Formal explorations in collective and individual rationality

by

Alexandru Marcoci

A thesis submitted to the Department of Philosophy, Logic and Scientific Method of the *London School of Economics and Political Science* for the degree of Doctor of Philosophy, 2 October 2017.

DECLARATION

I certify that the thesis I have presented for examination for the degree of Doctor of Philosophy of the *London School of Economics and Political Science* is solely my own work other than where I have clearly indicated that it has been used in joint work.

I confirm that Chapter 3 represents joint work (out of which my contribution is 75%) with James Nguyen (Marcoci and Nguyen, 2017). Chapters 4 and 5 represent joint work (out of which my contribution is 50%) with JN (Marcoci and Nguyen, Forthcoming), who was at the time all our joint work was done a PhD student in philosophy at the *London School of Economics and Political Science*. He is currently Postdoctoral Researcher in Philosophy of Science at the John J. Reilly Center at the *University of Notre Dame*. Finally, Chapter 11 represents my contribution (100%) to a joint project with Luc Bovens. LB is Professor of Philosophy at the *London School of Economics and Political Science* and my doctoral supervisor.

I confirm material in Section 7.2 has appeared in Marcoci (2015).

I declare that this thesis consists of 43,092 words.

ABSTRACT

This thesis addresses several questions regarding what rational agents ought to believe and how they ought to act.

In the first part I begin by discussing how scientists contemplating several mutually exclusive theories, models or hypotheses can reach a rational decision regarding which one to endorse. In response to a recent argument that they cannot, I employ the tools of social choice theory to offer a 'possibility result' for rational theory choice. Then I utilize the tools of judgment aggregation to investigate how scientists from across fields can pool their expertise together. I identify an impossibility result threatening such a procedure and prove a possibility result which requires that some scientists sometimes waive their expertise over some propositions.

In the second part I first discuss the existing justifications for a restricted principle of indifference that mandates that two agents whose experiences are subjectively indistinguishable should be indifferent with respect to their identities. I argue that all existing justifications rely on the same mistaken reasoning behind the 'staying' strategy in the *Monty Hall* problem. Secondly, I show this mistake is more widespread and I identify it in arguments purporting to show the failure of two reflection-like principles.

In the third part I look at a recent argument that fair policy makers face a dilemma when trying to correct a biased distributive process. I show the dilemma only holds if the correction has to happen in one-shot. Finally, I look at how we ought to design public restrooms

so that we reduce the discrimination faced by minority groups. I make the case for opening our public restrooms to all genders.

FOREWORD

This dissertation would have never happened without the relentless support of my supervisor, Luc Bovens. As I am writing this, on a Sunday evening, Luc is going through one of my chapters one final time. He has read and re-read every other chapter multiple times and throughout the years he has always accommodated my last-minute approach to deadlines, including the one for submitting this dissertation. Luc's support has extended far beyond research supervision and he has been a constant source of encouragement and optimism and I do not think I could have had a better supervisor. Thank you Luc for all your hard work!

The second source of support in writing this dissertation has come from my secondary supervisor, Richard Bradley. He has read all chapters and has offered invaluable feedback on each of them. Thank you, Richard.

I would like to thank James Nguyen with whom I have discussed almost all arguments in this dissertation, co-wrote numerous papers, travelled around the world to present our work and shared anxieties about our academic and personal lives. I am looking forward to many more conversations together and thank you for not pressuring me to finish my contribution to that joint manuscript that has been sitting on top of my to-do list.

This dissertation would have been significantly worse without the help of these three people. That being said, all mistakes (of which I am sure there are many) remain my own - although I leave the

resulting paradox for another time.

My time at LSE has been made both personally and academically more pleasant by the frequent interactions with a group of people too large to fit into this Foreword; among them are Miklos Redei, David Makinson, Nick Baigent, Christian List, Silvia Milano, and Goreti Faria. Finally, I owe special gratitude to the wonderful and tireless Bryan Roberts!

