Control and Its Limits

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the University of Pittsburgh

Leona Mollica

June 9, 2026

for
Douglas Vaaler

"In directing itself toward . . . and in grasping something, Dasein does not first go outside of the inner sphere in which it is initially encapsulated, but, rather, in its primary kind of being, it is always already 'outside' together with some being encountered in the world."

Martin Heidegger, *Being and Time*

Contents

Acknowledgements

By rights, this dissertation should never have been. That it exists at all is in spite of several personal setbacks I had no reason to expect I could surmount, and that it has what virtues it may is due overwhelmingly to the generous assistance and encouragement of others who were under no obligation to provide it. I therefore owe a considerable debt of gratitude to those who have made this thesis, against all odds and entitlements, possible.

I would like, first, to thank my parents, who both provided me with ample scope in my youth to explore the interests that have culminated in this project and have been an ongoing source of support and motivation. Thanks, mom and dad.

I also owe much to the small, idiosyncratic, but vibrant intellectual community at the late lamented independent Shimer College, who took me in as a weird bookish high school drop-out when other academic opportunities looked scarce. Without the unanticipated foothold they offered me in the academic world, any graduate studies here at Pitt would probably have been impossible.

I should single out a few members of that community, specifically. Among the faculty, special thanks to my undergraduate advisor James "JD" Donovan, who was supportive both personally and intellectually well past what his professorial obligations required; and to Adam Kotsko, whose two courses on Heidegger my junior year are reflected among other places in this dissertation's epigraph. Among the students, I am especially grateful in this context to the small group that first exposed me in a serious way both to the analytic tradition altogether and in particular to the output of the philosophical community at Pitt: Matt Kawahara, Sam Elalouf, and Joe Bradshaw. When there

were no classes about Frege or Russell or McDowell or Thompson on offer from the college itself, they worked on their own initiative to make those philosophical currents available nonetheless.

Also from my undergraduate days, I should make mention of my tutors during my brief time in Oxford: Daniel Came and Edward Kanterian. If it were not for their generous teaching, honest evaluation, and institutional advocacy, I doubt I would be here now. I should also thank Alex Skiles, whom I met as a wide-eyed senior undergraduate at the 2013 Society for Exact Philosophy conference and who kindly offered very helpful feedback during the graduate admissions gauntlet, and Alexander Pruss, who also generously provided feedback on what became my writing sample when applying.

I owe, of course, an invaluable debt to the graduate programme at Pitt, which both took me in despite my near complete lack of formal background in the analytic tradition and my oddball undergraduate pedigree and readmitted me after a long, unexpected, difficult leave of absence. Giving a list of faculty to whom I am individually greatly indebted would be more or less just giving a list of all the faculty whose courses I've taken (and then some), but a few names stand out for special gratitude. My advisor, James Shaw, and committee members Doug Blue and Sven Neth have been enormously considerate, stimulating, and helpful throughout this process (and, in James' case, well before). Mike Caie and (especially) Harvey Lederman also went way out of their way to assist me, well above and beyond their duties, and I am incredibly grateful for it. In many ways I think of the small seminar on vagueness in spring of 2017–for much of the duration attended just by Mike, Harvey, Pablo Zendejas Medina, and myself–as the high point of my whole graduate career. Erica Shumener, Dmitri Gallow, and Jessica Gelber were also notably helpful at various points in my time in the programme. The graduate student community has been similarly wonderful to me, again with too many to list them all. Special note, however, to Stephen Mackereth, Tom Breed, Laura Davis, and Douglas Vaaler–to whom this dissertation is dedicated.

I also owe a warm thank you to the USC philosophy department, who took me in as a visiting student in the 2018-19 year. Particular thanks to my host Jeremy Goodman and to my outside reader Andrew Bacon.

Outside the world of academic philosophy, I have a lot of friends to thank for their intellectual and moral advice and support. My brilliant

and beautiful partners, Elise Stalker (who provided helpful feedback on portions of the text) and Dorothy Scheurer (who introduced me, among other things, to the full potential of TikZ and to the main cast of the examples in Chapter 4); my dear friend, Victoria Trujillo; Taylor Friesen; Esther Kemball and Robert Friel; Stephanie Dalton; Ophelia Bauckholt (rest in peace); the former staff at the Berkeley REACH and current staff at Oakland's main public library branch; and several others, both from real life and on the internet, whom I do not have space to sufficiently list. Even when not specifically dedicated to philosophy, it has been a treasure beyond price to have so many generous and intelligent friends to bounce ideas off of and offer insight about where might be worth looking in intellectual pursuits, and to have provided the kind of social infrastructure that makes serious thought possible.

And, last but not least, an impersonal but no less heartfelt thank you to the heroes at Sci-Hub, Anna's Archive, and Library Genesis, without whose gallant efforts much of the research that went into this project would have been completely impossible. *Yo ho ho!*

Preface

Before beginning, I would like to make some remarks about the history and context of this project.

The ontogeny of this dissertation does not recapitulate its phylogeny; not even in reverse. The initial germ was the thought, which struck me on the final day of the 2018 Society for Exact Philosophy conference at the University of Connecticut, that the state of the art in ethics on "moral luck" could benefit from a combination of the argumentative methods deployed by Timothy Williamson against the sceptic in *Knowledge and Its Limits* and something like the general outlook developed by John McDowell in *Mind and World* in contrast to foundationalist faith in the "Given" and coherentist "spinning frictionlessly in the void" of one's own innerworldly mental contents. Much (perhaps all) of the professional writing on that problem seemed beset by assumptions about the logic of control very similar to those so trenchantly attacked by Williamson and McDowell in the logic of knowledge and perception, and thus liable to similar improvements from an analogous cure.

This logical assumption (of whose possible content at the time I had only a vague sense) was rarely if ever made explicit, but seemed essential to pretty much all published discussion. It somehow drove a wedge between, on the one hand, the belief that ordinary human agents are frequently responsible (that is, blameworthy and creditable) for ordinary, physical acts like spilling glasses of water or hitting someone in the face and, on the other, the general belief that one can only be responsible for things one controls. The tension this assumption generated between these two data of common sense defined the boundaries of all extant philosophical discussion of the matter, with (it seemed to

me) very ill effect.

This germ would come to eventual fruition in what is now Chapter 4, though not in quite the form I had first anticipated. The actual formal structure of the argument would, in the end, owe more to Delia Graff Fara's work on vagueness, as mediated by some unpublished work of Andrew Bacon's on epistemology, than to the anti-sceptical arguments of *Knowledge and Its Limits* (and their close kindred in that book, the arguments against luminosity and the KK principle) themselves. And, in drafting the essay, it became obvious that to make the main arguments clear either to readers or to myself I needed to work out in at least some preliminary detail an appropriate logic for the operative concept of control, together with its metalogic. The central thesis of that essay was a contentious one, since it contravened the entire basis of the literature on the topic to date, and an informal sketch of a proper argument wouldn't cut it with dialectical stakes that high, necessitating some level of formal development for the general logical framework in which the paper's main argument would take place.

The logical prolegomena to which the moral luck project directed me would take shape, with the extensive and very insightful aid of Andrew Bacon at USC, as Chapters 2 and 3, expanded versions of a publication in the *Journal of Philosophical Logic*. Assumptions very similar to those I hazily detected in the moral luck literature, which the logic I developed there gave me the terms to properly criticise, I also realised abounded in much formal work in practical philosophy and the theory of action, in particular decision theory and the so-called "STIT" tradition. As it evolved in the course of writing this dissertation, this work was divided among two chapters: Chapter 2 would set down the basic logic and its differences from that of the STIT programme, while Chapter 3 would extend and apply that logic to these questions in (especially) decision theory.

There were, naturally, developments outwards from this basic core, often at the prompting of others. §2.4, for example, I was inspired to write after some conversations with Douglas Vaaler about the work of Christine Korsgaard. Korsgaard, in *Self-Constitution* and related papers, gives pride of place to a very briefly exposited argument about action, choice, and modality, whose conclusions seemed almost tailor made for investigation in the logical framework developed in this dissertation. That section outlines some of the yield from those investigations. The conclusion, similarly, develops more explicitly certain

recurring motifs (about control, modality, and the split between meta- and object language) to have emerged throughout the writing of the main body of the text.

Thus, the structure of the project, as a product of its history: at the end, an intervention in the theory of responsibility; as groundwork for this at the beginning, an investigation into the logic of control and action; and, in between the two, additional applications of that preparatory investigation whose connections became clearer with the logical matters in view. The means first and the ends last, separated by byproducts and detours of interest.

I say all of this partly by way of simple autobiography, but also partly by way of trying to acquit my general project of a charge that could plausibly be laid against it, and to which I am at least a little sensitive: that it trivialises very serious questions about morality and action, and cavalierly reduces discussion of them to a sort of philosophers' parlour game. It is true that the current dissertation is, perhaps, slightly more formal in style and argumentative character than most dissertations in practical philosophy. Such formal methods can introduce distinctive philosophical temptations: towards a certain picture-thinking in terms of one's models, or taking the mere existence of a model with certain properties to *eo ipso* carry philosophical weight. Hercule Poirot's adage about the imagination in detective work applies with equal justice to formal models in philosophy: they are "a good servant, and a bad master."[1] At several junctures (e.g. §3.2.2, §3.3.3, §5.2.2), I have made efforts to push back on such temptations as they arise. And, as I have tried to explain in this preface, this formalism was adopted in the service of getting its ultimate arguments right, not for its own sake. The motivations of the project as a whole, despite its relatively formal execution, are more "humanistic" in orientation than narrowly technical.

I think the stakes of those motivating questions are real, and of concern well beyond the academy. The way in which most philosophers talk about "moral luck" seemed, and seems, to me to take for granted we must jettison one or the other of two dear standards of common sense, either of whose losses would be intolerable. On the one hand, we might reject the ubiquity of credit and guilt, or at least

[1] From *The Mysterious Affair at Styles*

that they attach to anything at all like the kinds of facts to which we ordinarily attribute them. On the other, we might accept that people can fairly be blamed for matters beyond their control, that "I couldn't help it!" might not be an indefeasibly exonerating cry from the accused (when true, at least). Together, I felt, these constituted the core of any notion of moral responsibility worth preserving: the first, our foothold as to the scope of the concept; the second, its overarching role in our practices of blaming and praising. This role without that foothold would be hollow, and the foothold without the role pointless; to retain the one in spite of losing the other would, in a sense, not even mean saving that much.

Both of these rejections appeared to me as live possibilities among educated, socially influential, and generally well-meaning sections of the public. Many of my peers, from my undergraduate days onward, openly embraced a view of right and wrong on which the principle that one is only to blame for what one controls is hopelessly retrograde and individualist. On this view, blame attaches not only to the perpetrators of certain sorts of wrongdoing, but also to their beneficiaries, however helpless they might be to resist these ill-gotten benefits. There was a similar tendency apparently at work in the way certain psychological predilections towards cruelty and abuse would be publicly castigated at once as both evil and uncontrollable, with no apparent sense of contradiction. All of these ways of thinking struck me as deeply morally dangerous, and without the control principle to appeal to there would be no way of articulating that danger.

On the other side, there was the way in which many people would champion social and legal reforms to excise all application of the notion of guilt from public life. Such thinking is particularly apparent in the way criminal punishment and other forms of institutional penalisation are sometimes discussed, as an antiquated or reprehensibly un-humanist practice. Crime specifically and social nuisance generally, the thinking goes, should be treated more along the lines of *disease* than those of *sin*, and the public measures to deter, limit, and sanction them should be undertaken with an eye exclusively to the broader social welfare and the welfare of the accused himself, rather than weighing the accused's fictitious "deserts" or "merits". This is the attitude that Elizabeth Anscombe so aptly criticised in some of her political writing:

> The reformatory theory [of criminal punishment] assimilates the state to a parent. In itself this is a monstrous impertinence. For it means that purely in virtue of the position of holding civil power, some men may claim to dispose of others in this way against their wills for their own good. Upholders of the 'reformatory' idea, or the idea of 'rehabilitation' are probably confused about this, not noticing the actual character of the claim; but they may be motivated by a certain sweetness. The good in their idea is the good of the injunction: "Do not forget, even in punishing him, that the convict is your brother, now in your power: do not become callous about his good." They take it for granted that they - or 'we' - have a right to decide 'what to do with criminals'.[2]

> Confusion arises on this subject [*viz.*, crime and criminal punishment] because the state is said, and correctly said, to punish the criminal, and "punishment" suggests "vengeance." Therefore many humane people dislike the idea and prefer such notions as "correction" and "rehabilitation." But the action of the state in depriving a man of his rights, up to his very life, has to be considered from two sides. First, from that of the man himself. If he could say "Why have you done this to me? I have not deserved it," then the state would be acting with injustice. Therefore he must be proved guilty, and only as punishment has the state the right to inflict anything on him. The concept of punishment is our one safeguard against being done "good" to, in ways involving a deprivation of rights, by impudent powerful people.[3]

Such attitudes towards wrongdoing and punishment, for all their purported humanity towards the subjects of what punishment they condone, thus threaten to collapse into a noxious paternalism about and utilitarian devaluation of the individual human being punished. Nor, I think, is this an idle threat. Often, when confronted with particularly galling cases of institutional mistreatment of some hapless

[2] From "The Source of the Authority of the State"
[3] From "Mr. Truman's Degree"

victim or another, something like the conception of the right to punish Anscombe is attacking in the above passages has seemed operative.

It is claims like these, and the courses of political and social action they suggest, that give the true stakes I saw in my project, little remarked upon though they are in the content of the main work itself. Because the function that work is performing on this front is merely preparatory, merely ground-clearing. I have tried to dispel certain confusions I believe have hindered philosophical discussion of responsibility and control, with which the way to a clear understanding of the relevant moral phenomena will be impassible, but without which that way is still long and difficult. Whether I have succeeded in even that much, of course, is not for me to say. But I hope, in articulating that task here, I have done something to justify my undertaking it, and undertaking it after the manner that I did.

University of Pittsburgh
August 2026

Chapter 1

Introduction

Progress in understanding properties, objects, and expressions of central philosophical interest emerges from rebutting wide-reaching doubts about their prevalence, existence, or meaningfulness. This is a sweeping and vague metaphilosophical pronouncement, and as such is subject to numerous counterexamples and ambiguities. But at least as remarkable as these problem cases for the claim are those cases it gets just right. We see the pattern, for example, in the role of the sceptic in debates within epistemology, who doubts that subjects like ourselves know much of what we ordinarily take ourselves to know; or of the unscrupulous amoralist in ethics, who doubts that we are subject to many, if any, of the moral norms that are usually assumed to bind us; or of the austere extensionalist in philosophising about modality and intensional attitudes, who doubts that intensional notions of possibility, necessity, or propositional attitudes are necessary to capture what is worth capturing in discourse about those topics, or perhaps that those intensional notions are even intelligible. Examples could clearly be multiplied. In each case, advances in our philosophical understanding come and have come largely, or perhaps even predominantly, from attempts to think through what is wrong-headed, or at least unpersuasive, in the radical doubts about the relevant subject matter.

In other domains, this is not so. Consider, for example, astronomy. I am aware of no prominent figures in the astronomical tradition to have seriously contended that the stars in the night sky are in fact all

or nearly all illusions, and that the scientific investigation into those bodies concerns an absent or nonsensical subject matter. Even were such a figure to rise to intellectual prominence, this would be purely a setback for the scientific pursuits of the astronomical community. Contending with the "celestial sceptic's" arguments would be important simply as a matter of putting the hurdle behind us that he would place on the road of scientific progress. His scepticism would be a hindrance to the *real* work in this area, and a successful rebuttal would be valuable principally as a way of excising this hindrance.

But the philosophical scepticisms just mentioned are different. Take, for example, the sceptic about knowledge. Any survey of highlights in the history of epistemology will involve a heavy showing on the sceptic's part. Not, only or primarily, as a *contributor* to the debates, but as a *target* of anti-sceptical responses. And these moments are significant, not simply as victories over a dialectical adversary to have cleared the way for fruitful subsequent inquiry now unencumbered by that adversary's nagging doubts, but as fertile sources of philosophical reflection and research themselves. (In many cases, the sceptic is not even present as a contributor, but as a purely imaginary interlocutor. Hardly to be expected if he is merely a shackle to be thrown off by refutation!) You may not even agree, for example, that Descartes' *Meditations*, the anti-sceptical portions of Kant's first *Critique*, or G. E. Moore's *Refutation* successfully rose to the challenge posed by the sceptic, but you will probably agree that (at least some of) these works are worthy of careful philosophical study. The same could be said of the responses to the cynical amoralist whose doubts about morally binding norms are answered in Plato's *Gorgias* and Korsgaard's *Sources of Normativity*, or the Quinean extensionalist who is the target of Lewis' and Kripke's rich early work in the logic of modality and quantification. In each case, the critic of platitudinous common sense about the subject matter in question "has done the adherents of the platitude a service: he has made them think twice" [Lewis, 1969].

The present dissertation is an attempt to bring this general principle of method to bear on a concept of wide-reaching philosophical significance: *control*. When I set my alarm, for example, I control (or at least seem to control) when I will wake up; my downstairs neighbour, who sets out the annual Halloween candy bowl, controls how many chocolate bars will feature in it; the mugger holding a gun to my

head in an abandoned alley controls whether I live or die. Control, so understood, is certainly an important philosophical concept, for there are many major philosophical questions that hinge on its nature and extent. Are we, for example, only morally responsible for those things we control? Can an agent acting in the present control facts about the past? How do we as mental beings control the movements of physical objects, like our bodies? Often these questions are taken to primarily concern other concepts involved in their formulation (moral responsibility, mental states, *vel sim*.) or related but distinct concepts (causation, intentional action, *vel sim*.). But they are just as much questions about control itself, whose answers will depend at least in part on who controls what and what it means for them to control it.

That establishes the first criterion from the opening pronouncement: that control is a concept of central philosophical interest. What of the second criterion: that it is subject to wide-reaching doubts about its prevalence, existence, or meaningfulness? There are, I believe, quite prominent lines of philosophical argument whose conclusion is that we control little or nothing of what we ordinarily take ourselves to. These arguments have often gone unrecognised as such, instead being treated primarily as arguments against the prevalence of related properties, like *moral responsibility*. Here is one:[1]

1. In his actual circumstances w, Joshua (who has a minor, rarely manifesting neuromuscular disorder) deliberately shoots Klaus in the head and kills him.

2. In the counterfactual circumstances w' that would have obtained had his hand frozen up, however, Joshua would not have succeeded in shooting Klaus, who would have survived.

3. In w and w', Joshua controls the same facts.

4. If an agent controls a fact in certain circumstances, the fact obtains in those circumstances.

5. Therefore, Joshua does not control in w the fact that Klaus dies.

6. An agent (in given circumstances) is only morally responsible for those facts the agent controls (in those circumstances).

[1]This a variant on the problem of "resultant moral luck" discussed at length in Chapter 4.

7. Therefore, Joshua is not morally responsible (in w) for the fact that Klaus dies.

I will have (much) more to say about the merits of this argument (and others like it) in Chapter 4, but for now let us consider it in the abstract, without addressing the plausibility of its specific conclusions. The argument, clearly, is a sceptical argument about moral responsibility, whose conclusion (that a man who deliberately shoots and kills another is not morally responsible for the victim's death) conflicts starkly with our pre-theoretical judgements of moral blame. And there may well be philosophical problems specifically around premise (6), which brings the concept of moral responsibility into the mix. But, before moral responsibility is even on the table, the argument raises similar sceptical worries purely about control. Step (5), that Joshua does not control whether Klaus dies, conflicts every bit as much with our untutored judgements as (7), albeit about control rather than guilt. So the argument is perhaps as potentially fruitful a foil for the defender of common sense about control as it is for the defender of common sense about responsibility. That is precisely the spirit in which I will ultimately be taking the above argument up.

So, control is a concept of considerable philosophical importance, which is subject to sceptical worries in extant philosophical discussion, and thus is a plausible candidate for improved understanding by way of addressing these sceptical worries. So far, so clear. Before embarking on the attempt to so improve our understanding, however, there are some general questions about the structure and aims of that attempt to get clear on. *First*, I would like to say a bit more about the operative concept of control, to head off some misunderstandings. *Second*, I would like to say something about the *kind* of understanding for which I will be trying, and to contrast it with some other possible approaches. *Third*, I would like to preview, in broad outline, the content of the analysis of control that will be reached by the end of the dissertation. I will tackle these three in turn.

I have already briefly tried to gesture towards the concept of control at work, by providing examples and connecting it to some familiar questions. But these do not fully specify the concept, in ways that might be misleading. It has often been suggested, for example, and I have no intention to contest, that the expression "control" might come

in several different senses, or that control itself might come in several importantly different kinds.[2] Michael Zimmerman has done perhaps more than any other philosopher to so taxonomise control, though related classification schemes are common throughout the contemporary literature on compatibilism in the debate over free will [Fischer and Ravizza, 1998] [Zimmerman, 2006] [Zimmerman, 2022]. So the reader may reasonably complain that the task I have set for myself is problematically vague or underspecified. We should be asking, not simply what control is *simpliciter*, and how it is touched by sceptical arguments, but what this or that *kind* of control is, and how these various kinds of control are affected (or not) by the sceptical arguments in question.

I think these distinctions matter less than they may first appear. As I explain towards the start of the next chapter, most of the assumptions made, arguments given, and conclusions reached will be sufficiently general that I think they should hold of more or less *any* kind of control (or under any reading of "control") that might be thought to be of interest in the associated debates. I try, at various points, to make good on this promise, indicating ways in which arguments that might seem to rest on reading "control" very liberally or considering particularly expansive kinds of control in fact can be reformulated in more conservative terms. Unless you think that the use of the expression "control" for all these different things is completely accidental, or that the various kinds of control at issue are united pretty much only by name, it is only to be expected that such general observations should be available about control as such, just as similar alleged polysemy and internal diversity in "knowledge" and knowledge is not generally taken to mean that no interesting claims can be made about knowledge or in terms of "knowledge" without further qualification. Getting clear on the features that unite the various modes of control that might interest philosophers seems as worthy and important a project as distinguishing and sorting them.

There is *one* alleged kind of control, however, or alleged use of the expression "control", which it *is* worth clarifying ahead of time is not of primary concern to us. I have in mind the kind of control that is often taken to go along with *freedom*, or *free will*, in a robust sense. This

[2] At least, no intention to contest for the most part. See §4.5.1 for an argument that one particular distinction sometimes drawn within control (between "direct" and "indirect" control) is illusory as it is usually understood.

is the kind of freedom many have worried is undermined by causal determinism about human action (even if ultimately deciding it is not), or by fatalist conclusions from the logic of past and future. On this reading, an agent controls some action of theirs only if that action is undertaken *freely*, in the demanding sense of "freely" that might be nullified by determinism or fatalism.

This is, clearly, a philosophically important kind of control, if it exists. It is not, however, the only such: there are ways of controlling the world not touched by these questions about determinism and freedom, which ways are still of philosophical interest. Investigation into the nature of these weaker grades of control might still, of course, shed light on the problems of free will and determinism. And there may be conclusions to be drawn generically about both these weaker grades of control and freedom-control, as species of a common genus. But it is possible also to draw conclusions narrowly about these weaker grades without weighing in on the thorny question of free will–as we will in fact do at various points in what is to come.[3]

Drawing this distinction matters, because it is an *extremely* vexed and *extremely* unsettled question whether *anyone* ever controls *anything* in the sense (if any) requiring deep freedom. It is thus very desirable to sidestep these fraught and ancient questions where possible, and to focus on kinds of control where we can be more confident the concept sometimes has actual application. At several crucial junctures, for example, we will consider cases in which we judge an agent like ourselves to control some things but not others. These judgements would be premature if the sense of control involved was freedom-control, for then we would need a resolution to one of the most intractable problems in all of philosophy before comfortably passing those judgements. So at least this limited restriction of the kinds of control being considered, in at least these specific portions of the argument, will be wise.

Are there forms of control that do not require freedom in this august and demanding sense? Quite clearly. Consider the following case.

> Disc Jockey: The disc jockey at the local radio station is assigned the task of selecting which single each morning

[3] If there is no such free will-involving species of control, of course, none of this bears on my arguments in the first place.

> to play first, as a wake-up record. Every morning, after arriving at work, he goes to the shelf housing all the records, picks a single, and broadcasts it from the station's booth.

The disc jockey in this simple story, obviously, controls which of the songs available from the shelf plays on the radio station first thing every morning. Any reading of the story that contradicted this would be perverse.

Reflecting on the disc jockey's place in a thoroughly pre-determined physical universe does nothing to obviate this fact. Such reflection may give us pause about passing judgement on the jockey's choice of record, or attributing a deep sense of ownership to him over his actions, but it does not make less perverse a reading on which he does not control which record gets played. Nor does this confidence depend upon the hope that a "compatibilist" reconciliation might be discovered between his freedom as a moral agent and his pre-determination as a component in a physical system: imagine that it has been shown conclusively that all such reconciliations are doomed, that physical determinism undermines his freedom every bit as much as it first threatens to, and you will still (I submit) not change your mind and decide the disc jockey does not control the first song of the morning.[4] To suggest as much is simply to misunderstand the role he plays at the station, whatever role he might also play in the universe at large.[5]

Here is another way of getting at the point. Sometimes, we talk about mere inanimate mechanisms or processes exercising control over their environments, just as we sometimes talk of them possessing knowledge of those environments. To take a philosopher's favourite example, a thermostat generally "knows" the approximate temperature of a room as part of its "controlling" the temperature in that room. And yet, *nobody* thinks a thermostat is "free" in the relevant sense; to the contrary, it is a paradigmatic instance of *un*freedom, of a mere fated mechanism. It may be, of course, that this talk of control in the thermostat's case is just metaphorical (though this strikes me as less plausible than in the case of knowledge talk). But still, this metaphor does not involve taking up some fanciful perspective on

[4] Much the same points could be made about *fatalistic* concerns about the logical fixity of the past, present, and/or future (see [Taylor, 1962] [Prior, 1967] [Fischer and Todd, 2015]).

[5] This argument is an expansion on similar points made in [Zimmerman, 2022].

which the thermostat is an autonomous moral agent with free will. The perfectly plain thralldom of the device to its surroundings and construction does not impugn its control one iota; thus, this control is of a sort that does not necessitate freedom in a robust sense, meaning that at least one such sort exists.

Exist it might, but are these less weighty forms of control still of any philosophical interest? Yes. The sceptical argument considered above purports to show that agents like ourselves do not control many of the things we take ourselves to, pre-reflectively, and by extension that they lack the moral responsibility we would expect of them. And yet, nothing in the argument hinged on the considerations you would expect were this the sort of control potentially undermined by determinism or fatalism. Nor, indeed, do prevalent variants on that argument concerning the same theme (in the literature called arguments about "resultant moral luck") rely, or take themselves to rely, on such a hefty notion of control-as-freedom. If the argument refutes the existence of much control in ordinary life, it is weaker form of control than that involving or identical with freedom in the above sense.

Our conclusions may have *some* bearing on questions about freedom-control. For, as a (putative) species of control, claims about control *generally* will also be claims about it, too. But we are here only concerned about it, if it exists, as one more species of the genus. Anything conceptually or metaphysically distinctive to it among species of control is irrelevant.

The restriction of attention away from freedom-control goes some way to explaining the relative absence in this dissertation of some of the most celebrated contemporary work in the philosophy of control, namely work on compatibilism in the tradition of Fischer and Ravizza, among whose predecessors I think we should count Frankfurt in his famous intervention on alternative possibilities [Fischer and Ravizza, 1998] [Frankfurt, 1969]. This absence is owing to a difference in subject matters: Fischer and Ravizza, and others in that philosophical tradition, are concerned ultimately to vindicate (or, in the opposite direction, show to be hopeless) a notion of freedom and moral responsibility available to agents living under determinism. But, since the sorts of control mainly under discussion in this dissertation will *not* be borne on this directly by the ultimate viability of compatibilism about free will, projects this centrally devoted to the question of that viability are not of direct concern.

This suffices for my first concern about the project. Now on to the second, about the kind of analysis of control here on offer (*not* the analysans, or the content of that analysis, which I will discuss next). There are at least two sorts of analysis of a concept which, while perhaps interesting, are not what will concern us in the ensuing chapters. First, there is one we have already seen: the method of distinctions, or taxonomy. I have explained why I am avoiding this style of investigation already, and so will not belabour the point. Second, there is the reductive project of "conceptual analysis" (or perhaps "metaphysical analysis", *vel sim.*), whose aim (at least traditionally) is an illuminating biconditional, whose left hand component reads something like "*S* controls whether *p*" and whose right hand component specifies some condition in terms other than those of control. In short, it is an attempt to *define* control in more basic terms (where the kind of basicness in play is left unspecified). Such biconditionals will feature below just as little as the kinds of taxonomy discussed above.[6]

There are a few reasons I avoid this style of philosophising here. Firstly, I am unimpressed with the track record, both in terms of success and in terms of philosophical depth, of these sorts of project in philosophy. This is so both for old school conceptual analysis and various other structurally similar styles of reductionism. Secondly, I suspect that, if any concept should be playing a basic role here, it will be control itself more than any others in terms of which it is likely to be explained.[7] But beyond either of these, such analyses are ill-suited to the task I expect of my eventual analysis of the concept. The point of the analysis, as I have explained here and in the preface, is to resolve certain sceptical puzzles that threaten our common sense understanding of the breadth of our control. What is important in combatting the sceptic here is some understanding of the general principles governing the interface of control and other relevant concepts (such as modality and the logical concepts of conjunction and negation). But, while such principles *can* be derived from a reductive account of control in other terms, the plausibility of these principles can also be assessed on its own, without a detour through (potentially quite dubious) conceptual reductions. That is the path I take in this dissertation.

[6]This lack of interest in reductive accounts of control marks another difference between myself and philosophers in the tradition of Fischer and Ravizza.

[7]This echoes a longstanding complaint of Williamson's about treatment of knowledge in the pre- and post-Gettier tradition [Williamson, 2000].

This gives some indication of what shape the analysis *will not* take; what of the shape that it *will*? It comes in two parts: one comparatively formal, the other comparatively "philosophical" (in the sense to be contrasted with "technical"). The former is a modal logic of control, interpreted as a propositional operator, and later on a bimodal logic of control and a modality of "practical possibility". This is a logic both in the proof theoretic sense of a class of sentences in a philosophically glossed object language, and in the model theoretic sense of a class of models defined by certain properties. As a framework for tackling the sceptic, this logic will allow us to formulate and assess certain general principles that will, I argue, be necessary for the sceptic's arguments to have any purchase. It provides, in other words, a more regimented context in which we might hope to pin the sceptic down.

The latter part is a *picture* of control to accompany this formal framework, which both emerges from arguments made possible by that framework and gives it philosophical significance. The picture is not at all promising as part of a reductionist programme: the way it describes control is not one that would help either to introduce the concept to one unfamiliar with it or to add it to a theoretical framework ultimately formulated in more "fundamental" terms. Instead, it gives a sense of the place of control in the interface between human agents and the world on which they act.

This characterisation of the eventual shape of the picture is a natural point of transition to its rough content. (The content of the logic eventually adopted is fully given by short enumerations of its axioms, inference rules, and model-defining properties in the technical appendices and so need not be repeated here.) The account is deliberately conceived, in many ways, as an analogue in the theory of control to McDowell's disjunctive account of perception in the theory of knowledge. An agent's control is a *flexible* capacity, which will extend and retract to different degrees depending on circumstances. Thus, just how many facts, and facts about which things, an agent controls will vary from possibility to possibility. It is not that there is some uniform subject matter controlled by an agent from possibility to possibility (say, its inner volitional state), from which different consequences may or may not follow. Rather, different subject matters (sometimes greater, sometimes lesser) will be subject to the agent's control in different possibilities; thus, for example, Joshua (from the

sceptical argument earlier) might control the state of Klaus in w, but not control it in w'. If this sounds platitudinous, that is because it is an attempt to restore platitudes in the area in the face of radical doubts.

Still, there are things to be said about the position to highlight its peculiarity. For one, it rejects the existence of *any* interesting domain of control by an agent not liable to this flexibility. A version of the view on which human agents will not vary between circumstances in the scope of their control over, say, basic bodily movements or inner volitions, is a version unsuited to the work that will be asked of it. There is, in this sense, no "interior castle" of actions not subject to the vicissitudes of fate. Philosophers' desire to insulate the objects of an agent's real, "direct" control from such vicissitudes sometimes leads them to posit a privileged area of control (one's bodily movements, one's inner volitions) not (or at least, less) so subject. The expanse of an ordinary human's control thus recedes from the sorts of "external" states and events with which they are ordinarily concerned in everyday life (preparing food, writing documents, etc.) in favour of more rarefied, "internal" objects of control which (it is suggested) cannot go awry in the same way as cooking a meal or writing a paper. Control over the latter somehow is supposed to take theoretical priority ahead of control over the former. Part of the point of my argument is to help secure these more familiar acts of control in their proper, central place in our picture of human agency.[8]

One question about the project, with all this in view, concerns the frequent appeal to common sense judgements about control in the course of the argument. Why so much reliance on the verdicts of common sense, and so little theorising in a broader sense? Why is there not more effort to integrate control into a larger philosophical picture of agency, or the functioning of biological and mechanical systems as they interact with their environment, or what have you?

Again, the anti-sceptical methodology explains this choice. The sceptical arguments rely less on theory than on the intuitive appeal of certain principles in the logic of control. The method employed here, of concentrating on which such principles fit with common sense

[8]This is the point of my use of the Heidegger quotation in the epigraph. That use is arguably somewhat disingenuous, since in that passage Heidegger clearly has more in mind certain forms of cognition than control as such. Still, the cognition he is discussing is (in his view) primarily *practical* and bound up with the manipulation of an agent's environment, so I think it at least somewhat apt.

verdicts about control, addresses these assumptions on the same level as that on which they were raised. Perhaps the ultimate account to be given of control will address in more detail the place these common sense verdicts have in a larger scheme, but in the initial fray with the sceptic it will be helpful to show that these verdicts need not give way immediately to paradox in the way threatened.

Now I will give a chapter by chapter summary of the remainder of the dissertation.

Chapter 2 introduces the logic of control, both axiomatically and semantically. While the *motivations* are not from semantic elegance, the semantics proposed has a certain elegance and ease of presentation to it. This chapter additionally considers, without endorsing, certain iteration principles one might add to the logic.

Chapter 3 moves from the logic of control proper to a discussion of *actions* in a more restricted sense. The discussion takes place in the framework afforded by the aforementioned semantics, enriched now by an additional modality of "practical possibility" (encompassing those possibilities the agent is "in a position" to be in). This turns out to provide a good model for talking about certain concepts in the foundations of decision theory, and our assessment of these concepts will provide a "dress rehearsal" for the moral philosophical considerations in the next chapter. Of particular interest is a widely accepted principle I call "partitionality", saying that the actions available to an agent are always pairwise incompossible, which I will argue we should reject.

Chapter 4 addresses, finally, an argument along the lines of the sceptical argument from earlier in this chapter. I argue that this argument implicitly rests on a principle regarding the logic of control much like partitionality that should be rejected for similar reasons. I sketch a picture of control to serve as an alternative to that motivating the sceptical argument, as well.

Chapter 5 concludes by examining a couple of recurring philosophical threads in the main body of the work.

Chapter 2

The Logic of Control

2.1 Introduction

This chapter proposes and motivates a logic of the concepts of *control* and *making so*; that is to say, a general theory of their interaction with the familiar operators of classical logic.[1] The senses in which I intend these expressions are familiar from everyday use: it is the sense in which, by placing a pot of water over a lit stove, one (seemingly, under ordinary circumstances) makes it so that the water boils, and controls *whether* it boils. This is not, of course, a new topic in philosophical logic, but the avenues pursued in our discussion will differ in measured but important respects from previous approaches.

The first section clarifies the target concept and scope of the investigation, as well as motivating an interest in it. The substantive section of the paper begins with an investigation of the basic principles of our operator; while it shares a number of motivations with extant treatments, the final theory is neither strictly stronger nor strictly weaker than its prevailing rivals, and the discussion in the previous section illuminates otherwise obscure aspects of old debates. The resulting logic, as proved in the appendix, has an appealing semantics for which it is sound and complete, even though the motivations given in the

[1] The choice of classical (over, say, intuitionistic) logic is a reflection of a background commitment to classicality in general. How far the results of this dissertation can be adapted by those who reject the underlying principles of classical logic is left as an open question.

main body of the paper are entirely non-semantic in nature, and there are natural extensions of the logic we may consider corresponding to natural conditions on the frames of the semantics.

2.2 Motivations

2.2.1 Conceptual Preliminaries

Our point of departure is a pair of equivalences:

1. Necessarily, one controls whether p iff one either makes it so that p or makes it so that not p

2. Necessarily, one makes it so that p iff p and one controls whether p

Given these principles, claims about what S controls are equivalent to certain claims about what S makes true, and *vice versa*. The principles thus allow us to pass freely from talk about control to talk about making true and back again, as we will in fact frequently do in what follows. Note that, as long as making it so that p entails p and controlling p is necessarily equivalent to controlling not p, either principle implies the other.[2]

There are two different ways to understand these principles: firstly, as stating certain substantive philosophical theses; secondly, as expressing certain definitions or abbreviations. On the first reading, the principles are to be read literally, as claims about entailments from facts about the scope of an agent's control/making-true to facts about the agent's making-true/control, as they are understood in natural language discourse. On the second reading, which this chapter and

[2]Suppose (1). Suppose S makes it so that p. Then, by factivity, p, and by disjunction introduction, S either makes it so that p or makes it so that not p. Suppose p and that S either makes it so that p or makes it so that not p; by factivity and disjunctive syllogism, S makes it so that p. So (2) follows.

Suppose (2). Suppose S controls whether p. Then either p or not p, by excluded middle. If p, then p and one controls whether p. If not p, then not p and one controls whether (not) p. Suppose either p and S controls whether p or not p and S controls whether not p. Either way, S controls whether p (since, necessarily, one controls whether p iff one controls whether not p). So (1) follows.

the following will in fact adopt, they are loose expressions in the material mode of what are properly stipulations in the formal mode. For example, the first principle (on this reading) indicates that statements of the form "S controls whether p" are to be understood as abbreviating sentences of the form "S either makes it so that p or makes it so that not p", just as (in standard presentations of first order logic) a sentence $\ulcorner \exists x\ \varphi x \urcorner$ abbreviates the sentence $\ulcorner \neg \forall x\ \neg \varphi x \urcorner$ (or *vice versa*); of course, whichever way the abbreviation goes (say, where we define $\ulcorner \exists x\ \varphi x \urcorner := \ulcorner \neg \forall\ \neg \varphi x \urcorner$), the statement of the other (say, $\ulcorner \forall x\ \varphi x \leftrightarrow \neg \exists\ \neg \varphi x \urcorner$) follows trivially. As with the case of quantifiers in first order logic, we are left with a choice of which way of talking (in terms of control or in terms of making true) to take as primitive, using the appropriate principle above to define the other way of talking (and letting the other principle fall out as an immediate consequence). Whichever of "control" or "make true" is left as non-primitive just so happens, on this reading, to be homophonic with a familiar English expression: its use in what follows is to be understood simply as an abbreviating shorthand for some more complex phrase.

To make matters more complicated still, there are two languages for control or making true talk with which we will be concerned. First is the informal philosophical metalanguage in which we will carry out our philosophical inquiry, second is the formal object language in which we will axiomatise our logic of control. It is possible to take different primitives in each case, and this is what we will in fact do. In the formal language, we will take the lexical item corresponding to making true (the box $\Box$) as primitive. This choice greatly disencumbers the theory: it will make our formulations less tedious in their syntax, more easily visualised in their semantics, and readily comparable in both respects to extant competitors. In the informal discussion, we will take control-talk as primitive. Interrogative uses of *control* are simply much more common in everyday speech than propositional uses of *make*, and the concept of control is more familiar within the larger philosophical tradition. While we will slide between control talk and making-true talk in what follows, it should be remembered that the latter is officially just philosophical shorthand for the former.

What of the substantive readings of the principles? I am inclined to endorse them, too. But the best argument for such a pair of philosophical theses is to show that, so understood, control and making true hang together in a theoretically satisfying way. That will, in a

sense, be the work of this entire dissertation.

In the ensuing discussion, therefore, we will treat the concept *S makes it so* as a property of propositions denoted semi-formally by a sentential operator $\Box$ (with sentences containing it to be ultimately understood, in accord with (1) and (2), as abbreviations of sentences about control). We will also carelessly conflate constructions like *S makes it so that X is* φ, *S makes it true that X is* φ, and *S makes X* φ; these are all to be analysed, again, in the natural way in terms of control. Thus, where p stands for "the water is boiling", $\Box p$ will stand for the sentence "*X* makes it so that the water is boiling" or, fundamentally, "The water is boiling and *X* controls whether it is".

This framing is somewhat committal. It rejects a view of agency in which *actions* as entities in their own right feature centrally for one treating it as a relation between agents and propositions, against certain prominent grains in action theory [Davidson, 1967] [Horty and Pacuit, 2017]. This is, however, not intended to convey any fundamentality or ultimate perspicuity of the agency-as-operator perspective. It is consistent with the arguments of this paper that this general framing be provisional, or be explained in other terms. What interests me is not the operator perspective itself, but what principles we should accept when taking it.

As a final point, nothing in this investigation hinges (or at least, is intended to hinge) on substantive assumptions about the scope of one's control, or the precise relevant sense of *making true* (and *controls*). It may be, for example, that one makes true such "worldly" propositions as that the water in the pot boils; or it may be the only things an agent ever makes true are "internal" facts about their will or mental state; or it may be that "controls" and "makes true" admit of stricter and weaker readings, on some of which an agent can be truthfully said to make true worldly propositions and on some of which she cannot.[3] While I will in later chapters go on to suggest firmer positions on some of these questions, there are general preliminary problems about the logic of these expressions independent of such ambiguities and substantive background views. It is these more general questions I intend to investigate here.

[3]These are analogous to similar positions taken and distinctions drawn in debates in epistemology between externalists and internalists.

2.2.2 Further Significance

These coordinate concepts of control and making so are of wide-ranging philosophical significance. They are closely connected, for one, to the understanding of an *action* present in both all versions of causal decision theory and Jeffrey-style evidential decision theory [Jeffrey, 1965] [Joyce, 1999] [Lewis, 1981]. Such versions of decision theory take the objects they recommend to a rational agent (given her preferences and prior doxastic state) to be *propositions* she is in a position to make true (as opposed to *lotteries* one is in a position to gamble on, as in [von Neumann and Morgenstern, 1947]). This is commonly taken to be an advantage of such theories: it renders these objects of recommendation more familiar, and unifies them with the objects of credence and desire. It does, however, raise the question of what is involved in making a proposition true, a question with both more substantive aspects in the theory of action proper and more abstract aspects, such as those discussed in this paper.

In less heavily technical areas, the concepts bear on debates in the study of free will between compatibilists and incompatibilists, and on debates over the scope of our moral responsibilities under the heading of "moral luck" [Lewis, 1981] [Inwagen, 1983] [Nagel, 1979] [Williams and Nagel, 1976] [Pritchard, 2006] [Hartman, 2019]. Debates in the former case centre precisely on the relation of making true to nomological necessity, and debates in the latter case have centered around the contentious principle that one is morally responsible for a fact p only if one controls (or has control over) whether p, which if true will mean that the logic of making true straightforwardly limits the logic of moral responsibility.[4] Getting clear on the logic of $\Box$, therefore, stands to illuminate a number of disparate areas.

2.3 Logical Principles

2.3.1 Normality

For our operator $\Box$ to be a *normal* one, the following must be true of it (taking some harmless liberties with use and mention):

[4] Though, as noted in Chapter 1, there will be limits to the bearing of our results on these questions.

Distributivity It must distribute across the material conditional, so that if one makes it so that, if p, q, then if one makes it so that p, one makes it so that q. $(\Box(p \to q) \to (\Box p \to \Box q))$

Necessitation If one can prove in classical logic, together with the aforementioned principle and this additional current rule of inference, that p, then one makes it the case that p (If $\vdash p$, $\vdash \Box p$)

Not only does common sense about what we can make so contradict these principles, they each fall afoul of it separately. To begin, consider that, in the presence of the **Necessitation**, **Distributivity** is equivalent to the following (sometimes called, respectively, M and C):

Weakening If one makes it so that both p and q, one makes it so that p and one makes it so that q

Agglomeration If one makes it so that p and one makes it so that q, one makes it so that both p and q

The first of these is open to obvious counterexamples. Suppose that I am handling a ball tethered to a pole by a robust 2m chain, which I freely place within 1m of the pole. Then I seem to have made true that (controlled whether) the ball is within 1m of the pole, but not to have controlled whether it is within 2m of the pole, contradicting **Weakening**. Note that this argument does not appeal to any overt, linguistic conjunction, disjunction, or quantification, thus sidestepping objections (such as those raised in [Chellas, 1992] to similar arguments) that it trades on syntactically introduced ambiguities and presuppositions.

