he does have enough money to pay the fine, and in the best of those worlds, he does pay it.

But we can adjust the example a bit, as we did in the previous two cases, to force these facts into the circumstantial background. Suppose the relevant law includes a financial hardship clause, according to which defendants whose net worth is below a certain low threshold automatically receive a 10% reduction of any legal fines, no matter how great. And suppose the Joker qualifies for such a reduction. In this case the Joker's legal dues are different: he must pay *90* million dollars, instead of 100 million. And his financial circumstances – at least the fact that his worth is below the threshold – must be among the relevant circumstances, because it is in virtue of these facts that he must pay 90 rather than 100 million.[77] To put it more clearly: suppose the circumstantial background B *didn't* include the fact that the Joker's worth is below the threshold. Then there would be worlds in the modal base **B** where the Joker's worth is above the threshold. How much does he pay in those worlds? Assuming there are any where he has enough, suppose he pays 90 million in some worlds, and others where he pays 100 million. Either way he fails to comply with the law. Suppose he pays 90 million: then it's true that he pays what he actually owes, but it's also true that he pays 10% less than the damages he has caused *and* he is worth too much to qualify for the reduction that would allow him to pay so little. Or suppose he pays 100 million – then he pays more than he actually owes. (We may

[77] Given our previous discussion, one might actually think that the fact that the Joker's worth is below the threshold belongs in the applicability conditions, rather than the circumstantial background, since it's in virtue of this fact that the reduction applies to him. Maybe so, but the modal base **B** would still be restricted to worlds where his worth is below the threshold, just with the restriction coming from a different source.

suppose, if we like, that overpayment is a crime too, so that he must not pay more than 90 million in this version of the scenario.) So the fact that his worth is below the threshold must be in B, because it is in virtue of this fact that he owes 90 million rather than 100 million. And if this fact is in B, then there will be no worlds in **B** where his worth is more than the threshold, and thus no worlds where he pays as much as he owes.

7.2.2 Bite the Bullet (The Joker Mustn't Pay)

The only other response, short of a significant change to the semantics, that seems worth considering is biting the bullet and denying that the Joker must pay 100 million dollars. Here again, experience shows that some will find this response at least tempting. I can only imagine that the temptation is rooted in a familiarity with actual legal procedures in the contemporary developed world, and a sense that this is somehow a result of necessity. It is true, at least in US law, that when one is unable to pay their legal dues as a result of their financial circumstances some sort of compromise is often reached. If I am fined more than I can pay as part of a sentence for a crime, I may have the option of community service instead. If I owe more than I can pay at the time the debt is assessed, I may be able to negotiate some sort of payment plan. And of course the legal status of bankruptcy is written into law in the US and many other places precisely to handle cases where debts greatly outrun the means of the debtors who possess them. But there are two important things to note about this. First, these sorts of negotiations are most appropriately characterized as a *change* in legal/financial obligations: one initially must pay more than they can, and some process is built into the legal system to modify these legal obligations in a way that makes them possible (or easier) to comply with. It would be unnatural to describe these situations as ones in which the obligation to pay the higher amount never

existed in the first place simply because it was more than the defendant could pay. Second, the possibility of these sorts of negotiations is highly contingent. It is a fact in many contemporary societies that the law provides processes for resolving conflicts between what is required of one and what one is able to do – but surely it needn't be! These sorts of laws had to be decided upon and written into the legal code. It is not a necessary feature of the law that it accommodates those who cannot meet their obligations. For the current response to make any sense, it would have to be the case that the law *cannot* require you to pay more than you can. But this just seems downright false. The claim "I don't have to pay it because I can't" is, unfortunately, rarely true.[78]

7.3 Assessment

I can't see how to handle this case without a drastic revision to the semantics for *must*. The trouble is that the Joker must do something he doesn't do at *any* of the worlds compatible with his circumstances, so he of course doesn't do it at any of the *best* worlds compatible with his circumstances. The theory is designed to read obligations and permissions off the details of the worlds compatible with our circumstances. When the circumstances preclude perfect compliance with the law in some way, the theory (by design) looks instead to worlds that fail to perfectly comply in that, but which otherwise minimize non-compliance. Whatever goes on at *those* worlds is what *must* be. But this structural feature of the view imposes a constraint that seems appropriate specifically for types of modality that are generally *maximizing*, i.e. which generally require that one do the best they can, given their options. It may be that preference- or rationality-based requirements work this way. And if you are a consequentialist about morality,

[78] Note that it's not actually clear *what* the Joker's obligation would be instead, i.e. what he does in the best worlds compatible with his financial situation. I suppose he pays whatever he can, which isn't much given that he's destitute. But surely his legal requirement is not simply that he pay whatever he can!

then you think that morality works this way. And there are no doubt other modal flavors that work this way. But the law, for one, appears not to. The law does not always care whether or not one is able in any sense to comply with it.[79] The law may be very unreasonable. Think back to the days of witch hunts and the Inquisition. There were all sorts of laws about witchcraft that required and forbade impossible things. Perhaps at some point, somewhere, there was a law requiring anyone found guilty of witchcraft to turn themselves into a newt. Nevermind the impossibility, metaphysical or otherwise – the law is the law!