My doctoral studies have been very generously supported by a British Society for the Philosophy of Science Doctoral Scholarship. Without their financial support, I wouldn't have embarked on this project. Moreover, I am grateful to LSE's Government Department who adopted me as a Fellow in the final stages of my doctoral studies.

Outside of my LSE life I had the great fortune to have around me the most understanding and loving group of people. They are (in no particular order): Sebastian (and Adam), Dragos, Ana, and Cristiana. I hope I can one day be as good a friend to you as you have been to me.

Throughout my life, my parents, Cristina and Doru and my grandmother, Ana, have been a source of support, love and, at times, financial stability. I can never thank them enough for what they have done for me and thinking back now to all that has happened in the last years, I can only say that I am incredibly lucky to have you in my corner.

Finally, the person who has contributed more to the writing of this dissertation than I have is my partner. I have shared the better part of the last 12 years with Diana and she has taught me more than I can recount within the wordcount of a PhD dissertation. I would

have been a different person without her in my life. This dissertation is dedicated to you!

London, 1 October 2017

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

INTRODUCTION

1

INTRODUCTION

This thesis addresses several questions regarding what rational agents ought to believe and how they ought to act. I divide my treatment of these questions into three parts: one on collective rationality, one on individual rationality and the final one on rationality in practice. The first section on collective rationality deals with puzzles of theory choice. The section on individual rationality deals with self-locating beliefs and protocols for belief updating. And, finally, the section on rationality in practice deals with a puzzle of fair distribution and presents an efficiency argument for the design of public restrooms.

COLLECTIVE RATIONALITY In a recent paper, Samir Okasha (2011) uses the formal framework of social choice theory, and Arrow's impossibility theorem in particular, to argue that there is no algorithm for using the information supplied by scientific virtues (e.g. simplicity, accuracy, scope) to rationally choose the best theory. If Okasha is right, then there is no function (satisfying certain desirable conditions) from 'preference' rankings supplied by scientific virtues over competing theories (or models, or hypotheses) to a single all-things-considered ranking. This threatens the rationality of science - in the sense in which irrespective of the method scientists employ to arrive at an all-things considered ranking from their rankings according to the individual virtues, sometimes they will end up with an inconsistent overall preference.

I contend Okasha's view of scientific rationality is too austere and I offer a two-fold response. To begin, I show that the threat to the possibility of rational theory-choice relies on an all-or-nothing understanding of scientific rationality and I articulate instead a notion of rationality by degrees. Such a move from all-or-nothing rationality to rationality by degrees will establish that theory choice can be rational *enough*. Further, I show that if Kuhn's claims about the role that subjective elements play in theory choice are taken seriously, and Okasha's framework is supplemented with an appropriate formal notion of ambiguity, then the threat dissolves. The outcome of this discussion is that there is a meaningful sense of scientific rationality beyond the mere impossibility of inconsistency.

This discussion provides several new insights. Firstly, it articulates a new (formal) concept of 'degrees-of-rationality'. Second, it provides a formal account of the Kuhnian notion of ambiguity. Thirdly, it represents one of the few discussions in the social choice context of what it means for a ranking to be ambiguous and provides, in turn, a proof of concept (via simulation) for a new way out of the Arrovian impossibility.

Taking Okasha's cue that social choice can inform philosophy of science, I investigate the commonly shared belief among scientists that: "[i]t is wrong to bemoan the high degree of specialisation in current research ... we should not forget that in a mature discipline only specialization gets things done." (Slaney, 1987, 1195) Now, academic disciplines are increasingly fragmented, and this naturally leads to diverse areas of expertise within them. But if the state of a discipline as a whole is supposed to depend on the beliefs of its experts in a certain way, then we arrive at an impossibility result similar to Sen's Liberal Paradox (1970). Just as Sen's Lewd and Prude were forced into inconsistency by their decisiveness over their own per-