Necessitation is even more absurd. One line of argument would appeal to the common intuition that ordinary agents do not make true any facts of logic (see [Kenny, 1976] [Belnap and Perloff, 1988]), in a possible worlds setting understood as the trivial proposition $\top$. But this thinking is vulnerable to methodological objections: it is a mainstay on formal theories of propositional attitudes that attitudes towards the trivial proposition should be seen as themselves degenerate and trivial instances of those attitudes, and our disinclination to report them explicable on merely pragmatic grounds [Stalnaker, 1984] [Lewis, 1986] [Stalnaker, 2006]. The defender of **Necessitation** is thus free to extend similar reasoning to making true (see again [Chellas, 1992]).

A better route appeals instead to intuitions about our control over *necessities* of varying strictness. There are some very restricted and weak species of necessity over which I exert control: I cannot (by reason of the ill-suited shoes I happen to be wearing) run a six minute mile, and yet I plausibly control whether I run a six minute mile, since the necessity involved is of a very weak kind. As the species of necessity increase in strictness, my intuitions of non-control grow increasingly stronger: I cannot (by reason of my overall build) run a four minute mile, cannot (by reason of human physiology) run a two minute mile, cannot (by reason of physical law) run a one nanosecond mile, and cannot (by conceptual necessity) run a zero second mile. In each case, I am more certain I do not control the relevant impossibility. Since controlling a proposition is controlling its negation, however, and since the negation of the last element in the sequence is the trivial proposition $\top$, this means I do not control $\top$. Failure to control $\top$ is thus a terminus in a progressively stronger series of intuitions, not an outlier treatable as an aberrant or degenerate case.

This is, obviously, a variant on the above argument against **Weakening**. It appeals directly to intuitions about what we do and do not control, rather than (as in [Horty and Belnap, 1995]) a version of the principle of alternative possibilities, or any other contentious general theses about control [Frankfurt, 1969] [Fischer and Ravizza, 1998] [Vihvelin, 2013]. Nor does it tie failure of control to any specific privileged species of necessity. These belong to a broader range of entanglements with modality our theory will seek to avoid.

2.3.2 A Basic Logic

While our target concept plainly cannot be analysed as a normal operator, it is illuminating to observe respects in which it does *not* violate normality. Unlike with many representational attitudes, for one, whenever p is logically equivalent to q it seems that one makes it so that p iff one makes it so that q: what one has made happen seems, as it were, insensitive to any differences of logically extraneous guise. Indeed, this seems true not just of logical equivalences, but equivalence under various narrower species of necessity. It seems to be a law of nature, for example, that a body has inertial mass of 5kg iff it has gravitational mass of 5kg, even though we can perhaps imagine the laws of nature requiring otherwise. And, for this reason, it seems that

if I make it so that a certain object has gravitational mass of 5kg (say, by adding material to it until it weighs 5kg on a scale), I have thereby made it so that it has 5kg inertial mass, and *vice versa*.

This suggests that the logic of action is, if not normal, at least *congruent* or *regular*, i.e. it is closed under the following rule of inference:

RE If p and q are provably equivalent in our logic, that one makes it so that p and that one makes it so that q are provably equivalent (If $\vdash p \leftrightarrow q$, $\vdash \Box p \leftrightarrow \Box q$)

Of course, **RE** requires not only that propositions equivalent modulo classical logic are made true when the other is, but that this holds of propositions equivalent modulo the *whole logic being proposed* (including the inference rule **RE**). Nevertheless, I think the theory's other principles are themselves suitably necessary to permit equivalence of what is made true holding them fixed. This assumption is ubiquitous among competitors for the reasons we have seen.

As an obvious first entry in our list of axioms (indeed, one that follows from our definition (2) of *makes that* in terms of control), making true is factive; what one makes true *is* true:

Factivity If one makes it so that p, p ($\Box p \rightarrow p$)

Next, while **Weakening** on inspection turned out to be implausible, its sister principle **Agglomeration** seems perfectly good: if I make my bread into a circle and make it pink, I thereby make it into a pink circle. We will therefore add the axiom to our logic.

On the subject of **Weakening**, with **RE** in place there is an additional difficulty posed by the principle. Together with **RE**, it implies that, when one makes it so that p, for any arbitrary q one makes it so that either p or q; that is, when one has made any given disjunct of a given disjunction true, one has made true the disjunction itself. Experience and imagination are replete with counterexamples: for example, when making it so that a chained ball is on the left side of the room (by pushing it, say), I need not have made it so that the chained ball is somewhere or other in the room; the chain strips me of any input on that matter. In other cases, the additional disjunct may even be true, but still unrelated to my doings. Even though the sun is shining as I turn on an overhead light, for example, it does not seem true that I

make it so that at least one of the sun or the overhead light shines, for that one or the other would shine is guaranteed independently of me.

In the first sort of counterexample, the agent determines one or another of some (not necessarily mutually exclusive) outcomes, where the range of options itself is fixed independent of the agent. In the second, among a range of such options, one of them is beyond the control of the agent, even if it still obtains. Such counterexamples cannot arise, however, when *both* disjuncts have been made true by an agent; in such a scenario, her control over the disjunction seems complete. I therefore propose the following axiom for our logic:

Modest Weakening If one makes it so that p and one makes it so that q, one makes it so that at least one of p or q is true $((\Box p \wedge \Box q) \rightarrow \Box(p \vee q))$

There is another broad class of cases where weakening seems licit. Suppose that I of my own volition choose to go walk, and moreover specifically to walk quickly to the store. Then it seems, not only that I must have controlled whether I walked or walked quickly to the store, but that I must have controlled whether I walked quickly *simpliciter*. While there are circumstances where it might seem reasonable to say I controlled whether I walked quickly to the store but deny that I controlled whether I walked quickly, the additional supposition that I controlled whether I walked at all seems to undermine this cotenability. This case is indicative of a broader trend: when I make true both p and some weaker proposition besides, I must also make true any proposition of intermediate strength (in the sense of being entailed by the former and entailing the latter).

The same intuition, in the current setting, can also be drawn out without the use of the modal or semantic concept of entailment, instead relying on the familiar logical concept of disjunction. Imagine Paris tossing the apple of discord to Aphrodite out of Aphrodite, Athena, and Hera. He might not control that the apple goes to one of the two A-goddesses, if he does not control that it goes to Aphrodite in particular; it could be he only controls that it goes to one or another of the three, without controlling any more specifically which one. And he might not control that it goes generically to an A-goddess, if he does not control that it goes to a goddess in the first place; it could be the apple is also forced beyond his control go to some A-goddess or

another, and he simply controls between the A-goddesses which one gets it. If he controls *both* that it goes to a goddess at all in the first place *and* that it goes in particular to Aphrodite, though, then it seems to come along for free that he controls it goes to an A-goddess.

We can formulate this principle in our logic straightforwardly.

Convexity If one makes it so that one of p, q, or r is true and one makes it so that p is true, one makes it so that one of p or q is true $(\Box(p \vee q \vee r) \rightarrow \Box p \rightarrow \Box(p \vee q))$

The theory thus axiomatised, together with modus ponens and the theorems of classical propositional logic, turns out to have an elegant semantics (which we will call *sandwich semantics*), for which soundness and completeness can be proved. The semantics is a possible worlds one, with sentences interpreted by subsets of a domain of worlds, equipped with a (possibly partial) function $(S_{\mathsf{min}}, S_{\mathsf{max}})$ (for short, S) from the domain to pairs of its subsets, where $w \in S_{\mathsf{min}}(w) \subseteq S_{\mathsf{max}}(w)$. The domain and function together constitute a *sandwich frame*. The classical connectives, as usual, are interpreted by their set theoretic analogues, while our operator $\Box$ is interpreted so that, at any world w, the propositions (if any) of which $\Box$ is true at w are those sandwiched between (inclusively) $S_{\mathsf{min}}(w)$ and $S_{\mathsf{max}}(w)$, i.e. the X such that $S_{\mathsf{min}}(w) \subseteq X \subseteq S_{\mathsf{max}}(w)$.

One may picture a model of the logic as a surface, with not necessarily every point associated with two concentric circles containing it, together with a function associating sentences of our language to regions of the surface. The circles around w (joined with their respective interiors) represent the strongest and weakest things the agent makes true at w, and the things one makes true are those subregions of the larger circle that entirely contain the smaller circle (see Figure 2.1).

Note that none of the principles here proffered *prima facie* require that one ever make anything true. This is confirmed by the semantics, which transparently allow for models with widespread or even universal failures of definedness for $(S_{\mathsf{min}}, S_{\mathsf{max}})$.

We call the logic generated by the axiom schemas and inference rules listed Conv, for "convexity logic"; the logic resulting from expanding that list of axiom schemas by some others we denote by appending the names of those additional schemas to the name Conv. The class of convexity logics is, clearly, a generalisation of the class

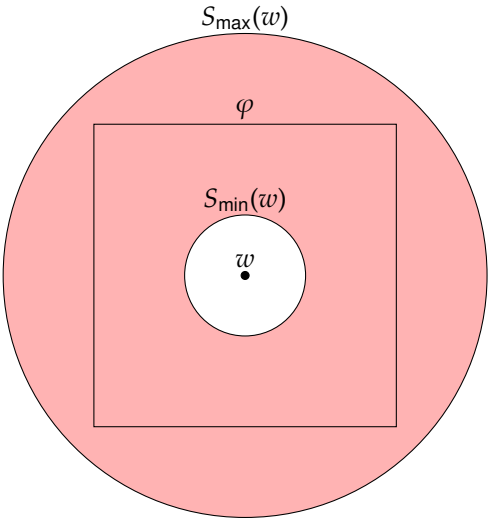

Figure 2.1: An illustration of the inner and outer circles at w with $w \in [\![\Box\varphi]\!]$ and the area "between" the two circles shaded

of factive normal logics, and the class of sandwich frames a generalisation of the class of reflexive Kripke frames: any normal logic is equivalent to a convexity logic with the axiom $\Box\top$, and every reflexive Kripke frame is equivalent to a sandwich frame $\langle W, (S_{\mathsf{min}}, S_{\mathsf{max}})\rangle$ whose S_{max} is the constant function to W. Convexity logics are, as a slogan, just like normal logics with locally restricted domains of worlds.

We may thus think of Conv as standing to sandwich semantics frames as the logic KT stands to reflexive Kripke frames. It serves as a minimal framework for theorising about a factive convex operator, just as KT with reflexive Kripke frames serves as a minimal framework for theorising about a factive normal operator.[5]

[5]Our models are, of course, a special case of neighbourhood or Scott-Montague models, and thus stand in an intermediate position between Kripke/relational and neighbourhood semantics for modal operators [Pacuit, 2017].

2.3.3 Further Axioms

There are two important extensions to our logic which, while we will not adopt them, will become salient later.

Positive Introspection If one makes it so that p, one makes it so that one makes it so that p ($\Box p \rightarrow \Box\Box p$)

Relative Negative Introspection If one makes it so that p but does not make it so that q, one makes it so that both p and one does not make it so that q ($\Box p \wedge \neg\Box q) \rightarrow \Box(p \wedge \neg\Box q$)

The first is, of course, the familiar axiom 4, which in normal logics corresponds to transitive Kripke frames. In the sandwich semantics, it corresponds instead to the condition of S not *narrowing*; that is, for any frame $\langle W, (S_{\mathsf{min}}, S_{\mathsf{max}})\rangle$ on which 4 is valid, and for any $w \in W$, for no $v \in S_{\mathsf{min}}(w)$ does $S_{\mathsf{max}}(v)$ lack some $u \in S_{\mathsf{max}}(w)$ nor does $S_{\mathsf{min}}(v)$ contain some $u \notin S_{\mathsf{min}}(w)$ (see Figure 2.2). In line with the picture given above, this means that, when moving to a point within the smaller of the two concentric circles associated with w, the new circles only differ (if at all) by strictly *increasing* the region between the two circles; either the inner circle contracts, the outer circle expands, or both (or neither). Conv4 is sound, complete, and definable on this class of frames.

Just as 4 prohibits "narrowing," so the latter (which we will call 5^*) prohibits "widening:" when moving within some $S_{\mathsf{min}}(w)$ to another world (with defined circles), the outer circle gains no worlds, and the inner circle loses none; again, we can prove the natural soundness, completeness, and frame definability results. Unlike its analogue 5, however, it does not in the presence of **Factivity** entail 4; the convex logic Conv45* with both axioms, instead, corresponds to those sandwich frames where no movement within the inner circle of w changes either inner or outer circle (that is, prohibits both narrowing and widening).[6]

There is a further axiom scheme one might adopt, a convex analogue to the Ver of normal logics:

[6] Conv5* is equivalent to Conv45*, however, as long as we consider only sandwich models where, for all w with $S(w)$ defined, $S(w')$ is defined for all $w' \in S_{\mathsf{min}}(w)$ (Corollary 2.A.16). That is, informally, 5^* fails to imply 4 only under the perhaps strange circumstance that it might be consistent with all one makes true that one makes nothing true.

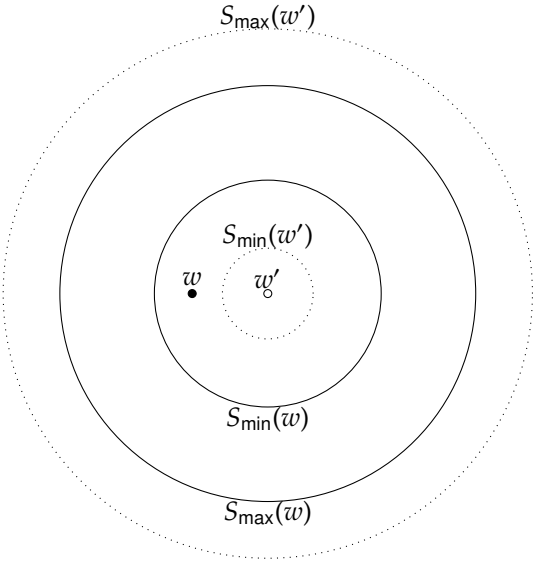

Figure 2.2: An illustration of the "expansion" permitted by **4** of the "sandwich" as moving from w to $w' \in S_{\mathsf{min}}(w)$

True Strengthening If one makes it so that p, and q, one makes it so that p and q $(q \wedge \Box p \rightarrow \Box(p \wedge q))$

Such a principle has some notable precedent. For example, Joyce seems to tacitly invoke it in the following passage from [Joyce, 1999]:

> "The situation in which a gambler takes himself to be gambling on S with prize W and loss L," [Jeffrey] wrote, "is one in which he believes the proposition G: $(S \wedge W) \vee (\neg S \wedge L)$." Since G is logically equivalent to $(S \rightarrow W) \wedge (\neg S \rightarrow L)$, jeffrey1965 can be read as claiming that the performance of an action always involves making a proposition of form $[\bigwedge_S (S \rightarrow O(A, S))]$ true.[7] This cannot be correct. While Jeffrey was quite right to think that an agent who chooses to stake the difference between W and L on the occurrence of S is making G true, he was wrong to imply that this is all she makes true. If S happens to be false, then the agent can make G true merely by making its second conjunct $\neg S \rightarrow L$ true. But if she can do this, then she can make any proposition of the form $(S \rightarrow X) \wedge (\neg S \rightarrow L)$ true.

The thought here seems to be that, if S is false, $S \rightarrow X$ will be vacuously true, and so by **True Strengthening** an agent making $(\neg S \rightarrow L)$ will make the conjunction G true as well.[8] It imposes on models, as is intuitively clear, the constraint that S_{min} always be a singleton where defined, though S_{max} may be as large or small as desired (subject to the constraint that it be true).[9]

To my mind, this principle is open to obvious, ubiquitous, and devastating counterexamples. Under ordinary conditions, for instance, a man flipping a fair coin will control exactly the first of the following, even in the event the coin lands heads:

- The coin will land.

[7] Here S ranges over worldly *states*, A over *actions* available to an agent, and $O(A, S)$ is the *outcome* of the action A given the state S.

[8] It could be that Joyce is instead appealing to some weaker principle, like **True Strengthening** save in that it only applies when q satisfies some further condition here satisfied by $(\neg S \rightarrow L)$. It is unclear, however, what this extra condition might be.

[9] There are sandwich *models* of **True Strengthening** without this property, but the axiom is not valid on the *frames* of these models. This is left as an exercise to the reader.

- The coin will land heads.

Since to land heads is just to land and land heads, this suffices to refute **True Strengthening**. Nevertheless, it is worth asking what might be initially appealing about it. I think the thought is something like the following.

There are two sorts of proposition, relative to an agent: the *acts* open to the agent (and their consequences), and the *state* of the world apart from her. The former are flexible and to-be-settled from the point of view of the agent, whereas the latter are already settled and so, in an important sense, necessary. (Specifically, a sense of necessity strong enough to make it so that a proposition is made true iff its necessary equivalents are.) Since such settled truths p are necessary in this sense, any fact made true is necessarily equivalent to its conjunction with p. So, since control is insensitive to distinctions among worldly states finer than necessary equivalence, control satisfies **True Strengthening**.

This picture holds a considerable grip on formal thinking about action. We have already seen one example in Joyce, where its failure exposed one sort of counterexample: cases where we do not control the *way* in which an action of ours plays out, even as this way is (as such) not a settled state independent of the action. The following section will consider another way of implementing this picture, liable to other kinds of counterexample.

2.3.4 Conv and STIT

The question of the true logic of making so–sometimes called the logic of *action* or of *STIT* (seeing to it that)–is not virgin philosophical territory. Early entries in the investigation of this logic include [Kenny, 1976] [von Wright, 1963], and the debate was raised to its current level of formal semantic sophistication by [Chellas, 1969] (for a comprehensive historical overview see [Segerberg, 1992]). Subsequent formal research has been dominated by the tradition of [Belnap and Perloff, 1988], and it is therefore worth noting where the current discussion agrees with and diverges from the main motivations and commitments of this tradition. We will describe two classes of STIT model: the traditional *branching* models and more minimalist *atemporal* models. While the latter are simpler and more comparable to our own sandwich models, a review of the branching semantics will be helpful in conveying

the philosophical motivations of the STIT project.

A *branching STIT frame* (with one agent) is a structure $\langle T, \leq, \mathrm{Ch}\rangle$, where T is some non-empty set, $\leq$ a partial order on T which branches only "upwards" and which has some common lower bound for any two elements, and Ch is a function from $m \in T$ to partitions on maximal chains (or "histories") running through m. Sentences are interpreted on such a frame not, as one might expect, as subsets of T, but as sets of pairs $\langle m, h\rangle$ where $m \in T$ and h is a history with $m \in h$ (an *historical pair*). A *branching STIT model* is, accordingly, a STIT frame together with a base interpretation function associating each atomic sentence with a set of historical pairs. Sometimes further structure is added to the underlying frame, such as a set of multiple agents or a contemporaneity relation on T, or further restrictions are imposed on the relation between the parameters, but for our purposes this minimal structure will suffice.[10]

The vocabulary of the language interpreted on such a frame minimally contains denumerably many atomic sentences, the classical connectives, and the unary modal operator [*dstit*] (the analogue to our $\Box$). The first two are interpreted as usual; a historical pair $\langle m, h\rangle$ will be in $[\![[dstit]\varphi]\!]$ just in case a) for all $h' \in \mathrm{Ch}(m)[h]$, $\langle m, h'\rangle \in [\![\varphi]\!]$, and b) for some $h'' \ni m$, $\langle m, h''\rangle \notin [\![\varphi]\!]$. This is not the only STIT operator studied, but it is both the best formally understood and the closest to ours in both formal and informal interpretation, and so we will restrict our attention to it.

Other operators sometimes included in the language are the tense operators P, F (for the past and future tenses, respectively) and the "historical necessity" operator $\blacksquare$. We have $\langle m, h\rangle \in [\![\mathsf{P}(\mathsf{F})\varphi]\!]$ iff $\langle m', h\rangle \in [\![\varphi]\!]$ for some $m' < (>)\, m$, and $\langle m, h\rangle \in [\![\blacksquare\varphi]\!]$ iff all $\langle m, h'\rangle \in [\![\varphi]\!]$.

From these technical remarks, much about the underlying philosophical picture should already be apparent. For STIT enthusiasts, time is a garden of forking paths: each moment has only one linear past, but an array of various possible future trajectories, of which only one will be taken. What is settled, or historically necessary, at a given point is what is bound to occur along any path thence is taken. At any given moment, these future trajectories are partitioned into the choices open to our agent, of which she chooses one. What she makes

[10] For extensive technical development of this framework see [Horty and Belnap, 1995] [Nuel Belnap and Xu, 2001] [Xu, 1995] [Xu, 1996] [Xu, 1998].

true are her choice and all of its consequences not already settled for her by her position in, and the overall structure of, time's branching garden path.[11]

While the branching time structure in the models is both traditional and well illustrative of the theory's guiding intuitions, atemporal axiomatisations and semantic structures are both available and formally well-understood [Nuel Belnap and Xu, 2001] [Philippe Balbiani and Troquard, 2008] [Lorini et al., 2011]. In these Kripke-style models, historical pairs are replaced with structureless worlds as points of evaluation, which together form the domain W of a model. An *atemporal STIT frame* $\langle W, \mathrm{Ch}, H\rangle$ consists of a non-empty domain W and two equivalence relations $\mathrm{Ch} \subseteq H$, representing respectively one's choice at a world and historical necessity, while an *atemporal STIT model* is such a frame plus a base valuation from atomic sentences to subsets of W. Atomic sentences and classical connectives are interpreted as usual; w is in $[\![\blacksquare\varphi]\!]$ just when $[\![\varphi]\!] \supseteq H(w)$, w is in $[\![[dstit]\varphi]\!]$ just when $[\![\varphi]\!] \supseteq \mathrm{Ch}(w), \not\supseteq H(w)$. This validates the same logic for these two operators as the old semantics; it is just a branching STIT frame shorn of its branching.

There is a close resemblance between $\mathsf{Conv45}^*$ and the logic of $[dstit]$. The axiomatic theory of the latter includes all our own axioms besides **Modest Weakening**, which plays the least significant role for us. The semantical picture is similar, too: in both cases, the extension of the operator at an index i is uniquely determined by two sets $S_{\mathsf{min}}(i), S_{\mathsf{max}}(i)$ containing it, one a subset of the other, the difference being that in the sandwich semantics its extension at i is all those sets between $S_{\mathsf{min}}(i)$ and $S_{\mathsf{max}}(i)$, in STIT semantics all those greater than $S_{\mathsf{min}}(i)$ but not greater than $S_{\mathsf{max}}(i)$. And while the STIT semantics bakes in the assumption of **Relative Negative Introspection** (and **Positive Introspection**), these can be eliminated fairly easily by replacing the co-domain of $\mathrm{Ch}(m)$ with *reflexive relations* on the histories through m (as is effectively done in [Chellas, 1992]).[12]

The entanglements between $[dstit]$ and historical necessity pose serious problems. The effect of the second, "negative" clause for $[dstit]$ is to invalidate **Weakening** in generality, but to only allow it

[11] Advocates of STIT thus typically accept a certain form of indeterminism, in that they generally deny that all facts about the future are historically settled.

[12] Or, in the atemporal semantics, letting Ch be merely reflexive rather than equivalent.

to fail under specific circumstances: where one sees to it that p, one does not see to it that either p or q only if it is historically necessary that either p or q. Under the analysis of making true or seeing to it as controlled truths, this seems open to obvious counterexamples. Suppose that I deliberately and freely place a die on my table to have six facing upward. In an hour, when I am asleep and unable to affect it, my friend will roll another, chancy die on the same table, which as it happens will land on a six. Then I seem to have controlled whether my first die would have a six face up on the table tonight, but not whether *any* die would have a six facing up on the table tonight (since one was going to land independent of me). But this latter fact, as I am making the first fact true, is unsettled and historically contingent if anything is; Conv (or even Conv45*), by contrast, does not impose this untoward result. The semantic analysis here on offer is begotten of an overly restrictive view of what it is we cannot make true, a view permitting only one possible source (historical settledness) of such failures of control over weak propositions.

2.4 Mere Action

In fn. 6, I mentioned the possibility of restricting attention to models in which, for any w, v with $S_{\mathsf{min}}(w)$ defined, $v \in S_{\mathsf{min}}(w)$, $S_{\mathsf{min}}(v)$ is defined as well. That is, models in which, whenever an agent makes anything true, it is entailed by what they make true *that they make something-or-other true*. One way to achieve this restriction, of course, would be by giving an axiom defining this condition in the object language. What follows is a brief reflection on how such an axiom might be formulated, and its philosophical import.

2.4.1 Undefinability of Mere Action

In order to formulate the axiom, we require some way of expressing the proposition that the agent makes something-or-other true (the *mere action* of the agent). That is, we need a sentence φ true (in any model) in exactly those worlds w on which S is defined, so that the axiom will read: $\Box p \rightarrow \Box(\varphi \wedge p)$. Informally, this says that, if an agent makes something p true, it makes true the conjunction of p and its mere action, which is true (given that it makes p true) iff φ obtains throughout the

actual value of S_{min}. In a convex setting with $S_{\mathsf{max}} = w \mapsto W$ where defined (what we might call a "quasi-normal" setting), the role of φ could of course be played by $\Box\top$. Might the general convex case allow for a similar sentence in the language $\mathcal{L}$?

No, it does not, as a straightforward example demonstrates (see Proposition 2.A.18). This means that we must, one way or another, expand our vocabulary if we are to express the performance of mere action by the agent.

2.4.2 Act and $\mathsf{C}_{\mathsf{Act}}$

One way to expand our language to include talk of whether the agent makes anything true at all would be to simply introduce quantification into sentence position, or "over propositions."[13] This would be in a sense the most flat-footed approach. It is flat-footed in that it takes the informal gloss motivating the expansion of the language most straightforwardly: on this reading, talk internal to the language of an agent making something-or-other true should take the form of an existential generalisation $\exists p \Box p$ of a sentence $\Box\varphi$ asserting that the agent makes some specific proposition true. It would thus be the most literal "translation" of the relevant English into our toy language.

And yet, despite this straightforwardness, it invites an uncomfortably large number of choice-points and technical questions about how to carry out the formalisation of the notion. Introducing a domain of sentence-level entities, for example, would force us to confront questions about how seriously to take the Boolean algebraic structure we have hitherto been innocently assuming on account of the regularity of our logic. It would also open vexed questions about the interplay between our modal operator and the quantifiers that would become involved; for example, should we accept the (converse) Barcan formula for the modal? Such questions would be plagued, not only by the difficulties common to the philosophical interpretation of (higher-order) quantified modal logic in all its forms, but by the non-normality of the modal operators at hand. Thus we would need to, for example, somehow distinguish the *propositional weakening* effect of

[13] This phrasing is a little tendentious, for some philosophers are keen to draw a sharp distinction between propositions proper, which are ranged over by variables in singular term position, and the entities ranged over by variables in sentence position. But for the present matter these concerns can be set aside without loss.

the standard sentential converse Barcan (which in normal settings is innocuous, as weakening is already ensured directly by the K axiom and rule of necessitation) from the *propositional necessitist* effect more commonly of interest in philosophical study of the schema and its converse [Williamson, 2013]. It would also require a recasting of our metalogical results, in terms of general (that is, Henkin) rather than first-order models. All interesting problems, no doubt, but somewhat overkill for the topic immediately at hand.

A simpler, if more artificial, solution is to simply introduce a new logical constant, Act, which we stipulate to have the desired interpretation. That is, given a sandwich frame $\mathcal{F}$, for all models $\mathcal{M}$ defined on $\mathcal{F}_\mathcal{M}$ we have $[\![\mathsf{Act}]\!]$ equal to the domain of the sandwnich function $\mathcal{F}_S$ a component of $\mathcal{F}$. Or, in simpler terms, Act is true at exactly those worlds where the function S is defined. This is equivalent to making $[\![\mathsf{Act}]\!]$ the generalised disjunction of all propositions to the effect that the agent makes p true (for arbitrary p).[14]

This formalisation, once added to our semantics for Conv, validates the obvious rule of inference we should want of the sentence: an "existential generalisation" to Act from any $\Box\varphi$. For, whenever we have $S_{\mathsf{min}}(w) \subseteq p \subseteq S_{\mathsf{max}}(w)$, we thereby also have $S(w)$ defined, and so $w \in [\![\mathsf{Act}]\!]$. As added to Conv4, we also get a valid rule of inference from any $\Box\varphi$ to $\Box(\mathsf{Act} \wedge \varphi)$. And, as added to Conv5*, we get a valid inference from any $\Box(\mathsf{Act} \wedge \psi)$ and any $\Box\varphi$ to the conclusion $\Box\Box\varphi$.

It should be kept in mind, however, that this formalisation brings with it weaknesses and somewhat artificial baggage of its own. To begin with, it imports certain heavy-handed commitments about the algebraic structure of the domain of propositions, well beyond those introduced by the logic of convexity proper. To draw this out, consider that $[\![\mathsf{Act}]\!]$ will be ill-defined when there is no join of all the propositions which state that the agent controls some particular fact. This enables us to construct algebraic frames of Conv (or Conv4 or Conv5*) in which the construction of $[\![\mathsf{Act}]\!]$ is unavailable, so that the the first two of the above inferences fail.[15] (On these alternative frames see §3.2.2,

[14] The logic of the operator is characterised in the appendix; it is obtained by adding an "existential generalisation" principle to Conv saying, intuitively, that if one makes it so that φ one makes something so.

[15] For example, take as the domain of propositions the Boolean algebra of intervals on the real line (closed under finitary intersection and finitary union) and, as the function $\mathcal{N}$ from a given member p of the interval algebra to the member $\mathcal{N}(p)$ asserting that

in particular fn. 14.) So there is no guarantee that the desired "mere action" proposition $[\![\mathsf{Act}]\!]$ will even exist.[16]

Even when existence *is* guaranteed, however, the proposition can be misleading. For example, in general possible world models of Conv where the neighbourhood function takes any world to either the null set or some (potentially unbounded) convex lattice in the powerset algebra of worlds, and thus where fully general versions of **Agglomeration** and **Modest Weakening** fail (unlike sandwich models), $[\![\mathsf{Act}]\!]$ can be entirely "sandwiched" between the infimum and supremum of facts an agent controls, while nevertheless not being controlled by the agent.[17] These limitations to our formalisation of mere action will become relevant in what follows.

In any event, we now have our axiom:

$\mathsf{C}_{\mathsf{Act}}\ \Box\varphi \rightarrow \Box(\mathsf{Act} \wedge \varphi)$

As noted in fn. 6, adding this axiom has the effect (in sandwich models) of collapsing $\mathsf{Conv5}^*$ into $\mathsf{Conv45}^*$, i.e. of requiring any $\mathsf{ConvC}_{\mathsf{Act}}\mathsf{5}^*$ model to make S constant throughout any defined $S_{\mathsf{min}}(w)$. We now turn to a set of philosophical problems hinging on $\mathsf{C}_{\mathsf{Act}}$.

2.4.3 Doomed to Action?

Christine Korsgaard's [Korsgaard, 2009] opens with the following pithy argument:

---

it is made true, the function that is the identity function on singletons of integers and the constant function to the null proposition elsewhere. Then it is easy to see that any interpretation (defined in the natural way) of this frame will be a model of Conv, $\mathsf{Conv4}$, and $\mathsf{Conv5}^*$, but the join of all makings true $[\![\mathsf{Act}]\!]$ undefined. In a formalisation along the quantificational lines discussed earlier, these failures would generally be ruled impossible by construction, even in general algebraic models.

[16] It is an interesting question, to which I have not determined an answer, whether there are any sentences *in the fragment of our language not containing* Act valid in all (algebraic) models on which $[\![\mathsf{Act}]\!]$ is defined but not theorems of Conv, i.e. whether Conv is incomplete with respect to the class of $[\![\mathsf{Act}]\!]$-defined algebraic models. My suspicion, which I cannot prove, is that the answer is Yes. The incompleteness of certain normal logics with respect to their general possible world models seems suggestive in this respect, but these questions extend beyond the scope of this dissertation [Gerson, 1975].

[17] Take, for example, a frame whose domain of worlds is the positive reals, and where for any $x < 0, \in \mathbb{R}$ the facts made true at x are just those sets in $\mathbb{R}$ containing $(0, x]$ and contained in some $(0, y)$ for $y > x$. It is easy to check that the axioms of Conv are all satisfied, that $[\![\mathsf{Act}]\!]$ is just $\mathbb{R}$, and yet that $\mathbb{R}$ is made true nowhere.

> Human beings are *condemned* to choice and action. Maybe you think you can avoid it, by resolutely standing still, refusing to act, refusing to move. But it's no use, for that will be something you have chosen to do, and then you will have acted after all. Choosing not to act makes not acting a kind of action, makes it something that you do. ...
>
> So action is necessary. What kind of necessity is this? Philosophers like to distinguish between *logical* and *causal* necessity. But the necessity of action isn't either of those. [...] It is our *plight*: the simple inexorable fact of the human condition. (emphasis retained)

Abstract, for the time being, from the *soundness* of this argument and focus simply on its conclusion. Korsgaard seems to be suggesting that, if an agent acts (surely not *just* a human agent, given the generality of the argument), then they are *condemned* to act, and act *of necessity*. The former, I take it, precludes *controlling* whether one acts. This preclusion perhaps reflects an assumed version of the principle of alternative possibilities in the background, requiring p to be possible (in some relevant sense) and possibly false if it is made true. But it need not: perhaps Korsgaard thinks mere action has some special property validating the principle of alternative possibilities in its case but not generally. Either way, the conclusion (as we are reading it) comes in two parts: an agent who merely acts, does not control that they merely act; and an agent who merely acts, necessarily (narrow scope) merely acts.

The extreme generality of the argument suggests that the conclusions are supposed to hold with broadly logical necessity. That is, at any possible world w, an agent who merely acts, does not control that they merely act; and an agent who merely acts, necessarily (narrow scope) merely acts. Of course, at a given world, this or that agent might not act, perhaps because they do not exist (say, because their parents never met). That is why the antecedent in both cases, *viz.* that the agent merely acts, needs to be included in the statement of the conclusion: at worlds where the agent fails to act (say, by failing to have been born), the conclusion should still be true.

Acting, I take it, always involves making something true, and *vice versa*. So we may treat the two as synonymous, for our (merely

intensional) purposes. Korsgaard's conclusions may then be summed up in two axioms: $\mathsf{C}_{\mathsf{Act}}$ and $\mathsf{Act} \rightarrow \neg\Box\mathsf{Act}$. The latter, clearly, can without loss be strengthened to $\neg\Box\mathsf{Act}$, since the consequent is also trivially true when the antecedent is false.

I suspect, further, that Korsgaard is assuming another feature of this necessity that condemns: necessarily, for an agent who is condemned to ϕ, their actions (that is, the things they make true) are collectively inconsistent with not ϕ-ing. It would be strange to describe ϕ-ing as "our *plight*: the simple inexorable fact of the human condition" were it fully consistent with all we make so that we not ϕ. This assumption would fit with a general pattern concerning the interaction of modality and control noted in arguing for the regularity or congruence of our logic: for many necessities, if p and q are necessarily equivalent, so are making p true and making q true. The assumption follows from Korsgaard's "condemnation" being such a necessity. Instantiating this schema in the case of mere action, we get: necessarily, if an agent is condemned to (merely) act, their actions are collectively inconsistent with not acting.

Now, if this is true, it has a rather striking consequence about the sandwich models validating all of Korsgaard's conclusions. As will be stressed over and over in what follows, in our semantics there are two ways an agent can fail in w to control some fact p. Firstly, p may be false somewhere in $S_{\mathsf{min}}(w)$; that is, the agent's actions may be jointly consistent with its falsehood. Secondly, p may be true somewhere outside $S_{\mathsf{max}}(w)$; that is, the agent's actions may not exhaust the ways in which p can obtain. The former possibility, however, is precisely what the assumption in the last paragraph rules out. Thus, Korsgaard's argument seems, in our logic, to deliver the following striking result: for any agent, there is some possibility where they make true *something*, but *nothing* is true that they *actually* make true.

These possibilities (call them *alien* possibilities for the agent at a given world) are somewhat strange. It is difficult, for one, in ordinary circumstances of ordinary agents to imagine what they would so much as involve. I can imagine, perhaps, of any given thing I am making true (that I am typing, that I am seated in my apartment, that I am whistling, etc.) that I could make it false. But in any imaginary scenario I can easily devise about my counterfactual circumstances, I will share

some things I am making true, however general.[18] They also mean, what is perhaps more philosophically notable, that there are no facts I *essentially* make true, if I make anything true at all. That is, for a given agent, there are no facts it makes true in all possibilities where it makes anything true. Thus, Korsgaard's conclusions have the rather surprisingly un-Kantian sounding consequence (in our models) that no actions are among the "conditions of possibility" for an agent's acting at all.[19]

How seriously should we take these results? And how much of a problem do they pose for Korsgaard's conclusions? As remarked in the previous section, the results are in at least one sense an artefact of our models, in that there are classes of algebraic and non-sandwich neighbourhood models for which the results just stated do not hold. There exist, for example, neighbourhood models on which the made true propositions at a world consist of not necessarily bounded convex lattices in the powerset of worlds, in which the main result (that, for any w, an agent has alien possibilities at w) fails.[20] The defender of Korsgaard's conclusions may take refuge in such models, as demonstrating the artificiality of our results about alien possibilities.

We noted two adverse consequences for Korsgaard's conclusions; the refuge offered by such models as just mentioned provides more solace in one case than in the other. The first of these consequences was that the required alien possibilities will be very difficult to imagine, for it is difficult to imagine away *all* the things an agent makes true at once, rather than imagining away each successively. This consequence is the product of what is in effect an infinitary version of **Modest Weakening** baked into the sandwich semantics. Given any collection P of propositions an agent makes true at a world w, in the sandwich semantics

[18] Things are worse than this: we will later on see reason to think Korsgaard's premise implies *all* possibilities inconsistent with everything one makes true are alien possibilities in which one acts! We shall also see reasons, however, to regard this prospect as less shocking than it might first appear. The most sustained discussion on this point is in §5.2

[19] She at various points seems to suggest that an agent a) only exists if they act and b) makes it so that they exist, whenever they do exist; thus, that the agent in question exists seems to be just such an "inevitably" made true fact.

[20] Let the domain of the model by $\mathbb{N}$, and the propositions made true at any n be the bounded sets containing n. It is easy to check that Conv is satisfied in the model, however its valuation is assigned. Nevertheless, there are no alien possibilities for any n.

the agent also makes true the join $\bigcup P$ of P at w. Without this feature of the semantics, the consequence will not follow. And, crucially, the more general neighbourhood semantics described above avoids this feature.

It is plausible to think, however, that this feature of the semantics is desirable. One way to spell this out would be to restate our logic and semantics in an infinitary language with infinitary disjunction and conjunction sentences, but it is possible to get at the basic point informally without needing to introduce these epicycles to the object language. Here is a warm-up. Intuitively, where there is some class of propositions P each member of which the agent makes true, they will also make true a (the) proposition true at exactly the worlds where each member of P is true. For example, if for each n the agent makes true that the water in the pot is below $x+1/n$ degrees Celsius, the agent thereby also makes true that the water is at or below x degrees Celsius. This is a sort of "infinitary" version of **Agglomeration**, and is validated on the natural way of understanding the sandwich semantics.

Why endorse this principle? Partly it is because, as long as we already endorse the finitary **Agglomeration**, there is something arbitrary in denying it. If, for any finite class P of propositions, the agent will make true the conjunction of P as long as it makes true each of P, why should this fail once P is infinite? In epistemic or doxastic logic, there is a possible reason to have the relevant operator (knowledge, belief) agglomerate finitarily but not infinitarily: epistemic subjects like ourselves might have essentially finite representational and reasoning abilities, such that we can be expected to be in a position to know (say) that the water is below $x + 1$ Celsius, and that is is below $x + 1/2$ Celsius, and that ..., but not thereby to collect all these infinitely many instances of knowledge into the stronger knowledge that the water is below or at x Celsius. But making true is not representational or reasoning-involving in this way, and thus there is not the same potential barrier between finitary and infinitary **Agglomeration**.[21]

Similar reasoning applies in the case of *co*-agglomeration, in the

[21]Something analogous seems to underwrite the kindness posterity has shown to the "Limit Assumption" in the logic of counterfactuals, both initially formulated and rejected by David Lewis [Lewis, 1973]. One way to read much of the subsequent literature on counterfactuals is as a gradual realisation of the folly in placing philosophical weight on failures of infinitary agglomeration in the logic of the counterfactual conditional [Warmbrod, 1982] [Caie, 2018] [Bacon, 2023] [Kocurek, 2026].

form of **Modest Weakening**. For the same reason it validates that axiom, the sandwich semantics bakes in an ω-entailment from (for each n) an agent making the water to be colder than $x - 1/n$ Celsius to the agent making the water to be colder than or at x Celsius, and this seems for the best. As with **Agglomeration**, it is already true that, for any finite class P of propositions, an agent making all the members of P true makes true a (the) proposition true at all worlds where any of P are true; why should this not hold of infinite P? But a semantics in which this holds will still validate the objectionable consequence for Korsgaard about alien worlds.

The other problem, that in our favoured models Korsgaard's conclusions run into conflict with the kinds of "inevitable" actions one would expect the Kantian to affirm, is less amenable to resolution by exotic models. For the result generalises beyond the sandwich models that caused the first problem to the more general setting of neighbourhood models. If there is some proposition p made true in such a model whenever an agent acts at all, $\Box p$ is equivalent in the model to Act (for it cannot be *weaker* than mere action).[22] So Korsgaard's assertion that an agent is condemned to act ($\neg\Box\mathsf{Act}$, for us) comes directly into conflict with any inevitable actions.

It would be exceedingly quick to take this as a refutation of Korsgaard. Several alternatives suggest themselves:

- It might demonstrate that the notion of action as controlling and making true is not adequate to the notion of action as employed in the kind of moral philosophy being practised by Korsgaard. In particular, the non-hyperintensionality of our concept might make it inappropriate for such work, at least without extensive caveats; more generally, perhaps Korsgaard is best read as rejecting the rule of congruence **RE**.

- The logical connections I have been attributing to Korsgaard between condemnation, necessity, and control might be poor interpretations of the text.

- "Inevitable" actions may not be as essential to her Kantian project as they appear.

[22] A similar argument applies in the case of algebraic models with $[\![\mathsf{Act}]\!]$ defined.

Nevertheless, even this much is of philosophical interest, for as a case study it helps clarify the potential connections and gaps between the intuitive concept axiomatised by Conv and the place action holds in much moral philosophising.

2.5 Conclusion

The current chapter has presented, and defended, a certain logic of control, in the sense promised in the introduction. The logic consists of general theses about (and *only* about) control and the familiar logical operators of classical sentential logic, generated by fundamental axioms about the same operated upon by a few simple rules of inference. That is, the logic defended is a theory in a narrowly *syntactic* sense, whose underlying principles are both motivated and articulated in the language of control itself.

Nevertheless, the theory (without being motivated by it) comes along with an elegant and simple model theory, which in an important sense can reduce questions about control to questions about these models. The reducibility is one way: allow ourselves talk in the metalanguage about this range of models, and there is no guarantee we can translate that talk back into the mere language of control. The following chapter will take up questions that naturally emerge from a philosophical consideration of these models, as applied in the representation of human agency. While these questions will be (provably) ineffable from the standpoint of the official object language of this chapter's Conv, Conv4, and Conv5*, they will draw out some of the philosophical import of these logics, considered as theories of agency.

2.A Metalogic

We here prove relevant results for the logics discussed above. Proofs are largely adapted from analogous ones in [Cresswell and Hughes, 1996] and [Pacuit, 2017].