In fact, this shortcoming of the Kratzer-style semantics is analogous to the shortcomings of the Classical View that motivated abandoning that view. The problems for the Classical View were cases where there were either no worlds compatible with both the laws and the circumstances (in the case of Violation-based Requirements), or no worlds compatible with the laws at all (in the case of Inconsistent Requirements). The solution was to rank the legally imperfect worlds compatible with the circumstances, and quantify over the best of those. In the Joker's case we have another scenario where there are no worlds compatible with both the laws, which say that one must pay for damages one causes to others' property, and the circumstances, which determine that the Joker cannot pay. But in this case ranking the legally imperfect worlds compatible with the circumstances doesn't help, because the Joker's requirement is not that he do the best he can, but rather that he do the thing he cannot.

So what should we make of this case? It isn't clear. Perhaps a new semantics for modals is needed. Perhaps possible worlds are insufficient, and something hypermodal is needed. Or perhaps we simply need to find some way within the Kratzerian framework, or something like it,

[79] The more common way of putting this, of course, would be to say that *ought* does not always imply *can*, for the legal *ought*, or at least that *must* does not imply *can*, if *must* and *can* turn out to be very different. The structure of Kratzer's modal framework *per se* seems to preclude this possibility. In fact, it seems to have the consequence that *must* ϕ implies *can* ϕ for every non-trivial instance of *must* ϕ.

to connect requirements to the actions they are about. I cannot see how to do that here, but I will not go so far as to claim it cannot be done.

8 Conclusion

I raised three novel problem cases for the standard Kratzer-style semantics for modals like *must* and *can*. The first case led me to conclude that the best bet for the Kratzerian was to make a minor amendment to the theory relativizing modals and orderings to agents. This seemed promising, but left some explanatory work to be done about how this more complicated ordering procedure would work. The second case led me to conclude that at least some modals should be treated as containing implicit conditional content, with the applicability conditions as the content of the antecedent. This coordinated nicely with Kratzer's independently motivated analysis of conditionals, it made the predictions we wanted, and it allowed us to make a useful distinction between *meeting* and *avoiding* a requirement, but it left us with a choice between two somewhat undesirable consequences, both of which were nonetheless better than the initial problem the amendment was meant to solve. The third case left me unable to see a path to a solution, since it seemed to require something other than quantification over the worlds available to the Kratzer-style semantics. If there is a solution that doesn't require abandoning or drastically revising her semantics, I can't see what it is at this point, but it is premature to rule it out.

Works Cited:

Bennett, Jonathan (2003). *A Philosophical Guide to Conditionals*. Oxford University Press.

Edgington, Dorothy (1995). On Conditionals. *Mind*, New Series, Vol. 104, No. 414: pp. 235-329.

Edgington, Dorothy (2008). Counterfactuals. *Proceedings of the Aristotelian Society, Vol.* cviii, *Part* 1: 1-21.

von Fintel, Kai (2001). Counterfactuals in a Dynamic Context. *Current Studies in Linguistics* series 36, 123-152.

Kratzer, Angelika (1977). What 'Must' and 'Can' Must and Can Mean. *Linguistics and Philosophy* 1 (3): 337-355.

Kratzer, Angelika (1981a) 'The Notional Category of Modality'. *Words, Worlds, and Contexts: New Approaches in Word Semantics*. Eds. Hans Jurgen Eikmeyer and Hannes Rieser. Berlin: W. de Gruyter, 1981. 38–74.

Kratzer, Angelika (1981b). Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic* 10 (2):201 - 216.

Kratzer, Angelika (1986). Conditionals. *Chicago Linguistics Society* 22 (2):1–15.

Kratzer, Angelika (1989). An investigation of the lumps of thought. *Linguistics and Philosophy* 12 (5):607 - 653.

Kratzer, Angelika (1991). Modality. In *Semantics: An International Handbook of Contemporary Research*. Eds.: Arnim von Stechow and Dieter Wunderlich. Berlin: W. de Gruyter. 639-650.

Kratzer, Angelika (2005). Constraining Premise Sets for Counterfactuals. *Journal of Semantics* 22 (2):153-158.

Lassiter, Daniel (forthcoming). Must, Knowledge, and (In)directness. *Natural Language Semantics*.

Lewis, David (1973). *Counterfactuals*. Oxford: Blackwell Publishers.

Lewis, David (1986). Counterfactual Dependence and Time's Arrow. Reprinted in *Philosophical Papers, Volume II*. Oxford University Press.

Lord, Errol. Violating Requirements, Exiting from Requirements, and the Scope of Rationality. *The Philosophical Quarterly* 61 (243): 392-399

Roberts, Craige (1989). Modal subordination and pronominal anaphora in discourse. *Linguistics and Philosophy* 12 (6):683-721.

Roberts, Craige (1997). Anaphora in Intensional Contexts. In Shalom Lappin (ed.), *The Handbook of Contemporary Semantic Theory*. Blackwell. 215-246.

Stalnaker, Robert C. (1968). A Theory of Conditionals. In Nicholas Rescher (ed.), *Studies in Logical Theory*. Blackwell.

Stalnaker, Robert C. (1981). A Defense of Conditional Excluded Middle. In William Harper, Robert C. Stalnaker & Glenn Pearce (eds.), *Ifs*. Reidel.

Swanson, Eric (2008). Modality in Language. *Philosophy Compass* 3/6: 1193-1207.

Szabo, Zoltan Gendler & Knobe, Joshua. Modals with a Taste of the Deontic. Forthcoming in *Semantics and Pragmatics*.

Yalcin, Seth. 'Epistemic Modals'. *Mind* 116 (2007): 983–1026.

Zvolenszky, Zsófia (2002). Is a Possible-Worlds Semantics of Modality Possible? A Problem for Kratzer's Semantics. In Brandon Jackson (ed.), Semantics and Linguistic Theory XII. Ithaca, New York: Cornell University Press, 20 (339–358).