sonal spheres, a discipline may be forced into inconsistency if experts are taken to be decisive over their respective areas of expertise. At least this will be the case if we require that the scientific consensus in the discipline – i.e what the discipline as a whole tells us about the world – be generated by a suitably-constrained function on the beliefs of scientists. I show this by importing Dietrich and List (2008)'s result from the context of aggregating individual judgements to that of aggregating scientific expert judgements. If this is the correct way of thinking about what fragmented scientific disciplines tell us about the world, it seems that science cannot protect itself from inconsistencies. Insofar as we think that Nature is free of inconsistency, and we want our scientific consensus to reflect this, something has to give. Building on the ideas of Gibbard (1974) I suggest that the best way of avoiding inconsistencies is for experts to waive their expertise and contribute beliefs on a par with everyone else's. As such I argue against the hegemony of experts.

This discussion provides several new insights: I offer a new formal framework for analyzing disagreement in science. Secondly, I translate and then prove Gibbard's result as applied to judgment aggregation. Finally, this discussion offers a first step towards a broader question which I do not develop in this thesis but follows naturally from the work here: can groups of scientists spanning multiple countries, institutions, and fields of expertise be a group agent? My answer is that they can indeed operate as a group agent as long as they are willing to waive their expertise over some propositions at least sometimes.

INDIVIDUAL RATIONALITY The second part of the thesis will analyse the justifications for several principles which are meant to offer a more substantial conception of rationality beyond the mere satisfaction of the Kolmogorovian axioms, viz. Elga's (2004) princi-

ple of indifference for self-locating belief (and Titelbaum's (2013) defense thereof), Christensen's (2010) principle of reflection and Mahtani's (forthcoming) opaque proposition principle. I find that all of the above arguments fail. I also show that the mistakes in these arguments are similar and result from how the above authors construe the way in which a rational agent updates her beliefs in the face of new information.

Assuming we wanted to build a Bayesian model for an agent who learns a new proposition, how should we construct the model? An old and often overlooked paper by Glenn Shafer (1985) argues that in order for Bayesian conditionalisation to be a good formal counterpart to learning, one ought to specify a partition of events which an agent can conditionalise upon. Luc Bovens and Jose Luis Ferreira (2010) interpret Shafer's insight as "[w]hen we are informed of some proposition, we do not only learn the proposition in question, but also that we have learned the proposition as one of the many propositions that we might have learned. The information is generated by a protocol, which determines the various propositions that we might learn." This point can be found in the formal literature on bayesian epistemology such as in the works of Pearl and Halpern, but is often ignored by philosophers. When we model the scenarios of Elga, Titelbaum, Christensen and Mahtani in light of Shafer's insight, their arguments become significantly weaker.

This works makes several contributions to the existing research in bayesian epistemology. Firstly, it shows how Shafer's insight affects many more problems than just *Monty Hall* and its analogues. Secondly, it explains why Elga's principle of indifference for self-locating belief fails and gauges the impact this has on several arguments in the literature on self-location.

RATIONALITY IN PRACTICE Finally, I conclude with a foray in practical rationality. Firstly, I discuss a recent argument due to McKenzie Alexander (2013) who presents a dilemma for a social planner interested in correcting an unfair distribution of an indivisible good between two equally worthy individuals or groups: *either* she guarantees a fair outcome, *or* she follows a fair procedure (but not both). I show that this dilemma only holds if the social planner can redistribute the good in question at most once. To wit, the bias of the initial distribution always washes out when we allow for sufficiently many redistributions. I show by means of simulation that for real, resource-bounded, agents at most 6 redistributions is enough.

Secondly, I explore what a rational way of arranging public restrooms would look like. I do so by engaging with the current debates on 'bathroom bills' and assess the economic benefits of introducing unisex restrooms. The idea is simple: The move to unisex restrooms increases economic efficiency due to the reduction in waiting times. Since everyone cares about economic efficiency, there is at least one argument in favour of unisex restrooms that all can agree on. I construct a model to evaluate the waiting times in unisex and single-sex restrooms and show by means of simulations, that, given certain plausible assumptions, unisex restrooms provide drastic reductions in the total waiting times. This translates into greater productivity at lower overhead costs in terms of estates for firms. The move to unisex restrooms will indirectly benefit people who identify as trans*, carers, as well as parents and children of different genders. Furthermore, it will lead to increased potty parity.