2.A.1 The Logic Conv

Our language $\mathcal{L}$ (or equivalently, its set of wff's) is as usual that generated by a countable set At of sentential variables $p, q, r, \ldots$ and the sentential operators $\neg$, $\vee$, $\wedge$, $\rightarrow$, and $\Box$.

$$\mathsf{At} \mid \neg\varphi \mid \varphi \vee \psi \mid \varphi \wedge \psi \mid \varphi \rightarrow \psi \mid \Box\varphi$$

Our logic itself, $\mathsf{Conv} \subset \mathcal{L}$, i.e. the φ such that $\vdash_{\mathsf{Conv}} \varphi$, is the set of wff's generated by all instances of the following axiom schemas

All classical tautologies

T $\Box\varphi \rightarrow \varphi$

M $(\Box\varphi \wedge \Box\psi) \rightarrow \Box(\varphi \wedge \psi)$

O $(\Box\varphi \wedge \Box\psi) \rightarrow \Box(\varphi \vee \psi)$

Conv $\Box(\varphi \vee \psi \vee \chi) \rightarrow \Box\varphi \rightarrow \Box(\varphi \vee \psi)$

with the following inference rules

MP if $\vdash_{\mathsf{Conv}} \varphi$ and $\vdash_{\mathsf{Conv}} \varphi \rightarrow \psi$, $\vdash_{\mathsf{Conv}} \psi$

RE if $\vdash_{\mathsf{Conv}} \psi \leftrightarrow \phi$, $\vdash_{\mathsf{Conv}} \Box\psi \leftrightarrow \Box\phi$

A model $\mathcal{M}$ of our logic is a triple $\langle W, V, (S_{\mathsf{min}}, S_{\mathsf{max}})\rangle$, with W some nonempty set, $V : \mathsf{At} \rightarrow \mathcal{P}(W)$, and the possibly partial $(S_{\mathsf{min}}, S_{\mathsf{max}}) : W \hookrightarrow \mathcal{P}(W) \times \mathcal{P}(W)$ (or just S) such that (where defined)

$$\{w\} \subseteq S_{\mathsf{min}}(w) \subseteq S_{\mathsf{max}}(w)$$

The clauses of the interpretation function $[\![\ ]\!]$ for the atoms and Boolean clauses are as usual, while

$$[\![\Box\varphi]\!] = \{w \in W : S_{\mathsf{min}}(w) \subseteq [\![\varphi]\!] \subseteq S_{\mathsf{max}}(w)\}$$

(Note that $w \notin [\![\Box\varphi]\!]$ when S is undefined on w, and that as components of a single function from W into the cartesian square of $\mathcal{P}(W)$ S_{min} and S_{max} are each defined iff the other is.) φ is *valid* in a model $\mathcal{M}$, i.e. $\vDash_{\mathcal{M}} \varphi$, iff

$$[\![\varphi]\!]_{\mathcal{M}} = W_{\mathcal{M}}$$

and is valid simpliciter, i.e. $\vDash_{\mathsf{Conv}} \varphi$, iff for all models $\mathcal{M}$,

$$\vDash_{\mathcal{M}} \varphi$$

We now prove a lemma that will be useful in proving completeness.

Lemma 2.A.1. *If* $\vdash_{\mathsf{Conv}} (\varphi \wedge \psi) \rightarrow \chi$ *and* $\vdash_{\mathsf{Conv}} (\neg\varphi \wedge \neg\psi) \rightarrow \neg\chi$, *then* $\vdash_{\mathsf{Conv}} (\Box\varphi \wedge \Box\psi) \rightarrow \Box\chi$.

Proof. Suppose the antecedent. Then by classical reasoning, we have

$$\vdash_{\mathsf{Conv}} (\varphi \wedge \psi) \rightarrow \chi$$

and

$$\vdash_{\mathsf{Conv}} \chi \rightarrow (\varphi \vee \psi)$$

By the axiom M we have

$$\vdash_{\mathsf{Conv}} (\Box\varphi \wedge \Box\psi) \rightarrow \Box(\varphi \wedge \psi)$$

By axiom O we have

$$\vdash_{\mathsf{Conv}} (\Box\varphi \wedge \Box\psi) \rightarrow \Box(\varphi \vee \psi)$$

But then by **Conv**, **RE**, and classical reasoning we have

$$\vdash_{\mathsf{Conv}} (\Box\varphi \wedge \Box\psi) \rightarrow \Box\chi$$

□

Corollary 2.A.2. *For any finite* $\{\varphi_1, \ldots, \varphi_n, \psi\}$, *if* $\vdash_{\mathsf{Conv}} (\varphi_1 \wedge \ldots \wedge \varphi_n) \rightarrow \psi$ *and* $\vdash_{\mathsf{Conv}} (\neg\varphi_1 \wedge \ldots \wedge \neg\varphi_n) \rightarrow \neg\psi$, $\vdash_{\mathsf{Conv}} (\Box\varphi_1 \wedge \ldots \wedge \Box\varphi_n) \rightarrow \Box\psi$.

2.A.2 Soundness

It is easy to see that our semantics is sound.

Theorem 2.A.3. *If* $\vdash_{\mathsf{Conv}} \varphi$, *then* $\vDash_{\mathsf{Conv}} \varphi$

Proof. We omit as routine the proofs of the validity of classical tautologies and **MP**. The validity of T is straightforward from the stipulation that (when defined)

$$\{w\} \subseteq S_{\mathsf{min}}(w)$$

The validity of M is straightforward from the fact that for $X, Y \supseteq S_{\mathsf{min}}(w)$,

$$X \cap Y \supseteq S_{\mathsf{min}}(w)$$

The validity of O is straightforward from the fact that for $X, Y \subseteq S_{\mathsf{max}}(w)$,

$$X \cup Y \subseteq S_{\mathsf{max}}(w)$$

The validity of **Conv** is straightforward from the fact that, for any X, Y, Z,

$$X \subseteq (X \cup Y) \subseteq (X \cup Y \cup Z)$$

By induction, for any φ, ψ provably equivalent and for which our theorem holds, in all $\mathcal{M}$,

$$[\![\varphi]\!] = [\![\psi]\!]$$

validating **RE** by the intensionality of the clause for $\Box$. □

2.A.3 Completeness

Let a *consistent set* Γ be a set of wff's in $\mathcal{L}$ such that for no $\varphi_1 \ldots \varphi_n \in \mathcal{L}$ (for finite n) do we have

$$\vdash_{\mathsf{Conv}} \neg(\varphi_1 \wedge \ldots \wedge \varphi_n)$$

Let a *maximal consistent set* Γ be a consistent set such that for all $\varphi \in \mathcal{L}$ either $\varphi \in \Gamma$ or $\neg\varphi \in \Gamma$. The proof that for any consistent set Γ there is a maximal consistent set Γ' such that $\Gamma \subseteq \Gamma'$ is as usual. We denote the set of maximal consistent sets in $\mathcal{P}(\mathcal{L})$ as $\mathcal{L}_{\mathsf{M}}$.

We define the *canonical model* $\mathcal{M}_{\mathsf{canon}}$, as per the above, as the triple $\langle W, V, (S_{\mathsf{min}}, S_{\mathsf{max}})\rangle$ such that

- $W_{\mathcal{M}_{\mathsf{canon}}} = \mathcal{L}_{\mathsf{M}}$
- For $p \in \mathsf{At}$, $V_{\mathcal{M}_{\mathsf{canon}}}(p) = \{w \in W_{\mathcal{M}_{\mathsf{canon}}} : p \in w\}$
- $S_{\mathsf{min}\mathcal{M}_{\mathsf{canon}}}(w') = \{w \in W_{\mathcal{M}_{\mathsf{canon}}} :$ for all $\Box\varphi \in w', \varphi \in w\}$ unless no $\Box\varphi \in w'$, in which case undefined
- $S_{\mathsf{max}\mathcal{M}_{\mathsf{canon}}}(w') = \{w \in W_{\mathcal{M}_{\mathsf{canon}}} :$ for some $\Box\varphi \in w', \varphi \in w\}$ unless no $\Box\varphi \in w'$, in which case undefined

Lemma 2.A.4. *Let Γ be some consistent set of wff's of the form $\Box\varphi$, $\Gamma_{-\Box}$ be the set obtained by exchanging φ for each $\Box\varphi \in \Gamma$, and $\neg\Gamma_{-\Box}$ the set obtained by negating each member of $\Gamma_{-\Box}$. Then if, for arbitrary $\neg\Box\psi$, we have $\Gamma \cup \{\neg\Box\psi\}$ consistent, we have either $\Gamma_{-\Box} \cup \{\neg\psi\}$ consistent or $\neg\Gamma_{-\Box} \cup \{\psi\}$ consistent.*

Proof. Suppose that $\Gamma \cup \{\neg\Box\psi\}$ is but neither $\Gamma_{-\Box} \cup \{\neg\psi\}$ nor $\neg\Gamma_{-\Box} \cup \{\psi\}$ also is consistent. Then there must be some finite $\{\gamma_1, \ldots, \gamma_n\} \subseteq \Gamma_{-\Box}$ where

$$\vdash_{\mathsf{Conv}} \neg(\gamma_1 \wedge \ldots \wedge \gamma_n \wedge \neg\psi)$$

and some (not necessarily disjoint) $\{\gamma_{n+1}, \ldots, \gamma_m\} \subseteq \Gamma_{-\Box}$ where

$$\vdash_{\mathsf{Conv}} \neg(\neg\gamma_{n+1} \wedge \ldots \wedge \neg\gamma_m \wedge \psi)$$

Indeed, by the monotonocity of inconsistency, we may simplify by saying we have both

$$\vdash_{\mathsf{Conv}} \neg(\gamma_1 \wedge \ldots \wedge \gamma_m \wedge \neg\psi)$$

and

$$\vdash_{\mathsf{Conv}} \neg(\neg\gamma_1 \wedge \ldots \wedge \neg\gamma_m \wedge \psi)$$

By classical reasoning and the first conjunct, we have

$$\vdash_{\mathsf{Conv}} (\gamma_1 \wedge \ldots \wedge \gamma_m) \rightarrow \psi$$

By classical reasoning and the second conjunct, we have

$$\vdash_{\mathsf{Conv}} (\neg\gamma_1 \wedge \ldots \wedge \neg\gamma_m) \rightarrow \neg\psi$$

But since by supposition $\Gamma \cup \{\neg\Box\psi\}$ is consistent, we have

$$\nvdash_{\mathsf{Conv}} (\Box\gamma_1 \wedge \ldots \wedge \Box\gamma_m) \rightarrow \Box\psi$$

But by Corollary 2.A.2, this is impossible. □

Theorem 2.A.5. $w \in [\![\varphi]\!]_{\mathcal{M}_{\mathsf{canon}}}$ *iff* $\varphi \in w$

Proof. The proof proceeds by induction. It is straightforward to see that it holds in the base step for all $p \in \mathsf{At}$, as the interpretation function $[\![\,]\!]$ in that case simply reduces to the assignment function V, which in turn by construction reduces in $\mathcal{M}_{\mathsf{canon}}$ to $\in$. We omit as routine the

induction steps for the Boolean connectives, leaving only our operator $\Box$.

We first prove right to left. Suppose $\Box\varphi \in w$. Now by hypothesis we have $w' \in [\![\varphi]\!]_{\mathcal{M}_{\text{canon}}}$ iff $\varphi \in w'$; by construction we then have for all $w' \in S_{\min\mathcal{M}_{\text{canon}}}(w)$ that $\varphi \in w'$, and thus

$$[\![\varphi]\!]_{\mathcal{M}_{\text{canon}}} \supseteq S_{\min\mathcal{M}_{\text{canon}}}(w)$$

Moreover, by construction of $S_{\max\mathcal{M}_{\text{canon}}}$ and by the inductive hypothesis,

$$[\![\varphi]\!]_{\mathcal{M}_{\text{canon}}} \subseteq S_{\max\mathcal{M}_{\text{canon}}}(w)$$

So by construction of $[\![]\!]$ we have

$$w \in [\![\Box\varphi]\!]_{\mathcal{M}_{\text{canon}}}$$

Now suppose $\Box\varphi \notin w$, i.e. (by maximality) $\neg\Box\varphi \in w$. Either $S_{\mathcal{M}_{\text{canon}}}(w)$ is defined or not. If not, then by stipulation

$$w \notin [\![\Box\varphi]\!]_{\mathcal{M}_{\text{canon}}}$$

Now let Γ be $\{\Box\psi : \Box\psi \in w\}$; then, if $S_{\mathcal{M}_{\text{canon}}}(w)$ is defined, by Lemma 2.A.4 and consistency of $w \supset \Gamma \cup \{\neg\Box\varphi\}$ we have either $\Gamma_{-\Box} \cup \{\neg\varphi\}$ consistent or $\neg\Gamma_{-\Box} \cup \{\varphi\}$ consistent. But then (as the domain $\mathcal{L}_{\mathsf{M}}$ includes all maximal consistent sets, and all consistent sets can be extended to some $v \in \mathcal{L}_{\mathsf{M}}$) there must exist either some $w' \supset \Gamma_{-\Box} \cup \{\neg\varphi\}$ or $w'' \supset \neg\Gamma_{-\Box} \cup \{\varphi\}$. If the former, then by the inductive hypothesis, the consistency of w', and construction of $S_{\min\mathcal{M}_{\text{canon}}}$ we have

$$w' \in (S_{\min\mathcal{M}_{\text{canon}}}(w) \setminus [\![\varphi]\!])$$

and thus

$$w \notin [\![\Box\varphi]\!]_{\mathcal{M}_{\text{canon}}}$$

If the latter, then by the inductive hypothesis, the consistency of w'', and construction of $S_{\max\mathcal{M}_{\text{canon}}}$ we have

$$w'' \in ([\![\varphi]\!] \setminus S_{\max\mathcal{M}_{\text{canon}}}(w))$$

and thus

$$w \notin [\![\Box\varphi]\!]_{\mathcal{M}_{\text{canon}}}$$

$\Box$

Corollary 2.A.6. $\vDash_{M_{\mathsf{canon}}} \varphi$ *iff* $\vdash_{\mathsf{Conv}} \varphi$

Proof. Suppose

$$\vdash_{\mathsf{Conv}} \varphi$$

Then we have $\{\neg\varphi\}$ inconsistent, and thus by maximality of $w \in W_{\mathcal{M}_{\mathsf{canon}}}$ each such w must contain φ, which by the above means

$$\vDash_{M_{\mathsf{canon}}} \varphi$$

Suppose

$$\nvdash_{\mathsf{Conv}} \varphi$$

Then we have $\{\neg\varphi\}$ consistent, meaning some $w \in W_{\mathcal{M}_{\mathsf{canon}}} = \mathcal{L}_{\mathsf{M}}$ must extend it, so by the consistency of $w \in W_{\mathcal{M}_{\mathsf{canon}}}$ this w must not contain φ, which by the above means

$$\nvDash_{M_{\mathsf{canon}}} \varphi$$

□

By familiar reasoning, proving completeness for our logic now only requires, in light of Corollary 2.A.6, that the frame $\langle W_{\mathcal{M}_{\mathsf{canon}}}, S_{\mathcal{M}_{\mathsf{canon}}}\rangle$ satisfies the requirements for a model frame given in the first section. (That it satisfies the requirements for its base valuation function $V_{\mathcal{M}_{\mathsf{canon}}}$ is immediate.)

Corollary 2.A.7. *If* $\vDash_{\mathsf{Conv}} \varphi$, *then* $\vdash_{\mathsf{Conv}} \varphi$

Proof. The conditions for a frame are just that, for all w with $S(w)$ defined,

$$w \in S_{\mathsf{min}}(w)$$

and

$$S_{\mathsf{min}}(w) \subseteq S_{\mathsf{max}}(w)$$

$S_{\mathcal{M}_{\mathsf{canon}}}(w)$, by construction, is defined only for w with some $\Box\varphi \in w$. Let w be such a world. Now by T and maximality, for all $\Box\varphi \in w$, also $\varphi \in w$, and so by construction of $S_{\mathsf{min}\mathcal{M}_{\mathsf{canon}}}(w)$ we have

$$w \in S_{\mathsf{min}\mathcal{M}_{\mathsf{canon}}}(w)$$

Since $S_{\mathcal{M}_{\mathsf{canon}}}(w)$ is defined only when there exists such a $\Box\varphi \in w$, the set of $w' \in W_{\mathcal{M}_{\mathsf{canon}}}$ with *some* $\psi \in w'$ such that $\Box\psi \in w$ extends the set of those containing all such ψ, i.e.

$$S_{\mathsf{min}\mathcal{M}_{\mathsf{canon}}}(w) \subseteq S_{\mathsf{max}.\mathcal{M}_{\mathsf{canon}}}(w)$$

□

Corollary 2.A.8. $\vDash_{\mathsf{Conv}} \varphi$ *iff* $\vdash_{\mathsf{Conv}} \varphi$

2.A.4 The Logics Conv4, Conv5*, and Conv45*

The logics $\mathsf{Conv4}, \mathsf{Conv5}^*, \mathsf{Conv45}^* \subset \mathcal{L}$ are those generated by the inference rules and axiom schemas of Conv (*mutatis mutandis*) plus each of the following and both, respectively.

4 $\Box\varphi \rightarrow \Box\Box\varphi$

5* $(\Box\varphi \wedge \neg\Box\psi) \rightarrow \Box(\varphi \wedge \neg\Box\psi)$

The models of Conv4 and Conv5* are just like those of Conv, with the added stipulation that, for the first,

$$S_{\mathsf{min}}(w') \subseteq S_{\mathsf{min}}(w)$$

and

$$S_{\mathsf{max}}(w') \supseteq S_{\mathsf{max}}(w)$$

for all $w' \in S_{\mathsf{min}}(w)$ where $S(w)$ is defined; for the second,

$$S_{\mathsf{min}}(w') \supseteq S_{\mathsf{min}}(w)$$

and

$$S_{\mathsf{max}}(w') \subseteq S_{\mathsf{max}}(w)$$

for all $w' \in S_{\mathsf{min}}(w)$ where $S(w)$, $S(w')$ are defined. Their canonical models are constructed just like that of Conv, save for maximal consistency in the respective logics replacing maximal $\vdash_{\mathsf{Conv}}$-consistency.

Soundness is routine.

Theorem 2.A.9. *If* $\vdash_{\mathsf{Conv4}} \varphi$, *then* $\vDash_{\mathsf{Conv4}} \varphi$

Proof. To extend our soundness proof for Conv, observe first that $w \in [\![\Box\Box\varphi]\!]$ iff for each $w' \in S_{\min}(w)$, $w' \in [\![\Box\varphi]\!]$ and for no $w'' \notin S_{\max}(w)$ is $w'' \in [\![\Box\varphi]\!]$ Now clearly, if $w \in [\![\Box\varphi]\!]$, by the at most restriction of $S_{\min}(w')$ and at most expansion of $S_{\max}(w')$ for $w' \in S_{\min}(w)$,

$$S_{\min}(w') \subseteq [\![\varphi]\!] \subseteq S_{\max}(w')$$

(i.e. $w' \in [\![\Box\varphi]\!]$). Moreover, if $w \in [\![\Box\varphi]\!]$,

$$S_{\min}(w) \subseteq [\![\varphi]\!] \subseteq S_{\max}(w)$$

so for no $w'' \notin S_{\max}(w)$ do we have $w'' \in [\![\varphi]\!]$, much less $w'' \in [\![\Box\varphi]\!]$. So we have $w \in [\![\Box\Box\varphi]\!]$ if $w \in [\![\Box\varphi]\!]$, i.e. we have 4 valid. □

Theorem 2.A.10. *If* $\vdash_{\mathsf{Conv5}^*} \varphi$, *then* $\vDash_{\mathsf{Conv5}^*} \varphi$

Proof. To see that 5^* is valid, we will show that, whenever

$$w \in [\![\Box\varphi \wedge \neg\Box\psi]\!]$$

we also have

$$w \in [\![\Box(\varphi \wedge \neg\Box\psi)]\!]$$

Consider that

$$w \in [\![\Box(\varphi \wedge \neg\Box\psi)]\!]$$

iff for all $w' \in S_{\min}$,

$$w' \in [\![\varphi]\!], \notin [\![\Box\psi]\!]$$

(i.e. not $S_{\min}(w') \subseteq [\![\psi]\!] \subseteq S_{\max}(w')$), and for all $w'' \notin S_{\max}(w)$, either

$$w'' \notin [\![\varphi]\!]$$

or

$$w'' \in [\![\Box\psi]\!]$$

But since for such $w' \in S_{\min}$,

$$S_{\min}(w') \supseteq S_{\min}(w)$$

and

$$S_{\max}(w') \subseteq S_{\max}(w)$$

(where $S(w')$ is defined), if we have

$$S_{\mathsf{min}}(w) \subseteq [\![\varphi]\!] \subseteq S_{\mathsf{max}}(w)$$

but not

$$S_{\mathsf{min}}(w) \subseteq [\![\psi]\!] \subseteq S_{\mathsf{max}}(w)$$

(i.e. if $w \in [\![\Box\varphi \wedge \neg\Box\psi]\!]$), we have the latter also for (defined) $S(w')$, so that

$$w' \notin [\![\Box\psi]\!]$$

and we have

$$w' \in [\![\varphi]\!]$$

trivially, as

$$w' \in S_{\mathsf{min}}(w) \subseteq [\![\varphi]\!]$$

Note that where S is not defined,

$$w' \notin [\![\Box\psi]\!]$$

trivially. And, since if

$$w \in [\![\Box\varphi \wedge \neg\Box\psi]\!]$$

we have

$$[\![\varphi]\!] \subseteq S_{\mathsf{max}}(w)$$

it follows that for all $w'' \notin S_{\mathsf{max}}(w)$ we have

$$w'' \notin [\![\varphi]\!]$$

□

It is easy to check that the further structural requirements just given define 4- and 5^*-validity as conditions on frames.

Theorem 2.A.11. *The sandwich frames on which* Conv4 *is valid are exactly the frames of models for* Conv4

Proof. It is clear from soundness that Conv4 is valid on any such frame. To see that no other frames validate Conv4, suppose that there are w, w' with $w' \in S_{\mathsf{min}}(w)$ and either $v \in S_{\mathsf{min}}(w'), \notin S_{\mathsf{min}}(w)$ or $u \notin S_{\mathsf{max}}(w'), \in S_{\mathsf{max}}(w)$. Interpreting p by either $S_{\mathsf{min}}(w)$ or $S_{\mathsf{max}}(w)$ respectively, we have $\Box p$ but not $\Box\Box p$ true at w. □

Theorem 2.A.12. *The sandwich frames on which* Conv5* *is valid are exactly the frames of models for* Conv5*

Proof. The proof is essentially that of the preceding, *mutatis mutandis*. □

Completeness is similarly routine.

Theorem 2.A.13. *If* $\vDash_{\mathsf{Conv4}} \varphi$, *then* $\vdash_{\mathsf{Conv4}} \varphi$

Proof. By our earlier result, it will suffice to show, where $\mathcal{M}'_{\mathsf{canon}}$ is the canonical model of Conv4, that for $w \in W_{\mathcal{M}'_{\mathsf{canon}}}$ with $S_{\mathcal{M}'_{\mathsf{canon}}}(w)$ defined,

$$S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w') \subseteq S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w)$$

and

$$S_{\mathsf{max}\mathcal{M}'_{\mathsf{canon}}}(w') \supseteq S_{\mathsf{max}\mathcal{M}'_{\mathsf{canon}}}(w)$$

for all $w' \in S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w)$.

Suppose there is some $w' \in S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w)$ and $v \in S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w'), \notin S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w)$, i.e.

$$S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w') \nsubseteq S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w)$$

Then by construction, there is some $\Box\varphi \in w, \notin w'$. But by 4 this is impossible, since for any $\Box\varphi \in w$, $\Box\Box\varphi \in w$, and so $\Box\varphi \in w'$. For this same reason, $S_{\mathcal{M}'_{\mathsf{canon}}}(w')$ is always defined when $S_{\mathcal{M}'_{\mathsf{canon}}}(w)$ is defined and $w' \in S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w)$.

Suppose next there is some $w' \in S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w)$ and $v \in S_{\mathsf{max}\mathcal{M}'_{\mathsf{canon}}}(w), \notin S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w')$ (i.e. $S_{\mathsf{max}\mathcal{M}'_{\mathsf{canon}}}(w') \not\supseteq S_{\mathsf{max}\mathcal{M}'_{\mathsf{canon}}}(w)$). Then again by construction, there is some $\Box\varphi \in w, \notin w'$, which we have just seen to be impossible. □

Theorem 2.A.14. *If* $\vDash_{\mathsf{Conv5^*}} \varphi$, *then* $\vdash_{\mathsf{Conv5^*}} \varphi$

Proof. Again it will suffice to prove that for $w \in W_{\mathcal{M}'_{\mathsf{canon}}}$ with $S_{\mathcal{M}'_{\mathsf{canon}}}(w)$ defined,

$$S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w') \supseteq S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w)$$

and

$$S_{\mathsf{max}\mathcal{M}'_{\mathsf{canon}}}(w') \subseteq S_{\mathsf{max}\mathcal{M}'_{\mathsf{canon}}}(w)$$

for all $w' \in S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w)$ with some $\Box\psi \in w'$ (where $\mathcal{M}'_{\mathsf{canon}}$ is now the canonical model of Conv5*).

Suppose there is some $w' \in S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w)$ with $S_{\mathcal{M}'_{\mathsf{canon}}}(w')$ defined and $v \in S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w), \notin S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w')$, i.e.

$$S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w') \not\supseteq S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w)$$

Then by construction (and definedness of $S_{\mathcal{M}'_{\mathsf{canon}}}(w')$), there is some $\Box\varphi \in w', \notin w$. Now since $S_{\mathcal{M}'_{\mathsf{canon}}}(w)$ is defined, there exists $\Box\psi \in w$, with $\psi \neq \varphi$. We then by maximality have

$$\Box\psi \wedge \neg\Box\varphi \in w$$

But then we have by $\mathsf{5}^*$ that

$$\Box(\psi \wedge \neg\Box\varphi) \in w$$

and so by construction and maximality

$$\psi \wedge \neg\Box\varphi, \neg\Box\varphi \in w'$$

which is by consistency of w' impossible.

Suppose next there is some $w' \in S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w)$ and $v \in S_{\mathsf{max}\mathcal{M}'_{\mathsf{canon}}}(w'), \notin S_{\mathsf{min}\mathcal{M}'_{\mathsf{canon}}}(w)$, i.e.

$$S_{\mathsf{max}\mathcal{M}'_{\mathsf{canon}}}(w') \not\subseteq S_{\mathsf{max}\mathcal{M}'_{\mathsf{canon}}}(w)$$

Then again by construction, there is some $\Box\varphi \in w', \notin w$, which we have just seen to be impossible. □

Corollary 2.A.15. $\mathsf{Conv4}$ *and* $\mathsf{Conv5}^*$ *are independent.*

Proof. For a model of $\mathsf{Conv5}^*$ invalidating $\mathsf{4}$, let $W = \{0, 1\}$, $S_{\mathsf{min}}(0) = S_{\mathsf{max}}(0) = W$, $S(1)$ be undefined. For a model of $\mathsf{Conv4}$ invalidating $\mathsf{5}^*$, let $W = \{0, 1\}$, $S_{\mathsf{min}}(0) = S_{\mathsf{max}}(0) = S_{\mathsf{max}}(1) = W$, $S_{\mathsf{min}}(1) = \{1\}$. For one satisfying both let $W = \{0\}$, $S_{\mathsf{min}}(0) = S_{\mathsf{max}}(0) = W$. □

Corollary 2.A.16. $\mathsf{Conv45}^*$ *is sound and complete on the class of frames where for any w with $S(w)$ defined, $S(v) = S(u)$ where $v, u \in S_{\mathsf{min}}(w)$.*

Proof. This is straightforward from the preceding proofs. □

Corollary 2.A.17. *Among models in which, for all w with $S(w)$ defined, $S(w')$ is defined for all $w' \in S_{\mathsf{min}}(w)$, the* $\mathsf{Conv5}^*$ *models are the* $\mathsf{Conv45}^*$ *models.*

Proof. Take any such model of $\mathsf{Conv5}^*$ and any w with $S(w)$ defined. For any $w' \in S_{\mathsf{min}}(w)$, $S_{\mathsf{min}}(w')$, which we know to be defined, must contain w, as

$$w \in S_{\mathsf{min}}(w) \subseteq S_{\mathsf{min}}(w')$$

So, by the defining property of $\mathsf{Conv5}^*$ models,

$$S_{\mathsf{min}}(w) \subseteq S_{\mathsf{min}}(w')$$

(as $w' \in S_{\mathsf{min}}(w)$) and

$$S_{\mathsf{min}}(w') \subseteq S_{\mathsf{min}}(w)$$

(as $w \in S_{\mathsf{min}}(w')$), i.e.

$$S_{\mathsf{min}}(w) = S_{\mathsf{min}}(w')$$

Similarly,

$$S_{\mathsf{max}}(w) \supseteq S_{\mathsf{max}}(w')$$

and

$$S_{\mathsf{max}}(w') \supseteq S_{\mathsf{max}}(w)$$

so that

$$S_{\mathsf{max}}(w) = S_{\mathsf{min}}(w')$$

Thus, for all $w' \in S_{\mathsf{min}}(w)$,

$$S(w) = S(w')$$

The reverse direction is trivial. □

2.A.5 The Logic $\mathsf{Conv}_{\mathsf{Act}}$

Let the language $\mathcal{L}^{\mathsf{Act}}$ be $\mathcal{L}$ closed under the introduction of the sentential constant Act. The logic $\mathsf{Conv}_{\mathsf{Act}}$ is that generated by the inference rules and axiom schemas of Conv (*mutatis mutandis*) plus Gen:

Gen $\quad \Box\varphi \rightarrow \mathsf{Act}$

A model $\mathcal{M}$ of $\mathsf{Conv}_{\mathsf{Act}}$ is just a model of Conv. The clauses for the interpretation function $[\![\ ]\!]$ are as above, with the additional clause

$$[\![\mathsf{Act}]\!] = \mathsf{dom}(S)$$

where dom indicates the domain function. $\vdash_{\mathsf{Conv}_{\mathsf{Act}}}$, $\models_{\mathsf{Conv}_{\mathsf{Act}}}$, etc. are defined in the obvious ways.

As indicated in §2.4.1, the constant Act is not definable in $\mathcal{L}$.

Proposition 2.A.18. *No formula in $\mathcal{L}$ defines the worlds at which S is defined.*

Proof. We construct a model in which the worlds in the domain all satisfy the same formulas, but only some have S defined. Let the domain W be $\{w_0, w_1\}$. $S_{\mathsf{max}} = S_{\mathsf{min}}$ is as the singleton function on w_0 and undefined on w_1. For all atomic formulas p, $V(p) = \emptyset$.

Proof is by induction. Base step, for the atomics, is trivial. Induction step for the Boolean connectives is straightforward. For the induction step with regard to $\Box\varphi$, by hypothesis either φ is universally false or it is universally true. Suppose the former. Then, by factivity of $\Box$, $\Box\varphi$ is everywhere false, and so everywhere the same. If the latter, then it is equivalent to $\Box\top$, which is true neither at w_0 (as $S_{\mathsf{max}}(w) \subset W$) nor at w_1 (as S is there undefined). Thus, again, $\Box\varphi$ is everywhere false, and so everywhere the same. So, for all φ, either φ is everywhere true or it is everywhere false. □

The proof clearly generalises to (potentially infinite) *classes* of formulas in $\mathcal{L}$, too, as these are all constant throughout the above model.

Corollary 2.A.19. *No class of formulas in $\mathcal{L}$ defines the worlds at which S is defined.*

Soundness for $\mathsf{Conv}_{\mathsf{Act}}$ is trivial.

Theorem 2.A.20. *If $\vdash_{\mathsf{Conv}_{\mathsf{Act}}} \varphi$, then $\vDash_{\mathsf{Conv}_{\mathsf{Act}}} \varphi$*

Proof. The only new requirement is to show that, if

$$S_{\mathsf{min}}(w) \subseteq [\![\varphi]\!] \subseteq S_{\mathsf{max}}(w)$$

then we have

$$w \in \mathsf{dom}(S)$$

But this is clear. □

Completeness again is proved using a canonical model $\mathcal{M}_{\mathsf{canon}}$.

- $W_{\mathcal{M}_{\mathsf{canon}}} = \mathcal{L}^{\mathsf{Act}}_{\mathsf{M}}$
- For $p \in \mathsf{At}$, $V_{\mathcal{M}_{\mathsf{canon}}}(p) = \{w \in W_{\mathcal{M}_{\mathsf{canon}}} : p \in w\}$
- $S_{\mathsf{min}\mathcal{M}_{\mathsf{canon}}}(w') = \{w \in W_{\mathcal{M}_{\mathsf{canon}}} : \text{for all } \Box\varphi \in w', \varphi \in w\} \cup \{w'\}$ unless $\mathsf{Act} \notin w'$, in which case undefined

- $S_{\max \mathcal{M}_{\text{canon}}}(w') = \{w \in W_{\mathcal{M}_{\text{canon}}} : \text{ for some } \Box\varphi \in w', \varphi \in w\} \cup \{w'\}$ unless $\mathsf{Act} \notin w'$, in which case undefined

The difference between the earlier canonical model for Conv and the current model for $\mathsf{Conv}_{\mathsf{Act}}$ crops up only in those $w \in W_{\mathcal{M}_{\text{canon}}}$ containing no sentences of the form $\Box\varphi$ while nevertheless containing Act, whose value for S is simply $(\{w\}, \{w\})$. By maximality of all w containing sentences of the form $\Box\varphi$, we have $w \ni \mathsf{Act}$, so that $S(w)$ is unaffected.

We will need the fact that, as usual, the Lindenbaum-Tarski lattice of the logic is atomless and thus (being Boolean) satisfies a splitting condition.

Lemma 2.A.21. *For any consistent φ, there are distinct $w, w' \in \mathcal{L}_{\mathsf{M}}^{\mathsf{Act}}$ such that $\varphi \in w \cap w'$*

Proof. By an easy induction on proof length, any proof of φ may be converted into a proof of $\varphi[p := \neg p]$ by replacing each step ψ in the proof of φ with $\psi[p := \neg p]$. Thus, where p does not occur in φ, if we have

$$\vdash_{\mathsf{Conv}_{\mathsf{Act}}} \varphi \to p$$

we have also

$$\vdash_{\mathsf{Conv}_{\mathsf{Act}}} \varphi \to \neg p$$

(and, by double negation elimination, *vice versa*). The infinity of sentential variables, moreover, guarantees the existence of such a p. Therefore, for any consistent φ, we have

$$\nvdash_{\mathsf{Conv}_{\mathsf{Act}}} \varphi \to p$$

and also

$$\nvdash_{\mathsf{Conv}_{\mathsf{Act}}} \varphi \to \neg p$$

Therefore there exist two distinct maximal consistent sets containing φ: one containing p and the other $\neg p$. □

Theorem 2.A.22. $w \in [\![\varphi]\!]_{\mathcal{M}_{\text{canon}}}$ *iff* $\varphi \in w$

Proof. The proof runs essentially as that of Theorem 2.A.5. The only changes are the addition of a further base step for Act and a slight alteration for one case in the left to right direction of the induction step for $\Box$.

$S_{\mathcal{M}_{\text{canon}}}$ is defined at w iff $\mathsf{Act} \in w$, by definition and straightforward reasoning. Thus the base step for Act is immediate.

The right to left direction for the induction step for $\Box$ is as before. The left to right direction in the case with $S_{\mathcal{M}_{\text{canon}}}(w)$ undefined is unaffected. The case with $S_{\mathcal{M}_{\text{canon}}}(w)$ defined may be split into two subcases: that in which (for some ψ) we have $\Box\psi \in w$, and that in which we do not. In the first case, the argument goes as before. (This relies on Lemma 2.A.4 still obtaining for the logic $\mathsf{Conv}_{\mathsf{Act}}$, which the reader may wish to verify.) In the latter case, $S_{\mathcal{M}_{\text{canon}}}(w)$ is simply $(\{w\}, \{w\})$, so that the conclusion follows by Lemma 2.A.21. □

Corollary 2.A.23. $\vDash_{M_{\text{canon}}} \varphi$ *iff* $\vdash_{\mathsf{Conv}} \varphi$

Proof. As before. □

Corollary 2.A.24. *If* $\vDash_{\mathsf{Conv}_{\mathsf{Act}}} \varphi$, *then* $\vdash_{\mathsf{Conv}_{\mathsf{Act}}} \varphi$

Proof. Again, we need only prove that the frame $\langle W_{\mathcal{M}_{\text{canon}}}, S_{\mathcal{M}_{\text{canon}}} \rangle$ satisfies the requirements for a model frame.

By construction,

$$w \in S_{\mathsf{min}\mathcal{M}_{\text{canon}}}(w)$$

wherever the function is defined. As for the proof that

$$S_{\mathsf{min}\mathcal{M}_{\text{canon}}}(w) \subseteq S_{\mathsf{max}\mathcal{M}_{\text{canon}}}(w)$$

for all w where the sandwich function is defined, in the case with $\Box\varphi \in w$ (for some φ) the argument is as before. Otherwise, it follows by construction, as $\{w\} \subseteq \{w\}$. □

Corollary 2.A.25. $\vDash_{\mathsf{Conv}_{\mathsf{Act}}} \varphi$ *iff* $\vdash_{\mathsf{Conv}_{\mathsf{Act}}} \varphi$

Thus, $\mathsf{Conv}_{\mathsf{Act}}$ is precisely the logic of Act, as introduced in §2.4.

Chapter 3

Control and Options

3.1 Introduction

Projects in philosophical logic of the same broad shape as that undertaken in the preceding chapter are often referred to as "logics of action", and the name is clearly somewhat appropriate. Making something true is a sort of action, and any action will involve making *something* true, even if it is as self-contained as that one's limbs move in a certain way, or that a certain inner volition occurs. Thus it makes sense, from one direction, to think of propositions made true generally as actions, and *vice versa*.

There is another way, though, in which we talk about actions that sits poorly with this understanding. For, unless there is exactly one proposition an agent makes true (in the terms of our semantics, unless $S_{\mathsf{min}}(w) = S_{\mathsf{max}}(w)$), in the above sense the agent will be performing *many* actions, as many as there are facts of strength intermediate between the strongest and weakest things they make true. But often philosophers will speak of *the* action (or choice, or option) taken by an agent in a given circumstance, and there is no seeming assumption in this talk that the agent in question is of the degenerate sort in whom S_{min} and S_{max} would coincide in actuality. There may be several such actions *available*, but only one in fact taken. If *the* action taken by an agent in this sense is a fact they make true (and the actions available to them propositions they *could* make true), the title of "action" will

be reserved only for a special such fact (or facts).

The current chapter is an exploration of how this notion might be articulated in the framework set out in the last chapter, and what might be said about action with that articulation in play. This kind of action has played a central role in much formal theorising about agency, whose formality has meant a certain explicitness in assumptions about actions' logical structure. We now turn to an assessment of one such widespread assumption.

3.2 The Idea of an Action

3.2.1 Actions as Propositions

A common feature of formal theories of agency that treat the actions open to an agent (or, commonly, *options*) as propositions is that these actions are treated as *pairwise incompossible*. This is true, for example, of evidential decision theories in the style of Jeffrey [Jeffrey, 1965], of the various available versions of causal decision theory [Lewis, 1981] [Gibbard and Harper, 1978] [Joyce, 1999], and of the aforementioned STIT models.[1] Where propositions are interpreted as subsets of a domain of worlds, this means that the set of actions available to an agent forms a *partition* on some subset of the domain. In this chapter I will try to raise doubts about this assumption of partitionality.

Before we can enter into any substantive discussion of partitionality, we need something most of its major extant expositions lack: a more regimented account of an agent's action propositions at a world. The definition we adopt will importantly take sides on an ambiguity in the literature, which it will be illuminating to examine before proceeding.

Our definition of an action proposition is: necessarily, p is an action proposition for A iff it is practically possible p be the strongest proposition A makes true.[2] We will also sometimes speak of *the* action

[1] For recent philosophical discussion of the nature of action propositions (or "options") in decision theory, see [Pollock, 2002] [Hedden, 2012] [Koon, 2020] [Schwarz, 2023]. None of these writers contest the pairwise incompossibility of an agent's action propositions at a world.

[2] Observe that this resolves a critical ambiguity in the tempting phrase, "The strongest propositions an agent is in a position to make true". For it may be, for all we have seen, that both p and $p \vee q$ (for $p \not\subseteq q$) are action propositions in our sense, even as p is

proposition of an agent at a world w, where this is simply the strongest proposition it makes true at w; context should disambiguate the two where necessary. Indulging in the ontology of possible worlds, and letting S_A denote the sandwich function for A, this can be clarified as follows.

Definition 3.2.1. *p is* an *action proposition for A at w iff there is some v practically possible from w such that $p = S_A\,\mathsf{min}(v)$*

Definition 3.2.2. *p is* the *action proposition of A at w iff $p = S_A\,\mathsf{min}(w)$*

The machinery of sandwich semantics gives us a straightforward means of modelling the relation between practical possibility and actions. An ordinary sandwich frame, as described above, consists of a non-empty domain W equipped with a (possibly partial) function $(S_{\mathsf{min}}, S_{\mathsf{max}})$ taking a member w in W to a pair of subsets in W, one containing the other and each with w as a member. To represent practical possibility (whose dual, practical necessity, we more formally denote by $\boxdot$), we may add as a further parameter a reflexive relation R on W (equivalently representable as a function from W to its power set), to be interpreted standardly as in Kripke semantics for normal logics at least as strong as KT. Under the construal of actions as the possibly strongest propositions made true, for all $w \in W$, the actions at w will be those $V \subseteq W$ such that for some $v \in R(w)$, $S_{\mathsf{min}}(v) = V$. In pictorial terms, every world is associated with a set (including itself) of worlds practically possible from it, and its action propositions are the inner circles of that set.[3]

We will appeal, in the ensuing argument, to a further schematic principle relating making true to practical possibility:

Practical Strengthening If it is not practically possible that not p, and one makes true q, one makes true that p and q $((\boxdot p \wedge \Box q) \rightarrow \Box(p \wedge q))$

stronger than $p \vee q$. Indeed, it is trivial to construct models in which there is an infinitely descending chain possible actions propositions, ordered by strength, so that there is *no* strongest proposition available to an agent in this other sense.

[3] A natural question at this juncture is whether partitionality of a model is definable in our object language. The answer is No (see Theorem 3.A.5 in the appendix). For this reason, arguments about partitionality are ineliminably arguments about models conducted in the metalanguage, rather than arguments about control in and of itself conducted in the object language.

As an axiom schema, this corresponds semantically (check the appendix for details) to the constraint that $S_{\mathsf{min}}(w) \subseteq R(w)$.[4] We call the logic resulting from joining the axioms and inference rules of Conv for $\Box$ with those of KT for $\boxdot$, plus **Practical Strengthening**, Conv+, and it is provably sound and complete with respect to just the class of models described (call them *supplemented sandwich models*).

While this helps to clarify the abstract *representation* of practical possibility in the model theory, it does little to explicate its intuitive content. What is involved in some state of affairs being practically possible, as we are using the expression?

There are a few things we can say about the modality to help pin it down. First, in one form or another it is familiar from the informal motivation of formal models of agency, where it regularly plays more or less the role articulated in Definition 3.2.1. This is so, for example, in the STIT models described above, where the role is played by the "historically settled" modality; and, as we shall shortly review, a similar role is played in David Lewis' classic exposition of causal decision theory by an unanalysed modality denoted by "can". In particular, where these models of agency are *normative*, the (maximally specific) things one can make true will be the candidates for rational recommendation. This gives our first criterion for practical possibility: the actions facing an agent should be the maximally specific things the agent is practically able to make true.

Complementing this structural role, we can ostend with concrete examples. If I need to get from downtown Oakland to UN Plaza in San Francisco in under thirty minutes, I can assess different options. I might take the BART; I might hail a taxi; I might even, if I am particularly energetic or desperate, seriously exert myself in biking. I will not, however, rationally assess my running there in that time, for I know that it is not possible for me to make that so. Indeed, it is not just a matter of my not being able to *make* that happen: it is impossible for me to run that distance in that time altogether. The former options, unlike the latter, are practical possibilities for me. That last option might, of course, be possible in some weaker sense, by not contravening logic or the laws of physics. And the earlier options will be impossible in some senses: failing to take the train when it is available might be

[4] We will not in the upcoming argument need to appeal to this constraint on S and R applying universally in our model, just for one designated $\alpha \in W$, allaying possible worries about its modal status.

incompatible with my aesthetic or lifestyle preferences, say, and all but one option will (degenerately) be incompatible with the actual future course of events. But there is still a sense of possibility, relevant to rational deliberation, that cuts along the indicated lines.

The relevant modality, finally, should satisfy **Practical Strengthening**. There are several reasons to think there will exist a modality meeting the earlier criteria and satisfying **Practical Strengthening**. For one, in assessing the rational value (in terms of conditionally expected utility, counterfactually expected utility, harmony with my set of practical reasons, *vel sim.*), it makes sense to ignore ways those ways the options could (*sensu latu*) come to pass that are not (in some relevant *sensu strictu*) possible for me. Only the Red and Yellow Lines, for example, are available as BART transit between downtown Oakland and San Francisco. It thus makes sense to only consider the pros and cons of my taking either of those lines across the Bay, and not those for the Blue Line, which I cannot take that way. Thus, **Practical Strengthening** fits well the structural role on practical possibility in normative theories of agency.

Practical Strengthening, moreover, has considerable intuitive appeal insofar as the concept is motivated by cases. Again, for example, if I make it so that I take the BART to UN Plaza, I will specifically be making myself take the Red or Yellow Line there, for there is no other line I *could* take. Similarly, suppose that Riko makes her mother's whistle squeal (by blowing it, say), and that by dint of its resonance it is beyond her power for it to squeal at any tone besides $C^\sharp$; it seems to follow she has made it squeal in $C^\sharp$, as that is the only practical way for it to squeal at all. So **Practical Strengthening** is natural given the kinds of cases used to ostensively define the concept.

Finally, the principle follows from an instance of a broader pattern noted earlier: when p and q are necessarily (for some sufficiently strong brand of necessity) equivalent, one makes p true iff one makes q true. **Practical Strengthening** follows from the claim that practical necessity is such a necessity.[5] The converse entailment, notably, does not hold, so that the assumption of **Practical Strengthening** is weaker and therefore strictly dialectically preferable over outright assuming the

[5]Suppose $\boxdot(p \leftrightarrow q) \rightarrow (\Box p \leftrightarrow \Box q)$ for all p, q. Suppose, further, $\boxdot p_0$. Then (by normality of $\boxdot$) $\boxdot(q \leftrightarrow (p_0 \wedge q))$, so by the first supposition if $\Box q$ then $\Box(p_0 \wedge q)$. Thus, if $\boxdot p_0$ and $\Box q$, then $\Box(p_0 \wedge q)$.

congruence-like principle.[6] It therefore inherits, and indeed improves upon, the plausibility of this general pattern.

This Ramsey-like definition still leaves the concept somewhat indeterminate, of course. It may be that there are several modalities meeting the above bill, in which case we will need to arbitrarily fix one. And it still leaves somewhat opaque, in the manner typical of Ramseyfication, what practical possibility is "in itself," or *why* it unites these theoretical roles under one heading. I will have more to say about this later on (see §5.2.2, where the link to congruence-like principles plays a starring role), but for the time being this should be grist enough for the philosophical mill.

Practical Strengthening does require for plausibility a constraint on the joint interpretation of making true and practical possibility: the weaker our action propositions, the more things must be practically possible, lest the latter unduly strengthen the former. But this constraint has an intuitive basis: we should not interpret the practical modality more restrictively than need be, and naïvely the weaker the strongest things we are in a position to make true, the easier it is to make them true, and so the less restricted practical possibility has to be to accommodate the practical possibility that we make them true.

A word here on our extended semantics' relation to STIT semantics.[7] Practical necessity (or its corresponding parameter R) plays much the same role in our theory as historical necessity (and its corresponding accessibility relation) does in the logic of STIT. Whereas it is a ubiquitous assumption in the STIT literature that this is an S5 modality (i.e. its accessibility relation is equivalent), however, our own theory has only stipulated that practical necessity is factive. Why this weakening?

There are a couple of reasons not to assume at this juncture quite so strong a logic for practical necessity. To begin, that logic has deep motivations within the traditional STIT theory that do not extend to ours. In the traditional "branching time" version of the STIT semantics, it will be remembered, the points of evaluation are not *moments* in a branching time structure but *historical pairs* $\langle m, h \rangle$ of such a moment plus a *history* (maximal chain) in the structure running through that

[6] To see this, consider a model satisfying the above articulated constraint on the practical accessibility relation R in which, for some w, $R(w) \subseteq S_{\max}(w) \subset W$. Then $S_{\max}(w)$ and W are practically equivalent at w, though the latter is not made true.

[7] Thanks to anonymous reviewer for raising this question.

moment. In such a traditional presentation, historical necessity does not even have its own distinctive model parameter: its accessibility relation is the one obtaining between $\langle m, h\rangle, \langle m', h'\rangle$ just in case $m = m'$, which of course must be an equivalence. This fits with a committal philosophical interpretation of the modality: for p to be historically necessary, at a moment m and given a history h, is for p to be true at m given any other history h', too; historical necessity (at a historical pair) is, as it were, truth (at a moment) *irrespective* of any particular history, just as metaphysical necessity is truth (at a time) irrespective of any particular world and eternality is truth (at a world) irrespective of the time. Even in the study of atemporal STIT models, the influence lingers of this guiding vision. It is this metaphysical picture that inspires (indeed, when articulated formally in terms of branching time structures, *forces*) the assumption of S5 for the operator.

Our own theory does not come along with such a specific philosophical motivation. It is agnostic about what this practical necessity (or, dually, practical possibility) fundamentally amounts to, and instead makes only the relatively weak commitment that it satisfy **Practical Strengthening** and be factive (which seems to be part and parcel of practical possibility being a species of *possibility* in a more than merely technical sense).[8] As long as there is *some* relevant kind of possibility satisfying **Practical Strengthening**, our arguments will be of interest concerning this modality. And there is no guarantee that this modality will fit the picture described in the last paragraph.

This leads into the second reason: the desirability of weak assumptions. One reason for introducing talk of practical necessity at all in this paper, beyond just its intrinsic interest, is to prove that there are no partitional supplemented sandwich models validating certain natural assumptions about realistic cases of agency (see below). It is therefore beneficial to give the weakest such assumptions possible, as this allows us to prove a stronger result. This non-existence proof still goes through, *a fortiori*, if one imposes stronger constraints on the models (such as a logic of S5 rather than KT for practical necessity). Even if practical necessity *does* have the logic of S5, it is interesting to see that our anti-partitional results do not hinge on this fact about it.

[8]Note that, as long as a modality satisfies **Practical Strengthening**, factivity comes along for free, conditional on the agent making some p true: if it makes p true and q is necessary, p and q both follow by **Practical Strengthening** and the factivity of making true.

Another ubiquitous line of thinking about action propositions and practical capacity, and its similarly ubiquitous conflation with the thought formalised in Definition 3.2.1, finds especially clear expression in [Lewis, 1981]. In a revealing passage, Lewis defines the action partition for an agent like so:

> Suppose we have a partition of propositions that distinguish worlds where the agent acts differently. . . . Further, he can act at will so as to make any one of these propositions hold; but he cannot act at will so as to make any proposition hold that implies but is not implied by (is properly included in) a proposition in the partition. The partition gives the most detailed specifications of his present action over which he has control. Then this is the partition of the agents' alternative *options*. (emphasis retained)

There are two thoughts going on here. Evidently, there is ours: the action propositions are those the agent "can act at will so as to make. . . hold" with no stronger such propositions contained in them. But there is another, too; two worlds will be distinguished by this partition when the agent *acts differently* in them. Acting, I take it, should here be understood as making something true, and an agent will then "act differently" in w and v when it makes something true in w it does not make true in v, or *vice versa*. So understood, partitionality is immediate.[9]

These are not, from our perspective, the same proposal at all. For there is nothing in our basic formalism of sandwich models to compel partitionality when action propositions are understood in the first way, and yet it follows trivially from the second. If, for example, the strongest fact an agent controls in w is that they emit a pitch in the bass-baritone range, and there is a world w' in which the strongest fact they control is that they emit a pitch in the baritone-alto range, the cells $[w], [w']$ of their Lewis partition will be stronger than any facts

[9] There are hints at a similar conflation between enact*ing* and enact*ed* in other *loci classici* of action partitions, though less clearly stated. Thus Joyce can, in summarising and endorsing Jeffrey, identify actions "with propositions that a decision maker can make true or false as she pleases" even as both authors exclusively use sentences *describing an agent making something true* to denote action propositions [Joyce, 1999]. I take the Lewis passage as my main focus of interpretation because of both the foundational role of the discussion containing it and its relevant explicitness.

they control in either. Why, then, does Lewis slide between the two thoughts so readily? And why, if the cells of his partition need not be controllable by the agent, are they recommended to the agent in his decision theory? Why propose "choices" one cannot choose?

The most natural response I can give on his (and, by extension, the tradition's) behalf, though he nowhere states it overtly, is that for him *we control what it is we make true*. There is (besides the tautologous interpretation) a way of reading this as saying the same as 4, which as we have seen merely requires that one's action proposition contract while moving within it and thus brings us no closer to our solution. The intended reading is that at w the agent controls (the big conjunction specifying) *exactly* which propositions it makes (and does *not* make) true at w. Given this, the equivalence of the two definitions follows immediately.

This assumption is in effect just to accept our logic Conv45*, for the constraint it imposes on a sandwich frame is the very same (see Proposition 3.A.1). Now accepting Conv45* is not required for partitionality, since as formulated partitionality unlike Conv45* places no constraints on S_{max}. But it is the best way I can see to motivate the view from the understanding of action propositions we have adopted, and as we will see our reasons for rejecting partitionality will hinge on questioning it. But before passing to these reasons, an excursus on the semantics.

3.2.2 The Meaning of S

One problem to present itself more clearly, with the notions of action proposition and practical possibility now in view, is the intuitive interpretation of our semantics' distinctive parameters, S_{min} and S_{max}. To what real analogues do the sets they select at an index correspond? And given this, what further constraints should we impose on the interaction between S and R in supplemented frames?[10]

One captivating proposal links all three parameters S_{min}, S_{max}, and R together with a *type of actions*. In such a framework, in addition to and more basic than the *(action) propositions* an agent can *make true*, there is a type α of actions (considered as *events*) an agent can *perform*. Indeed, at any given world w there is some unique action α_w the agent

[10]Thanks to an anonymous reviewer for pressing these issues.

performs, and the proposition *that* an agent performs α_w (represented semantically by $\{w' : \alpha_{w'} = \alpha_w\}$) will be its action proposition at w; partitionality, notably, falls out of this proposal directly, given that = is an equivalence relation. Such a vision of action weds the Davidsonian belief in the primacy of events with the STIT presentation of action as making true, and is naturally captured by STIT logics directly involving act-names.[11]

This accounts for S_{min}; to get S_{max} we once again invoke practical possibility and its corresponding accessibility relation R. Clearly, at any given w, there will be some disjoint set $\mathcal{A}_w$ of action propositions available to the agent at w, corresponding to the members of α the agent can practically possibly perform at w. $S_{\mathsf{max}}(w)$ is the join of $\mathcal{A}_w$, representing the *scope* or *limits* of the agent's control at w; this guarantees it will always be a superset of the agent's action proposition, and if we assume additionally and plausibly that practical possibility is an equivalence (or even just transitive) relation we also guarantee that the frames will validate $\mathsf{Conv45}^*$, which as we saw above fits especially cleanly with partitionality.[12] Thus, we derive a neat hierarchy in the metaphysics of action, in which α underlies S_{min}, and S_{min} in turn with R underlies S_{max}, predicting the constraints desired.

Despite its tidiness, I reject this reading of the semantics. We will see in the ensuing section reason to reject partitionality altogether, but even more damningly it requires that one's S_{max} at a world be very strong, since it can include no practically impossible worlds. I seem, currently, to be making my knee rest by my shoulder, but this comes up against obvious counterexamples if we assume that practical possibility has at least the logic of $\mathsf{S4}$. After all, my knee could (in a

[11] See e.g. [Lorini et al., 2011] [Boudou and Lorini, 2018]; note that the use of action labels is in these papers developed in the context of multi-agent and collective STIT logics, which introduce subtleties well beyond our scope.

[12] Suppose the agent performs α_w at w. Clearly, for $w' \in S_{\mathsf{min}}(w) = \{w' : \alpha_{w'} = \alpha_w\}$, $S_{\mathsf{min}}(w') = S_{\mathsf{min}}(w)$ by partitionality. Moreover, by transitivity of R and $S_{\mathsf{min}}(w) \subseteq R(w)$, there are no new $v \in R(w'), \notin R(w)$, so by construction of $S_{\mathsf{max}}(*)$ as $\bigcup \mathcal{A}_*$, $S_{\mathsf{max}}(w') \subseteq S_{\mathsf{max}}(w)$. Since $S_{\mathsf{min}}(w) = S_{\mathsf{min}}(w') \subseteq R(w')$, similarly by transitivity $S_{\mathsf{max}}(w') \supseteq S_{\mathsf{max}}(w)$. Thus, $S_{\mathsf{max}}(w') = S_{\mathsf{max}}(w)$ and $S(w) = S(w')$.

Note that this validates a constraint like **Practical Strengthening** for S_{max}. For a given w and $w' \in R(w)$, $S_{\mathsf{min}}(w') \subseteq R(w')$, so by transitivity of R we have $R(w) \supseteq S_{\mathsf{min}}(w')$ for all $w' \in R(w)$, meaning also $R(w) \supseteq \mathcal{A}_w$. So, $S_{\mathsf{max}}(w) \subseteq R(w)$; in fact, if it is practically impossible at w the agent perform *no* action, $S_{\mathsf{max}}(w) = R(w)$, since then $u \in S_{\mathsf{min}}(u) \subseteq \mathcal{A}_w$ for each $u \in R(w)$. This is, recall, the same structure exhibited by the STIT analogues Ch, H of $(S_{\mathsf{min}}, S_{\mathsf{max}})$.

very weak sense) have rested by my shoulder were I to make the sun go supernova, which means S_{max} must include such worlds, even as none are practically possible for me.[13] My joint conviction both that I have no power over the sun going supernova and that I make my knee rest by my shoulder is stronger than any theoretical attractions of the above picture.

My preferred reading of the semantics rejects the original contention of action-uniqueness. If we do countenance the type α of event-actions, we should replace its corresponding *function* with a *relation* the agent bears at a given world to (possibly) *many* event-actions. This sits better with our understanding of realistic agents, like you or me. At any given world w, there are several actions we will be performing at once (typing, sitting, moving ones eyes, etc.), and we will in turn *make true* that these actions *occur*. Given our earlier axioms for Conv, these enacted propositions (about the occurrence of event-actions) determine further enacted propositions (closing under conjunctions, disjunctions, and logical intermediates), which together constitute the *convex sublattice* (in the Boolean lattice of propositions, i.e. $(\mathcal{P}(W), \subseteq)$) of facts made true at w.[14] A truly single-minded agent, who performed only one action whatsoever, would make true nothing beyond the fact of the action's occurrence.

This picture upends the hierarchy of the last. S_{min} and S_{max}, rather than playing any central role themselves, fall out as just the meet and join respectively of the "generating" action-occurrence propositions; this point is made vivid by observing that, as long as this "generating" set is infinite, the finitary closure rules corresponding to our axioms for Conv need not determine any such bound propositions,

[13] Indeed, I think no world in which I make the sun go supernova is now *practically possibly* practically possible for me, which would undercut any need for $\mathsf{S4}$ in this argument.

[14] Proving soundness and completeness for this semantics is straightforward: the soundness proof very nearly writes itself (as is common) and the completeness proof comes along for free with that for the sandwich semantics, since all sandwich models are convex lattice neighbourhood models and the new semantics therefore introduces no new validities.

The same point can be made of *algebraic* models, models whose frames each consist of a(n arbitrary) Boolean algebra representing the propositions and an operator taking each proposition p to the proposition that the agent makes p so. The construction of the semantics, as well as the soundness proof and proof that the new class of models generalises the sandwich models (thus proving completeness), is left as an exercise to the reader.

and indeed it is easy to see how our semantics might be generalised without technical loss by replacing the codomain of S with the (possibly unbounded) convex sublattices of $(\mathcal{P}(W), \subseteq)$ instead of the pairs $X_1 \subseteq X_2$ in $\mathcal{P}(W)^2$. Generalised infinitary versions of these rules, which would guarantee bounds on the sublattice at a given world, are indeed appealing independently, but again it is the closure rules giving rise to the bounds and not *vice versa*. This is a vision on which it is the axioms themselves explaining the value of the semantics, not fundamentally semantic intuitions underwriting the plausibility of the axioms.

Before proceeding, we should make note of one further link between practical necessity and control, which follows from certain suggestions made earlier. Above, we saw that **Practical Strengthening** is plausibly explained by a congruence-like principle about practical necessity, one of a family of such plausible principles: necessarily, if p and q are practically necessarily equivalent, one controls that p iff one controls that q. Note that, if a model validates this principle, it guarantees a certain connection between the sandwich function S and the accessibility relation R. Specifically, not only must $S_{\mathsf{min}}(w) \subseteq R(w)$, the complement of $S_{\mathsf{max}}(w)$ must also be contained in $R(w)$. For, if there is some $v \notin S_{\mathsf{max}}(w), \notin R(w)$, the proposition $S_{\mathsf{min}}(w) \cup \{v\}$ will coincide with $S_{\mathsf{min}}(w)$ within $R(w)$, i.e. at w the two will be practically necessarily equivalent. Thus, by the congruence-like principle, the agent will control both propositions at w, a contradiction. We will return to this point briefly in the conclusion.

3.3 Against Partitionality

3.3.1 A Problem Case

We can now proceed to the case against partitionality. As a warmup, here is a toy case that intuitively conflicts with the principle. Suppose that Alex, a man with typical human fine motor control and under ordinary circumstances, places his pen idly somewhere on a 300cm-wide square sheet of paper, leaving behind a spot of ink. Assuming for the example that the only propositions he is practically in a position to make true are those about where the pen falls on the sheet, partitionality is deeply counterintuitive here. There should be, given some minimal assumptions, a smallest region (set of points on the sheet) R

of the paper such that Alex makes the pen fall centred somewhere in R. And for every different way Alex could have acted, there will be some other, counterfactual smallest region in which he makes the pen fall. These regions are, in effect, his action propositions in the toy case.

What partitionality (as restricted, again, to propositions about where the pen lands) requires is that these regions form a partition on (a subregion of) the sheet's surface. There is something deeply strange about this. Intuitively we would think his "action region" in a given practically possible world should contain a buffer of ϵ cm around the point where the pen actually falls; ordinary motor control is not good enough to forestall minuscule deviations from one's actual path in a certain direction. And yet this would require, under partitionality, that there be some points (namely those in a given action region less than ϵ cm from its border) on the sheet Alex simply cannot touch with the pen.[15] What force would prevent him from so placing it? And even absent these margin of error considerations, whence comes this particular partition on the sheet? The whole proposal smacks of arbitrariness.

This is so far just a picture, not an argument. Alex makes many more facts true than just ones about the location of his pen on the paper, and nothing we have said precludes these stronger enacted propositions from ensuring his action propositions satisfy partitionality. Thus, for example, while the interior of the 1cm circle around a point x might be (in the world where he hits x) the smallest region in which he makes his pen fall, perhaps at each such world he *chooses* that it will fall inside the 1cm circle about the struck point (and thus *makes himself make* the pen fall there); thus, at two worlds x_1 and x_2 within 1cm of one another, while his action *regions* overlap, his action *propositions* at the two are disjoint, for one entails, and the other precludes, *that he makes the pen fall within 1cm of* x_1. This is particularly natural if we distinguish, as in the first proposal from the last section, his event-action at a world from the proposition(s) it determines; as, on this view, his event-action at a world is unique, his event-actions at x_1 and x_2 *must* be distinct (and his action propositions disjoint) if his action regions at both are not identical, since his action fixes what he makes true. It thus remains to convert the above picture into a tighter

[15]This framing is, of course, partly adapted from the discussion of "luminosity" in [Williamson, 2000]. We will not in the main argument, however, need to appeal to the kinds of margin of error principles central to that discussion.

argument.

The tip of Alex's pen, we can stipulate, is centred on a point η near the middle of the sheet. In such a case, the following all seem true:

3. It is practically necessary that, if the pen falls within 500,000 μm of η on the sheet, Alex makes it appear somewhere on the sheet
$$\boxdot(q_{500,000} \rightarrow \Box p)$$

4. Alex makes the pen fall on the sheet less than 500,000 μm from η
$$\Box(p \wedge q_{500,000})$$

5. Alex does not make the pen fall on the sheet less than 1 μm from η [16]
$$\neg\Box(p \wedge q_1)$$

Plus the following family of premises indexed to $0 < i \leq 500,000$:

C$_i$ Alex does not make it so that he both makes the pen fall less than i μm from η on the sheet and does not make it fall less than $i-1$ μm from η on the sheet[17]
$$\neg\Box(\Box(p \wedge q_i) \wedge \neg\Box(p \wedge q_{i-1}))$$

(4) and (5) require little comment. (5), in particular, simply follows from the description of the case, plus a lack of superhuman precision. (3) is justified by supposing that only Alex would put the pen on the sheet. The real work, clearly, is to be done by (C$_i$), which imposes a limit on Alex's *higher order control*, his control over *what* he controls.

It is a trivial consequence of (4) and (5) that there is some i such that Alex makes the pen fall less than i, but not less than $i-1$, μm from η; assuming that, for all $j < k$, if it falls less than j μm it falls less than k μm (which follows by **Convexity** if their disjunction is equivalent to the second disjunct), this i will indeed be *unique*. $\Box(p \wedge q_i)$ and $\neg\Box(p \wedge q_{i-1})$ then are a (the) limit to Alex's fine motor control over where the pen falls on the paper, up to micrometres from η. What (C$_i$)

[16] For the argument immediately at hand we could replace 1 with 0, leaving the uncontrolled proposition contradictory, though this (slightly) stronger concession will be dialectically useful in discussing the models of the next section.

[17] This assumption bears a close and nonaccidental resemblance to the gap principles known in the theory of vagueness due to the late Graff Fara [Fara, 2003] [Bacon, 2022].

says is that the degree of Alex's fine motor control thus expressed is not *itself* something he controls. This claim, that ordinary humans like Alex do not fix the precise limits of their own motor control in action, seems eminently plausible.

Note that (C_i), as an expression of higher order non-control, is (when conjoined with Alex's limit $\Box(p \wedge q_i) \wedge \neg\Box(p \wedge q_{i-1})$) simply a negated instance of $\mathsf{5}^*$. This provides a deeper insight into $\mathsf{Conv5}^*$: it is the logic of agents with perfect higher-order control, and logics imposing $\mathsf{5}^*$ (like that of [*dstit*]) can only describe such idealised agents.

These assumptions are collectively incompatible with partitionality:

Proposition 3.3.1. *There is no supplemented sandwich model in which (3 - 5) and all (C_i) are true at a world w while the actions at w are pairwise disjoint.*[18]

A simple proof of this fact is given in the appendix, but the thrust behind it should be easy to see. Throughout the inner circle of w, p is made true, and thus when propositions stronger than p contain this inner circle (assuming the inner circle function is constant throughout the inner circle of w, which is forced by partitionality), they are made true throughout the circle. This will include both $\Box(p \wedge q_i)$ and $p \wedge \neg\Box(p \wedge q_{i-1})$, and so also their conjunction, which directly conflicts with (C_i).

There is an obvious objection to this argument. Many philosophers who appeal to the concept of an action proposition think that, in the relevant sense of *making true*, we make true only much weaker, more "internal" propositions, about (say) the state of our volition and not what appears on sheets of paper.[19] Thus the objector will not concede (3) or (4), undermining the argument.

[18] As to why the argument is formulated in the metalanguage in terms of models rather than directly in our object language itself, see fn. 3.

[19] Saith Joyce: "My inclination is to [...] use the term *act* narrowly to denote *pure, present exercises of the will*. Many things we ordinarily call acts do not count as such under this reading. Walking to work, for example, is not really an act one can choose to perform because, unfortunately, one cannot simply will it to be the case that one's legs function properly, that one will not be shot by a madman on the way out the front door, that a great chasm will not open in the earth to prevent one from reaching one's destination, or even that one's 'future self' will carry through on one's choice." See also Ford's discussion of "volitionalism" in [Ford, 2018] and the discussion in [Lederman and Holguín, 2023] of "tryings" in decision theory.

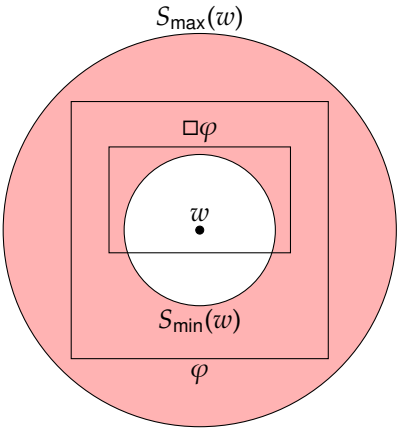

Figure 3.1: An illustration of a model where making $[\![\varphi]\!]$ true (i.e. $[\![\Box\varphi]\!]$) is stronger than $[\![\varphi]\!]$, and is not itself made true at w. For exactly those $v \in S_{\min}(w)$ "below" the rectangle intersecting the inner circle, $S_{\min}(v) \not\subseteq [\![\varphi]\!]$

But there is a simple fix. If we grant, for example, that Alex makes true that the pen *would* fall less than 500,000 μm from η, were the ordinary and limited causal dependence between the state of the paper and Alex's will to hold. With these intuitions in mind, letting o stand for *the ordinary and limited causal dependence obtains* and $\Box\!\!\rightarrow$ for the subjunctive conditional, replace (3), (4), (5), and (C_i) with

- $\boxdot((o \mathrel{\Box\!\!\rightarrow} q_{500,000}) \rightarrow \Box(o \mathrel{\Box\!\!\rightarrow} p))$
- $\Box(o \mathrel{\Box\!\!\rightarrow} (p \wedge q_{500,000}))$
- $\neg\Box(o \mathrel{\Box\!\!\rightarrow} (p \wedge q_1))$
- $\neg\Box(\Box(o \mathrel{\Box\!\!\rightarrow} (p \wedge q_i)) \wedge \neg\Box(o \mathrel{\Box\!\!\rightarrow} (p \wedge q_{i-1})))$

respectively. Under minimal assumptions about the subjunctive conditional $\Box\!\!\rightarrow$ (in particular that $o \mathrel{\Box\!\!\rightarrow}$ is a normal operator, so that the counterfactual consequences of o are closed under weakening, agglomeration, and necessitation), the argument will proceed just as cleanly as before.

Why accept that Alex makes true that the pen would fall less than 500,000 μm from η, were the ordinary and limited causal dependence between the state of the paper and Alex's will to hold? Might this not be objectionably "external" to Alex? We should accept it because, as long as this dependence hypothesis is sufficiently detailed and strong, the counterfactual claim should be *equivalent* to Alex's having thus-and-such a volitional state. We can interpret the counterfactual conditional here as holding fixed similarity of inner volitional state, so that the claim simply states the consequences of his actual volitional state relative to whatever condition is specified in the antecedent of the conditional. The claim thus will follow from some relatively uncontentious congruence-like principle for control, specified in terms of some very strong form of necessary equivalence in the antecedent (making the principle itself very weak).[20]

It might be wondered whether this feature of Alex's situation somehow depends upon the infinitude or non-discreteness of his space of possible actions. It does not. To see this, replace him putting his pen on a sheet of paper with turning a certain dial between the minimum and maximum settings, fixing the volume of some audio output in the natural way. Once turned at all, it will output the audio at the specified volume; if not touched, it will not output audio at all. The dial, we may stipulate, can be initially adjusted to any real valued angle between its minimum and maximum angles, but will then immediately move forward or backward to rest at the nearest of very but finitely many evenly spaced increments between the minimum and maximum. The number of these increments and the size of the dial, we may suppose, are large and small enough, respectively, that Alex cannot control at exactly which of the increments the dial comes to rest, though it inevitably will come to rest at one. It is easy to adjust our premises to this alternative setup, and we shall occasionally refer back to this discrete alternative.

3.3.2 Modelling Clumsy Agency

Of course, this on its own is only half of an argument against partitionality; to dispel worries about the assumptions going into it, we must

[20] Another very different line of response to a similar objection to a similar argument will be described later in §4.4.2.

Model 3.1: Naïve representation of Alex

W	$[0,300]^2_{x,y} \subset \mathbb{R}^2$
R	$\{(w,v) : w,v \in W\}$
$S_{\mathsf{min}}(w)$	$\{v : \lvert w_x - v_x\rvert^2 + \lvert w_y - v_y\rvert^2 < 1\}$
$S_{\mathsf{max}}(w)$	W
$V(p\,;\, q_i^w)$	$W\,;\, \{v : \lvert w_x - v_x\rvert^2 + \lvert w_y - v_y\rvert^2 < i/10,000\}$

provide an at least somewhat realistic supplemented model validating those assumptions.

Our initial discussion of Alex suggests one such. Have the domain of the model be the real coordinates on the sheet, the accessibility relation R the universal relation, any point's inner circle the points with 1cm of it, and the outer circle the whole sheet.

The model then interprets worlds as coordinates on the sheet of paper on which to place the pen, takes it to be practically possible to place the pen anywhere, and has it that Alex's action proposition at w is the disc of worlds less than 1cm from w while he makes it true everywhere that the pen falls somewhere.[21] V gives our propositional constants p and q_i^w (for $w \in W, i \in \mathbb{N}$) their natural interpretations, with q_i^w standing for the analogue with respect to w of q_i for η.

While this indeed validates all our assumptions, it suffers from serious problems of unrealism. For while it does validate ($\mathrm{C}_{10,000}$) (the relevant such premise) at η, it affirms his perfect higher order control in a deeper sense. In this model, the disjunction

$$\mathbf{D}_{10,000} \bigvee_{w \in W} (\Box(p \wedge q^w_{10,000}) \wedge \neg\Box(p \wedge q^w_{9,999}))$$

is true everywhere, and as $S_{\mathsf{max}}(w) = W$ for all w, Alex makes true this disjunction, effectively controlling his precise degree of motor control considered independently of where he actually places the pen (his *location-neutral* such degree of control). It is tempting to respond that, as a toy model, we should expect such artificialities, but this would be to selfishly excuse the very flaw we of which we accused the friends of partitionality. Denying ($\mathrm{C}_{10,000}$) is bad; this looks even worse.

[21] The choice of 1cm is of course simply for convenience; feel free to replace it with some more plausible distance if desired.

Model 3.2: Adding a degree-of-control parameter

W	$([0,300]^2_{x,y} \times (0,300]_d) \subset \mathbb{R}^3$
R	$\{(w,v) : w,v \in W\}$
$S_{\min}(w)$	$\{v : \lvert w_x - v_x\rvert^2 + \lvert w_y - v_y\rvert^2 < 1, (d - \frac{1}{2}h) \leq v_d \leq (d + \frac{1}{2}h)\}$
$S_{\max}(w)$	W
$V(p\,;\, q_i^w)$	$W\,;\, \{v : \lvert w_x - v_x\rvert^2 + \lvert w_y - v_y\rvert^2 < i/10,000\}$

An initial impulse might be to add a third parameter d of *degree of control* into the indices in the domain, represented by $(0,300] \subset \mathbb{R}$ as the radii of the action propositions for each degree. Thinking of the domain as a 300cm^3 cube with the point-thick top layer shaved off, at a given point (x,y,d) one's action proposition would be the interior of a cylinder of radius d of height h centred about its vertical axis (plus its top and bottom).

This would make good on the idea that Alex's is but one possible degree of motor control among many. But, however implemented, it introduces more problems than it solves.[22]

For one, it guarantees by the same reasoning Alex's yet higher order control over his *control over his degrees of control itself*, control even *more* dubious than the higher order control the model was just revised to deny, thus simply kicking the problem up one level. Let us be clear about what this means. It does *not* mean that, if one's degree of control is d, that one thereby controls that one's control is d; this will fail whenever $h > 0$. It also does not mean that we must reject "margin of control" principles for d, *à la* (C_i). Indeed, such principles are guaranteed to hold, again, whenever h in a model is nonzero. Instead, the objectionable consequence is an analogue to $(\mathrm{D}_{10,000})$, letting r^w_z stand for the agent's degree of control falling within z centimetres of their degree of control at w and ϵ for some suitably small number:

$$d\mathbf{D}_{\frac{1}{2}h} \bigvee_{w \in W} (\Box(p \wedge r^w_{\frac{1}{2}h}) \wedge \neg\Box(p \wedge r^w_{(\frac{1}{2}h)-\epsilon}))$$

This follows because, again, the disjunction is true everywhere, and the agent controls the universal proposition W.

[22] The drawbacks to be discussed presently provide nice illustrations of the perils of "overfitting" by multiplication of degrees of freedom denounced in [Williamson, 2024].

We could, of course, introduce higher levels of control infinitely or indefinitely, but even barring the obvious problem of his control over his *infinitely* higher order control, this method of amending the model has the perverse quality of making yet more unreasonable claims of Alex at each finitary step. It abandons the frying pan by way of an infinitely nested worsening series of fires.

Second, implementing it usefully requires abandoning the very intuitions supporting (C_i) in the first place. Our first suspicion about Alex is that his practical possibilities are limited to ones with his current degree of control, that he has no higher control over this degree, and that these facts are mutually reinforcing rather than in tension. But here they are placed precisely in *conflict*. If Alex's practical possibilities all share his actual degree of motor control (that is, wRv only if $w_d = v_d$), then by **Practical Strengthening**, **Convexity**, and $S_{\mathsf{max}}(w) = W$, Alex controls his exact degree of control (in sandwich speak, $S_{\mathsf{min}}(w) \subseteq \{v : v_d = w_d\} \subseteq S_{\mathsf{max}}(w)$; pictorially, his cylindrical action proposition would be flattened to a disc). A less toyish model will of course require multiple degrees of control available at a given coordinate of the page, but this crude introduction of them does more harm than good.

We could also let h vary from world to world, so that the "height" of the cylinder for w might differ from that of w'. Represent the function from worlds to heights by f. Then we still get

$$d\mathbf{D}'_f \bigvee_{w \in W} (\Box(p \wedge r^w_{f(w)}) \wedge \neg\Box(p \wedge r^w_{f(w)-f^*(w)}))$$

where f^* is chosen suitably.[23] It is, firstly, unclear that this marks a great improvement: it still means that the agent controls what their level of control is at each possibility, even if this level of control is permitted (by them!) to fluctuate from possibility to possibility. And, what is more, it starts to look like multiplying degrees of freedom unduly. If there is a better alternative, we should take it.

A better, and simpler, approach indeed starts by adding an additional world representing the possibility of simply not placing the pen on the sheet at all. We take crucial advantage of the convex non-normality of our operator: this world's singleton is its own inner and outer circle, while for the rest we have $S_{\mathsf{max}}(w) = [0, 300]^2$ and S_{min} as

[23] This will be possible as long as h is always nonzero; if it is ever zero this introduces new problems, discussed immediately above.

Model 3.3: Adding a paper-avoiding possibility

W	$[0,300]^2_{x,y} + 1$
R	$\{(w,v) : w,v \in W\}$
$S(1)$	$(\{1\},\{1\})$
$S_{\mathsf{min}}(w \in [0,300]^2)$	$\{v : \lvert w_x - v_x\rvert^2 + \lvert w_y - v_y\rvert^2 < 1\}$
$S_{\mathsf{max}}(w \in [0,300]^2))$	$[0,300]^2 \subset \mathbb{R}^2$
$V(p\,;\, q_i^w)$	$[0,300]^2 \subset \mathbb{R}^2\,;\, \{v : \lvert w_x - v_x\rvert^2 + \lvert w_y - v_y\rvert^2 < i/10,000\}$

before. When hitting the sheet, that is the least he controls; when not hitting it, the same. R is, again, universal.

We still have, of course, Alex controlling ($\mathrm{D}_{10,000}$) as long as he hits the sheet, but in this setting we should take pains to distinguish ($\mathrm{D}_{10,000}$) from the following conditional claim (for some appropriate kind of conditional $\dashrightarrow$, be it material or counterfactual or strict).

$$\dashrightarrow \mathbf{D}_{10,000}\ \ p \dashrightarrow \bigvee_{w \in W} (\Box(p \wedge q^w_{10,000}) \wedge \neg\Box(p \wedge q^w_{9,999}))$$

In our earlier model, these were both simply the trivial proposition, because p was true everywhere, but here only the latter is trivial, and only the former is made true (a distinction made possible by the peculiarities of our logic for $\Box$). ($\dashrightarrow \mathrm{D}_{10,000}$) is equivalent to the trivial $p \dashrightarrow p$, because its consequent ($\mathrm{D}_{10,000}$) is equivalent to p. Alex does not control it because it is too weak and necessary for him to do so, not because it is too strong and contingent.

It is this disjunction, I think, that represents in this setting the location-neutral degree of Alex's motor control considered as a *fixed capacity* of his, the sense in which his controlling it strikes us as absurd. He of course controls the degree of control he exercises in placing the pen on the sheet (in the sense of making true ($\mathrm{D}_{10,000}$)), but this is simply because for an agent with his particular limited motor capacities to place a pen on the sheet is just to place it while exercising exactly those limited capacities; for a helplessly clumsy thrower, in the same vein, to toss a ball just is to toss it clumsily, and indeed to toss it with their exact kind of clumsiness. Over his background musculature and neurology, over the dispositional sensitivity of his pen to mild pressure and sudden movements, Alex thereby need exercise no control at all.

Model 3.4: Degrees of control varying by location

W	$[0,300]^2_{x,y} + 1$
R	$\{(w,v) : w,v \in W\}$
$S(1)$	$(\{1\},\{1\})$
$S_{\mathsf{min}}(w \in [0,300]^2)$	$\{v : \lvert w_x - v_x\rvert^2 + \lvert w_y - v_y\rvert^2 < f(w)^2\}$ for $f : [0,300]^2 \subset \mathbb{R}^2 \to [1,1.5] \subset \mathbb{R}$
$S_{\mathsf{max}}(w \in [0,300]^2))$	$[0,300]^2 \subset \mathbb{R}^2$
$V(p\,;\, q_i^w)$	$[0,300]^2 \subset \mathbb{R}^2$; $\{v : \lvert w_x - v_x\rvert^2 + \lvert w_y - v_y\rvert^2 < i/10,000\}$

If the claim that Alex controls his exercised degree of control down to the micrometre still bothers you, we can introduce still more relevant verisimilitude by rejecting its constancy across the sheet. We could instead have, say, the inner circle of w be all points within a random $f(w)$ between 1cm and 1.5cm. $f(w)$ represents the motor control Alex *would* exercise at w specifically.[24]

This respects the intuition that the degree of motor control Alex would exhibit in placing his pen down somewhere on the paper is, to some extent, itself variable; we in our bodily motions like placing down a pen leave up to chance, not just where the pen lands, but how much control we exert over this. This leaves, of course, him controlling the more varied disjunction

$$\mathbf{D}'_{10,000} \bigvee_{w \in W} (\Box(p \wedge q^w_{10000f(w)}) \wedge \neg\Box(p \wedge q^w_{10000f(w)-1}))$$

but there is something less strikingly impressive about this feat. He controls his (location-neutral, exercised) control down to the micrometre, but not to a *consistent* micrometre.

Do these remarks undermine the plausibility of (C_i)? No. For what renders unthreatening Alex's control over ($D_{10,000}$) is its practical equivalence with an obviously easy proposition for him to make true, namely that he hits the paper at all. For him to make true the exact boundary of his control in hitting η (that is, for the relevant instance

[24] This assumes a restricted version of counterfactual excluded middle. For detractors who take it the counterfactual degree of control exercised at a given coordinate is indeterminate, add to the domain for each relevant counterfactual possibility a copy w of that coordinate with the appropriate $S_{\mathsf{min}}(w)$.

of (C_i) to be false), however, would correspondingly be for him in hitting η to possibly control down to a micrometre's variation *where exactly he actually hits the paper*, which has little pre-theoretic intuitive plausibility (whence our conviction in (5)).

Model 3.3 straightforwardly satisfies for each candidate for η all our earlier assumptions. For Model 3.4, for an evenly distributed finite set X of points within $S_{\mathsf{min}}(w)$ the probability that, for all $x \in X$, $S_{\mathsf{min}}(x) \subseteq S_{\mathsf{min}}(w)$ approaches zero as $|X| \to \infty$ given natural assumptions about the probability distribution on the values of $f(w)$, so that almost certainly our assumptions from the last section are true almost everywhere.

3.3.3 Whither 4?

The models of Alex just considered patently fail to validate **4**, the principle that making true iterates. This failure raises two natural questions, given the conceptual importance of **4**: (a) do there exist (supplemented sandwich) models validating both **4** and our premises, and, (b) if so, are these models realistic? The answers are, respectively: unambiguously Yes, and tentatively No.

First, the narrowly formal question of model existence. It will be easiest to see this by considering the discretised dial version of Alex's situation (a slightly more complicated model in the same spirit for the original pen-and-paper setup is detailed in the appendix).

Theorem 3.3.2. *There is a model satisfying both* **4** *and the premises (for arbitrary η).*

Proof. Where n is the number of increments on the dial past the minimum, we let the domain be the Cartesian product of the sets of naturals beneath n inclusively and strictly, respectively, plus a further point 1 outside. We begin by partitioning the product, such that (m, m) belongs to a cell containing (m, l) for all $m \leq l < n$ and $(m + 1, k)$ for all $m \geq k \geq 0$ (this is readily seen to in fact be a partition). At each pair (i, j) in the product (aside from those in the furthest right cell $[(n-1, n-1)]$), $S_{\mathsf{min}}(i, j)$ is the set of pairs in $[(i, j)]$ with the second coordinate greater than or equal to j (for $(i, j) \in [(n-1, n-1)] = [(n, n-1)]$, it is those in the cell with second coordinate equal to or *less* than j), and $S_{\mathsf{max}}(i, j)$

the entire product.[25] The upper and lower limits of the sandwich at 1 are just {1}, and the accessibility relation for practical possibility is universal.

That the model satisfies the desiderata is easy to check. □

This verbal presentation is somewhat cumbersome, but a visual illustration of the model is presented as Figure 3.2. Since (as restricted to the Cartesian product component) making true is effectively a normal operator in the model, it can be represented by an "accessibility relation" on the domain; since this accessibility relation will be guaranteed transitive, rather than putting an arrow between any two worlds in which one "sees" the other, one can use arrows to indicate the relation by depicting a weaker relation whose transitive closure is the desired quasi-accessibility relation.

The product, of course, represents those possibilities in which Alex turns the dial and there is any audio output, and the further point the possibility in which he does not. (The latter is not necessary for the logical result at issue, but it helps in trying to interpret the model realistically.) The first coordinate represents the position of the dial (and the volume of the audio output), so that at (i, k) the dial is at the ith increment. At the pairs on the diagonal, Alex makes the dial to rest at either its actual increment or one higher, and makes nothing stronger true about the dial's position. It is easy to check that, on the diagonal, our desired principles all come out true, plus the iteration axiom **4**. In particular, along the diagonal, Alex does not control the limit of his control over the dial, for it is consistent with his action proposition that he control the dial's position exactly.

While this model certainly validates the desiderata (at least on the diagonal), it has no clear realistic interpretation. What, if anything, is the second coordinate in the product supposed to measure?[26] Not

[25] The strange "reverse ordering" feature of the rightmost cell $[(n-1, n-1)]$ is absent in the more complicated model given in the appendix, but is essential in the current construction for the premises to all be satisfied throughout the diagonal. For, were the "ordering" of $[(n-1, n-1)] = [(n, n-1)]$ not "reversed," at $(n-1, n-1)$ we would have Alex's S_{min} and S_{max} equal, so that we would at the uppermost member of the diagonal violate the relevant instance of the analogue of (C_i). Indeed, in this non-"reversed" construction, we would have the relevant (C_i)-analogue false *throughout* $[(n-1, n-1)] = [(n, n-1)]$, and thus at (n, i) for all i.

[26] Again, as in the discussion above of the appropriate model for Alex's placement of the pen in §3.3.2, Williamson's warnings in [Williamson, 2024] are especially apt.

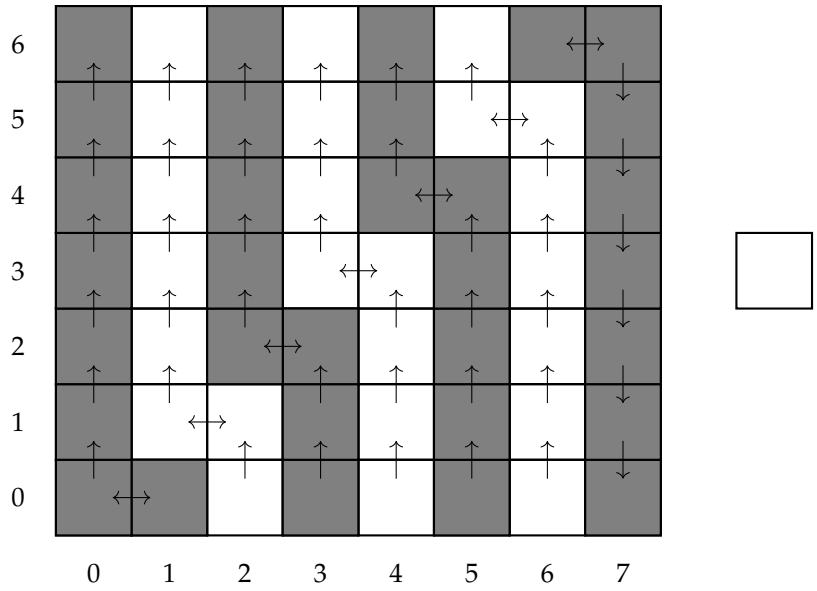

Figure 3.2: The 4 model for the dial for $n = 7$, with the cells of the partition on the Cartesian product component in alternating grey and white, and the transitive "accessibility relation" for making true as confined to the product component indicated with arrows.