Part II

COLLECTIVE RATIONALITY

2

OKASHA'S ARROVIAN IMPOSSIBILITY RESULT

Suppose a scientist is faced with a collection of competing scientific theories, models or hypotheses. Moreover, suppose that she cares about a number of distinct scientific virtues – accuracy, simplicity, and scope for example. How is the scientist to rationally choose the 'best' alternative, all-things-considered? One would like to think that, whatever the details of how the choice is made, some rational procedure is followed to make it.

According to Kuhn (1972), scientists faced with such a choice, even if they agreed about which theoretical virtues should guide their choice, could still rationally disagree about which is the 'all-things-considered' best competing theory. In this sense, there is no 'unique algorithm' that takes how well theories fare according to the scientific virtues and delivers a 'winner'. This is not to say that theory choice is 'a matter of mob psychology' (Lakatos, 1970, p. 178), but rather that the shared and objective virtues do not determine by themselves a winner, or unique ranking of the theories.

Okasha (2011) uses the formal framework of social choice theory to argue that, rather than there being no unique algorithm for using the objective information supplied by the scientific virtues to rationally choose the best theory, there is *no* such algorithm whatsoever: 'Where Kuhn saw an embarrassment of riches, Arrow tells us that

there is nothing at all' (Okasha, 2011, p. 93).

Okasha employs a simple but persuasive argumentative strategy. Let $\mathfrak{V}$ be a finite set of m scientific virtues, and $\mathfrak{T}$ a finite set n of competing theories. Each scientific virtue $i \in \mathfrak{V}$ provides an ordinal ranking of the elements of $\mathfrak{T}$, from most to least virtuous according to i. These rankings are transitive, reflexive, and complete binary relations.[1] That this simple framework does justice to all scientific theory choice contexts is debatable and Okasha offers a few examples where it isn't plausible. However, Kuhn appears to be somewhat sympathetic to the impossibility of obtaining cardinal information regarding the relative strength of theories according to the preferred list of scientific virtues:

> All historically significant theories have agreed with the facts, but only more or less. There is no more precise answer to the question whether or how well an individual theory fits the facts. But questions much like that can be asked when theories are taken collectively, or even in pairs. It makes sense to ask which of two actual and competing theories fits the facts better. (Kuhn, 1970, p. 147)

When theory x is preferred to theory y by virtue i I write $y \prec_i x$. A theory choice situation is a profile of rankings of theories by virtues, where a profile is an ordered tuple $\langle \prec_1, ..., \prec_m \rangle$ for virtues 1 through m. A theory choice function maps profiles to an all-things-considered binary relation $\preceq$ defined over $\mathfrak{T}$. A theory x is strictly preferred (all-things-considered) to theory y, i.e. $y \prec x$ if and only if $y \preceq x$ and it's not the case that $x \preceq y$.

1 In this paper I restrict the focus to *strict* rankings associated with scientific virtues for simplicity. Where relevant the approach articulated in the next pages can be extended to accommodate rankings allowing for ties. This is of no conceptual importance.

The discussion above presupposes that a scientist's endorsement of a theory, hypothesis and model is completely determined by the way in which he deems that theory to fare with respect to a given set of virtues. Nothing else matters. This seems to capture the flavour of Kuhn's portrayal of theory choice as described in Kuhn (1970, 1972). However, one might object, that this variety of virtues is implausible in science, and that, at least in many theory choice contexts, it is the accuracy or fit to data of a theory that should determine its value. On this criticism scientific theory choice isn't a matter of multi-criterion decision making, but the result of the application of a single criterion, i.e. accuracy, to rank all theories under consideration. Morreau in one of his replies to Okasha puts this challenge to rest quite persuasively:

> Theories can be brought into agreement with observations and experimental results by adjusting parameters. Scientific data are generally noisy, though. So if we single-mindedly pursue fit to available data, without regard to other criteria, we will end up preferring overly complicated theories, whose many parameters we can tune to fit the noise. These theories might agree with every error in the data; but they will fit underlying facts and future data less well than do other, simpler theories. They will overfit the data. To avoid this, acceptable choice rules must balance fit against simplicity. And they must do so not in spite of the special importance of accuracy in science but precisely because of it, because balance is what secures accuracy in the long run. (Morreau, 2014, p. 1259)

What requirements should one impose on a rational theory choice function? Okasha (2011, pp. 92-93) argues that the Arrovian conditions on preference aggregation have clear analogues in the context of theory choice. Unrestricted domain (UD) requires that a theory choice function be applicable irrespective of how theories are ranked

by virtues. Weak Pareto (WP) requires that, for all $x, y \in \mathfrak{T}$, if x is ranked above y according to every virtue, then x should be ranked above y all-things-considered. Independence of Irrelevant Alternatives (IIA) requires that the all-things-considered relation between two theories takes into account only how those two theories are ranked by the scientific virtues, i.e. the overall comparison between the two be insensitive to how virtues rank them with respect to a third theory. Non-Dictatorship (ND) demands that there is no virtue i such that for every pair of alternatives $x, y \in \mathfrak{T}$, whenever i prefers x to y, x is ranked above y all-things-considered. Finally, Overall Rationality (OR) demands that a theory choice function deliver a transitive, complete ranking for every element in its domain.[2] With theories replacing social alternatives, and scientific virtues replacing voters, it is immediate to see that Arrow's (1951) impossibility result applies. In other words, there is no theory choice function that satisfies UD, WP, IIA, ND and OR. Okasha's challenge is, that assuming rational theory choice requires the existence of such a function, rational theory choice is impossible.

> 'If we agree that U, P, N, and I are conditions on reasonable theory choice, then it is obvious that an Arrovian impossibility result applies. So long as there are at least three alternative theories, there exists no theory choice rule that satisfies all four conditions. This spells bad news for the possibility of making 'rational' theory choices' (Okasha, 2011, p. 93).[3]

2 Note that this condition is usually build into the definition of an aggregation function. In this paper I include it as a separate condition in light of the fact that I focus later on aggregation functions that map to intransitive/incomplete binary relations. I will use the term 'intransitive' to refer to the intransitivity of the entire binary relation ($\preceq$), note that this is compatible with $\prec$ being transitive.

3 Note that his 'U' is my 'UD'; his 'P' my 'WP'; his 'N' my 'ND'; his 'I' my 'IIA'; and where he assumes our OR in the definition of a theory choice rule I pull it out as a further condition.

The sense of the impossibility is important to gauge from the very onset. Arrow's result (and implicitly, Okasha's), doesn't preclude any aggregating function from delivering a transitive overall ranking in some theory choice situations. Rather the impossibility precludes any function from *always* being able to deliver such a 'nice' output. It is easy to see that any function satisfying the first four conditions would deliver a trivial and 'nice' output applied to a situation in which all virtues rank the given theories in the same way. What the result tells us is that there are (logically) possible situations for every aggregation function we were to adopt in which it will fail to output a transitive overall ranking. However, the result is silent with respect to how likely it is for such situations to arise in science.

Okasha's argument has generated much discussion. Morreau (2014, 2015) suggest restricting UD. Rizza (2013) and Stegenga (2015) follow up on Okasha's (2011) own suggestion of enriching the informational basis of scientific virtues by providing a common cardinal scale allowing for inter-virtue comparisons (thereby dropping IIA). Whether all scientific virtues provide such information is questionable as discussed in Okasha's (2015) response to Stegenga and indeed in Kuhn (1970, see above) . Relatedly, Gaertner and Wuethrich (2016) suggest *imposing* a cardinality via a scoring rule in a way that captures the spirit of IIA in a cardinal framework. Finally, Bradley (forthcoming) suggests that rationality only requires *ruling out* certain alternatives, not a transitive and complete ranking of theories.