Alex's degree of control, surely, for this not only would mean that exact control would become harder to exert the further right Alex tilted the dial, but would make it practically possible for him to exert *exact* control – which for large enough n and a small enough dial will be very implausible. (And this is setting aside the peculiarity of the inverted order of the final cell in the partition.) These obstacles give less the impression of artefacts of a simple model to be overcome with more sophisticated refinements and more the impression of clues that, as a realistic presentation of Alex's situation, the model is a dead end.

Of course, on its own this failure tells us little. That one model validating **4** should be inadequate to the task does not mean all such must be. Perhaps, with sufficient ingenuity, a plausible candidate along different lines could be produced.

But I think this is unlikely, because of the following argument. Return to the case of the pen and paper, and suppose Alex makes the pen fall within 1cm of η but does not make it fall within $x$cm of η for $1 > x$ (i.e. 1cm is the limit of his control). Then any possibility within his action proposition must be one where he also makes the pen fall within 1cm of η, on account of **4**. But it cannot be that at every such possibility he fails to make it fall within .9999cm of η, else throughout his action proposition it would be true that the pen falls on the paper and that he fails to make it fall within .9999cm of η, and thus he would make true the conjunction: that the pen falls on the paper and he fails to make it fall within .9999cm of η. Therefore, as he also makes true that he makes true that it falls within 1cm of η (by **4**), by agglomeration he would violate the relevant (C_i). So, there must be some (practical) possibility w within his action proposition where he makes the pen fall within .9999cm of η. If the (C_i) are all true at w, this in turn implies the existence of a practically possible w' within his action proposition at which he makes the pen fall within .9998cm of η, and so on and so forth. To avoid contradiction, there must therefore be some practical possibility, not excluded by his action, in which some (C_i) is violated.

This result (that some (C_i) is practically possibly violated) strikes me as implausible. And, by replacing the (C_i) with analogues formulated in terms of ever smaller units of measurement, the above

Of all degrees of freedom in a model introduced to salvage some general principle or case judgement, thoroughly uninterpretable such degrees are surely the most suspect. Similarly ill-motivated additional parameters appear in the **4**-validating models of the pen and paper presented in the appendix.

argument can easily be generalised to one that his action is compatible with *arbitrarily* fine-grained higher order control.[27] Perhaps I should not trust my intuitions about the (C_i) and like principles to extend to their practical necessity, but these results still strike me as serious costs. Combined with the artificiality of the obvious models validating it, they spell poor prospects for the 4 axiom as a principle in the logic of action.

Indeed, it is possible to extend this reasoning to an argument that, under ordinary circumstances, one's action proposition *never* contains the action propositions of its member worlds all nested within (stronger than) it. In the case about Alex and the pen just described, for example, Alex makes true that he makes true p, the proposition that the pen falls on the paper. Now, in the above argument, all the relevant propositions he is successively shown to make true (that the pen falls within $(1 - .0001j)$cm of η, for $0 \leq j \leq 10000$) are stronger than p (trivially so, for the final one). Thus, if every w within his actual action proposition $A_@$ has its action proposition A_w contained in $A_@$, if he makes one of these propositions about the pen true he makes true that he makes it true, as for every $w \in A_@$ this proposition will be intermediate in strength between A_w and p. This allows the *reductio* to be run again, deriving a contradiction. So, if the above reasoning holds, we should also conclude that our action propositions in ordinary circumstances *never* nest in the way 4 says they *always* do.

What is at stake here? Partly 4 is simply of interest intrinsically: its truth (and necessity) are among the most obvious questions once we allow the notions of control and making true to iterate. Partly it is of interest in pinning down the structure of an agent's space of action propositions. Were 4 to hold, as a matter of the logic of making true, it would mean that action propositions must be, if not all mutually incompossible, at least *nested* in logical space: for any action proposition p and any $w \in p$, the action proposition of w would be entirely contained in p, guaranteeing some nontrivial logical structure to the space of an agent's possible actions.

There are more interesting consequences still from the extension of our argument to the claim that under ordinary circumstances, one's action proposition never contains nested within it the action propo-

[27] The appendix gives models allowing for both infinite and merely indefinite precision compatible with any action proposition.

sitions of its constituent worlds. It shows, for one, that we cannot use our arguments for the *overlap* of action propositions to let *nested* action propositions play certain theoretical roles. Jeffrey, for example, suggests in passing we might understand the trivial proposition $\top$ as a degenerate sort of action, inaction-as-action:

> An act is then a proposition which is within the agent's power to make true if he pleases, and the necessary proposition would correspond to not acting: to letting what will be, be. But deliberate inaction is a form of action in an attenuated but useful sense of that term, and accordingly we might think of $[\top]$ as one of the available acts; certainly, noninterference with the status quo may be one of the courses open to the agent, and to call that course a course of action is not to stretch meanings beyond ordinary usage. [Jeffrey, 1965]

This goes hand in hand with a vision on which one can influence the outcome of a process with more or less specificity (a perfectly normal situation), such that one's action proposition in the case where one influences the process more precisely is *contained in* (entails, is stronger than, etc.) the action when one influences it more crudely. That is, the familiar and mundane phenomenon of being able to exercise stricter or weaker control over a process is adequately modelled by a lattice of action propositions describing the outcome of that process. This lattice is ordered by strength, and terminates in the very weakest proposition to entail that the process occurs at all. If our argument against "nesting" is correct, this role is unavailable to action propositions.

Beyond these first-order issues, however, there is an important dialectical point. The existence of models for Conv4+ validating the premises of our argument against partitionality shows how little the argument requires: one does *not* need a fully general argument against the "MM" principle to draw the desired conclusion.[28] This marks an important structural difference between our argument and similar ones against contentious introspection conditions in epistemology.

[28] On analogy with the "KK" principle: that, if an agent knows p, it knows itself to know p.

3.4 Non-Partitionality and Decision Theory

Supposing the arguments of the last section are correct and the action propositions of an agent do not partition any subset of logical space, the question arises as to what effect this failure of partionality might have on the standard machinery of those varieties of decision theory in which the concept an action partition traditionally plays a part. We now turn to some potential such effects.

3.4.1 Representation Theorems

It is first worth noting one area where the failure of partitionality is *not* directly formally relevant to decision theory as normally understood: the representation theorems for evidential and causal decision theory due to Bolker and Joyce respectively [Bolker, 1966] [Bolker, 1967] [Joyce, 1999]. Neither proof rests on any special structure of the *action* propositions specifically; rather, in each case the proof considers them merely as elements in a larger class of *conditions* over which the agent may have preferences, and which is not assumed to possess a partitional structure. The proofs then move from qualitative rankings of comparative likelihood and desirability on the conditions and states of the world, respectively, to the existence of a *quantitative* pair of credence and utility measures. At no point does the partitionality or lack thereof among an agent's actions enter formally into the proof.

Thus, the rejection of partitionality is shown to be insignificant from the perspective of the representation theorems. This supplements our earlier conclusions, by indicating the limited relevance of partitionality to the crowning technical jewels of decision theory, in both its evidential and causal variants. To the extent that these representation theorems are taken to lend authority to the theories of practical rationality whose formal adequacy they in some sense purport to guarantee, this does not extend to the partitionality of actions, even in the case of causal decision theory where these partitions have played so large a historical role in the theory's motivation and exposition. If there are reasons to endorse partitionality, they must be provided elsewhere.

3.4.2 Grand and Small Worlds

An important distinction in the practical application of decision theory, introduced in [Savage, 1954], is that between the *grand world* (or, we might say, the *global decision problem* for an agent) and any number of *small worlds* (or, the *restricted decision problems* for the agent). We will come to a more formal definition momentarily, but for now it will be worthwhile to motivate the distinction intuitively.

The grand world, or global decision problem, consists of the full range of maximally specific acts available to an agent, some partition of event space into maximally specific "states" of the world, a specification of the consequences to attach to any available act in any such world-state, and the desires and beliefs of the agent, which jointly are supposed to determine a preference ordering among the acts. One of the small worlds, by contrast, will be a restriction or coarsening of the grand world, in terms of both its acts and its world-states, while preserving the grand world's beliefs and desires and (if all goes well) likewise inducing a preference ordering on the coarser acts. The former represents the problem of practical rationality, as it confronts a given agent, unrestrictedly speaking: what, in the widest possible sense and in the most specific terms, ought an agent to do, given all they believe of and want from the world in its entirety? The latter represents the problem of practical rationality confronting the agent in some narrower domain, as in (for example) the stock examples of the extorted driver in the dodgy neighbourhood, the beneficiary of the Newcombe predictor-cum-philanthropist's generosity, etc. It is the distinction between the agent's true situation, as a practical reasoner, and some relatively narrow perspective on the same.

There is another way to think of the grand-small world relation, emphasised by Joyce [Joyce, 1999]. Rather than thinking of the grand world as basic and the small world derived from it, one thinks of successively "grander" worlds as extensions of a given small world, as the agent considers more and more about the facts facing him. Where once he saw one possibility, now he sees several different ways it might be realised. Hence why the new "grander" state and action partitions will refine the old "small" ones. The grand world proper will be a sort of limit of this process, an "end of inquiry" representing a fully thought-through decision problem.

As recognised by Savage, the techniques of decision theory, how-

ever construed, are pretty poor equipment to tackle the problem of the grand world.

> Carried to its logical extreme, the "Look before you leap" principle demands that one envisage every conceivable policy for the government of his whole life (at least from now on) in its most minute details, in the light of the vast number of unknown states of the world, and decide here and now on one policy. This is utterly ridiculous, not as some might think because there might later be cause for regret, if things did not turn out as had been anticipated, but because the task implied in making such a decision is not even remotely resembled by human possibility. It is even utterly beyond our power to plan a picnic or to play a game of chess in accordance with the principle, even when the world of states and the set of available acts to be envisaged are artificially reduced to the narrowest reasonable limits.
>
> Though the "Look before you leap" principle is preposterous if carried to extremes, I would none the less argue that it is the proper subject of our further discussion, because to cross one's bridges when one comes to them means to attack relatively simple problems of decision by artificially confining attention to so small a world that the "Look before you leap" principle can be applied there.

It is no accident that the standard examples of decision theory are all alike small worlds, presenting some small range of coarsely specified options (say: {take one box, take two boxes}) and their would be consequence in some similarly widely individuated and small class of world-states ({the first box is full, the first box is empty}). Both the full ambitions and natural formulations of the decision theorist, however, are generally nevertheless about the grand world itself rather than any small world derived from it. It is in this sense that the decision theorist is standardly a theorist about practical rationality *as such*, and not about rationality in some special domain or domains. This poses a foundational problem for the decision theorist: to specify the relationship between the global decision problem (which in an important sense is his true object of study) and its restricted decision problems

(which are where his formalisms will be really applicable).

Now, in the initial machinery due to Savage, the mediation between grand and small worlds is a quite heavy-duty affair, in which small worlds are constructed logically from a base grand world. Not only are the relevant definitions very involved and the proofs guaranteeing their propriety largely nontrivial, but these proofs depend upon assumptions about the space of an agent's potential actions which are not in fact realistic. For example, they require as part of the background theory of grand worlds, for any possible outcome X (understood in Savage's framework as one of a range of states independent of the world-states relative to which an act will *result* in a given outcome), a "constant" act $\mathbf{f}_X$ (understood as a function from states to worlds) returning X for every world-state in the state partition. These difficulties are downstream of his account of acts as functions on states, and his attendant sharp distinction among acts and states and consequences, which both complicate the formulation of the theory and invite awkward assumptions about the structure and range of these three components of a given decision problem.

There is other clunkiness in the theory besides. For example, while it is possible to talk for any small world S relative to the grand world G about a "small small" world S' relative to which S will play the role of grand world, S' will, somewhat counterintuitively, *not* directly be a small world relative to G.[29] (This is particularly ill-fitting with the "end of inquiry" understanding of the grand world described earlier.) And after all, is it so obvious there *must* be a grand world at the bottom of such a regress, as Savage's construction requires? Could it not be that there is an infinite hierarchy of ever grander, more finely granular worlds extending any given small world?[30] Certainly it would be nice not to inject into one's decision theory any premature assump-

[29]This is because the small world outcomes are defined as grand world acts (that is, functions $f : S_G \to O_G$ from global states to global outcomes), and the small world acts functions from small world states to grand world acts (that is, functions $f' : S_R \to (S_G \to O_G)$ from restricted states to functions from global states to global outcomes).

Thus, the "type" (as it were) of "small small" world acts is different from that of small world acts: not functions $f' : S_R \to (S_G \to O_G)$ from restricted states to functions from global states to global outcomes, but functions $f'' : S_{R'} \to (S_R \to (S_G \to O_G))$ from yet further restricted states to functions from restricted states to functions from global states to global outcomes.

[30]In this context see the literature on "possibility semantics" [Humberstone, 1981] [Holliday, 2024].

tions about such metaphysical questions. While he deserves credit for recognising the problem of the grand/small world distinction, in many ways Savage's attempts to address it obscure as much as they illuminate.

It was one of the most striking advances by Jeffrey's interpretation of acts as propositions that it opened the way to a far simpler and more elegant understanding of this distinction. As Joyce explains the matter in [Joyce, 1999], it will suffice to think of small worlds as ones whose acts and states are simply *joins* of grand world acts and grand world states. Put differently, in such a framework we may consider the state and action sets of the grand world simply as (the atoms in) independent atomic subalgebras $\mathcal{A}$ and $\mathcal{S}$ of the Boolean algebra of all possible events, modelled as (say) a σ-algebra restricting the power set algebra of all possible worlds, whose atoms are the action propositions and states, respectively. A small world, considered as a pair of an action set and state set, will be just a pair of algebras $(\mathcal{A}', \mathcal{S})'$, where $\mathcal{A}'$ and $\mathcal{S}'$ are subalgebras of $\mathcal{A}$ and $\mathcal{S}$ respectively.[31] In a way, this is simply generalising an earlier insight of Savage's: in his theory, the *states* of small world are grand world events to form a partition, and the uniform treatment of states, acts, and outcomes enables us to treat the relation between small and grand world acts the same. This gloss on the distinction neatly resolves all the problems just mentioned plaguing Savage's own more baroque version.

This happy resolution, however, is also particularly prone to trouble from the rejection of partitionality. For, if the action propositions do not satisfy partitionality, the class of acts cannot in general be understood as a Boolean subalgebra of the Boolean algebra of events, and it will make no sense to talk of subalgebras of that subalgebra.[32] It

[31] If we assume, as we already have in earlier sections, and against the counsel of Joyce himself, the resurgently popular *conditional excluded middle* as a law of the logic of the counterfactual conditional, we may also interpret the outcome or consequence of an action A in a state B simply asme way the full counterfactual consequences of $A \wedge B$. On conditional excluded middle, see [Lewis, 1973] [Stalnaker, 1980] [Williams, 2010] [Goodman, MS].

[32] If the logic of action contains **4**, the arbitrary joins of the action propositions will at least be guaranteed to form a *Heyting* algebra–a straightforward consequence of the well-known fact that the up-sets of a preorder form a Heyting algebra, and an expression of a deep and long-recognised connection between the (normal) logic **S4** on the one hand and intuitionistic logic on the other [Esakia, 1974]. (See Proposition 3.A.2.)

However, we have already seen reasons to reject **4** in the logic of action, so this fact is

will not do, absent this guarantee of algebraic structure on the agent's space of acts, to simply require that a small world action will be a union of potentially overlapping action propositions. Or at least, such a manoeuvre will not clearly allow the theorist to recover as small worlds the familiar motivating cases that are the common stock of debates in decision theory. For in these cases the action sets really *do* form a partition, and there is no guarantee that such a partition will be available as an appropriate set of joins of grand world action propositions.

Take, say, a Newcombe case, in which the subject's action set consists of two options: taking the first box only, or taking both boxes. Now, in an even remotely realistic case of this sort, the subject's motor control and ability to control their environment will be at least somewhat limited, as with Alex and his pen. And, if our models of Alex's pen-placing are any guide, there will be some long sequence of possibilities $(w_0, w_1, \ldots, w_n)$ where in all $w_{i \leq n}$ the subject takes either the first box only or both, in w_0 the subject takes one box, and in w_n the subject takes both boxes, such that for all $i < n$, w_{i+1} is in the subject's action proposition at w_i. We may picture this as a sorites-like sequence, at one end of which the subject clearly indicates exactly one box, and at the other end of which the subject clearly indicates both.[33] At some middling point, the subject gestures relatively loosely, without controlling which side of the divide (indicating one box, indicating two) the indication stands on.

There will therefore be a (grand world) action A straddling the two (small world) actions, with no guarantee *a priori* that A is not the unique rational action in the subject's grand world. Similarly, if there are no (maximally) rational grand world actions, we do have no guarantee that every one- or two-boxing action outcompeted by a straddling action but not *vice versa*. That is, the small world may not merely conflate the rational course of action with irrational ones, it may omit consideration of the rational one altogether.[34]

unlikely to be of great importance.

[33] For potential concerns about vagueness, see §4.4.2.

[34] It may be objected that, in a Newcombe problem, the relevant actions are in fact *decisions* (understood along the lines of inner volitions), about which the argument for non-partitionality is less compelling. After all, the predictor is generally understood as a penetrating *psychologist*, not a prognosticator specifically of future physical box-takings. I have two replies. First, we can make all the same points about cases where

Note that, if the class of action propositions in the Newcombe case has roughly the same structure as that of Alex's action propositions, they will not even correspond in a natural way to *any* algebra with the familiar logical operations of disjunction, negation, implication, etc. Where the action propositions form the cells of a partition, they correspond in an obvious way to to a certain complete atomic Boolean algebra. As noted in fn. 32, when the action propositions are all appropriately "nested" within one another, the class of them corresponds in the same way to a certain *Heyting* algebra. But in a case like Alex's–where there are pairs of action propositions $\langle p, q\rangle$ where p and q overlap but neither contains the other, and any world $w \in p \cap q$ has its action proposition overlap both p and q–this is not possible. Conjunction, for example, will not be interpretable by intersection, since the class of arbitrary joins of action propositions is not closed under intersection. It may be possible to contrive some very weak variety generalising that of Boolean algebras in which the action propositions will still constitute an algebra of that variety, but without natural interpretations of such basic operations as conjunction these constructions seem unlikely to be of interest or theoretical value. Thus, failures of partitionality cause the "algebraic" understanding of decision problems to break down at a very early stage.

In an ordinary setup to a Newcombe problem, of course, the subject will generally be in a position to make true that they one-box and in a position to make true that they two-box, so that suitable joins of action propositions can be constructed from the grand world actions to play the role of the one-boxing and two-boxing small world actions. But now it is not guaranteed that the grand world rational action (if such there be) will be a refinement of either small world action. It is one thing to admit that the grand world might deliver a *different* verdict on which of two small world acts is rational than the small world itself does; it is another to admit the grand world might deliver a verdict whose terms the small world cannot even interpret.

Let us sum up. Partitionality about an agent's action propositions is widely assumed among decision theorists, though it cannot draw succour specifically from the standard representation theorems. We

the external physical actions really are what do the heavy lifting in fixing the stakes in the decision matrix, like Jeffrey's case of the extortioner and the car. Second, we will later see reasons specific to inner volitions to think them prone to similar arguments as that given about Alex and his pen (see §4.4.2).

have, however, seen that there are good reasons to doubt this partitionality. One area where these doubts raise serious problems for standard accounts of decision theory is in the mediation between "small" and "grand" worlds or decision problems, first noted but only rather clumsily theorised by Savage. In particular, our arguments against partitionality make a hash of the usual post-Savage way of distinguishing and relating small and grand worlds, as articulated by Joyce. They therefore present a problem for the decision theorist, or perhaps confront him anew with an old problem: in what does this distinction consist, if we cannot avail ourselves of partitionality in the manner of Joyce, and how do we connect the two types of world?

3.5 Conclusion

This chapter has been an elaboration on the concept of an action proposition, as it appears in formal models of agency. More importantly, however, it has mounted a case that the logic of control must violate certain principles, principles widely enough assumed for their failure to be felt across much of the extant theorising on the subject. The sense of "logic of control" is slightly different from the one to have dominated the previous chapter, admittedly. Where Chapter 2 investigated control in terms of which *principles in* the language of control are *true*, here our attention has been on which *models of* that language are *realistic*. Undefinability results, as demonstrated in the appendix, show this gap to be technically insurmountable: no clever formulations will allow us to reduce the questions we have been asking about models of control to questions simply about control itself, nor even about control and practical possibility. Still, the investigation has remained (merely) logical in an important sense: it has been restricted to questions about the shape that control takes in logical space. In this way, the conclusions reached are only formal, only "thin," not substantive or "thick".

It is as yet unclear that such purely formal investigations could shed any light on philosophical questions not themselves posable in such extremely general terms. The last chapter opened by promising relevance of the logical inquiries to come to big ticket philosophical matters of freedom, morality, and choice. How could these questions be borne upon by deliverances as abstract as the preceding? The

following chapter will try to answer that question in one particularly weighty case.

3.A Models and Countermodels

Here we prove various novel metalogical results appealed to in the course of the argument.

3.A.1 Assorted Facts about Sandwich Models

We can now prove the correspondence between Conv45* and the principle attributed to Lewis in §3.2.1.

Proposition 3.A.1. *Take an arbitrary sandwich frame* $\langle W, S\rangle$. *Let* $\mathcal{N}(p \subseteq W)$ *be* $\{w \in W : S_{\mathsf{min}}(w) \subseteq p \subseteq S_{\mathsf{max}}(w)\}$. *Let* O_w *be the intersection of all* $p \ni w$ *such that for some* $q \subseteq W$ *either* $p = \mathcal{N}(q)$ *or* $p = W \setminus \mathcal{N}(q)$.

If for all w *(with* $S(w)$ *defined),* $S_{\mathsf{min}}(w) \subseteq O_w \subseteq S_{\mathsf{max}}(w)$, *then for all* v *and* $u \in S_{\mathsf{min}}(v)$, $S(v) = S(u)$, *and* vice versa.

Proof. **Left to right:** Clearly, given the antecedent, for all w (with $S(w)$ defined),

$$O_w = S_{\mathsf{min}}(w)$$

as by definition

$$O_w \subseteq \mathcal{N}(S_{\mathsf{min}}(w)) \subseteq S_{\mathsf{min}}(w)$$

It is now easy to see that for any $v \in O_w$, neither is there $u \in O_v, \notin O_w$ nor $u' \in O_w, \notin O_v$. For if the former, then

$$\mathcal{N}(S_{\mathsf{min}}(w)) \ni w, \not\ni v$$

which is prohibited by construction of O as long as

$$v \in O_w = S_{\mathsf{min}}(w)$$

and if the latter, either

$$\mathcal{N}(S_{\mathsf{min}}(v)) \ni v, \not\ni w$$

or no $\mathcal{N}(p) \ni v$ at all, both of which are prohibited by construction of O as long as

$$v \in O_w = S_{\mathsf{min}}(w)$$

since

$$w \notin \mathcal{N}(p) \text{ iff } w \in (W \setminus \mathcal{N}(p))$$

Therefore

$$O_w = O_v$$

and thus

$$S_{\mathsf{min}}(w) = S_{\mathsf{min}}(v)$$

Moreover, whenever

$$O_w = S_{\mathsf{min}}(w) = O_v = S_{\mathsf{min}}(v)$$

we also have

$$S_{\mathsf{max}}(w) = S_{\mathsf{max}}(v)$$

as otherwise w, v will be distinguished by some p in that

$$w \in, v \notin \mathcal{N}(p)$$

(or *vice versa*). But again by definition of O, this is impossible. Therefore,

$$S(w) = S(v)$$

Right to left: Suppose for all w (with $S(w)$ defined) and $v \in S_{\mathsf{min}}(w)$,

$$S(w) = S(v)$$

Then

$$w \in \mathcal{N}(p) \text{ iff } v \in \mathcal{N}(p)$$

so we have

$$O_w = O_v$$

So by $v \in O_v$ for all $v \in S_{\mathsf{min}}(w)$,

$$S_{\mathsf{min}}(w) \subseteq O_w$$

But

$$O_w \subseteq \mathcal{N}(S_{\mathsf{max}}(w)) \subseteq S_{\mathsf{max}}(w)$$

Therefore,

$$S_{\mathsf{min}}(w) \subseteq O_w \subseteq S_{\mathsf{max}}(w)$$

□

Now we prove the result mentioned in fn. 32.

Proposition 3.A.2. *If* 4 *is valid on a sandwich frame, the arbitrary joins of action propositions form a complete Heyting algebra.*

Proof. Let $v \preceq w$ iff, where A_w is the action proposition at w, $v \in A_w$. Clearly, $\preceq$ is reflexive. Moreover, as the validity of 4 ensures $A_v \subseteq A_w$ if $v \preceq w$, $\preceq$ is transitive. Thus it is a preorder.

Let A_X be $\bigcup\{A_u : u \in X\}$ ($X \subseteq W$). Then, for any $w \in X$ and $v \preceq w$, $v \in A_w$ and so $\in A_X$. Thus, any arbitrary union is an up-set. Consider an up-set X. For each $w \in X$, for all $v \preceq w$ we have $v \in X$, so that $A_w \subseteq X$. And, clearly, $\bigcup\{A_w : w \in X\} \supseteq X$. Therefore, X is an up-set iff it is an arbitrary join of action propositions, and so these joins form a Heyting algebra. That the algebra is complete is clear from the arbitrariness of the joins in the definition. □

3.A.2 The Logic Conv+

Let the language $\mathcal{L}_\boxdot$ be $\mathcal{L}$ closed under the introduction of formulas $\ulcorner\boxdot\varphi\urcorner$. Let the logic Conv+ $\subset \mathcal{L}_\boxdot$ be that generated by the axiom schemas and inference rules of Conv (*mutatis mutandis*), together with the additional axiom schemas

$\mathsf{K}_\boxdot$ $\boxdot(\varphi \to \psi) \to (\boxdot\varphi \to \boxdot\psi)$

$\mathsf{T}_\boxdot$ $\boxdot\varphi \to \varphi$

PS $(\boxdot\varphi \wedge \Box\psi) \to \Box(\varphi \wedge \psi)$

and the additional inference rule

Nec if $\vdash_{\mathsf{Conv+}} \varphi$, $\vdash_{\mathsf{Conv+}} \boxdot\varphi$

A model $\mathcal{M}^+$ of Conv+ is just like a model of Conv, together with an extra parameter $R : W \to \mathcal{P}(W)$; R obeys the further constraints that for all $w \in W$, $S_{\mathsf{min}}(w) \subseteq R(w)$ (where defined), and $w \in R(w)$. As usual, $[\![\boxdot\varphi]\!]_{\mathcal{M}^+} = \{w \in W : R(w) \subseteq [\![\varphi]\!]_{\mathcal{M}^+}\}$. The canonical model $\mathcal{M}^+_{\mathsf{canon}}$ of Conv+ is constructed as for Conv, with the obvious extensions for $R_{\mathcal{M}^+_{\mathsf{canon}}}$ (particularly that $w' \in R_{\mathcal{M}^+_{\mathsf{canon}}}(w)$ iff $\varphi \in w'$ for all $\Box\varphi \in w$).

It is easy to see the logic is sound.

Theorem 3.A.3. *If* $\vdash_{\mathsf{Conv+}} \varphi$, *then* $\vDash_{\mathsf{Conv+}} \varphi$

Proof. The only distinctive case to show is for PS. By the stated constraint on R and S_{min}, whenever

$$w \in [\![\boxdot\varphi]\!], \in [\![\Box\psi]\!]$$

it follows that

$$S_{\mathsf{min}}(w) \subseteq [\![\psi]\!] \subseteq S_{\mathsf{max}}(w)$$

(by construction of $[\![\Box\psi]\!]$) and

$$S_{\mathsf{min}}(w) \subseteq [\![\varphi]\!]$$

(by the above constraint), and thus

$$S_{\mathsf{min}}(w) \subseteq [\![\varphi]\!] \cap [\![\psi]\!] = [\![\varphi \wedge \psi]\!] \subseteq [\![\psi]\!] \subseteq S_{\mathsf{max}}(w)$$

and so by construction of $[\![\Box(\varphi \wedge \psi)]\!]$, it follows that

$$w \in [\![\Box(\varphi \wedge \psi)]\!]$$

Whence the conclusion follows trivially. □

The bulk of the completeness proof can rely on those given above.

Theorem 3.A.4. *If* $\vDash_{\mathsf{Conv+}} \varphi$, *then* $\vdash_{\mathsf{Conv+}} \varphi$

Proof. It is easy to check that the proofs for Theorem 2.A.5 and Corollary 2.A.6 can be extended to proofs of their analogues for Conv+. Thus all that is required is a proof that our canonical model $\mathcal{M}^{+}_{\mathsf{canon}}$ satisfies the criteria for a model.

For the constraints on R by itself the proofs are well known. We thus need only demonstrate that, when defined,

$$S_{\mathsf{min}\mathcal{M}^{+}_{\mathsf{canon}}}(w) \subseteq R_{\mathcal{M}^{+}_{\mathsf{canon}}}(w)$$

Suppose there is some $w' \in S_{\mathsf{min}\mathcal{M}^{+}_{\mathsf{canon}}}(w), \notin R_{\mathcal{M}^{+}_{\mathsf{canon}}}(w)$ (where these are defined). Then by construction there are $\varphi \in w', \psi \notin w'$ (that is, by maximality, $\varphi, \neg\psi \in w'$) such that

$$\Box\varphi, \boxdot\psi \in w$$

But then we have (by consistency of w')

$$\varphi \wedge \psi \notin w'$$

and so by construction of $S_{\mathsf{min}\mathcal{M}^{+}_{\mathsf{canon}}}$ we have

$$\Box(\varphi \wedge \psi) \notin w$$

which contradicts the hypothesis given PS (and maximality). □

Say that a model is *partitional* just in case, for all w, for all $v, v' \in R(w)$, $S_{\mathsf{min}}(v)$ and $S_{\mathsf{min}}(v')$ are (when defined) either identical or pairwise disjoint.

Theorem 3.A.5. *No formulas define the partitional models.*

Proof. It will suffice to show that, for two models $\mathcal{M}, \mathcal{M}'$ with $\mathcal{M}$ partitional but not $\mathcal{M}'$, the φ valid on $\mathcal{M}$ are just those valid on $\mathcal{M}'$.

Let $W_{\mathcal{M}} = \{w_0, w_1\}$; $S_{\mathsf{min}\mathcal{M}}(w_0) = S_{\mathsf{max}\mathcal{M}}(w_0) = \{w_0\}$; $S_{\mathcal{M}}(w_1)$ undefined; and for all atomic p, $V_{\mathcal{M}}(p) = \{w_0\}$. Let $W_{\mathcal{M}'} = \{v_0, v_1, v_2\}$; $S_{\mathsf{min}\mathcal{M}'}(v_0) = S_{\mathsf{max}\mathcal{M}'}(v_0) = \{v_0, v_1\}$, $S_{\mathcal{M}'}(v_1) = (\{v_1\}, \{v_0, v_1\})$; $S_{\mathcal{M}'}(v_2)$ undefined; and for all p, $V_{\mathcal{M}'}(p) = \{v_0, v_1\}$. For both, R is the universal relation (the constant function to W). Clearly, $\mathcal{M}$ but not $\mathcal{M}'$ is partitional, so it remains to show $\vDash_{\mathcal{M}} \varphi$ iff $\vDash_{\mathcal{M}'} \varphi$.

First we prove by induction on complexity that $w_1 \in [\![\varphi]\!]$ iff $v_2 \in [\![\varphi]\!]$. For the atomic base case, this is given. For the Boolean connectives, this follows by classicality of the logic. Moreover, as $S(w_1)$ and $S(v_2)$ are undefined, the step for $\Box\varphi$ is likewise trivial.

Next, $w_0 \in [\![\varphi]\!]$ iff $v_1 \in [\![\varphi]\!]$. Again, the base step and Boolean inductive step are trivial. For $\Box\varphi$, if $w_0 \notin [\![\varphi]\!]$ and $v_1 \notin [\![\varphi]\!]$, it follows by T that $v_1 \notin [\![\Box\varphi]\!]$ and $w_0 \notin [\![\Box\varphi]\!]$; if $w_0 \in [\![\varphi]\!]$ and $v_1 \in [\![\varphi]\!]$, either $v_2 \in [\![\varphi]\!]$ and (thus, by the above) $w_1 \in [\![\varphi]\!]$ or not. If so, $v_1 \notin [\![\Box\varphi]\!]$ and $w_0 \notin [\![\Box\varphi]\!]$, since $v_2 \notin S_{\mathsf{max}}(v_1)$ and $w_1 \notin S_{\mathsf{max}}(w_0)$. If not, then $S_{\mathsf{min}}(v_1) \subseteq [\![\varphi]\!] \subseteq S_{\mathsf{max}}(v_1)$ (meaning $v_1 \in [\![\Box\varphi]\!]$) and $S_{\mathsf{min}}(w_0) = [\![\varphi]\!] = S_{\mathsf{max}}(w_0)$ (meaning $w_0 \in [\![\Box\varphi]\!]$).

Finally, $w_0 \in [\![\varphi]\!]$ iff $v_0 \in [\![\varphi]\!]$. Base and Boolean steps are trivial. For $\Box\varphi$, by the inductive hypothesis we have $w_0 \in [\![\varphi]\!]$ iff $v_0 \in [\![\varphi]\!]$. As (by T) the result is trivial if $v_0 \notin [\![\varphi]\!]$ and $w_0 \notin [\![\varphi]\!]$, suppose $v_0 \in [\![\varphi]\!]$ and $w_0 \in [\![\varphi]\!]$. Again, either $v_2 \in [\![\varphi]\!]$ (and thus $w_1 \in [\![\varphi]\!]$) or not. In the former case, the conclusion follows as before. Otherwise, by the preceding paragraph, we have $w_0 \in [\![\varphi]\!]$ and $v_1 \in [\![\varphi]\!]$, so $v_0 \in [\![\varphi]\!]$ and $v_1 \in [\![\varphi]\!]$, and thus, as $S_{\mathsf{min}}(v_0) = S_{\mathsf{max}}(v_0) = \{v_0, v_1\}$ and $v_2 \notin [\![\varphi]\!]$, $v_0 \in [\![\Box\varphi]\!]$. But (since we are assuming $w_1 \notin [\![\varphi]\!]$), as $S_{\mathsf{min}}(w_0) = S_{\mathsf{max}}(w_0) = \{w_0\} = [\![\varphi]\!]$, also $w_0 \in [\![\Box\varphi]\!]$.

So the $\varphi \in \mathcal{L}$ true at w_0 are just those true at v_0, v_1, and the $\varphi \in \mathcal{L}$ true at w_1 are just those true at v_2 (extending this to $\varphi \in \mathcal{L}_{\boxdot}$ is straightforward). The desired result–$\models_{\mathcal{M}} \varphi$ iff $\models_{\mathcal{M}'} \varphi$–is now immediate. □

Say that a model $\mathcal{M}$ satisfies **Parity** just in case, for all $w, v \in W_{\mathcal{M}}$, $v \in S_{\mathsf{min}}(w)$ iff $S(w) = S(v)$ (we have encountered this already in our discussion of Lewis' principle). The above proof gives the following immediate consequence.

Corollary 3.A.6. *No formulas define **Parity**.*

We now prove Proposition 3.3.1.

Proof. Suppose our model is partitional and that (3 - 5) and (C_i) are true (at $w \in W$), where i is the least number such that $\Box(p \wedge q_i)$ and $\neg\Box(p \wedge q_{i-1})$ are true. As (by (4))

$$S_{\mathsf{min}}(w) \subseteq [\![q_{500,000}]\!], R(w)$$

so (by (3) and PS) for all $w' \in S_{\mathsf{min}}(w)$, $S_{\mathsf{min}}(w')$ is defined and

$$S_{\mathsf{min}}(w') \subseteq [\![p]\!] \subseteq S_{\mathsf{max}}(w')$$

By partitionality,

$$S_{\mathsf{min}}(w) = S_{\mathsf{min}}(w')$$

We thus have for all $\Box(\varphi \wedge p)$ true at w, i.e. such that

$$S_{\mathsf{min}}(w) \subseteq [\![\varphi]\!] \cap [\![p]\!] \subseteq [\![p]\!] \subseteq S_{\mathsf{max}}(w)$$

with S_{min} defined and fixed throughout $S_{\mathsf{min}}(w)$, that

$$[\![\Box(\varphi \wedge p)]\!] \supseteq S_{\mathsf{min}}(w)$$

and thus

$$S_{\mathsf{min}}(w) \subseteq [\![\Box(p \wedge q_i)]\!] \subseteq [\![p]\!] \subseteq S_{\mathsf{max}}(w)$$

and therefore we have $\Box\Box(p \wedge q_i)$ true (at w).

Since we also have $\neg\Box(p \wedge q_{i-1})$ true (at w), we by similar reasoning have $\Box(p \wedge q_{i-1})$ true at no $w' \in S_{\mathsf{min}}(w)$, and thus

$$S_{\mathsf{min}}(w) \subseteq [\![p]\!] \cap [\![\neg\Box(p \wedge q_{i-1})]\!] \subseteq [\![p]\!] \subseteq S_{\mathsf{max}}(w)$$

and thus $\Box(p \wedge \neg\Box(p \wedge q_{i-1}))$ true (at w).

We by these and M therefore have

$$\Box(\Box(p \wedge q_i) \wedge p \wedge \neg\Box(p \wedge q_{i-1}))$$

true at w, or equivalently

$$\Box(\Box(p \wedge q_i) \wedge \neg\Box(p \wedge q_{i-1}))$$

true (at w), which contradicts (C_i). □

Theorem 3.A.7. *There is a model M everywhere validating* Conv4*, and for any chosen η we have all of (3 - 5) and (C_i) true somewhere in M.*

Proof. For its domain W our model has the union $\bigcup_i [0,1]^2_i$ of a continuum-large set of disjoint copies $[0,300]^2_i$ of the (closed, real) 300-square indexed by $i \in I = [0,1]^2 \cup \{\lambda\}$, where λ is an arbitrary element not in $[0,300]^2$.

We begin by defining a strict linear order $\prec$ on the $[0,300]^2_i$ according to a natural order on I with order type $(1+\theta+1)^2+1$, where θ is the order type of the reals. For $(x,y),(x',y') \in [0,300]^2$, $(x,y) \prec (x',y')$ (and so $[0,300]^2_{(x,y)} \prec [0,300]^2_{(x',y')}$) iff either $y < y'$ or both $y = y'$ and $x < x'$; for all $i \neq \lambda$, $i \prec \lambda$ (and so $[0,300]^2_i \prec [0,300]^2_\lambda$). $\preceq$ is defined naturally from $\prec$ and $=$.

Next, for $(x,y) \in [0,300]^2$, define

$$D(x,y) = \{(x',y')_{(x,y)} \in [0,300]^2_{(x,y)} : |x-x'|^2 + |y-y'|^2 < 1\}$$

and define

$$D^+(x,y) = D(x,y) \cup \bigcup_{i \in I} \{(x,y)_i \in [0,300]^2_i : \forall(x',y')\ (x,y)_i \notin D(x',y')\}$$

It is easy to see that the $D^+(x,y)$ form the cells $[(x,y)_i]$ of a partition D^+ on W.

For $w = (x,y)_i$, let $S_{\mathsf{max}}(w) = W$, let $S_{\mathsf{min}}(w)$ be the D^+ cell $[w]$, and let $R = W^2$. Let $\{(x,y)_i : i \in I\}$ represent the pen being located at coordinates x and y on the paper in cm. It is now easy to check the desiderata are met. □

Note that, given that the intersection of any $D^+(x,y)$ and $[0,300]^2_\lambda$ is simply the singleton $\{(x,y)_\lambda\}$, the model just given is one in which

any action proposition is compossible with maximal control over the location of the pen. This can easily be altered to instead give merely arbitrarily high control.

Corollary 3.A.8. *There is a model satisfying the above criteria in which every action proposition is compossible with arbitrarily high, but not fully, precise control over the coordinates of the pen; that is, for any $z \in \mathbb{R}, 0 < z < 1$, and any $S_{\mathsf{min}}(w)$, there is a $v \in S_{\mathsf{min}}(w)$ satisfying $v \in [\![\Box(p \wedge q_{10,000z})]\!]$.*

Proof. Consider W from the last proof, and index continuum many disjoint copies of W by $z \in \mathbb{R}, 0 < z \leq 1$, with an order on $W^* = \{W_z : 0 < z < 1\}$ induced naturally by the order of inverse greatness on the indices, stipulating $W_1 = W$. Introduce the $D(x, y)_z \subset W_z$ as before, but with z^2 taking the role of 1. Let $D^+(x, y)_z$ be the set containing all members of $D(x, y)_z$ and all $((x, y)_i)_z$ for $i \in I$. The domain of the model will then be $\bigcup_{0<z\leq 1} W_z$, and the action propositions $D^*(x, y)$ for all x, y will just be the unions $\bigcup_{0<z\leq 1} D^+(x, y)_z$. It is easy to check that, as naturally interpreted, the model satisfies the desiderata. □

Chapter 4

Control and Responsibility

4.1 Introduction

Consider the following simple story.

> Reckless Driver: Yukari is driving fast and incautiously with Chiyo in the back seat when she accidentally runs over a pothole, thrusting Chiyo up against the car roof and giving her a nasty bruise on her head. It could easily have been, however, that (for reasons beyond Yukari's control) there was no pothole, in which case Yukari would have still driven equally recklessly but left Chiyo unscathed.

Many philosophers have felt a tension between our judgements about agents' culpability in cases like Reckless Driver and certain attractive principles relating control and moral responsibility. On the one hand, we judge that Yukari has a different level of culpability than she would have had there been no pothole; on the other, we believe an agent's culpability can only be different if what it controls is different (the *control principle*).[1] Some have taken it that we should uphold principle

[1]For the initiating papers see [Nagel, 1979] [Williams and Nagel, 1976], also as forerunners [Kant, 1784] [Feinberg, 1970]. For major contributions in the literature

over intuition, others intuition over principle, and some an intermediate or selective position, but few have contested that there is a tradeoff here to be made.

In my view, this is all mistaken, for there is no conflict between the principle and the cases. These intuitions and principles are logically consistent; moreover, the auxiliary principles philosophers tacitly rely on to produce an inconsistency are implausible. Once we reject these implicit principles, the way is open to an alternative understanding of control that eliminates the alleged paradox. When it comes to Yukari and her guilt, we can have our cake and eat it too.

It will be helpful before beginning to gesture, at least in broad strokes, at the philosophical picture that will motivate this argument. On my view, Yukari controls different sets of facts in her actual and counterfactual circumstances. In particular: in the former she does, and in the latter she does not, control whether Chiyo's head is bruised. The scope of her control is *flexible* as between the two situations, expanding in the one and receding in the other. This stands in contrast to the principles I take to underly the supposed paradox around Reckless Driver, on which her level of control over her environment is uniform across both cases. And once it is in place, there is nothing paradoxical about Reckless Driver at all: her culpability can vary between her possible circumstances because she controls different things in each.

To see the draw of this position, it will be necessary to first of all indicate the limits of the implicit alternative. To this end, I begin by isolating, and then attacking, the implicit assumption to undergird the claim that cases like Reckless Driver present a paradox. I argue that this assumption, which I call **Parity**, is incompatible with compelling claims about the limits of our higher-order control, our control over *what it is* we control. With these difficulties besetting **Parity** in view, the way will be clearer to the positive vision just alluded to.

---

see [Zimmerman, 1987] [Greco, 1995] [Wolf, 2001] [Zimmerman, 2002] [Pritchard, 2006] [Levy, 2011] [Hanna, 2014] [Hartman, 2017]. For surveys and essay collections see [Williams, 1981] [Statman, 1993] [Nelkin, 2021].