3

A DEGREE NOTION OF RATIONALITY

In this paper we draw attention to Okasha's assumption that scientific rationality is an all-or-nothing notion and motivate the move to a degrees notion of rationality, instead. By considering rationality in *degrees*, rather than in the all-or-nothing sense, we can precisely gauge the threat Okasha poses, and we show that this is highly sensitive to the number of alternatives and virtues under consideration. Whether or not theory choice is 'rational' depends on where one sets a threshold. This a substantial decision and different scientists may reasonably disagree. We do not offer a solution to this problem, nor do we believe there is one. However, we argue the constraint on this threshold is sufficiently weak in order to make it reasonable to accept that choosing among competing theories based on their relative ranking according to a set of scientific virtues is rational *enough*.

But rather than questioning any of the conditions, our strategy in this paper is to investigate alternative ways of construing theory choice as a rational enterprise which don't require identifying rational theory choice with the existence of a function that satisfies the five Arrovian conditions. The way we do this is by focusing on the extent to which pairwise majority voting is rational. But we begin small.

3.1 MINIMAL RATIONALITY

In this section we introduce a minimal all-or-nothing notion of rationality, call it 'minimal rationality'. This is the building block which will be used to construct the notion of rationality by degrees that interests us in this paper. Begin with a trivial social choice example. Suppose Bill, Albert, and Chloe are trying to choose, as a group, between watching a football match, going to the cinema, or visiting a restaurant. Suppose they have the following preference rankings:

Albert:	*Football*	$\prec_A$	*Cinema*	$\prec_A$	*Restaurant*
Bill:	*Football*	$\prec_B$	*Cinema*	$\prec_B$	*Restaurant*
Chloe:	*Football*	$\prec_C$	*Restaurant*	$\prec_C$	*Cinema*

In this social choice context, the Arrovian conditions are supposed to supply constraints on what the group should do. Some of these conditions, such as IIA for example, put inter-profile constraints on the behaviour of an aggregation function. Others, such as OR, put constraints on the behaviour of a function that apply profile by profile. What we call 'minimal rationality' supplies a way of thinking about whether an aggregation function that satisfies the Arrovian conditions of WP, IIA, ND and UD, e.g. pairwise majority voting as we define it below, is normatively acceptable on a profile by profile basis. The conditions on this are those supplied by OR: the value of the function when applied to that profile should itself be a preference ranking, i.e. transitive and complete. Our point then, is that some aggregation functions can be normatively acceptable at some profiles, and yet normatively unacceptable at others.

At the profile displayed above pairwise majority vote supplies the following:

Group: *Football* $\prec$ *Cinema* $\prec$ *Restaurant*

and is thus normatively acceptable with respect to this profile. However in a scenario where Bill, Albert and Chloe held preference rankings such that pairwise majority voting delivers an intransitive all-things-considered value, then the function is not normatively acceptable with respect to that profile. This would be the case if they had held preference rankings that generated Condorcet's paradox. A weaker requirement that might be of particular relevance in the context of theory choice is that the function deliver a Condorcet winner (an alternative preferred to all other alternatives in the all things considered ranking), rather than a preference ranking. With three alternatives, these conditions are equivalent.

When applying the machinery of social choice theory to theory choice, Okasha's strategy is to take the question of whether or not an aggregation function is normatively acceptable to correspond to the question of whether or not a (theory choice) aggregation function is rational. And he takes the conditions on whether or not this is the case to be the Arrovian ones. Again, some of these conditions but inter-profile constraints on the behaviour of such functions, and others apply profile-by-profile. So again, we can ask of a given aggregation function which satisfies the conditions of WP, IIA, ND and UD whether or not it is 'minimally rational' on a profile-by-profile basis. And again, the conditions on this are supplied by OR. These observations provide the following definition:

MINIMAL RATIONALITY A theory choice function f is minimally rational with respect to a profile $\mathcal{P} \in \mathfrak{D}_m^n$, if and only if it

meets UD, WP, IIA, ND (with respect to $\mathfrak{D}_m^n$) and takes $\mathcal{P}$ to a transitive and complete ranking.[1]

A natural weakening of minimal rationality is to demand a Condorcet winner, rather than a transitive complete ranking. This gives rise to an analogous notion of minimal Condorcet rationality in the obvious way.