4.2 The Control Principle

According to usual presentations of the alleged paradox, the puzzle arises from something like the following two claims, which are taken to be individually plausible but jointly inconsistent:

> **Control Principle**: For all agents S and possibilities w, w': if the moral grade of S in w is different from its moral grade in w', S does not control exactly the same facts in w and in w'
>
> **Culpability**: Yukari's actual moral grade is not the moral grade she would have were there no pothole

This is, of course, a supervenience thesis. If we parse "dependence" as supervenience, it says simply that any agent's moral grade depends upon what it controls. This comports with both many more casual statements of the principle in the literature, and the way in which contrasting pairs of possibilities are generally taken to bear upon it.

Control Principle's reference to distinct possibilities is relevant because Reckless Driver itself discusses two possibilities (circumstances, situations) for Yukari, to which I will return constantly: her actual situation, in which Chiyo is bruised; and the counterfactual situation without the pothole, in which Chiyo is unbruised. I will be arguing that philosophers generally believe that **Control Principle** and **Culpability** conflict because they accept certain mistaken assumptions about the logic of control. But before we can get into the details of this argument, some comments on the claims themselves.

These formulations refer to an agent's *moral grade*, which is left undefined and unspecified. An agent's moral grade is some relevant aspect of its overall moral responsibility (in the retrospective sense of guilt or credit). While most agree the principle says *something* about an agent's responsibility must be the same if what it controls is the same, there is little consensus about what that something is. Proposed candidates include, for example, the *degree* (as opposed to scope) of an agent's responsibility, the agent's *moral worth*, and the ways in or degree to which the agent is *morally assessable*.[2] The current formulation is neutral as between these alternatives.

[2]For the degree/scope version see e.g. [Zimmerman, 2002] [Peels, 2015] (also [Demirtas, 2026] for recent criticism); for moral worth, sometimes called *aretaic* assessability,

Which of these candidates is correct is irrelevant to the present work. While philosophers disagree about what aspect of an agent's responsibility the control principle applies to, they generally agree that it says this aspect can only differ if *what the agent controls* differs. As long as Yukari controls different things in the two possibilities from Reckless Driver, the principle is therefore irrelevant to **Culpability**. I will be arguing that we have no reason not to think her control differs between them, irrespective of any moral grade, so what this grade involves exactly is immaterial.

Not all suggested versions of the control principle take quite this form. Some, for example, simply say that (for any fact p) an agent is responsible for p only if it controls p. Others, rather than comparing a single agent's grade and control across possibilities, compare grade and control between (say) distinct actual agents. I will address such rival versions towards the end. The operative concept of *control* also requires some unpacking. As I am using the expression, the objects of control are propositional, the denotations of *that-* clauses, where controlling (the fact) that p just amounts to simultaneously controlling whether p and being such that p.[3] This is in contrast to the literature's usual understanding of control as over *events*. But this difference does not affect the dialectic, aside from making the relevant claims slightly more tractable. In particular, it is orthogonal to the distinctions sometimes drawn in the literature between relevantly different grades of control (for discussion see in particular [Zimmerman, 2006]).

For all the description of Reckless Driver says, Yukari might control wildly different things in the two possibilities. So, there is no *logical* inconsistency between **Culpability** and **Control Principle** as stated, even adding in the description of the case. Standardly, philosophers tacitly rely on an extra assumption to generate the paradox, which is (unlike **Control Principle** and **Culpability**) rarely or never contested:

> **Equal Control:** Yukari actually controls just the facts she would have were there no pothole

This suffices to derive a contradiction from the other two, since it establishes that she controls no different facts in the two possibilities,

see e.g. [Richards, 1986] [Greco, 1995]; for assessability generally see e.g. [Nagel, 1979] itself.

[3] Note that this trivially makes *controls that* factive.

which (by **Control Principle**) means her moral grade is the same in each (contradicting **Culpability**). Indeed, without **Equal Control** it is hard to see why **Control Principle** would bear on **Culpability** at all. In what follows, I identify the most plausible rationale for **Equal Control** and argue against it. If this argument is correct, it undermines the basis for treating **Control Principle** and **Culpability** as inconsistent.

Before starting on this investigation, however, it is worth taking stock of just how radical the denial of **Culpability** is. If **Equal Control** is plausible as a claim about Reckless Driver, analogous claims of equal control should presumably also be plausible about most of what we ordinarily think of as human action. It will turn out, for example, that I plausibly do not control whether the lights are on in my apartment, since there is an alternative possibility in which my city suffers a freak power outage and nobody's lights are on, and I do not control whether such a power outage occurs, and actually control the same things as I would in that possibility. If these claims are true, the sorts of things in virtue of which we ordinarily treat our fellow humans as morally assessable–that they negligently left the lights on and thereby pointlessly increased the power bill, say–turn out not to bear on their moral grades, since they do not control them. Thus the morally significant share of our lives very severely contracts. There may be *some* kinds of acts whose moral significance these claims will not overturn; perhaps, for example, purely internal benevolent or malevolent volitions will be untouched. But the possible presence of such stragglers of moral relevance does not mean the resulting view is not very revisionary.

This radicalism is often skated over in the literature, beginning with [Nagel, 1979] itself. As Nagel sees things, the control principle also means that, for example, ordinary contingently existing agents will not be morally assessable for any actions they take, because they do not control whether they came into existence and so (presumably) there is an alternative possibility w in which they never existed and where they control the very same things as in actuality (that is, there is a true analogue of **Equal Control** about actuality and w). It is a common, on some tellings in fact the consensus, combination of views that the control principle holds in such cases as to contradict claims like **Culpability** but not in such cases as to contradict much weaker claims of differing moral grades across possibilities, like that

the contingent agent actually has a moral grade it lacks in w.[4] The more radical, wide-reaching consequences of the control principle in these cases, that is, are taken to indicate that they are where the real trouble for the principle lies, whereas in cases like Yukari's its initially counterintuitive consequences (like the failure of **Culpability**) can be taken more or less in stride.

In my view, this way of looking at the matter gets things backwards. If certain reasoning delivers disastrous results in cases with bigger stakes and less mundane subject matter, where similar reasoning delivers merely wildly revisionary results in cases with smaller stakes and more mundane subject matter, this is an indication that something has already gone wrong in its application in the latter case. Moreover, that latter reasoning will be a more promising case for diagnosing whatever is going wrong in both, since it screens off difficulties potentially imported by the former case's stakes and subject matter. Rescuing Yukari's culpability from paradox, that is, might show us how to do the same in these more extreme variations on the same theme.

4.3 From Equal Control to Parity

If Yukari actually controls some relevant fact that she does not in her counterfactual situation, it is presumably that Chiyo is bruised. What about Reckless Driver would tempt us to deny that she actually controls this? In general terms, we feel that the description demonstrates some kind of parity between her actual and counterfactual circumstances, which requires her to control the same things in both; since, in the counterfactual, she does *not* control whether (or that) Chiyo is bruised, she must not control it actually. But what does this parity consist in? Before I give my own answer, let us consider a few obvious responses to see why they fail.

Perhaps the parity *just is* that she seems to control the same things in both. In other words, maybe **Equal Control** is simply a basic intuition about Reckless Driver, not underwritten by an independently specifiable general condition on such cases.The reasons behind our

[4]Hartman describes "the most common positions in the moral luck debate" as "the positions that affirm some kinds of moral luck and deny others," and makes clear that he has this pattern of claims chiefly in mind [Hartman, 2019].

intuitions must hit bottom somewhere, and maybe **Equal Control** is at that bottom.

This is unsatisfying for a couple of reasons. First, it is question-begging: **Culpability** and **Control Principle** are both extremely *prima facie* compelling and jointly inconsistent with **Equal Control**, so why not just scrap it in light of them? We need *some* reason not to reject the claim besides its sheer intuitive plausibility, since the other two claims have their own competing raw appeal. Second, the fact that we can reliably and systematically produce new problem cases like Reckless Driver suggests there is some general pattern underlying them. The literature is replete with such puzzle cases, and undergraduates can easily be induced to devise new ones for term papers. This seems to indicate we have some independent grasp on what makes **Equal Control** (*mutatis mutandis*) plausible about a case, which we use in constructing new problem cases; or, in other words, that there is some common repeatable feature to the cases that we take as evidence for their analogues of **Equal Control**. So, again, what is this shared feature?

One unsatisfying answer is that her *actions* are the same in each possibility. This merely kicks the question up a level: why can't her actual actions include the act of bruising Chiyo, while her counterfactual actions do not? Again, intuitively, because the description demonstrates some parity between the two. For similar reasons, it is unhelpful to say that the "only differences" are factors she does not control, like the presence of the pothole. If she controlled Chiyo's bruise, this would be obviously false, since the bruise would itself be such a factor. The explanation must go deeper.

Maybe it is that the counterfactual possibility without the bruise is *nearby* to actuality. But this faces two problems. First, some distant possibilities where Yukari does not control Chiyo's bruise (such as ones involving miraculous intervention stopping her bad driving from harming Chiyo) also diminish (if not eliminate) our confidence in her actual control over it; when such distant possibilities are pointed out, we become less certain that she actually controls the bruise.[5]

[5]Except among those who believe or suspect cases like Yukari's very widely undermine our moral responsibility, it is unpopular to suggest that these kinds of distant possibilities *actually demonstrate* Yukari's lack of the relevant kind of control [Nagel, 1979] [Levy, 2011]. But I think it is psychologically undeniable they dispose us towards some level of doubt.

Second, there are nearby possibilities where Chiyo is not bruised that, when pointed out, do *nothing* to diminish our confidence, such as in Temperamental Driver.

> Temperamental Driver: Yukari drives incautiously, bruising Chiyo as usual. However, whether she drives cautiously or incautiously is a fickle, temperamental matter; she could have easily decided to drive cautiously instead, in which case Chiyo would have gone unbruised.[6]

One could, of course, maintain that the counterfactual situation in Temperamental Driver is not nearby because nearness of w to w' (in the relevant sense) requires sameness of action or control in w and w'. But this is just to bring us back to where we started: why assume that Yukari's actions or controlled facts are the same in the two possibilities from Reckless Driver but not in those from Temperamental Driver?

What does Temperamental Driver crucially lack that Reckless Driver has? Let us say that *p distinguishes* two possibilities just in case p obtains in exactly one of them. Similarly, a proposition *excludes* or *rules out* w (or another proposition q) when it does not obtain in w (or any possibility where q obtains). In the non-actual possibility from Temperamental Driver, Yukari *does* (or at least naïvely seems to) control a fact ruling out her actual circumstance: that she drives cautiously. In the unrealised alternative from Reckless Driver, by contrast, Yukari avoids her actual, injurious history only by the good grace of God; that is, she controls no facts in that possibility that distinguish it from actuality. For all that she controlled, things might still have gone as they actually did.

If it is this fact about the pair of actual and counterfactual circumstances in Reckless Driver that guarantees she controls the same things in both, it suggests the following principle:

> **Parity**: For all agents S and possibilities w, w': if S in w controls no fact distinguishing w and w', S controls exactly the facts in w it controls in w'

Tacit belief in **Parity** is a natural explanation for the assumption of **Equal Control**. The antecedent, as just noted, is intuitively satisfied in

[6] Ch. 4 of [Hartman, 2017] raises concerns much like this against the closely related "modal conception of luck."

Reckless Driver. Not only does this straightforwardly imply **Equal Control**, it also reflects what one would casually say in explaining why the description implies Yukari controls the same things in both possibilities. One is tempted to say something like, "Well, if there had been no pothole, it would have been dumb luck things didn't go like they actually did;" if dumb luck is or involves lack of control, this is exactly what **Parity** predicts.[7] So **Parity** captures the intuitive reasons for believing **Equal Control** about Reckless Driver.

There are more general reasons to think **Parity** underwrites such judgements about this case and others like it. The principle gives a good recipe for generating new puzzle cases; to try it yourself, imagine some situation w where an agent seemingly does something culpable, then concoct another situation w' in which the agent obviously controls no facts ruling out w. This procedure reliably produces case descriptions for which (the analogues of) **Culpability** and **Equal Control** both seem plausible, which suggests that we take it as evidence that an agent controls the same things in w and w' when nothing the agent controls in w' rules out w (i.e. we assume **Parity**). Finally, the principle is generally validated in (or strongly suggested by) formal models of agency where it can be formulated, evincing widespread acceptance. These include causal decision theory [Lewis, 1981] [Joyce, 1999], (standard presentations of) evidential decision theories with propositional control [Jeffrey, 1965], and the so-called logic of STIT or seeing-to-it-that [Belnap and Perloff, 1988]. In all of these, as for me, an agent's actions or objects of control are understood as propositions, and control accordingly as a propositional operator.[8] So **Parity**

[7] The identification of luck with noncontrol is traditional but controversial: while universally accepted at the outset of the present debate, in recent years there have emerged dissidents who argue against the identification [Driver, 2012] [Levy, 2011] [Peels, 2015]. Philosophers who frame the puzzle in terms of a control principle, however, are especially likely to accept the identification of luck with control [Hartman, 2017] [Hartman, 2019] [Statman, 2019]; and even among philosophers to reject the identification, the view that luck does not *entail* noncontrol remains very unpopular [Lackey, 2008]. See [Coffman, 2009] for general discussion and [Hartman, 2017] for an opinionated survey.

[8] This does require some assumptions about the intended philosophical interpretation of these theories. In particular, it requires that what one controls is exactly what one *makes true* (in the case of decision theory as usually presented) and *sees to it that* (in the case of STIT logic), or at least that an agent controls the same facts in two possibilities iff it makes the same facts true (or sees to the same facts) in each, or that the same logic given for these operators can be given again for control for similar reasons. These further bridge claims are, when not explicitly endorsed, at least strongly suggested in

is widely believed independently, and explains intuitions like **Equal Control** in ways that fit well with what we naturally say about the puzzle cases. It thus seems as strong a candidate for explaining the appeal of claims like **Equal Control** as anything could be.

This reasoning remains somewhat conjectural, of course. There are any number of other principles that could, from a purely logical point of view, equally support **Equal Control**. But the proponent of **Equal Control** needs *some* basis for it, on peril of begging the question, and **Parity** would appear to be as plausible as any other. If the friends of **Equal Control** object, it remains up to them to provide a more satisfactory alternative.

4.4 Against Parity

4.4.1 Parity vs. Higher-Order Control

Now for the argument against **Parity**. To begin, consider the following case:

> Discus Throw: Sakaki stands on the edge of an empty 100m × 100m square field on a clear, windless day and throws a discus roughly towards its centre. As it happens, the discus falls centred around **c**, the field's midpoint.

I will (for each n) use d_n to stand for the claim that the discus falls centred somewhere within $n$cm of **c**.

The d_n usefully represent the *degrees of control* Sakaki may have exerted over where the discus fell. It is natural, for example, to assume that she does not control d_1, but does control d_{50000}.

> **Graded Uncontrol**$_1$: Sakaki does not control whether the discus falls centred within 1cm of **c**[9]

the informal philosophical presentations of these formalisations.

[9]One might question this assumption on the grounds that, since Sakaki could (presumably) have not thrown the discus at all, she does control whether it falls within 1cm of **c**, since she easily could have made it otherwise (namely by making it not fall anywhere). But if we accept this reasoning–that an ability to make it otherwise than that p means one controls whether p–there is no reason to think in the first place that Yukari does not control, actually or counterfactually, whether Chiyo is bruised. She could (presumably) have simply not driven the car at all, or driven carefully, or insulated Chiyo's

> **Graded Control**$_{50000}$: Sakaki controls whether the discus falls centred within 50m of **c**

By classical logic, this means there must be some first $0 < i \leq 50000$ such that Sakaki controls d_i but not d_{i-1}, which I will assume is unique. This conjunction (that she controls d_i and does not control d_{i-1}) represents the *precise degree* of Sakaki's control over the discus, up to centimetres. When the conjunction is true, I call i the *limit* of her control.

Note in particular that **Graded Uncontrol**$_1$ does not require *in general* that, if the discus falls at some point **c'**, Sakaki cannot control that it falls within 1cm of **c'**. It is entirely consistent that she *might* have controlled it so exactly, and all the premise requires is that in this particular case she did not. My reasoning here therefore requires no principle in the theory of control analogous to the much-contested margin of error condition on knowledge in [Williamson, 2000], to which it bears a superficial resemblance. This is significant, for many critics of Williamson's argument concede that plenty of instances of a given margin of error principle (say, that if Mr. Magoo knows a tree is n inches or shorter, then it is not $n-1$ inches or taller) might be true, while denying that the principle holds with the full generality required for it to do the work Williamson expects of it [Berker, 2008] [Stalnaker, 2009]. **Graded Uncontrol**$_1$ is therefore immune to analogous criticisms.

I also assume, crucially, that Sakaki does not control down to the centimetre what her control's limit is. Letting m be the limit of her control (that is, supposing she controls d_m and does not control d_{m-1}), this means

> **Uncontrolled Limit**: Sakaki does not control that she both controls d_m and does not control d_{m-1}

Unlike the other premises, this is a claim about her *higher-order* control, her control over what she controls. **Uncontrolled Limit** is hard to dispute; while some amount of higher-order control is plausible, controlling the boundaries of one's control down to such an exact measure

---

head, *vel sim.*, thereby making it otherwise than that Chiyo gets bruised. And if we allow that she controls whether Chiyo is bruised, the original puzzle vanishes anyway. So I will set this consideration aside. Thanks to an anonymous reviewer for pointing this out.

seems superhuman.[10] Her control is in this sense *imprecise*.

Note that this invocation of limited higher-order control is very different from that in another recent intervention [Carlson et al., 2026]. In that discussion, limits on higher-order control appear in an exotic scenario involving a mind-controlling third party, and involve radical deficits in higher-order control on the part of the agent in question. Here, by contrast, the circumstances described are entirely banal, and the deficits in higher-order control are only marginal. The use to which putative limits are put here and in [Carlson et al., 2026] are quite different too (the upshot of the paper is opposed to the control principle rather than friendly to it), but most important to note here is the difference in how limits in higher-order control are approached.

I will also be assuming two (by now familiar) constraints on the logic of control: necessarily, when an agent controls some facts, it also controls their conjunction; and, if it controls both p and q, it controls any fact both entailed by p and entailing q, where entailment is understood as a consequent obtaining in every possibility where the antecedent does.

> **Agglomeration**: For all S, w: if, in w, S controls the facts that p_0 and that p_1 and that ..., S controls the fact that p_0 and p_1 and ...
>
> **Convexity**: For all S, w: if, in w, S controls that p and that q, and r is true in every possibility where p is and q is true in every possibility where r is, S in w controls that r

Both **Agglomeration** and **Convexity** are on firmer footing than **Parity**. On the one hand, **Agglomeration** sounds almost tautological. On the other, it is basic to our everyday reasoning about control that controlling a fact p implies controlling certain logical consequences of p. **Convexity** explains this intuition straightforwardly and modestly, without making such dubious predictions as that an ordinary agent will control necessary truths (which are not strictly *intermediate* in strength between pairs of other facts). **Parity**, by contrast, is murkier, despite its popularity. If one of the three must go, **Parity** is the obvious choice.

[10] This bears a close structural relation to the "gap principles" of [Fara, 2003] in the logic of vagueness, see also [Bacon, MS] for discussion of analogous principles in the theory of knowledge.

And, holding fixed the stated assumptions about Discus Throw, one of them must in fact go. **Graded Control**$_m$, **Graded Uncontrol**$_{m-1}$, **Uncontrolled Limit**, **Convexity**, and **Agglomeration**[11] are jointly inconsistent with **Parity**. Here is the proof.[12]

To begin, **Parity** has two general consequences about how control iterates: (CC) if Sakaki controls that p, necessarily she controls that she controls that p; and (CNC) if Sakaki controls that p but does not control that q, necessarily she controls the conjunction: p and she does not control that q.

To see that (CC) follows, suppose she controls p. She also, by **Agglomeration**, controls the conjunction of every fact she controls (call it *the Big Conjunction*), which of course entails p. Since control is factive, that Sakaki controls that p entails p; since (by **Parity**) what she controls in every possibility w that is compatible with every fact she controls (that is, every w where the Big Conjunction holds) is the same as what she actually controls, her controlling that p is entailed by the Big Conjunction. So she controls the Big Conjunction, which entails that she controls that p, which in turn entails p. Since she also controls p, by **Convexity** she controls that she controls that p.

To see (CNC), suppose she controls p but does not control q. By **Parity**, in every possibility where the Big Conjunction holds she does not control that q, i.e. the Big Conjunction entails that she does not control that q. But, since she controls that p, the Big Conjunction also entails p. So the Big Conjunction entails the conjunction: p and she does not control that q. But this conjunction entails p, which she also controls. So, by **Convexity**, she controls the conjunction of p and her not controlling q.

Now, the limit of Sakaki's control is m, so she controls that d_m but not that d_{m-1}. Suppose **Parity**. By (CC), she controls that she controls that d_m. By this and (CNC) (substituting "she controls d_m" for p and substituting d_{m-1} for q), she controls that: she controls that d_m but does not control that that d_{m-1}, contradicting **Uncontrolled Limit**.[13]

There is nothing special about Discus Throw required for this re-

[11] Plus the factivity of control.

[12] That this argument must quantify over possibilities directly rather than being neatly conducted entirely in terms of control is inevitable, see Corollary 3.A.6.

[13] Note that nothing in this argument hinges on the uncontrolled d_{m-1} being *true*.

sult. Premises just like **Graded Control**$_{10000}$, **Graded Uncontrol**$_1$, and **Uncontrolled Limit** hold for almost any instance of controlled physical activity, as long as you can find a sequence of propositions–like $(d_0, \ldots d_n, \ldots d_{50000})$–describing progressively less narrow ways the activity could have deviated from its actual outcome. We could similarly, for example, list a sequence of circumstances in which Yukari thrusts Chiyo ever and ever slightly more forcefully up through incautious driving. If **Parity** does not hold for Sakaki as she lands the discus, for the same reasons it does not hold for any of us as we do anything to our environment (like bruising a child through reckless driving). The principle is not just false, but catastrophically and ubiquitously false. Since **Parity** was the motivation for assuming **Equal Control** to begin with, this means there is no longer any reason to believe **Equal Control**. And without **Equal Control**, there is no reason not to accept both **Control Principle** and **Culpability**.

Let us review the argument thus far. Philosophers have traditionally taken **Control Principle** and **Culpability** to be inconsistent because they have assumed that the description of Reckless Driver requires that Yukari controls the same facts in both her actual and counterfactual circumstances (i.e. **Equal Control**). This, I argued, is because in the counterfactual circumstance in Reckless Driver she controls no facts ruling out actuality, and philosophers have assumed that an agent controls in w exactly what it does in w' if, in w, the agent controls no facts ruling out w' (**Parity**). But in cases like Discus Throw, where we do not control the limits of our control over what happens (a very common occurrence), highly plausible principles about the logic of control imply that **Parity** must be false. Therefore, we should reject **Parity**, leaving us with no reason to deny that **Control Principle** and **Culpability** are consistent and, indeed, both true.

4.4.2 Objections

I will now consider some objections to the above argument.

Control Internalism

Objection: Facts about where the discus falls are too external to Sakaki for her to control them in the relevant sense. One way to draw this out is that the discus' landing in a certain location cannot be a *basic*

action of Sakaki's, as an agent's basic actions are widely agreed to only include its bodily motions, and **Control Principle** is often taken to conflict with **Culpability** only when the sense of *control* is restricted to that over basic actions [Davidson, 1971] [Zimmerman, 1987] [Greco, 1995] [Coffman, 2007] [Levy, 2011]. Thus, for trivial reasons Sakaki controls no d_i, depriving the argument of any interest.

Reply: In the case of bodily motions, the argument can be modified very easily. Suppose that, instead of throwing a discus, Sakaki is partly extending her arm. Her control over the angle at which her elbow is bent will be imprecise in the same way as with her discus throw, so that the argument can be run again.

A slightly more difficult version of this objection comes from *volitionalists* and their kindred, according to whom the objects of our control (of the relevant sort) are internal, purely mental states or acts like decisions, tryings, or willings. While this position is marginal in the present debate (though see [Khoury, 2018]), it is in line with a minority internalist tradition in action and decision theory (see [Hornsby, 1980] [Ginet, 1990] [Hedden, 2012] [Lederman and Holguín, 2023], also discussion in [Ford, 2018]). The volitionalist is less likely to grant the analogue of **Graded Uncontrol**$_1$ for Sakaki's volitional state about the discus' location, since this volition arguably picks out its target destination with an arbitrary degree of precision.

Even the volitionalist, however, should grant there exist some cases of imprecise control. Suppose that Kaorin is trying to mix red paint, with as clear a target colour as she can. In particular, she is trying for an orangeish red paint, but lacks (for reasons beyond her control) the colour discrimination required to draw a more precise boundary in colour space (say, vermillion). So she controls that she is trying for orangeish red paint, but not that she is trying for vermillion (since she lacks the discrimination required to try for it, and control is factive). This conjunction represents a limit to Kaorin's control over her trying to mix paint, just as Sakaki's controlling d_m but not d_{m-1} represents a limit to her control over the discus. Since it essentially involves a colour (vermillion) which she cannot delineate, the volitionalist should grant she does not control this conjunction, either.

The argument can then proceed as before, replacing d_m with "Kaorin is trying for orangeish red paint" ($t_{\text{orangeish}}$) and d_{m-1} with "Kaorin is trying for vermillion paint" ($t_{\text{vermillion}}$). Suppose **Parity**. By the principle (CC) from the earlier argument, she controls that she controls

$t_{\text{orangeish}}$. By this and (CNC), since she does not control $t_{\text{vermillion}}$, she controls the conjunction: she controls $t_{\text{orangeish}}$ but does not control $t_{\text{vermillion}}$. Which is exactly the limit she does *not* seem to control, as just stated.[14]

Note that her noncontrol over $t_{\text{vermillion}}$ and the analogue for Kaorin of **Uncontrolled Limit** (that she does not control: that she controls $t_{\text{orangeish}}$ and doesn't control $t_{\text{vermillion}}$) have special appeal for the volitionalist. Such internalist accounts of the objects of control or choice are often motivated by "action-guiding" requirements that the agent have epistemic access to what it controls (e.g. [Hedden, 2012]); as Kaorin lacks understanding of vermillion, she is particularly ill-placed to have such access either to whether she tries to mix vermillion paint, or to limits involving vermillion on her exercised control. And, once again, given the pervasiveness of such poverty in discriminatory power, this **Parity**-undermining imprecision is ubiquitous.

Hyperintensionality

Objection: **Convexity** has untenable consequences. As a limiting case, it implies that, when an agent controls p and p is necessarily equivalent to q, it also controls q. This is controversial: prominent strains in the literature posit an epistemic condition on the relevant sort of control, where an agent might control p but not the necessarily equivalent q as long as it lacks epistemic access to their necessary equivalence [Zimmerman, 1997] [Rosen, 2004] [Levy, 2011]. So the principle leaves control insufficiently "hyperintensional."

Reply: The potentially inaccessible necessary truths these philosophers have in mind are normative and moral ones to which individuals may be blinded by circumstance and have to work to appreciate, like ones about the wrongness of racial prejudice. So, for example, I might control that I am uttering racist invectives but not that I am being immoral in uttering those invectives, even while these facts are necessarily equivalent, because racist invectives are necessarily wrong to utter. I am aware of no philosophers to suggest that agents' control might be affected by ignorance of non-moral, broadly logical princi-

[14] This instance of the argument against **Parity** presented in the last subsection, relying as it does on no "frog boiling" reasoning about gradually less plausibly controlled facts, illustrates just how little the core of that argument depends on "margin of error"-like principles *à la* [Williamson, 2000].

ples like **Parity** and **Agglomeration**, to which we can stipulate Sakaki has access anyway. Since the argument only appeals to such principles, and involves no purported moral necessities, it can therefore ignore control's alleged hyperintensionality.

Local vs. Global Parity

Objection: **Parity** is a much stronger condition than required for predicting anti-control intuitions about Reckless Driver. It is important to our intuitions about these cases that the alternate possibilities be *nearby*.[15] That is to say, in place of **Parity** we should endorse

> **Local Parity**: For all S, w, w': if S in w controls no fact distinguishing w and w' and w' is nearby to w, S controls exactly the facts in w it controls in w'

But weakening the principle in this way prevents immediately obtaining the desired result, which should therefore have no purchase over us.[16]

Reply: In order for my argument against **Parity** to fail against **Local Parity**, at least one of two things must be true. Either it must be compatible with the Big Conjunction that Sakaki does not control d_m, or it must be compatible with the Big Conjunction that Sakaki controls d_{m-1}. Reject both, and the contradiction still follows. For the sake of concreteness, without loss of generality, let us suppose it is compatible with the Big Conjunction that Sakaki does not control d_m; that is, there is a possibility compatible with the Big Conjunction in which Sakaki does not control d_m. Given **Local Parity**, this possibility must be *remote*, rather than nearby.

1. Given all that Sakaki controls, it could (hardly) be that Sakaki does not control d_m, but could not easily be that she does not control d_m; that is, there is a (remote) possibility compatible with the Big Conjunction on which she does not control d_m, but no nearby such possibility compatible with the Big Conjunction.

[15] The natural support for this complaint would come from the champions of so-called modal accounts of luck. See [Pritchard, 2005] [Pritchard, 2006] [Levy, 2011] [Driver, 2012] [Peels, 2015].

[16] To see that it no longer logically follows, suppose each w is only near itself.

In addition to the intuitive pull of **Uncontrolled Limit**, Discus Throw inspires a similarly strong sense that Sakaki could easily have failed to control d_m.

2. It could easily be that Sakaki does not control d_m; that is, there is a nearby possibility on which she does not control d_m.

Finally, since d_m is among the facts she controls, the only way for **Uncontrolled Limit** to be true is that the Big Conjunction is compatible either with her not controlling d_m or with her controlling d_{m-1}; again, for concreteness, let us stipulate the former.

3. Given all that Sakaki controls, it could be that Sakaki does not control d_m; that is, there is a possibility compatible with the Big Conjunction on which she does not control d_m.

Our confidence in (2) and our confidence in **Uncontrolled Limit** (and thus, by extension, (3)) are connected. The plausibility of one sustains that of the other. Were we to learn that it could not easily be that Sakaki fail to control d_m, that this fact about her control is somehow safely guaranteed by her circumstances, it would undermine our intuition of **Uncontrolled Limit** (i.e. of (3)). Now here is a straightforward way for this connection to work: the possibility referenced in (3) *just is* that referenced in (2). Our belief in the latter possibility founds a belief in the former because they are the very same (suspected) possibility.

But this is precisely what **Local Parity**, by way of (1), rules out. Because the d_m-uncontrolled possibility in (2) is nearby, it cannot be the d_m-uncontrolled possibility in (1) or (3). So, if **Local Parity** is what saves the appearances, (2) must found a confidence in **Uncontrolled Limit** and (3) only in an indirect and convoluted way. The nearby presence of a possibility in which Sakaki does not control d_m will somehow suggest a *distinct*, and *remote*, such possibility compatible with the Big Conjunction. This strains credulity: what could account for such a link between the (modally) near and far? Better to stick with the obvious story, on which both (2) and (3) refer to the same possibility, and leave behind the spooky action at a logical distance forced upon us by **Local Parity**.

This proposal, that the intuited possibility of (2) is also that of (3), in fact helps explain what might strike us as appealing about **Local Parity** as over and above (mere) **Parity** in the first place. If we assume

Parity proper, and the following principle linking control with modal proximity, **Local Parity** falls out trivially.

> **Proximate Strengthening**: If it cannot easily be that not-p, and one controls q, one controls that p and q

Proximate Strenghtening, as is clear from last chapter, is a member of a genus of widely intuitive principles governing control and modality, and thus seems likely to underwrite the appeal of **Local Parity**. It helpfully identifies the possibilities of (2) and (3), since the possibility in the latter must be nearby if it exists at all. And it offers an easier means of discerning the possible circumstances of an agent across which the agent will control just the facts it actually does: since only nearby such possibilities could be compatible with the agent's Big Conjunction, only those possibilities are candidates. What it does not do is support **Local Parity** independently of **Parity** itself.

Vagueness

Objection: The argument is soritical. It first of all assumes a sharp break in a sorites sequence about the degree of Sakaki's control: that she controls d_m but not d_{m-1}. It then leverages the intuitive arbitrariness of the break to motivate **Uncontrolled Limit**, which serves as the crux of the main argument: it is *prima facie* implausible that Sakaki controls that the limit of her control is thus-and-such, because we are wary to assert of *any* thus-and-such that it is the limit of her control, let alone that she controls this fact. The argument as a whole thus relies on a style of reasoning known to lead to paradox, and should be rejected.

Reply: This objection comes in two parts: first, that the argument assumes a sharp break (*viz.*, that Sakaki controls d_m but not d_{m-1}) in a sorites sequence (the sequence of propositions whose nth member is that Sakaki controls d_n); second, that the arbitrariness of such a break is what makes **Uncontrolled Limit** seem plausible. I agree that, if the argument relied for its plausibility on either of these moves, it would be suspiciously soritical. But it does not.

Firstly, nothing in the argument requires the assumption of a *sharp* break in the sequence, as opposed to a break *simpliciter*. As originally presented, it relies on a bit of classical reasoning to derive the conclusion that there must exist *some* i such that Sakaki controls d_i but not

d_{i-1} and stipulates that the variable m refer to this i. The formulation can thus give the impression that it requires some true sentence of the form ⌜Sakaki controls d_ι but not $d_{\iota-1}$⌝ (for ι a numeral expression), expressing some true proposition about the border of Sakaki's (vaguely extensive) control over the discus. This would be objectionable to the many theorists of vagueness who reject bivalence for borderline statements involving vague expressions.

This impression, however, is illusory. Instead of referring to a specific d_m as the limit of Sakaki's control, we could have classically derived the lengthy disjunction "Either Sakaki controls d_1 but not d_2 or she controls d_2 but not d_3 or. . . " and then conducted the remainder of the argument as a proof by cases, though there are obvious presentational hindrances posed by such a strategy. The stipulative reference for the expression m is an informal gloss on a classical consequence of our stated assumptions, and so is a consequence of them also in theories of vagueness to preserve the theorems of first-order classical logic. These include the extremely popular *supervaluationist* theories of vagueness to derive from [Fine, 1975], among whose primary selling points is this preservation of classical theorems. There are thus at least very prominent non-bivalent accounts of vagueness on which this stipulation for m will not in and of itself cause trouble, so that the putative non-bivalence of vague sentences is not as such an obstacle to the stipulation.

These remarks about supervaluationism, whose informal motivation relies essentially on quantifying over the *precisifactions* of a vague concept, naturally lead into my answer to the second part of the objection. According to the objector, **Uncontrolled Limit** relies for its plausibility on the arbitrariness or borderline status of the limit m, and is thus objectionably akin to the usual suspect in a sorites argument: the assumption that, for all n, if the fire station is nearby the cafe if within n metres of it, it is similarly nearby if within $n + 1$ metres (or whatever). We judge that Sakaki cannot control that she controls d_m but not d_{m+1} simply because we decline to judge that she controls that conjunction, because we are reluctant to affirm any such borderline conjunctions to begin with. The negation of the conjunction is classically equivalent to an instance of the typically soritical schema (that, for all n, if n is φ, $n + 1$ is φ), and control of the conjunction entails its truth. Thus a known fallacy emerges as an obvious basis for our initial credence in **Uncontrolled Limit**.

There is reason to doubt this explanation of **Uncontrolled Limit**'s appeal, however. Such soritical reasoning relies specifically on the vagueness of the operative concept (nearness, heap, hirsuteness, *vel sim.*). Replacing the vague concept as it features in the sorites sequence with a more (even if not *perfectly*) precise concept, accordingly, in general makes this culprit premise *less* plausible. (If vagueness is just whatever it is about some concepts that gives rise to the sorites paradox about them, and precision whatever it is about some concepts that blocks the emergence of such a paradox, this generalisation is in a sense an essential truth about vagueness and precision as such.) For example, replacing the concept of being nearby with the concept of being within fifteen minutes' walking distance, the above premise about the proximity of the cafe to the fire station is correspondingly less intuitively compelling. This provides a helpful criterion by which to judge the soriticality of a given intuitive assumption: when swapping out the relevant concept for a preciser one, the intuitive appeal diminishes.

How does this criterion hold up in the case of **Uncontrolled Limit**? Note first of all that the concept of control appears *twice* (really, thrice) in the assumption as presented: both in the main clause ("Sakaki does not *control* that. . . ") and the subordinate clause (". . . she both *controls* d_m and does not *control* d_{m-1}"). In replacing control with some relative precisification, it is the second, and not the first, appearance that qualifies, for it is only in the second that the concept of control features as the relevant vague concept of a sorites sequence. So, replacing the second (and third) but not first appearance of the concept with some more precise concept control* (say, an operationalisation of the concept suited for experimental or clinical study of Sakaki's motor skills), we get the schema

> **Uncontrolled Limit***: Sakaki does not control that she both controls* d_m and does not control* d_{m-1}

However we fill in the blank left by control*, I submit, this assumption is if anything even more intuitively plausible than the original. In this case, as in many others, it becomes less, not more, convincing that we control a phenomenon down to some fine grain of detail when that phenomenon is understood in the obvious candidates for more precise terms. In ordinary circumstances, similarly, a painter painting a

large picture of a gradient-coloured sunset even less plausibly controls down to the square millimetre how much of the canvas reflects light at a wavelength between 625nm and 750nm than he controls down to the square millimetre how much of the canvas is red: the former is a more specific and apparently demanding sort of control (for human painters under ordinary conditions). Thus the above criterion for soriticality is not met, suggesting that the relevant reasoning is not objectionably soritical.[17]

In fact, considerations about vagueness arguably give support to **Uncontrolled Limit** from a new direction. It is a commonplace in the literature on vagueness that, when it is borderline whether p, an agent will not know that p.[18] A *borderline* answer to whether p, that is, demands a *negative* answer to whether a given agent knows that p. Given the parallels between control and knowledge or perception, might a similar principle hold for control? A thorough investigation of this question would fall well beyond the scope of this dissertation, but it strikes me as at least potentially plausible. And in that case, the borderline status of the subordinate clause from **Uncontrolled Limit** would be positive reason to endorse the premise, on theoretical grounds.

4.5 Further Connections

In this section, I consider some ways of taking the argument of the last section further. The first places the argument in the context of a larger picture of control, modelled on certain disjunctivist views from epistemology and the philosophy of perception. The second takes up the question of *unfairness* in moral grades, and its relation to the control principle. The third considers variations on the control principle and the puzzle about Yukari, and how the main argument might be adapted to them.

[17] This is similar to a reply by Williamson to an accusation of soriticality against his anti-luminosity argument in [Williamson, 2000].

[18] For a qualified dissent from this consensus, see [Goodman, 2026].

4.5.1 Disjunctive Control

To complete the case against **Parity**, we still need an idea of how it could possibly be false, and how we could have been tempted to think it was true. In this section, I will sketch a picture of control that allows **Parity** and **Equal Control** to fail, and contrast it with competing, popular, **Parity**-friendly theories. These provide, respectively, an understanding of how to reject the principle, an explanation of what might have rendered it initially plausible anyway.

If **Parity** fails in the case of Yukari and Chiyo, what does Yukari control in actuality that she does not in the pothole-free counterfactual (or *vice versa*)? The natural candidate is: whether Chiyo is bruised. Thus, Yukari actually controls a fact (that Chiyo is bruised) inconsistent with her luckier counterfactual circumstance, in which she controls *no* facts inconsistent with actuality.

Here is a thought that accommodates this. An agent acts through capacities for control that vary drastically in their reach, depending on the conditions of the agent. Under more constraining conditions, for example (like her counterfactual situation), Yukari's operative capacity reaches only to her body and car; under more enabling conditions (like her actual situation), it reaches all the way out to Chiyo. So the topic or subject-matter of her control is different in each possibility: in the counterfactual, where her capacities only reach to the car, she only directly controls facts about her driving; in actuality, where her capacities reach as far as Chiyo, she directly controls facts *both* about her driving *and* about the state of Chiyo's head.

This way of talking is inspired by (epistemologically) disjunctivist accounts of perception in the tradition of [McDowell, 1994].[19] On such views, often presented as solutions to (or dissolutions of) sceptical worries, an agent's visual experience (say) when looking at a red wall under normal conditions is not the same as its experience when looking at a white wall under trick red lighting. In the former, circumstance permits the experience to reach all the way out to the wall itself, genuinely seeing it to be red; in the latter, the experience is a mere abortive *seeming* seeing that the wall is red. The accurate agent's, but not the inaccurate agent's, experience involves facts about the wall's colour. Thus the accurate agent's experience gets as far as

[19] On epistemological disjunctivism see also [McDowell, 1983] [Williamson, 2000] [Haddock and Macpherson, 2008] [Byrne and Logue, 2009] [Pritchard, 2012].

seeing something (namely, that the wall is red) ruling out the trickily lit possibility; under tricky lighting, the agent would see no facts about the wall's colour ruling out actuality. The capacity at work is flexible in how much it can grasp, depending on the agent's conditions, just as with Yukari's control.[20]

McDowell puts the matter well in the epistemic case:

> If that [*viz.*, getting into states that consist in having a certain feature of the objective environment perceptually present to one] is what a capacity is a capacity to do, that is what one does in a non-defective exercise of it. . . .
>
> For instance, a capacity to tell whether things in one's field of vision are green is a capacity – fallible, by all means – to get into positions in which the greenness of things is visibly there for one, present to one's rationally self-conscious awareness. If something's greenness is visibly there for one, one has conclusive warrant for believing that it is green. One's perceptual state *leaves no possibility that it is not green*. . . .
>
> To say that a perceptual capacity is fallible, or that anyone who has it is fallible in its exercise, is to say that the capacity does not ensure that its possessor is always *in a position to discriminate* defective exercises from non-defective exercises. (emphasis added) [McDowell, 2011]

In a "non-defective" exercise of a perceptual capacity, therefore, a subject is in a perceptual state *incompatible* with things not being as she perceives them. In a "defective" exercise of her capacity, however, she will be in a perceptual state *indiscriminable from* (and therefore, I take it, *compatible with*) her being in that "non-defective" state. This is possible because her perceptual state is, not merely different, but *weaker* in the defective case. It is just so with control, too. In her actual, damaging exercise of her control, she is in an agentic state that is incompatible with Chiyo being unbruised; in her counterfactual, innocent exercise of it, she is in an agentic state compatible with being in that more

[20] Note that this is not the same as the analogy, often invoked by disjunctivists (among others) about perception, between perception and *intentional action* [Williamson, 2017]. It indeed contributes to the intuitive pull of **Equal Control** that Yukari does not *intentionally* bruise Chiyo.

damaging state. **Parity** fails for the same reason the perceptive subject can rule out the possibility of illusion, but the illuded subject cannot rule out the possibility of true perception.

There is an incidental, but potentially disorienting, point of failure in this parallel. In perception, it is generally the *more* expansive exercises of the relevant capacities that are regarded as preferable to their corresponding more limited exercises (such as seeing that an object is green vs. merely seeming to see it is green). In control, it is generally the *less* expansive exercises that enjoy this privilege. Thus, it is preferable for Yukari to be in her counterfactual, non-bruising agentic state than her actual, bruising one. To carry over in full the analogy with McDowell's description of perception, one would need to say that her failure to bruise is a "defective" exercise compared with her bruising.

This disanalogy results from differences of emphasis among epistemologists and moral philosophers. As a rule, the former take more of an interest in cases where knowledge and seeing are preferable to error and illusion, whereas the latter take more of an interest in cases where responsibility and control are less desirable than their absence (because the responsibility involved is guilt rather than credit).[21] But there are cases in each domain where things go the other way. One is better off, morally, in a situation where one controls the (counterfactual) fact that Hitler dies in 1933 than in a situation where one does not control whether he dies, because one then enjoys the credit for killing him, just as it is better, visually, to really see that the object is green than to not. In the other direction: a jury, as a trier of fact, is legally defective if it sees certain bits of evidence not admissible in court, just as Yukari is personally defective as a driver for injuring her passenger. There are "good" epistemic cases where a subject is ignorant rather than perceptive, and "bad" moral cases where one is deficient in control and responsibility rather than excessive.

There is another difference between the disjunctivism about control

[21] There may well be general reasons for these patterns of interest. Perhaps, for example, it takes fewer factors going wrong in an action to render the action blameworthy than it does factors going right to make the action praiseworthy, so that from a sparse hostile description of an action we are reasonably more comfortable passing negative judgement than we are passing positive judgement given a similarly sparse friendly description of a similar action. There are even more obvious reasons to concentrate on cases where cognition is preferable to its absence, given the importance of knowledge and the dangers of ignorance to everyday life. Even if this is so, it does not undermine the general point being made in this paragraph.

just sketched and usual ways of presenting disjunctivism about perception. It is easy to take away the impression from those presentations–though it is not, I think, always part of the considered content of the philosophical position(s) being thus presented–that the relevant flexibility in the scope of an agent's perception only emerges through uncommon and abnormal circumstances in which the agent might find itself.[22] The flexibility will manifest, for example, in pairs of possibilities where one member is an exotic instance of visual trickery, like the specially-lit white wall made to falsely appear red or the cleverly painted mule passed off as a zebra. That is, on such a reading of the perceptual disjunctivist, they take the relevant flexibility to specifically be between normal and abnormal possible perceptual conditions.

Cases like Discus Throw, on the other hand, demonstrate that failures of **Parity** crop up *wherever* an agent exercises limited higher-order control, which need not involve such abnormal possibilities. There is no apparent inconsistency, for example, between the premises of the main argument in the last section and the stipulation that in no possibility are Sakaki's conditions for tossing the discus outside their ordinary, normal bounds. Perhaps if her possible tossing conditions were always *exactly* the same, or *absolutely* favourable, this would undermine our intuitions about her limited first-order and higher-order control. But we are allowing *some* possible variation and unfavourability, just variation and unfavourability within a normal range, without abnormal possible conditions for tossing the discus corresponding to the trickily lit wall or disguised mule as conditions for visual perception. If the flexibility of control posited by the disjunctivist is what explains failures of **Parity**, including in cases like Sakaki's, then this flexibility is just a consequence of our limited higher-order control, even when the only possibilities being considered are pretty narrowly limited. Disjunctivism is needed to account for our frailty (as manifest in our limited higher-order control) *as such*, not just our frailty in the face of possible conditions outside the norm.

The disjunctivist picture of control is opposed to views on which the topic of an agent's control is the same across all its possible cir-

[22] It may *sometimes* be part of the content of those positions. Theories of knowledge on which "normality" plays an important role, for example, may fall into this category when normality is treated as an all-or-nothing, though possibly contingent, state of a possibility. Theories of knowledge centrally relying instead on *comparative* normality less obviously fall into the category [Goodman and Salow, 2023].

cumstances. Good examples can be found in recent work on *options* in normative decision theory. According to several theorists, what an agent fundamentally controls is a *complete specification* of some state of itself. This may be, for example, a complete specification of what it does and does not intend [Hedden, 2012], of what atomic bodily movements at a given moment it and is not undertaking [Koon, 2020], or of all of its conditional policies [Pollock, 2002]. This full specification is always the strongest fact the agent controls; whatever else it controls will just be (some, not necessarily all) facts entailed by it. Since these specifications are complete, they thereby exclude one another; if any two were pairwise consistent, this would be because they leave some potentially distinguishing question about the agent's state unsettled. The subject-matter of the agent's control–the agent's complete state–thus partitions logical space, and is always the same.

Such views of control are incompatible with my disjunctivist picture, with the arguments about partitionality of the last chapter, and with the broader claim that Yukari actually controls that Chiyo is bruised but counterfactually controls no facts ruling out actuality. If the strongest fact Yukari controls in w entails that Chiyo is bruised, and Chiyo is not bruised in w', such theories demand she controls some fact in w' ruling out w. On its own this is not quite as strong as **Parity**. It leaves open that there is some w'' where Yukari has the same complete state S as in w, but that she controls different *consequences* in w and in w'' of the fact she is in S (the strongest fact she controls in both). But it does make **Parity** seem very natural (why should which consequences she controls be different in w and in w''?), and it rules out the way I have suggested of rejecting **Parity** in the case of Reckless Driver. Given the earlier argument against **Parity** itself, and how the disjunctivist picture helps reconcile our intuitions about Yukari's guilt and control, these are strikes against such "fixed subject-matter" theories.

This picture of control is, so far as I can tell from the extant literature on the subject, novel. There are some similarly disjunctivist, McDowellian views advanced about *intentional action*, but that is an importantly different matter [Levy, 2021]. The form of control over a fact that intuitively affects one's level, or scope, or other dimension of moral responsibility with respect to the fact does not, *prima facie*, necessarily consist in intentionally acting so as to ensure that fact. After all, events that have befallen through negligence sometimes seem

controlled in the requisite sense, since we adjust the moral grades of agents in light of them–as, for example, Yukari's bruising of Chiyo. But these events are, almost by definition, unintentional. Disjunctivism about control, therefore, has a wider (or, at least, different) scope than disjunctivism just about intentional action.

The *prima facie* plausible mere intensionality of control offers another important distinction between the concepts of control and intentional action. As noted early in Chapter 2, and as appealed to frequently throughout this dissertation thereafter, when p is necessarily equivalent to q, control of p amounts relatively uncontroversially to control of q; indeed, this is true not just for the broad or metaphysical necessity, but for a range of weaker modalities, too. Intentional action, on the other hand, may well be intensional in the ultimate analysis of things, but if so this will be a hard won insight into the nature of representational content. The representational character of intentional action renders it at least *non-obvious* that, say, intentionally drinking water amounts to intentionally drinking H_2O, or intentionally either drinking water and flying a kite or drinking water and not flying a kite. Control, in a phrase, relates the agent to *world states*, not *representational contents*.

Note that this does not bar one's cognitive state from being *relevant* to one's control. For example, a man who desperately wants to get into a locked room will have his level of control concerning whether the door is unlocked explained (in ordinary circumstances) by his state of knowledge about the lock's combination. But this does not show his control of the door's locked state to be a cognitive or representational or guise-sensitive matter by way of the representational character of that knowledge. After all, his state of knowledge about the combination similarly explains, in ordinary circumstances, whether he will soon be in the room. But this does not show his soon being in the room to be a representational state, nor his imminent location to be a representational attitude. His soon being in the room is as non-representational a state, and his imminent location as non-representational a relation, as any about him might be. This explanatory role of cognitive, representational states may explain away certain appearances that control itself is (or at least certain kinds of control are themselves) so cognitively laden and representational.[23]

[23]See, for example, the main motivating examples from [Levy, 2011].

Nothing in what I said about the distinction between control and intentional action excludes the possibility intentional action might be necessary for the relevant sort of control in other ways, either. Perhaps, for example, the inadvertent control exercised over Chiyo's head through Yukari's negligence must rest upon witting control exercised over some other subject matter through Yukari's intentional activity [Zimmerman, 2022]. Her driving the car recklessly, say. But this at most proves that the relevant variety of control depends upon the same subject *also* acting intentionally. This possibility no more demonstrates that control is intentional action (and disjunctivism about control just disjunctivism about intentional action) than analogous theses about knowledge always being downstream of perception, if correct, would demonstrate that knowledge is sensory perception (and disjunctivism about knowledge just disjunctivism about sensory perception).

So, while there are similar familiar views in *structure*, it is important to note that they differ from the one articulated in this section in their *topic*. But it is equally, if not more, important to distinguish this view from an established and superficially similar one that shares this topic (control, not intentional action) but lacks the same genuinely disjunctivist character. I have in mind a reading of the slogans of the control disjunctivist that affirms them by distinguishing *direct* from *indirect* control, and which would go something like this:[24]

> In the circumstance where Yukari bruises Chiyo, she controls something that she does not in the circumstance where she does not bruise Chiyo: to wit, that Chiyo is bruised. For in virtue of what she controls in a *direct* sense, namely her driving badly (or perhaps, her moving her hands and feet on the wheel and pedals erratically; or perhaps, issuing in haphazard driving-related volitions), certain relevantly connected causal effects issue, namely Chiyo bruising up, which she thereby controls in an *indirect* sense. In the circumstance where Chiyo is *not* bruised, however, while she *directly* controls the same facts (*viz.* that she drives badly), she does *not* control that Chiyo goes unbruised even in an indirect sense, for her bad driving in no sense is the, or a,

[24]The distinction being drawn on, between direct and indirect control, is from Zimmerman, though the envisaged argument is not [Zimmerman, 2006] [Zimmerman, 2022].

cause of Chiyo not bruising.

Similarity in direct control, in other words, does not guarantee similarity in (or even similarity in *extent of*) indirect control, which is included in the control covered in the control principle. This gap between direct and indirect control is how **Parity** fails, and thus how the control principle is reconcilable with differing moral grades for Yukari across her two circumstances.

This reading respects the letter but not the spirit of the key claims of the disjunctivist about control. For it is crucial to the disjunctivist vision, in control as in perception, that there is *no common factor* between the good and bad cases, that might be rescued as the "true" content of one's control or perception across those cases. The control or perception of an agent is flexible *in its entirety*, not just in some comparatively superficial aspect while the deeper component remains inflexible. So the reading just offered, which quarantines the flexibility at hand to the merely indirect control of the agent, violates the basic insight (or, perhaps, delusion) of the disjunctivist position.

Similar readings could (though to the best of my knowledge, historically have not) beset the disjunctivist about perception. Such a reading would agree that the subject really seeing the red wall to be red differs in what he sees from the illuded subject merely seemingly seeing the trickily lit wall to be red. But they would do so only in the sense that there is some second-rate, indirect, but still factive sense of "seeing" in which the former but not the latter "sees" the wall to be red by way of seeing directly the same thing the latter sees: that the wall looks red. (I think I can perhaps detect such a second-rate sense of "seeing", as in distinguishing the way in which the young man "directly" sees that his lover's shoes are in his rival's foyer and the way in which he thereby "indirectly" sees that he has been sexually betrayed.) This interpretation, clearly, is watering down the perceptual disjunctivist doctrine to the point of banality, in just the same way as the above-outlined hypothetical reading of control disjunctivism does.

Certainly the proposed reading would be one on which the parallel between perceptual and control disjunctivism would be illusory. But is it for that reason wrong? After all, there are plenty of pairs of philosophical positions superficially similar but deeply dissimilar, and

this fact on its own does not show either member of any of them to be wrong or ill-motivated.

Yes, it is, for on this reading disjunctivism cannot underwrite the arguments against **Parity** in the last section in the same way that disjunctivism about certain cognitive states underwrites similar arguments in epistemology. If this reading, call it that of the *pseudo-disjunctivist*, is correct, then failures of **Parity** are possible because of cases where there are *indirectly* controlled propositions in one possibility that can distinguish it from another even as there are no controlled propositions in that other possibility distinguishing it from the first. For example, in two possibilities w and v, the agent may in w indirectly control some p that is false in v, even as all the facts the agent controls (directly or indirectly) in v are true in w as well. **Parity** can fail according to the pseudo-disjunctivist, that is, in virtue of indirectly controlled propositions.

Parity, however, fails not just in virtue of indirectly controlled propositions. Were that so, then restricting attention to just direct control would validate **Parity**, and yet it does not. Recall that in §4.4.2 we saw that the argument from Discus Throw can be straightforwardly adjusted so that the controlled facts are just pure bodily movements, or even mere tryings or volitions. The problems with **Parity** emerge, that is, even when specifically restricting the controlled propositions under consideration to those controlled in the most direct manner you please. So, to accommodate these arguments, pseudo-disjunctivism will not do. There can be no residual core of control for which disjunctivism fails and **Parity** holds, if the arguments from the last section are correct and so motivate a disjunctivist solution to the original puzzle. A disjunctivism that so hedges its bets is one that cannot deliver on those points where it promises.

The right lesson to draw, I think, from the disjunctivist solution to the puzzle about Yukari is quite the opposite from that of the pseudo-disjunctivist. Far from relying for its plausibility on a distinction between direct and indirect control, the upshot of disjunctivism about control is that there is no such distinction of interest. *All* genuine control is "direct" control, including control over things like Chiyo's bruise. Just as, on McDowell's view, in seeing that the tree outside is blossoming we are open simply and immediately to *the fact that the tree is blossoming*, so in controlling that Chiyo is bruised Chiyo's head is vulnerable simply and immediately to *Yukari*. The perceptive agent is

subject nakedly to the world itself; the environment of the controlling agent is subject nakedly to the agent herself. She moves her body, and the rest is (*pace* Davidson) fatefully up to *her*.

This slogan is consistent with certain truisms about control and action which it might seem at first to be denying. Yukari controls that Chiyo is bruised *by* controlling that her driving is shoddy. That she controls whether her driving is shoddy *enables* her to control whether Chiyo is bruised.[25] And, in general, controlling agents will control some facts by controlling other facts enabling them to control the former. But these former facts, controlled by controlling other facts, are still directly controlled. The agent does not first control the facts by which it controls those former facts, and only then control those facts with an extra push or as a merely derivative excrescence. That some things (say, that Chiyo is bruised) an agent (say, Yukari) controls the agent is enabled to control by controlling other facts (say, that her driving is shoddy), does not make the enabling objects of control more "direct" in any important sense (on the disjunctivist view).

How can an agent control some fact by controlling another, without making the latter control more direct than the former? A parallel with perception will be instructive. Suppose that a perceptive subject, in ordinary conditions, sees a 1m-by-2m canvas before them to be uniformly green. Now, this is clearly going to be enabled, even according to a sensible disjunctivist, by their seeing the left half of that canvas to be uniformly green. Could they not see as much (if, say, the left half were obscured by an interposing object), they could not see the canvas to be uniformly green in turn. The same goes for any sufficiently large portion of the canvas of sufficiently simple shape. (Very small or very complicatedly shaped portions might raise trickier questions.) But, clearly, it is not as though the subject, by the disjunctivist's lights, *first* sees the left half of the canvas to be uniformly green, and *then* sees the canvas overall to be uniformly green by dint of that former seeing (perhaps in conjunction with other seeings, such as seeing the right half to be uniformly green). It is not so chronologically, and it is not so in any interesting sense "logically." Nor is the seeing of the canvas to be uniformly green merely derivative, or second-rate (if anything, it is the seeing of the left half to be so that is). The enabling seeing

[25] Thus I am not denying the existence of "basic actions," i.e. actions one performs not by performing any other actions, nor insisting that all actions will be basic. Or at least, I am not denying it simply for this reason.

here (of the left half) is part of the enabled seeing (of the canvas), but it is *immediately* such a part, rather than befalling the subject first and only then allowing them to go the extra step required to undergo the enabled seeing of the whole canvas, or issuing in that whole seeing as a mere perceptual byproduct. So, on my view, it is with control in cases like Yukari's: to enable is not in any deep sense thereby to precede.

This contrast with the pseudo-disjunctivist helps explain why **Parity** might initially have been appealing. When there is some hope that there might be a domain where the extent of our control is inflexible, it can foster the sense that our control over this domain is our *real* control. That is, there is a natural temptation to identify control with the pseudo-disjunctivist's "direct" control (though not necessarily accepting the pseudo-disjunctivist's further, more specific claims). This, in turn, makes **Parity** look plausible as a claim about this restricted, privileged form of control. If, as seems natural for such a fundamental ethical principle, the control principle also concerns this privileged form of control, the route to paradox is paved.

The allure of **Parity** in any form is diminished by seeing that the obvious candidates for this privileged, inflexible form of control are still subject to the main argument of the last section. Genuine disjunctivism about control, meanwhile, is consistent with the conclusions reached in the last section, and presents a coherent account of how paradox can be avoided in cases like Reckless Driver. The obvious reason to reject the disjunctivist picture, that control proper should be inflexible in the relevant sense, dissolves once the prospects for discovering such a form of proper control are revealed to be bleak. The initial paradox is thus a product of false hopes dashed in the last section, and disjunctivism about control a way of embracing that hopelessness and avoiding that paradox.

Many participants in the debate around the control principle, going back to [Nagel, 1979] itself, have proposed that the puzzle about cases like Reckless Driver is importantly related to sceptical arguments and conclusions. Usually these philosophers mean that the puzzle rests on real or apparent limits to an agent's (or its assessors') knowledge of its actions, and so our answers to sceptical questions about such limits will directly concern our answer to the puzzle [Pritchard, 2006] [Enoch and Marmor, 2007]. The current approach links the puzzle to sceptical arguments more indirectly: the solution (disjunctivism

about control) to one worry is an image, reversing direction-of-fit, of a solution (epistemological disjunctivism) to the other. Our more solid understanding of the problem of scepticism acts, not as a key, but as a guide to our understanding of control.

4.5.2 Unfairness

It is sometimes suggested that, were agents in positions like Yukari's to enjoy different moral grades across the relevant pairs of counterfactual circumstances, this apportionment of credit and blame might be objectionably "unfair." In general, where this is debated, it is not in cases *quite* like Reckless Driver, but rather more like the following:

> Reckless Would-Be Driver: Chiyo is waiting for a ride home from school. As it happens, Yukari does not happen upon her, and she instead picked up by a friend and driven home safely. Were Yukari to happen upon her, however, she would insist on driving Chiyo home and, driving incautiously, bruise Chiyo on the head.
>
> Would-Be Reckless Driver: Yukari drives Chiyo cautiously and safely, as is her wont. However, it could have been that her inborn disposition and formative early childhood experiences instead led her to be a habitually incautious driver and temperamentally insistent about driving those like Chiyo around, in which case she would have driven Chiyo recklessly and given her a bruise

In each case, it is again widely believed that assigning Yukari different moral grades across the pairs of circumstances presented would undermine the control principle.[26]

Still, the context of such debates generally makes clear that, if there is a doubt as to the fairness of differing moral grades for Yukari in the bad and good cases from, say, Reckless Would-Be Driver, this is a doubt that transfers as much or more to differing moral grades for

[26]These three cases–Reckless Driver, Reckless Would-Be Driver, and Would-Be Reckless Driver–are examples of what Nagel, in the founding documents of the debate, called *resultant*, *circumstantial*, and *constitutive* moral luck, respectively [Williams and Nagel, 1976] [Nagel, 1979]. This classification scheme has been widely adopted in the subsequent literature.

her in the paired cases from Reckless Driver. For the *reason* for the emphasis in these debates on cases like Reckless Would-Be Driver and Would-Be Reckless Driver, as is commonly made explicit, is that it is generally assumed that constancy of moral grade in cases like Reckless Driver is a much more plausible proposal than in cases like the two above, and that the problem of unfairness is therefore simply less likely to arise. As Philip Swenson writes (adjusting somewhat for our examples):

> Many of us have made peace with rejecting resultant luck. (When I use terms like 'resultant luck', 'circumstantial luck', etc., I mean resultant moral luck, circumstantial moral luck, etc.) [Yukari in her actual circumstances] and [Yukari in her counterfactual circumstances] are clearly both [blame]worthy. And perhaps it isn't too revisionist to say that they are equally [blame]worthy. Perhaps it is even intuitive. However, rejecting other sorts of moral luck poses greater difficulties. [Swenson, 2022]

If, for the reasons given, we *do* attribute differing moral grades to Yukari in the two cases from Reckless Driver, any doubts about fairness in the other two cases should therefore translate to Reckless Driver as well, if they have any purchase to begin with.

There is one way in which, if the way of thinking outlined above is correct, attributing different moral grades across the paired circumstances will straightforwardly *not* be unfair. To wit, it will not *eo ipso* contravene the control principle. But, of course, this is not the only way assigning moral grades might be unfair: we frequently castigate moral judges for blaming agents, even when the object of this blame is controlled by the agent. It is, for example, unfair to blame someone for clearly praiseworthy deeds. So nothing in what I have said on its own rules out the possibility that blaming Yukari differently in her two circumstances is somehow unfair.

What this does demonstrate, however, is that rejection of such purported unfairness cannot *underwrite* our commitment to the control principle. This goes against what seems to be a suggestion of Swenson's in [Swenson, 2022] (and earlier Roger Crisp in [Crisp, 2017]), that opposition to unfair allocations of blame in cases like Yukari's is what explains our commitment to the control principle. For if, our

analysis is right, there is no reason to think that different moral grades for Yukari across her circumstances will contradict the control principle; thus, no objectionable features in an uneven moral grading across the circumstances can explain the appeal of the control principle as applied in Yukari's case. If unfairness means something beyond contradicting the control principle, its rejection will not be more basic than acceptance of the control principle; if it simply means contradicting the control principle, we have already just shown differing judgements in Yukari's case to be perfectly fair.[27]

4.5.3 Alternative Control Principles

My argument so far has been directed against a specific understanding of the control principle. On this formulation, it connects differences between two possibilities in what an agent controls to differences between those possibilities in the agent's moral grade. It is worthwhile to consider, in overview, other versions of the principle, and prospects for extending the current argument to cover them. These remarks will only be cursory, but they provide, I hope, some basis for hope that this can be done successfully.

Before beginning in earnest, it will be worth clarifying one potential red herring. Formulations of the control principle often are given in terms of, not *what an agent controls* at a given world, but differences between possibilities *depending (or not) on factors the agent controls*. Such a principle will go something like: when no differences between two possibilities depend on factors an agent controls, the agent will have the same moral grade in each. So perhaps the preceding arguments will need to be refined.[28]

[27] I should be clear that neither Swenson nor Crisp takes differing moral grades in Reckless Would-Be Driver to actually *be* unfair (though things are presumably different for Reckless Driver). But their route to this conclusion is very different and significantly more elaborate than that I have outlined here, on which unfairness (as violation of the control principle) does not come up in assigning different grades to Yukari across her circumstances simply because she may control different things in her different circumstances, so that the difference between the two is not mere "luck." In particular, Swenson (building on Crisp) effectively develops a distinctive account of the relevant moral grade–initial expected desert–on which Yukari's moral grade is uniform across her relevant possible circumstances, allowing it to remain the same and the control principle thus preserved even supposing that she controls the same things in these various possibilities.

[28] Note that the universal negative quantifier here is essential. Suppose the principle

I do not think, however, that this difference in formulation poses substantively new problems. Now, I take it that a difference between two possibilities is just a proposition distinguishing the two. But, if that is so, then the prospect that there might be no differences between Yukari's possibilities depending on factors she controls in those possibilities is every bit as vulnerable to the argument as given as is **Control Principle**. For, degenerately, every proposition "depends" upon itself, so if (as this chapter has argued) she controls in one of two possibilities some proposition p (say, that Chiyo is bruised) that distinguishes the two possibilities, there is thereby a "difference" (namely: p) depending degenerately upon a "factor" (again: p) that she controls. Thus, such differences in formulation raise no trouble for our argument as stated.

This assumes that the relevant species of dependence is reflexive. What if it is, instead, irreflexive, or at least non-reflexive in some cases? Note first that there is something strange about this presentation of the control principle, so understood. "It was I who controlled whether p, and yet I did not control enough to be blameworthy for p" reads very strangely, but it is the sort of excuse that this understanding of the control principle would permit.

Setting this aside, though, this version still should not cause trouble, given everything we have seen, for the claim that Yukari's ac-

instead stated: when there is *some* difference between two possibilities *not* depending on factors an agent controls, the agent will have the same moral grade in each. Then the principle would be subject to immediate, trivial counterexamples. There are, for example, two possibilities, in the first of which Cain kills Abel and it is raining in Fiji, in the second of which Cain spares Abel and it is shining in Fiji. Clearly, Cain's moral grade should differ across the possibilities, even as he does not control any determinants of the weather in the South Pacific. For similar reasons, it is dangerously imprecise to speak of "the" differentiating factor between two possibilities, since in general there will be *many* such factors. This is the significance of formulating the principle as a supervenience thesis.

This is an important point, for philosophers sometimes seem to slip into talking as though it is these more exonerating principles that govern control and responsibility. Carlson et al., for example, write that "if Alf's and Bo's degree of blameworthiness depends on their degree of first-order control [because it varies across cases where their degree of first-order control varies], it depends on something beyond their control [because what they happen to control is controlled by the third party Celia]. And to acknowledge that an agent's degree of blameworthiness depends on a factor beyond their control is to affirm the existence of moral luck," which violates the control principle [Carlson et al., 2026]. But, reading the argument here at face value, this does not follow, for the presence of *some* differentiating factor not under the agent's control between the two differently morally graded possibilities does not show the principle to be violated.

tual and counterfactual moral grades differ. After all, the arguments given for her actually controlling that (say) Chiyo is bruised are easily adapted into arguments that she actually controls other facts on which that seems to immediately depend, such as that Chiyo hit her head hard on the car roof. Similar moves can be made in other cases; for example, in homicidal cases like that from the introduction, that the victim dies (playing the role of Chiyo getting bruised) will depend upon a bullet striking them (playing the role of her head hitting the roof). Thus, given that Yukari controls one distinguishing fact between her actual and counterfactual circumstances (say, that Chiyo is bruised), there should be other facts (say, that Chiyo hits her head on the roof) on which the first fact depends and also under Yukari's control. Since whether Chiyo is bruised distinguishes the possibilities, there is a difference between them depending on factors Yukari controls. Thus, even without reflexivity for the relevant dependence relation, this version of the control principle leaves the main point unaffected.

A similar point can be made about another minor difference in language between **Control Principle** and other standard formulations of the control principle: the distinction between *controlling* something and that something being *under* one's control. A common way of putting the control principle will go something like: when no differences between two possibilities depend on factors under an agent's control, the agent will have the same moral grade in each. But, again, this introduces no new problems, for it is pretty clear that things one controls are under one's control (even if it is not entirely clear that the reverse inclusion holds). So, the same reasoning as above can be applied, and all the prior argument goes through *a fortiori* just as well as before.

Proceeding to more substantive matters, there are at least three alternative (and overlapping) kinds of formulation of the control principle worth discussing from the literature. First, there are "direct" versions that say an agent only is responsible for a fact or event p if the agent controls p.[29] Second, what I will call "topical" versions relativise the moral grade to one topic and only compare which facts the agent controls *about that topic* (e.g. [Greco, 1995] [Hartman, 2017]). For

[29] For perhaps the purest, classic statement of such a version, see [Zimmerman, 1987]. It is commonly stated casually or alongside comparative versions.

example, such a principle might say that if Yukari controls in w and w' the same facts *about how bruised Chiyo is*, her moral responsibility *regarding how bruised Chiyo is* is the same in w and w'.[30] Finally, there are "cross-agent" versions that compare two *distinct* agents' responsibility and control. Here is one example:

> If persons S_1 and S_2 are [...] exactly alike with respect to some event X, except regarding factors that are outside of their control, then S_1 and S_2 are equally praiseworthy or blameworthy with respect to X. [Hartman, 2017]

I believe that my main argument can be adjusted to deal with each of these variations. Let us tackle them in turn.

We have already encountered "direct" versions above. To begin with, most philosophers who distinguish them from "comparative" formulations like **Control Principle** take the latter to be harder to reconcile with intuitions like **Culpability** [Greco, 1995] [Hanna, 2014] [Hartman, 2017]. So, if I have shown how to accept both **Control Principle** and **Culpability**, this is a *more* noteworthy accomplishment than doing the same for direct versions.

The usual stated puzzle about the direct version is that it conflicts with Yukari being responsible for Chiyo being bruised. But the only reason one might believe these two claims conflict is that one believes the description of the case shows Yukari does not control whether Chiyo is bruised, a claim itself requiring justification (just like **Equal Control**). **Parity**, as we have seen, provides a natural, widely accepted means of deriving this further claim from the description of the case: since Yukari in her counterfactual situation controls nothing ruling out actuality, and since in the counterfactual she does not (by factivity of control) control that Chiyo is bruised, so (by **Parity**) she

[30] Often, the topic of the controlled facts is described as an "event", see e.g. Hartman's formulation below. This cannot be meant in the ordinary way, however; after all, in Yukari's counterfactual without Chiyo's bruise, the event of Chiyo being bruised simply does not occur for her to control anything about it.

This is even more stark in the "cross-agent" versions described below. These versions are formulated as though Twin Yukari, by recklessly driving with Twin Chiyo, controls the same event as Yukari does by recklessly bruising Chiyo. The best way to understand these topics or "events," I think, is as *shared qualitative dimensions* along which agents' situations can vary, such as how battered the head of a child in one's backseat is, or where in one's field one's discus falls. This is how I will understand the term in what follows.

must not actually control it. If **Parity** is indeed the basis of this variant on the puzzle, the earlier anti-**Parity** arguments should resolve it as well.

Next, the "topical" versions. As a schematic such principle, we might write

> **Topical Control Principle**: For all agents S and possibilities w, w' and topics T: if the moral grade of S regarding T in w is different from its moral grade regarding T in w', S does not control exactly the same facts about T in w and in w'

Then we would weaken **Culpability** to say Yukari's grade *regarding how bruised Chiyo is* is different in the two possibilities. We do not need **Equal Control** in full to generate a conflict between these two claims. Instead, all we need is a weakened, topical **Equal Control** (saying Yukari controls the same facts about Chiyo's bruise in each possibility). This follows from a topical **Parity** (saying an agent controls the same facts about a topic T in both w and w', as long as in w it controls no facts about T excluding w'), a natural extension of the original principle, which does not (at least straightforwardly) require **Parity** proper. So merely rejecting the unrestricted **Parity** is not enough to resolve the puzzle.

But the restricted, topical **Parity** is wrong for the same reasons as the original **Parity**. All the d_i from the discussion of **Discus Throw** share a topic (where the discus falls), so the argument can be run again against the topical **Parity**. If this variation on the puzzle in fact rests on this topical **Parity**, it offers no new obstacle to the solution I have proposed.

Like topical formulations, "cross-agent" versions of the control principle concern only an agent's control over a restricted range of facts.[31] In particular, for an agent S, they concern S's control over facts of the form *that S is* φ. That is, they concern the agent's control over what properties it has. Say that agent S *gives itself* a property F iff S controls the fact that S is F (understanding "properties" liberally as the sorts of things denoted by simple or complex predicates). Then reformulating a cross-agent version in my terms would look something like

[31] Indeed, cross-agent formulations are generally topical as well. For good reason, as we will see shortly.

> **Cross-Agent Control Principle**: For agents S, S' and topics T: if the moral grade of S regarding T is different from the moral grade of S' regarding T, S does not give itself the same T-properties as S' does[32]

Principles like **Cross-Agent Control Principle** are thought to cause trouble for intuitions about cases like

> TWIN DRIVERS: Yukari is driving incautiously when she accidentally runs over a pothole, bruising Chiyo. Twin Yukari, under similar circumstances, drives equally recklessly but with no potholes, leaving Twin Chiyo unscathed.

where Yukari and Twin Yukari both intuitively have different moral grades regarding the wellbeing of children in their backseat *and* intuitively give themselves the same properties about the wellbeing of children in their backseat.[33] The former intuition is obviously on the pattern of **Culpability**, the latter on that of **Equal Control**. These three are inconsistent in the same way as before.

Why would anyone believe this analogue of **Equal Control**? A topical **Parity** in full is not needed, not even replacing talk of possibilities with talk of possibility-agent pairs and of controlled facts with self-given properties. All it takes is

> **Agent Parity**: If S gives itself no T-properties not had by S', S and S' give themselves the same T-properties

[32]Some philosophers distinguish a further *subjunctive* version of the control principle: if the moral grade of S *in* w regarding T is not that of S' *in* w', then S in w and S' in w' give themselves different T-properties (a form of *strong individual* supervenience). These are supposed to cause trouble for intuitions in cases like TWIN DRIVERS (see below) where Yukari and Twin Yukari are not worldmates, so to speak. When S is the same as S', this reduces to **Control Principle** as a special case, so it is easier to adjust my earlier argument from DISCUS THROW here. Hence my focus on the trickier **Cross-Agent Control Principle**.

[33]The topicality of **Cross-Agent Control Principle** is crucial, because if it instead said Yukari and Twin Yukari must give themselves the same properties *tout court* if Twin Yukari gives herself none distinguishing her from Yukari, it would be obviously false. Twin Yukari clearly gives herself plenty of properties lacking in Yukari, such as driving around *Twin* Chiyo rather than Chiyo. So the principle must be topical if it is to be remotely plausible.

Since Twin Yukari gives herself no properties about the bruisedness of the child in her back seat not shared by Yukari, the two must give themselves the same such properties. So the paradox emerges again.

The argument from Discus Throw against **Parity** does not affect **Agent Parity**. That argument crucially relies, to begin with, on the lemma that if S controls that p, S controls that S controls that p. Nothing like this follows from **Agent Parity**: since **Agent Parity** only involves what other agents *actually* control, instead of quantifying over all possibilities, it says nothing about entailment relations between facts about control. So my earlier argument is not quite enough.

Agent Parity, however, still has implausible consequences about Sakaki and her discus. Suppose again it falls centred around the field midpoint **c**, and suppose (plausibly enough) that the strongest fact she controls about where her discus' centre falls on her field is that it is within 1m of **c**. This fact specifies the smallest region R of the field such that she controls that the discus centre is in R (call it her *controlled region*). Suppose that Twin Sakaki, with her own similar field, under similar conditions, throws her own similar discus, falling (for some small ϵ) 1-ϵm to the left of the field midpoint **c'** (another plausible stipulation). What could Twin Sakaki's controlled region be?

If **Agent Parity** is true, it must be the points within 1m of **c'**. For she must control the same properties about her discus' relative location in her field as Sakaki does about *her* discus' location in *her* field. But this is clearly ridiculous. Why should they be exactly the same? What tight bearing should Sakaki's control over her toss have on Twin Sakaki's control over hers? How does she manage to prevent a deviation of ϵm to the left, for arbitrarily small ϵ, but not of almost 2m to the right? **Agent Parity** is, if anything, even more straightforwardly refuted by considering Discus Throw than **Parity** itself.

Much confusion is sowed by speaking as though Yukari and Twin Yukari somehow *jointly* control or do not control factors distinguishing their individual situations. Hanna, for example, states the principle as, "Two people ought not to be morally assessed differently if the differences between them are due only to factors beyond *their* control" (emphasis added) [Hanna, 2014]. But the two do not collectively control this. Rather, Yukari controls certain features of her car trip (that is, gives herself properties about her trip), and Twin Yukari certain features of (that is, gives herself properties about) hers. Whether Yukari controls any features distinguishing her from Twin Yukari, and

whether Twin Yukari controls any distinguishing her from Yukari, are two separate questions.

These comments on alternative versions of the control principle are only an initial survey. The list in this section is not exhaustive, the treatment given them is only preliminary, and more deserves to be said on how and how far the basic argument of §4.4.1 can be extended to other formulations of the relevant puzzle. Still, the above discussion suggests promise.

Finally, while I have framed the main philosophical concern about Reckless Driver as one around what Yukari *controls*, some philosophers contend that the relevant principle is not about control but about *luck*.[34] According to them, it is not (merely) that responsbility requires control, but that (more or less) luck precludes responsibility [Pritchard, 2005] [Pritchard, 2006] [Driver, 2012] [Levy, 2011] [Peels, 2015]. Without entering into the subtleties of this expansive literature, I will say merely that the compatibility of the control principle itself (however understood) and Yukari's guilt is independently interesting, whether or not one thinks that we should endorse anti-luck constraints on responsibility besides.[35]

4.6 Conclusion

This chapter has developed a response to puzzles about Yukari's blameworthiness in Reckless Driver along different lines than traditional approaches, inspired instead by the logical investigations of the previous chapter. Rather than revising either our intuitions about Yukari's responsibility or the link between responsibility and control, or even claims about the relation of control to other concepts like epistemic access or alternate possibilities, my argument has focused on the pure logic of control itself. If correct, this offers promise for a more

[34]For recent entries in the dispute on the noncontrol side see [Hartman, 2017] [Hartman, 2019] [Statman, 2019]; see also [Coffman, 2015] for a general overview.

[35]Most of these philosophers still agree Yukari does not control Chiyo's bruise; for example, Pritchard concedes in passing and without contest that the reckless driver has a "lack of control over [her] consequences" (like passengers' bruises) even as these consequences may not be *lucky* (and thus are possibly blameworthy) [Pritchard, 2006]. Thus, Pritchard still resolves the paradox as stated by rejecting the control principle, rather than harmonising it with the intuition about Yukari's culpability by rejecting **Equal Control**.

general investigation of the links between the logic of control or action on the one hand, and moral philosophy on the other–a traditionally neglected area of cross-fertilisation.

One traditional question I have not addressed is whether my view ascribes *moral luck* to Yukari.[36] On the present view, the question is critically ambiguous: since the control principle and pre-theoretic intuitions about guilt are generally taken to be incompatible, philosophers (or at least, philosophers who identify luck with failure of control) have usually conflated both rejection of the principle and embrace of intuition under that heading. Once we dispense with the principles required to derive this incompatibility, the category of "moral luck" thus loses both coherence and interest.

4.A Agglomeration

My argument against **Parity** relies on a generalised (infinitary) agglomeration principle for control, **Agglomeration**. One might wonder whether this strong a principle is necessary for the result, or if it could instead be replaced with a finitary version (perhaps at the cost of argumentative expediency):

> **Finitary Agglomeration**: For all S, w: if, in w, S controls that p and S controls that q, S in w controls that p and q

It could not.

Proof. To see this, suppose our domain is $\mathbb{R}$ and the actual world 0. For $w \in [0,1]$, let the controlled facts of Sakaki be the filter on $[0,2]$ generated by the prefilter of all $[0,x]$ for $1 < x \leq 2$; for all other w', let the unique controlled fact be $\{w'\}$. For all $0 < n \leq 50000$, let $[\![d_n]\!]$ be $[0,n]$.

The limit of Sakaki's control will be d_2. Clearly, **Graded Uncontrol**$_1$, **Graded Control**$_2$, **Parity**, **Convexity**, and **Finitary Agglomeration** are all satisfied. So too is **Uncontrolled Limit**: $[\![$ Sakaki controls d_2 but does not control $d_1]\!]$ is just $[0,1]$, which is nowhere in her control. □

[36] "Resultant" moral luck in particular, (un)lucky responsibility for the consequences of one's actions, usually regarded as the most dubious variety of alleged moral luck [Ferrante, 2009] [MacKenzie, 2017] [Hartman, 2019]. On our view, of course, it is misleading to call Chiyo's bruise a mere "consequence" of Yukari's action at all.

The philosophical significance of this limitation will be discussed in the conclusion. More formally, it demonstrates that the metalogical results of Chapters 2 and 3 concerning the logics Conv4 and Conv5* with respect to the sandwich semantics cannot be extended straightforwardly to a more general semantics replacing sandwich frames with pairs $\langle W, S\rangle$ where (possibly partial) S can map $w \in W$ to any convex sublattice of $(\{X : w \in X \subseteq W\}, \subseteq)$. This straightforwardly generalises similar rudimentary results in the Scott-Montague semantics for S4 and S5.

Chapter 5

Conclusion

The preceding chapters have presented and applied a theory of the control agents like ourselves exercise over our environments. The main purpose of this theory and its applications, as advertised in the introduction, has been to address certain sceptical worries about the extent of our control. I take these worries to have been dealt with by the theory in the immediately preceding two chapters.

In addition to this, however, in the course of developing and applying the theory there have emerged a number of recurring threads of philosophical interest, which we have at times been forced to pass by. In this conclusion, I would like to take the opportunity to note and partially develop some of them.

5.1 The Inadequacy of the Object Language

Two of the central arguments of this dissertation–the argument against partitionality in Chapter 3 and the argument against **Parity** in Chapter 4–share a common overarching structure, though they differ slightly in implementation. In both cases, it is shown that the thesis to be refuted (partitionality, **Parity**) forces the acceptance of certain axioms for making true, which then are noted to conflict with the premise that the agent's higher order control is limited in a certain way.

It is clear, in one way, that such arguments will inevitably not take place entirely within the object language of our logic of action. For

the theses to be refuted, as we have seen, are impossible to formalise within the object language (see Theorem 3.A.5 and Corollary 3.A.6). Thus, it is necessary to appeal to linguistic devices not available in that object language to so much as state the conclusion, let alone to conduct the argument. It is not for *this* reason a drawback to this dissertation's argumentative strategy that it has relied on such devices (direct quantification over possibilities, extra operators, infinitary logical operations) in the course of arguing its central theses, for these terms are forced upon it intrinsically by its opponents.

There is, however, a further sense in which both arguments depend upon these linguistic devices, which is *not* a simple consequence of the terms of the debate, and thus is a potential rhetorical drawback. In both cases, some principle not liable to formulation in our official unimodal object language is required to move from partitionality or **Parity** to the bad axioms. In the former case, it is the bridge principles linking making true and practical necessity; in the latter, as just shown, it is specifically infinitary agglomeration for making true. This is, I think, a somewhat unappealing quality to the arguments, for it in a precisely specifiable way means that they must avail themselves of conceptual resources beyond the logic of control discussed and motivated in Chapter 2. I think these are defensible further conceptual resources, to be clear, but that does not mean (rather, it indicates the contrary) that they are not nevertheless still in need of defending.

Part of the defence to be made is simply that the additional principles are plausible in their own right. I gave some intuitive motivations for the bridge principles concerning control and practical necessity already in Chapter 3. As for infinitary agglomeration, for one it is hard to see why agglomeration might hold, but only finitarily. That would seem rather arbitrary. There is also the intuitive appeal of the Barcan formula in the case of control: that if, for each thing, the agent makes it φ, the agent makes it so that everything is φ. As usual with the Barcan, the intuitive appeal is somewhat dampened by raising questions about necessitism and contingentism. But, when these questions are bracketed, it is very hard not to simply infer from each thing (where the quantifier is perhaps suitably restricted) having been made φ by the agent to the agent having made everything φ. Infinitary agglomeration supplies an explanation for this intuition. In other words, the "extra" principles in the logic are justifiable in the same way as the original ones from Chapter 2.

It is also perhaps telling, however, that there are two separate such routes to the refutation of **Parity**. Much the same basic insight is clearly doing most of the work in both of the big arguments. And yet, this same big idea can be made to work by way of either of two apparently very different additions to our set of premises: a schema in a bimodal language relating modality to control, and a statement in the informal metalanguage governing the relation between control and a generalised conjunction. Were it just one such path to the desired conclusion that was available, it might increase apprehension that the dialectical move was a sort of mere trick. The presence of both paths helps assuage this concern.

5.2 Control and Modality

As noted in §2.3.4, an important distinguishing feature of our core formal theory of control is its independence from certain "thick" modalities, like that of historical necessity. Still, the theory as developed has not been *completely* void of modal concepts and considerations. For one, as the logic of a congruent operator (recall from §2.3.2), its natural models come equipped with an important, if "thin," modality: that of unrestricted or broadly logical necessity, the modality represented by the universal accessibility relation on the domain of worlds. This broad modality, and control's congruence or intensionality in respect of it, also featured centrally in the reasoning of Chapter 4. And at several points in the argument, it has been useful to point the to intuitive plausibility of stronger congruence-like principles in terms of stronger factive modalities, like that expressed in **Practical Strengthening**, which say (or are entailed and explained by the antecedently plausible claim) that (necessarily, for all p and q), if p and q are necessarily equivalent (for the relevant kind of necessity), controlling that p and controlling that q are materially equivalent. As observed early on, it makes sense to see the principle expressed in the rule of congruence as a special case (in the broadest modality) of this wider phenomenon (in a number of modalities). Is there anything more to say in general about this link, or class of links, between control and modality?

One obvious question it raises is what, if anything, connects the

modalities for which such principles hold (call them *quasi-congruent*).[1] Clearly, not every species of necessity satisfies such a principle. The failure of the principle of **True Strengthening** (from §2.3.3), for example, goes along with a failure of a congruence-like principle about *fatalistic* necessity, in which to be necessary is to be true.[2] Is there any plausible deeper account of what ensures the congruence-like principles about just those modalities in which they hold?

There are a couple of different ways to think of the way in which this class of modalities is unified. The first is a more formal, model theoretic perspective. On this way of seeing things, given a certain sandwich frame $\mathcal{F}$, the quasi-congruent modalities are a certain class of accessibility relations definable on the world domain of $\mathcal{F}$, which comes with desirable properties. Significantly, the sandwich function S on the frame fixes this class in a neatly specifiable way. Thus, in a sense it is *what one controls* (throughout logical space) that explains the unity of this lattice.

The other way to look at the matter is more philosophical, and takes the reverse perspective. Rather than this range of different species of necessity unified by the scope (possible and actual) of one's control, it is these necessities which collectively help restrict what that scope can be in any given world. The aforementioned class of modalities then falls out as an epiphenomenon of no fundamental philosophical import. We will examine these two ways of considering the problem in turn.

5.2.1 Quasi-Congruence in Sandwich Models

Before embarking, let us state the main result upfront. The quasi-congruent modalities in a given sandwich frame $\mathcal{F}$ (or model $\mathcal{M}$ based upon $\mathcal{F}$) constitute a convex bounded sublattice of the lattice of reflexive relations on the domain of $\mathcal{F}$, as ordered by set inclusion. The upper bound of this lattice is just the universal relation, corresponding to the unrestricted modality. The lower bound, meanwhile, is

[1] This expression is not entirely to my liking, as it would be more accurate to say that control is quasi-congruent in respect of a given modality than that the modality is itself quasi-congruent. But I know of no better succinct way to express this property of certain modalities.

[2] Thus, in an arbitrary frame its logic is Ver and it is represented by the diagonal accessibility relation $\{(w, w) : w \in W\}$ on the domain W.

definable in a straightforward way from the sandwich function S of $\mathcal{F}$; in particular, what is necessary in this sense at w is given simply by what the agent controls at w. Thus, for any such frame (model), we can speak of the lattice $\mathcal{F}_Q$ ($\mathcal{M}_Q$) of quasi-congruent modalities on it, abbreviating to simply Q where the frame or model is given by context.

For readability, several propositions will be indicated without formal statement or proof. The formal elaboration may be found in the appendix.

Let us first render our earlier definition of a quasi-congruent modality slightly more formal. Modalities are understood in this context as represented by accessibility relations of a domain of worlds, and one will restrict (extend) another if it is a subset (superset) of the other.[3] A modality $\boxdot_1$ is *quasi-congruent* just in case, necessarily, for all p, q: if p and q are $\boxdot_1$-necessarily equivalent, the agent controlling p and its controlling q are materially equivalent (see Definition 5.A.1).

Note that the quasi-congruent modalities are clearly closed under extension, since if

$$R(w) \cap p = R(w) \cap q$$

then for any R' restricting R you have

$$R'(w) \cap p = R'(w) \cap q$$

so that if R' is quasi-congruent you will thereby have

$$S_{\mathsf{min}}(w) \subseteq p \subseteq S_{\mathsf{max}}(w)$$

iff you have

$$S_{\mathsf{min}}(w) \subseteq q \subseteq S_{\mathsf{max}}(w)$$

It is natural to ask why the discrepancy, in the definition given above, between the strength of equivalences in the antecedents and consequents of a congruence-like principle. That is, why formulate it as in the preceding, rather than on the pattern of "Necessarily, for all p, q: if p and q are $\boxdot_1$-necessarily equivalent, the agent controlling p and their controlling q are $\boxdot_1$-necessarily equivalent"? After all, that is the natural "translation" of the rule of congruence **RE** into

[3] For a much more involved exploration in a more general setting of the idea of a hierarchy of necessities as ordered by such a relation of extension, see [Bacon and Zeng, 2022].

a sentence of natural language (treating the metalinguistic $\vdash$ like a necessity operator), and also applies in the case of the (in general false) rule for fatalistic necessity, in which necessary and material equivalence coincide. Why should the same pattern not also hold for e.g. practical necessity? How, if not, are these instances of the same phenomenon?

There are three interconnected points to observe here. First, the "mixed" version in the extremal cases of the unrestricted and fatalistic necessities is equivalent to the "unmixed" version just gestured at. This is obvious in the fatalistic case (since material and fatalistically necessary equivalence are the same), the unrestricted case is only slightly more complicated. (See Proposition 5.A.2.)

Given the assumption that the unrestricted modality satisfies at least S4, this means that it cannot satisfy the "mixed" principle without also satisfying the "unmixed" one. Indeed, Corollary 5.A.3 shows this is a general feature of S4 necessities. The proof also does not rely undischargedly on the congruence of the operator $\Box$, so that the equivalence is not simply a consequence of the truth of both. Thus, in these limiting cases, the mixed and unmixed versions are equivalent. The fact that the principles are naturally given for them in their unmixed form thus says nothing of depth.

Second, if $\boxdot_2$ extends $\boxdot_1$, the unmixed congruence principle for $\boxdot_1$ will not in general entail that for $\boxdot_2$. This puts it in contrast to the mixed principles, as we have seen quasi-congruence to be closed under inclusion. And finally, the mixed version for a given modality does *not* in general entail the unmixed version. Both these points can be illustrated by simple countermodels. In fact, we can give a simple countermodel for both simultaneously, see Proposition 5.A.4.

The hierarchy of quasi-congruent modalities exhibits a striking feature in our models. The existence of such a most restricted quasi-congruent modality $\boxdot$ is guaranteed in any standard sandwich model, where modalities are represented by possible accessibility relations. To extract this relation R from a sandwich model, first add to R all reflexive pairs (w, w) for w in the model domain. Next, for any such w, add (w, v) iff S is defined at w, and either $v \in S_{\mathsf{min}}(w)$ or $v \notin S_{\mathsf{max}}(w)$. A world is possible in this sense, intuitively, if either *none* or *all* of one's (actual) actions occur there. R is, additionally, a quasi-congruent

modality.[4] R is therefore the most restricted possible modality in the model to satisfy a (mixed) congruence-like principle, i.e. is the most restricted quasi-congruent modality.

Note that this necessity will not thereby be *definable* in terms of control and the Boolean operators, as a simple model illustrates (Proposition 5.A.5). Since quasi-congruence is closed under extension, the quasi-congruent modalities are all and only those to extend R. This suffices to prove the original contention about the lattice structure of these modalities.

One can think of this lattice Q as generated by the minimal quasi-congruent modality R (or, to signal its status as greatest lower bound of the lattice, $\bigwedge Q$), itself generated in turn by the sandwich function. Thus, what unifies the quasi-congruent modalities, on this picture, is the way in which they can be extracted from what it is an agent controls in each world.

What is striking is not the mere existence of this modality. As we will see in the ensuing section, that the quasi-congruent modalities relative to a (properly) congruent operator like "makes true" have such an infimum is a result of general properties of quasi-congruence, rather than a peculiarity of control or its sandwich semantics. Rather, it is the simple formulation in terms of the sandwich function: these minimal possibilities are exactly those either ruled out by none of the agent's actions or ruling out all of them. It is a captivating thought that this species of possibility might play a special role in understanding control philosophically.

5.2.2 The Meaning of $\bigwedge Q$

Just as with the sandwich function S, the lattice Q of quasi-congruent modalities raises questions about its philosophical import. In particular, what, if anything, should the infimal quasi-congruent modality in a model be taken to stand for, informally speaking? Should we take it to have some special significance in understanding the ubiquity of congruence-like principles for modalities of independent interest?

[4] We saw in §3.2.2 that every quasi-congruent modality extends this modality. Suppose, moreover, that p and q coincide within $R(w)$, i.e. that they are $\boxdot$-necessarily equivalent. If the restriction of them to $R(w)$ is not S_{min} (if defined), then $\Box p$ and $\Box q$ are both false at w. If the restriction is S_{min}, then $\Box p$ and $\Box q$ are both *true* at w. If S is undefined at w, they are both false trivially.

This is reminiscent of the problem addressed in §3.2.2: to find, if any exists, an intuitive interpretation of S_{min} (and S_{max}). Once again, we are confronted (at each world w) in a given sandwich model by a bounded convex sublattice of a larger lattice given simply by the unstructured domain of the model, with that sublattice (and so its lower bound) induced by the model's sandwich function, and tasked with discerning any importance of the bounds of that lattice to the underlying phenomenon of philosophical interest. In the previous case, we eventually decided against attributing any special significance to the lower bound $S_{\mathsf{min}}(w)$ of the sublattice at a given world, leaning instead towards treating the sublattice as a whole, or perhaps some class of generators not given by the sandwich function, as representing the true relevant things of interest. Might a similar conclusion apply here?

We should note, to begin, that the arguments from §3.2.2 do not translate perfectly to this case. For one, whereas we saw that in a generalised neighbourhood semantics for Conv (rather than the Kripke-esque sandwich semantics) the lattice of controlled propositions at a world will not be guaranteed to have a lower or upper bound, the analogue of Q in a generalised semantics will still necessarily be bounded, because it is still *complete* as a lattice. The lattice as a whole will therefore still have a lower bound, and will be guaranteed an upper bound because a generalised possible worlds setting still has a canonical unrestricted modality.

To see this point, first adjust the definition of a quasi-congruent modality to the generalised case. Rather than *relations*, modalities will now be represented by *functions* from worlds to *filters* on the powerset of the domain. To preserve harmony with the terminology from the Kripke-like setting in which these functions to filters are represented by relations on the domain (i.e. in which the filters are principal), we will consider the lattice of them as ordered by *reverse* inclusion at every argument. That is, for two modalities M_1 and M_2, M_1 will restrict M_2 (M_2 will extend M_1) iff, for all w, $M_1(w) \supseteq M_2(w)$ (see Definitions 5.A.7 and 5.A.8).

This definition clearly generalises the old one, and corresponds in the generalised semantic setting to the same congruence-like principles as before. That the (factive) modalities on a frame form a complete distributive lattice, with the lower bound the natural function from worlds to principal ultrafilters and upper bound the unrestricted

modality (the constant function to $\{W\}$), is left as an exercise to the reader.[5]

Even in the generalised, neighbourhood setting, there exists a privileged minimally extensive (or, if one prefers, maximally strong or restricted) quasi-congruent modality, given any pattern of the scope of an agent's control across all possibilities (see Proposition 5.A.9). Still, the question remains as to just how philosophically significant this infimal quasi-congruence really is, even granting it must exist.

There is another problem raised by the infimal $\bigwedge Q$, already foreshadowed to some extent in §2.4. On the picture developed in the preceding section, one's $\bigwedge Q$-possibilities at a w where one acts are deeply heterogeneous. On the one hand, there are those possibilities so close and familiar as to be consistent with *all* one's actual actions; on the other, there are those so distant and exotic as to be consistent with *none* of them. The latter are reminiscent of the "alien possibilities" of §2.4.3; indeed, if the necessity that (according to Korsgaard) dooms agents to act is quasi-congruent, then *each* of these remote possibilities will be an alien possibility. What is more, this infimal modality is supposed to restrict many other familiar ones. Why, then, do all these modalities include possibilities so far as well as so near?

In answer to the first problem: I think not very. First off, the preceding result about the completeness of the lattice of quasi-congruent modalities arguably casts doubt on the significance of the infimum $\bigwedge Q$. Suppose, for example, that the real story about the quasi-congruent modalities goes like this. For deep, independent philosophical reasons, there are two incomparable modalities that are each quasi-congruent. The other quasi-congruent modalities come along "for free" given these two–including the infimal such necessity $\bigwedge Q$ and the unrestricted necessity. This is a story on which $\bigwedge Q$ exists, but is clearly of no deep philosophical import. Perhaps the real story is something like this.

Here is another consideration. Of the quasi-congruent modalities we have considered, none are really suitable candidates for $\bigwedge Q$. Clearly the unrestricted necessity is not, for then an agent would only ever control one fact at a given world. But neither is practical necessity, nor the "modal proximity" modality expressed by "It could not

[5]Hint: Prove for an arbitrary w that the filters with members all containing w form a complete distributive lattice, then define the lattice of modalities pointwise.

easily be that not-φ". On practical necessity, consider Alex and his paper-and-pen, from §3.3. Were practical necessity $\bigwedge Q$, this would require that, if he both makes the pen fall on the paper and makes it fall of some proper part of the paper's surface, it is not practically possible that it fall on the paper outside that part. This is, clearly, stretching the concept of practical necessity beyond plausibility! On modal proximity, suppose that Alex places the pen deftly but fickly, so that he makes it fall on the paper, specifically making it fall within 1cm of η, but could easily have instead made it fall within 1cm of η', 5cm away from η. Once again, the identification of practical necessity and $\bigwedge Q$ then leads to contradiction.[6] Thus, both of these initially plausible candidates for a philosophically significant $\bigwedge Q$ turn out to be non-starters.

If $\bigwedge Q$ does not play a fundamental role in explaining the ubiquity of congruence-like principles for modalities of interest, what does? Here is an alternative story. Drawing on the remarks about the objects of control as world states in §4.5.1, it is important to the concept of control in our understanding of other agents that we treat their control (in a given world) as not distinguishing among environmentally indistinguishable states. If, for example, we treat as a fixed fact of a lizard's desert environment that for it to be exposed to the sun is for it to be warm, then in understanding it as controlling its warmth we should thereby understand it as controlling its exposure to the sun. Its control should not be understood as respecting distinctions not respected by the background environment against which it acts.

The congruence-like principles give formal expression to this norm of understanding. Various dimensions D of the environment (such as the relation between warmth and sunlight) go along with facts about D fixed by the environment (such as that the lizard is warm iff it is sunlit). The collection of the D-fixed facts at each world w determines pointwise the quasi-congruent D-modality. One way of describing recognition of such quasi-congruence informally is as *restricting* attention among all the possibilities to just the D-possible ones. But I prefer to talk of it instead as *individuating* possibilities coarsely, though not in the sense of replacing the possibilities with equivalence classes of the same. The various D-impossible ways of realising some non-maximal

[6] It will do no good to interpret the relevant notion of proximity or ease as proximity or ease *given how one in fact acts*, for then (impossibly for a quasi-congruent modality) it will exclude all the possibilities inconsistent with each thing one makes true.

possibility are *collapsed* in this coarsening, thus screening them off as irrelevant. Crucially, this coarsening is world-relative: different possibilities (in the unrestricted sense) may be collapsed differently depending on one's perspective in logical space. The formalisation in terms of our models, of course, cashes out in both cases to the same thing, namely quasi-congruence in the sense formalised above. But I find the coarsening a more illuminating metaphor.

This helps explain the second puzzle mentioned, that of the seeming heterogeneity of $\bigwedge Q$ and other quasi-congruent modalities. The possibilities to exclude all things the agent makes true at w might at first seem objectionably alien or exotic (to w). But to be alien or exotic to w is a roundabout way of saying they are D-impossible at w, for some dimension D of the agent's environment! Thus, such D-impossibilities will be the very ones to be screened off by the coarsening, and so included in some (coarsely individuated) object of the agent's control at w. So the initial temptation, to make them out as very distant given our actions just as the members of S_{min} are thereby very close, is the reverse of reality.[7]

The variety of dimensions to one's environment we treat as fixed determine the various quasi-congruent modalities of interest to understanding control. Each such dimension itself, of course, will correspond to a quasi-congruent modality of its own. But the lattice structure of these modalities also entails that these modalities will have a greatest lower bound. That bound is practical necessity, which thereby amalgamates *all* the quasi-congruent modalities of independent interest. This is not to say that practical necessity is $\bigwedge Q$, however. Further such modalities are induced by the particular action propositions taken by an agent among those practically possible at each world, with their infimum $\bigwedge Q$. But these modalities are mere byproducts of those choices of action on the agent's part, amongst the (environmentally individuated) possibilities open to it. They represent, at most, what could further have been taken as fixed without thereby distorting one's understanding of the agent's scope of control in a given world.

So far the picture has only really described the mechanics of the *attribution* of control. On the picture thus far, a special attitude towards

[7] It is not clear to me that they must be close enough that one will be acting in all of them–might not inertness be a D-possibility, for each D? So the proposed requirement that practical necessity or some other quasi-congruent modality make it necessary one act may still run into the kinds of implausibility mentioned in §2.4.3.

certain modalities (treating them as fixed) restricts what someone interpreting an agent can think that agent controls, by taking their control to satisfy congruence-like principles in respect of those specially attended modalities. But then, which modalities will be *actually* fixed, so that an agent's control is thereby *actually* congruent with respect to them?

I think the most plausible answer is along broadly interpretivist lines. Control features heavily in our understanding of agents because it plays an important role in explaining the way in which they navigate and influence their environment. The actually fixed facts about an agent's environment, along various dimensions, are accordingly just those facts such that the agent's (say, the lizard's) navigating and influencing its environment (say, the desert) is actually well explained by taking its control not to distinguish among possibilities collapsed by that class of facts. Note that the agent navigating and influencing this environment (at all) is a weaker thing than the agent navigating and influencing this environment *exactly as it actually does*, so fewer facts need be treated as fixed to explain it than all those facts according to which the agent's control does not actually discriminate, i.e. fewer facts than the value of $\bigwedge Q$ at actuality.[8] This is admittedly sketchy, because it will still require a theory as to the role control plays in the explanation of agency. But it does, at least, point the way to a bridge from an account of control *attributions* to one of control *simpliciter*. And, to the extent the main contents of this dissertation provide rudiments of such a theory, they make a start on traversing that bridge, as well.

[8]One might object: This still leaves the relevant modalities underdetermined! After all, this at most only speaks as to which *filters* of propositions are the values of fixed modalities (considered as functions from worlds to such filters) at *this* world. What of the value of the D-modality at another world w?

The answer depends on how w is specified. If it specified in generic, non-rigid terms as a counterfactual possibility (say, as the world that would obtain were the lizard from above in a forest rather than the desert), then it is easy to generalise the answer given here: the *counterfactually* fixed facts about D are just those facts such that the agent's navigating and influencing the D-aspects of its environment *would be* well explained by taking its control not to distinguish among possibilities collapsed by that class of facts. If w is instead specified rigidly, this already determines everything true-at-w, including which facts about D are fixed there. So in that case the question is redundant.

5.3 The Lessons of Higher-Order Impotence

At the heart of this dissertation's basic argument is a simple, seemingly innocuous assumption. As frail, finite agents, we wield only imperfect control over our behaviour. For any bit of behaviour of ours you pick, we control how it plays out only to a limited degree. Moreover, we do not perfectly control *what* that limit is; this is a further limit in our *higher-order control* over our behaviour, beyond the first-order limits just mentioned. This has the air of a truism. On its own, nothing could seem more innocent.

Appearances, however, are deceiving. From arbitrary instances of this general expression of our agency's finitude, several revisionary consequences follow. We learn, for example, that our actions cannot partition logical space in the way they are usually assumed to. We learn from that, in turn, that fundamental problems in decision theory are not amenable to the easy solutions traditionally offered for them. And we learn that a supposed deep paradox about the nature of moral responsibility is no paradox at all. Again and again, higher-order impotence turns out to overturn our familiar ways of thinking about these topics.

We may view this, of course, exactly in the direction stated. Higher-order impotence is a fertile source of philosophical insight, touching on several seemingly distant topics and areas. But we may also consider the matter contrapositively. Insofar as the assumption of higher-order impotence has far-flung corollaries going against the grain, this indicates that its negation, a counterintuitive assumption of perfect higher-order control, is equally widely engrained. This assumption, that where our control over our behaviour is limited we control its limits, unifies common strands of philosophical thought across several subdisciplines and phenomena of philosophical interest. In the view developed here, of course, this is unification by falsehood. But it is a unity worth observing all the same.

5.A The Logic of Quasi-Congruence

In this appendix we prove several results about quasi-congruent operators mentioned and relied upon in the preceding.

Definition 5.A.1. *Given a sandwich frame $\mathcal{F}$ (or convexity model $\mathcal{M}$ based upon $\mathcal{F}$), a* quasi-congruent modality *R is a reflexive relation on the domain $\mathcal{F}_W$ such that, for all $w \in \mathcal{F}_W$ and $p, q \subseteq \mathcal{F}_W$, if $R(w) \cap p = R(w) \cap q$, $S_{\mathsf{min}}(w) \subseteq p \subseteq S_{\mathsf{max}}(w)$ iff $S_{\mathsf{min}}(w) \subseteq q \subseteq S_{\mathsf{max}}(w)$*

On the standard interpretation of a modal operator by a relation on the domain of a model, this will validate the exact kind of congruence-like principle described earlier for the modality interpreted by R.

Proposition 5.A.2. *Let $\blacksquare$ satisfy the logic* S4*. Then $\blacksquare(\blacksquare(\varphi \leftrightarrow \psi) \rightarrow \blacksquare(\Box\varphi \leftrightarrow \Box\psi))$ is equivalent to $\blacksquare(\blacksquare(\varphi \leftrightarrow \psi) \rightarrow (\Box\varphi \leftrightarrow \Box\psi))$.*

Proof. By T and necessitation, the left to right direction is trivial.

For the right to left, take an instance

$$\blacksquare(\blacksquare\alpha \rightarrow \beta)$$

(where α and β abbreviate $\varphi \leftrightarrow \psi$ and $\Box\varphi \leftrightarrow \Box\psi$ respectively). Then, by 4, we have

$$\blacksquare\blacksquare(\blacksquare\alpha \rightarrow \beta)$$

Reasoning (by K and necessitation) within the outermost modal, we get

$$\blacksquare(\blacksquare\blacksquare\alpha \rightarrow \blacksquare\beta))$$

by an application of K. Applying the S4-equivalence of $\blacksquare$ and $\blacksquare\blacksquare$, we get

$$\blacksquare(\blacksquare\alpha \rightarrow \blacksquare\beta))$$

Which gives the desired result. □

Corollary 5.A.3. *Let $\blacksquare$ and $\boxdot$ satisfy the normal bimodal logic with the* S4 *axioms for each modal and the further axiom $\blacksquare\varphi \rightarrow \boxdot\varphi$. Then $\blacksquare(\boxdot(\varphi \leftrightarrow \psi) \rightarrow \boxdot(\Box\varphi \leftrightarrow \Box\psi))$ is equivalent to $\blacksquare(\boxdot(\varphi \leftrightarrow \psi) \rightarrow (\Box\varphi \leftrightarrow \Box\psi))$.*

Proof. The argument is as above, except that after deriving

$$\blacksquare\blacksquare(\boxdot\alpha \rightarrow \beta)$$

derive

$$\blacksquare\,\boxdot\,(\boxdot\alpha \rightarrow \beta)$$

by applying the bridge axiom within the scope of the outermost modal. Then proceed as before. □

Proposition 5.A.4. *There is a model satisfying an unmixed congruence-like principle for a factive normal modal $\boxdot_1$, but satisfying only a mixed congruence-like principle for a modal $\boxdot_2$ extending $\boxdot_1$*

Proof. Let the domain of the frame be $W = \{w_1, w_2, w_3\}$. Let the sandwich function S be undefined on w_1, $S_{\mathsf{max}}(w_2) = S_{\mathsf{max}}(w_3) = W$, $S_{\mathsf{min}}(w_2) = \{w_2\}$, and $S_{\mathsf{min}}(w_3) = \{w_2, w_3\}$. Let the accessibility relation R_1 for $\boxdot_1$ be such that $R_1(w_1) = \{w_1\}$, $R_1(w_2) = \{w_2\}$, and $R_1(w_3) = \{w_2, w_3\}$. Let the corresponding R_2 be $R_1 \cup \{(w_1, w_3)\}$.

Clearly, R_2 extends R_1 as a relation, which is reflexive, so that $\boxdot_2$ extends $\boxdot_1$ as a modality. The unmixed congruence principle for $\boxdot_1$ is trivially true at w_1, and is easily checked to be true at the remaining w. As $\boxdot_2$ extends $\boxdot_1$, for which a mixed congruence-like principle is true, it also satisfies a mixed such principle. But, at w_1, the propositions $[\![p]\!] = \{w_3\}$ and $[\![q]\!] = \{w_2, w_3\}$ are $\boxdot_2$-necessarily equivalent, even as $\Box q$ but not $\Box p$ is true at the R_2-accessible w_3. Thus, at w_1, $\boxdot_2(p \leftrightarrow q)$ but not $\boxdot_2(\Box p \leftrightarrow \Box q)$ is true, contradicting the relevant unmixed congruence-like principle. □

Proposition 5.A.5. *The modality $\boxdot$ corresponding to the most restricted quasi-congruent possible accessibility relation R is not in general definable in the language of* Conv

Proof. Consider the model with domain $W = \{w_1, w_2, w_3\}$, with $S_{\mathsf{min}} = S_{\mathsf{max}}$ undefined everywhere but w_3, where it is the singleton function. Let $[\![p]\!] = \{w_2\}$. We will show that there is no open sentence φ with a single free variable (in potentially many instances) with $\varphi(p)$ true exactly at the w such that wRw_2, so that $\neg \boxdot \neg$ (and so $\boxdot$) is not definable in the language.

Under the usual Boolean operations, $\{w_2\}$ generates the Boolean lattice $B = \langle \subseteq, \{\emptyset, \{w_2\}, \{w_1, w_3\}, W\}\rangle$. Moreover, for all $b \in B$, b is "made true" at no w, so that B is closed under application of the corresponding "makes true" operator. Thus, for any singly free φ definable in the language, $[\![\varphi(p)]\!] \neq \{w_2, w_3\} = [\![\neg \boxdot \neg p]\!]$. □

Theorem 5.A.6. *For any sandwich frame, the quasi-congruent modalities form a complete distributive lattice Q. $(w, v) \in \bigwedge Q$ iff either $v = w$ or $S(w)$ is defined and either $v \in S_{\mathsf{min}}(w)$ or $v \notin S_{\mathsf{max}}(w)$*

Proof. Mostly clear from the preceding. That Q is distributive follows trivially from the distributivity of $\langle \mathcal{P}(W^2), \subseteq\rangle$. □

Definition 5.A.7. *A* generalised sandwich frame *consists of an inhabited domain W and a (possibly partial) function S from W to convex sublattices of $\mathcal{P}(W)$, such that, for all $X \in S(w)$, $w \in X$*

A generalised sandwich model on a given such frame is defined in the natural way.

Definition 5.A.8. *Given a generalised sandwich frame $\mathcal{F}$ (or generalised sandwich model $\mathcal{M}$ based upon $\mathcal{F}$), a* quasi-congruent modality *M is a function from the domain $\mathcal{F}_W$ to filters on $\mathcal{P}(\mathcal{F}_W)$, such that, for all w, $w \in F$ for all $F \in M(w)$ and, for all $p, q \subseteq W$, if $((W \setminus p) \cap (W \setminus q)) \cup (p \cap q) \in M(w)$, $p \in \mathcal{F}_S(w)$ iff $q \in \mathcal{F}_S(w)$*

Proposition 5.A.9. *The quasi-congruent modalities on a given frame form a complete distributive convex sublattice of the lattice of factive modalities on the frame under reverse inclusion at every argument, whose supremum is the unrestricted modality*

Proof. First observe that the quasi-congruent modalities are closed under extension. This is clear for essentially the same reason as in the non-generalised case: if

$$((W \setminus p) \cap (W \setminus q)) \cup (p \cap q) \in M(w)$$

(as an abbreviation: $p \equiv q \in M(w)$) implies for all w, p, q that $p \in S(w)$ iff $q \in S(w)$, then if M' extends M, for any w, p, q, if $p \equiv q \in M'(w)$ then $p \equiv q \in M(w)$ already, and so $p \in S(w)$ iff $q \in S(w)$.

That the unrestricted modality is quasi-congruent is clear, and thus it is clearly also the supremum if the class constitutes a lattice.

The join $\bigvee M^*$ of some (arbitrarily large) class M^* of quasi-congruent modalities is at each w the intersection $\bigcap\{M(w) : M \in M^*\}$, which by the above is clearly also a quasi-congruent modality.

The meet $\bigwedge M^*$ of M^* is at each w the filter generated by the $M(w)$ for $M \in M^*$. That is, $m \in \bigwedge M^*(w)$ iff, for some $M_1, \ldots, M_n \in M^*$ and $m_1 \in M_1(w), \ldots, m_n \in M_n(w)$, we have

$$m \supseteq \bigcap\{m_i : 1 \leq i \leq n\}$$

Now suppose $p \equiv q \in \bigwedge M^*(w)$; that is, suppose

$$p \equiv q \supseteq \bigcap\{m_i : 1 \leq i \leq n\}$$

for some $m_1 \in M_1(w), \ldots, m_n \in M_n(w)$ with $M_1, \ldots, M_n \in M^*$. This is to say that

$$m_1 \cap \cdots \cap m_n \cap p = m_1 \cap \cdots \cap m_n \cap q$$

By quasi-congruence of M_1, we have

$$m_2 \cap \cdots \cap m_n \cap p \in S(w) \text{ iff } m_2 \cap \cdots \cap m_n \cap q \in S(w)$$

(or, if $n = 1$, $p \in S(w)$ iff $q \in S(w)$, in which case we are done). By quasi-congruence of M_2, we have

$$m_2 \cap \cdots \cap p \in S(w) \text{ iff } m_3 \cap \cdots \cap p \in S(w)$$

(or iff $p \in S(w)$, in case $n = 2$). Repeating this procedure $n - 1$ times and chaining the material equivalences, we get

$$m_2 \cap \cdots \cap m_n \cap p \in S(w) \text{ iff } p \in S(w)$$

Clearly, the same process can be applied to derive

$$m_2 \cap \cdots \cap m_n \cap q \in S(w) \text{ iff } q \in S(w)$$

Thus, we have

$$q \in S(w) \text{ iff } m_2 \cap \cdots \cap m_n \cap q \in S(w) \text{ iff } m_2 \cap \cdots \cap m_n \cap p \in S(w) \text{ iff } p \in S(w)$$

and thus

$$p \in S(w) \text{ iff } q \in S(w)$$

Thus, $\bigwedge M^*$ is quasi-congruent.

Distributivity follows from distributivity of the lattice of modalities generally. □

Note that the logic of $\Box$ (or, equivalently, the structural constraints on S) do not enter into the proof, aside from the bare congruence of the operator (the function returning values in $\mathcal{P}(\mathcal{P}(W))$). Thus the result is not a consequence of the convexity logic of the control operator.

Bibliography

[Bacon, 2022] Bacon, A. (2022). *Vagueness and thought*. Oxford University Press.

[Bacon, 2023] Bacon, A. (2023). Counterfactuals, infinity and paradox. In Faroldi, F. L. G. and Putte, F. V. D., editors, *Kit Fine on Truthmakers, Relevance, and Non-classical Logic*, pages 349–388. Springer Verlag.

[Bacon, MS] Bacon, A. (MS). Stalnaker on the kk principle. https://andrew-bacon.github.io/papers/KK%20principle.pdf.

[Bacon and Zeng, 2022] Bacon, A. and Zeng, J. (2022). A theory of necessities. *Journal of Philosophical Logic*, 51:151–199.

[Belnap and Perloff, 1988] Belnap, N. and Perloff, M. (1988). Seeing to it that: A canonical form for agentives. *Theoria*, 54(3):175–199.

[Berker, 2008] Berker, S. (2008). Luminosity regained. *Philosophers' Imprint*, 8:1–22.

[Bolker, 1966] Bolker, E. D. (1966). Functions resembling quotients of measures. *Transactions of the American Mathematical Society*, 124(2):292–312.

[Bolker, 1967] Bolker, E. D. (1967). A simultaneous axiomatization of utility and subjective probability. *Philosophy of Science*, 34(4):333–340.

[Boudou and Lorini, 2018] Boudou, J. and Lorini, E. (2018). Concurrent game structures for temporal stit logic. In *Proceedings of the Seventeenth International Conference on Autonomous Agents and MultiAgent Systems*. ACM Press.

[Byrne and Logue, 2009] Byrne, A. and Logue, H. (2009). Introduction. In Byrne, A. and Logue, H., editors, *Disjunctivism: Contemporary Readings*. MIT Press.

[Caie, 2018] Caie, M. (2018). Benardete?s paradox and the logic of counterfactuals. *Analysis*, 78(1):22–34.

[Carlson et al., 2026] Carlson, E., Johansson, J., and Nyman, A. (2026). Higher-order control: An argument for moral luck. *Australasian Journal of Philosophy*.

[Chellas, 1969] Chellas, B. F. (1969). *The Logical Form of Imperatives*. PhD thesis, Stanford University.

[Chellas, 1992] Chellas, B. F. (1992). Time and modality in the logic of agency. *Studia Logica*, 51(3-4):485–517.

[Coffman, 2007] Coffman, E. J. (2007). Thinking about luck. *Synthese*, 158(3):385–398.

[Coffman, 2009] Coffman, E. J. (2009). Does luck exclude control? *Australasian Journal of Philosophy*, 87(3):499–504.

[Coffman, 2015] Coffman, E. J. (2015). *Luck: Its Nature and Significance for Human Knowledge and Agency*. New York, USA: Palgrave Macmillan.

[Cresswell and Hughes, 1996] Cresswell, M. J. and Hughes, G. E. (1996). *A New Introduction to Modal Logic*. Routledge.

[Crisp, 2017] Crisp, R. (2017). Moral luck and equality of moral opportunity. *Aristotelian Society Supplementary Volume*, 91(1):1–20.

[Davidson, 1967] Davidson, D. (1967). The logical form of action sentences. In Rescher, N., editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press.

[Davidson, 1971] Davidson, D. (1971). I. agency. In Marras, A., Bronaugh, R. N., and Binkley, R. W., editors, *Agent, Action, and Reason*, pages 1–37. University of Toronto Press.

[Demirtas, 2026] Demirtas, H. (2026). Against the degree-scope response to moral luck, or a farewell to responsibility for consequences. *Journal of Philosophy*.

[Driver, 2012] Driver, J. (2012). Luck and fortune in moral evaluation. In Blaauw, M., editor, *Contrastivism in Philosophy: New Perspectives*. Routledge.

[Enoch and Marmor, 2007] Enoch, D. and Marmor, A. (2007). The case against moral luck. *Law and Philosophy*, 26(4):405–436.

[Esakia, 1974] Esakia, L. (1974). Topological kripke spaces. *Soviet Mathematics - Doklady*, 15:147–151.

[Fara, 2003] Fara, D. G. (2003). Gap principles, penumbral consequence, and infinitely higher-order vagueness. In Beall, J. C., editor, *New Essays on the Semantics of Paradox*. Oxford University Press.

[Feinberg, 1970] Feinberg, J. (1970). *Doing & Deserving; Essays in the Theory of Responsibility*. Princeton: Princeton University Press.

[Ferrante, 2009] Ferrante, M. (2009). Recasting the problem of resultant luck. *Legal Theory*, 15(4):267–300.

[Fine, 1975] Fine, K. (1975). Vagueness, truth and logic. *Synthese*, 30(3-4):265–300.

[Fischer and Ravizza, 1998] Fischer, J. M. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

[Fischer and Todd, 2015] Fischer, J. M. and Todd, P., editors (2015). *Freedom, Fatalism, and Foreknowledge*. Oxford University Press, Oxford New York.

[Ford, 2018] Ford, A. (2018). The province of human agency. *Noûs*, 52(3):697–720.

[Frankfurt, 1969] Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66(23):829.

[Gerson, 1975] Gerson, M. (1975). The inadequacy of the neighbourhood semantics for modal logic. *Journal of Symbolic Logic*, 40(2):141–148.

[Gibbard and Harper, 1978] Gibbard, A. and Harper, W. (1978). *Foundations and Applications of Decision Theory*, chapter Counterfactuals and Two Kinds of Expected Utility. Number 13a in University of Western Ontario Series in Philosophy of Science. Dordrecht: D. Reidel.

[Ginet, 1990] Ginet, C. (1990). *On Action*. Cambridge, England: Cambridge University Press.

[Goodman, 2026] Goodman, J. (2026). Penumbral knowledge. Forthcoming in *The Philosophical Quarterly*.

[Goodman, MS] Goodman, J. (MS). Consequences of conditional excluded middle. *Manuscript*.

[Goodman and Salow, 2023] Goodman, J. and Salow, B. (2023). Epistemology normalized. *Philosophical Review*, 132(1):89–145.

[Greco, 1995] Greco, J. (1995). A second paradox concerning responsibility and luck. *Metaphilosophy*, 26(1-2):81–96.

[Haddock and Macpherson, 2008] Haddock, A. and Macpherson, F. (2008). *Disjunctivism: Perception, Action, Knowledge*. Oxford University Press.

[Hanna, 2014] Hanna, N. (2014). Moral luck defended. *Noûs*, 48(4):683–698.

[Hartman, 2017] Hartman, R. J. (2017). *In Defense of Moral Luck: Why Luck Often Affects Praiseworthiness and Blameworthiness*. Routledge.

[Hartman, 2019] Hartman, R. J. (2019). Moral luck and the unfairness of morality. *Philosophical Studies*, 176(12):3179–3197.

[Hedden, 2012] Hedden, B. (2012). Options and the subjective ought. *Philosophical Studies*, 158(2):343–360.

[Holliday, 2024] Holliday, W. H. (2024). Possibility semantics.

[Hornsby, 1980] Hornsby, J. (1980). *Actions*. Routledge and Kegan Paul.

[Horty and Pacuit, 2017] Horty, J. and Pacuit, E. (2017). Action types in stit semantics. *Review of Symbolic Logic*, 10(4):617–637.

[Horty and Belnap, 1995] Horty, J. F. and Belnap, N. (1995). The deliberative stit: A study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24(6):583–644.

[Humberstone, 1981] Humberstone, I. L. (1981). From worlds to possibilities. *Journal of Philosophical Logic*, 10(3):313–339.

[Inwagen, 1983] Inwagen, P. V. (1983). *An Essay on Free Will*. Oxford University Press.

[Jeffrey, 1965] Jeffrey, R. (1965). *The Logic of Decision*. New York: McGraw-Hill.

[Joyce, 1999] Joyce, J. (1999). *The Foundations of Causal Decision Theory*. New York: Cambridge University Press.

[Kant, 1784] Kant, I. (1784). *Grundlegung zur Metaphysik der Sitten*.

[Kenny, 1976] Kenny, A. (1976). Human abilities and dynamic modalities. In *Essays on Explanation and Understanding: Studies in the Foundations of Humanities and Social Science*. D. Reidel Publishing Company.

[Khoury, 2018] Khoury, A. (2018). The objects of moral responsibility. *Philosophical Studies*, 175(6):1357–1381.

[Kocurek, 2026] Kocurek, A. W. (2026). Is stalnaker's semantics complete? *Erkenntnis*, pages 1–9.

[Koon, 2020] Koon, J. (2020). Options must be external. *Philosophical Studies*, 177(5):1175–1189.

[Korsgaard, 2009] Korsgaard, C. M. (2009). *Self-Constitution: Agency, Identity, and Integrity*. Oxford University Press, New York.

[Lackey, 2008] Lackey, J. (2008). What luck is not. *Australasian Journal of Philosophy*, 86(2):255–267.

[Lederman and Holguín, 2023] Lederman, H. and Holguín, B. (2023). Trying without fail. Unpublished manuscript.

[Levy, 2011] Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford University Press UK.

[Levy, 2021] Levy, Y. (2021). Disjunctivism about intending. *American Philosophical Quarterly*, 58(2):161–180.

[Lewis, 1969] Lewis, D. (1969). *Convention: A Philosophical Study*. Wiley-Blackwell, Cambridge, MA, USA.

[Lewis, 1973] Lewis, D. (1973). *Counterfactuals*. Blackwell, Malden, Mass.

[Lewis, 1981] Lewis, D. (1981). Are we free to break the laws? *Theoria*, 47(3):113–21.

[Lewis, 1986] Lewis, D. K. (1986). *On the Plurality of Worlds*. Blackwell.

[Lorini et al., 2011] Lorini, E., Longin, D., and Mayor, E. (2011). A logical analysis of responsibility attribution: emotions, individuals and collectives. *Journal of Logic and Computation*, 24(6):1313–1339.

[MacKenzie, 2017] MacKenzie, J. (2017). Agent-regret and the social practice of moral luck. *Res Philosophica*, 94(1):95–117.

[McDowell, 1983] McDowell, J. (1983). Criteria, defeasibility, and knowledge. In *Proceedings of the British Academy, Volume 68: 1982*, pages 455–79. Oxford University Press.

[McDowell, 1994] McDowell, J. (1994). *Mind and World*. Cambridge: Harvard University Press.

[McDowell, 2011] McDowell, J. (2011). *Perception as a Capacity for Knowledge*. Marquette University Press.

[Nagel, 1979] Nagel, T. (1979). *Mortal Questions*. Cambridge University Press.

[Nelkin, 2021] Nelkin, D. K. (2021). Moral Luck. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition.

[Nuel Belnap and Xu, 2001] Nuel Belnap, M. P. and Xu, M. (2001). *Facing the Future: Agents and Choices in Our Indeterminist World*. Oxford University Press on Demand.

[Pacuit, 2017] Pacuit, E. (2017). *Neighborhood Semantics for Modal Logic*. Short Textbooks in Logic. Springer.

[Peels, 2015] Peels, R. (2015). A modal solution to the problem of moral luck. *American Philosophical Quarterly*, 52(1):73–88.

[Philippe Balbiani and Troquard, 2008] Philippe Balbiani, A. H. and Troquard, N. (2008). Alternative axiomatics and complexity for deliberative stit theories. *Journal of Philosophical Logic*, 37(4):387–406.

[Pollock, 2002] Pollock, J. L. (2002). Rational choice and action omnipotence. *Philosophical Review*, 111(1):1–23.

[Prior, 1967] Prior, A. (1967). *Past, Present and Future*. Clarendon P., Oxford,.

[Pritchard, 2005] Pritchard, D. (2005). *Epistemic Luck*. Oxford University Press UK.

[Pritchard, 2006] Pritchard, D. (2006). Moral and epistemic luck. *Metaphilosophy*, 37(1):1–25.

[Pritchard, 2012] Pritchard, D. (2012). *Epistemological Disjunctivism*. Oxford University Press.

[Richards, 1986] Richards, N. (1986). Luck and desert. *Mind*, 95(378):198–209.

[Rosen, 2004] Rosen, G. (2004). Skepticism about moral responsibility. *Philosophical Perspectives*, 18(1):295–313.

[Savage, 1954] Savage, L. (1954). *The Foundations of Statistics*. New York: John Wiley and Sons.

[Schwarz, 2023] Schwarz, W. (2023). Objects of choice. *Mind*.

[Segerberg, 1992] Segerberg, K. (1992). Getting started: Beginnings in the logic of action. *Studia Logica*, 51:347–378.

[Stalnaker, 1984] Stalnaker, R. (1984). *Inquiry*. Cambridge University Press.

[Stalnaker, 2006] Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies*, 128(1):169–199.

[Stalnaker, 1980] Stalnaker, R. C. (1980). A defense of conditional excluded middle. *Ifs*, page 87?104.

[Stalnaker, 2009] Stalnaker, R. C. (2009). On hawthorne and magidor on assertion, context, and epistemic accessibility. *Mind*, 118(470):399–409.

[Statman, 1993] Statman, D., editor (1993). *Moral Luck*. Albany: State University of New York Press.

[Statman, 2019] Statman, D. (2019). The definition of "luck" and the problem of moral luck. In Church, I. M. and Hartman, R. J., editors, *The Routledge Handbook of the Philosophy and Psychology of Luck*, pages 195–205.

[Swenson, 2022] Swenson, P. (2022). Equal moral opportunity: A solution to the problem of moral luck. *Australasian Journal of Philosophy*, 100(2):386–404.

[Taylor, 1962] Taylor, R. (1962). Fatalism. *Philosophical Review*, 71(1):56–66.

[Vihvelin, 2013] Vihvelin, K. (2013). *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. Oup Usa.

[von Neumann and Morgenstern, 1947] von Neumann, J. and Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton University Press.

[von Wright, 1963] von Wright, G. H. (1963). *Norm and Action*. New York: Humanities.

[Warmbrod, 1982] Warmbrod, K. (1982). A defense of the limit assumption. *Philosophical Studies*, 42(1):53–66.

[Williams, 1981] Williams, B. (1981). *Moral Luck: Philosophical Papers 1973?1980*. Cambridge University Press.

[Williams and Nagel, 1976] Williams, B. A. O. and Nagel, T. (1976). Moral luck. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 50:115–151.

[Williams, 2010] Williams, J. R. G. (2010). Defending conditional excluded middle. *Noûs*, 44(4):650–668.

[Williamson, 2000] Williamson, T. (2000). *Knowledge and Its Limits*. Cambridge University Press.

[Williamson, 2013] Williamson, T. (2013). *Modal Logic as Metaphysics*. Oxford University Press, Oxford, England.

[Williamson, 2017] Williamson, T. (2017). Acting on knowledge. In Carter, J. A., Gordon, E. C., and Jarvis, B. W., editors, *Knowledge First: Approaches in Epistemology and Mind*, pages 163–181. Oxford: Oxford University Press.

[Williamson, 2024] Williamson, T. (2024). *Overfitting and Heuristics in Philosophy*. Oxford University Press, New York, NY.

[Wolf, 2001] Wolf, S. (2001). The moral of moral luck. *Philosophic Exchange*, 31(1).

[Xu, 1995] Xu, M. (1995). On the basic logic of stit with a single agent. *Journal of Symbolic Logic*, 60(2):459–483.

[Xu, 1996] Xu, M. (1996). An investigation in the logics of seeing-to-it-that.

[Xu, 1998] Xu, M. (1998). Axioms for deliberative stit. *Journal of Philosophical Logic*, 27(5):505–552.

[Zimmerman, 2022] Zimmerman, M. (2022). *Ignorance and Moral Responsibility*. Oxford University Press.

[Zimmerman, 1987] Zimmerman, M. J. (1987). Luck and moral responsibility. *Ethics*, 97(2):374–386.

[Zimmerman, 1997] Zimmerman, M. J. (1997). Moral responsibility and ignorance. *Ethics*, 107(3):410–426.

[Zimmerman, 2002] Zimmerman, M. J. (2002). Taking luck seriously. *Journal of Philosophy*, 99(11):553–576.

[Zimmerman, 2006] Zimmerman, M. J. (2006). Moral luck: A partial map. *Canadian Journal of Philosophy*, 36(4):585–608.