Minimal rationality can then be built up to the full blown notion of rationality that Okasha requires, in the sense of meeting all of the Arrovian conditions everywhere in a domain of profiles. If a function f is minimally rational for every profile in $\mathfrak{D}_m^n$ then choosing the most suitable theory out of n alternatives using m virtues by means of f is always rational. In other words, there is nothing more to being rational with respect to a domain than being rational with respect to every element in that domain. This is the requirement Okasha argues is not met by any theory choice function. Indeed, by Arrow's theorem there is no f satisfying this (for any $n \geq 3$ and $m \geq 2$). In section 5 we discuss a weakening of this requirement and show that with it in place, the prospects of rational theory choice improve.

3.2 FIXING THE THEORY CHOICE FUNCTION

One thing to note before introducing the notion of rationality by degrees is that in this paper we restrict our attention to a particular theory choice function. Arrow's result is general, it entails that there is *no* function that is rational for domains where $n \geq 3, m \geq 2$. As such, discussion of any particular function can be suppressed. This is not so when discussing minimal rationality and rationality by degrees. Therefore for the purposes of this paper we focus on pairwise majority voting, which is defined as follows:

1 $\mathfrak{D}_m^n$ represents the class of all profiles that can be defined over m virtues and n theories.

PAIRWISE MAJORITY VOTING For a set of virtues $\mathfrak{V}$, let $\Delta^{+} =_{df} \{i \in \mathfrak{V} : y \prec_i x\}$, $\Delta^{-} =_{df} \{i \in \mathfrak{V} : x \prec_i y\}$. Then: y $\prec$ x if and only if $|\Delta^{+}| \geq |\Delta^{-}|$.[2]

There are other functions that satisfy UD, WP, IIA, and ND, including Pareto dominance and extension procedures, both of which violate completeness for various profiles, see (List, 2013, §3.2.2). We nevertheless ignore these other functions for the remainder of this paper. The reason is that the focus here is not on which aggregating method is most suitable for theory choice. We are, instead, trying to show the importance of the way in which we construe scientific rationality for the purposes of evaluating how rational theory choice is given Okasha's challenge. For this purpose it is enough to look at a single aggregation method and it seems natural to use the most well-studied one for our proof of concept.

3.3 RATIONALITY BY DEGREES

In Section 3 it was noted that it is not always the case that a theory choice function will lead to an intransitive (or incomplete) all-things-considered ranking. For example, pairwise majority voting is minimally rational with respect to at least some profiles in $\mathfrak{D}^{3}_{3}$. Suppose that the scientific virtues of accuracy, simplicity and scope provide the following profile of preference rankings over $\mathfrak{T} = \{x, y, z\}$: $\langle x \prec_{si} z \prec_{si} y, x \prec_{ac} z \prec_{ac} y, x \prec_{sc} y \prec_{sc} z \rangle$, and that x,y and z and simplicity, accuracy and scope exhaust the alternatives and virtues under consideration. Then majority voting yields the all-things-considered ranking of $x \prec z \prec y$. But if this was the only profile in $\mathfrak{D}^{3}_{3}$ with respect to which majority voting were minimally rational, then scientists using this function would succeed in rationally

2 For the remainder of this paper we will assume there is an odd number of virtues. This means that the output of the aggregation, under pairwise majority, will always be a strict ordering (if an ordering, at all).

choosing in only 1 out of 216 of the possible cases.[3] Suppose, however, again in $\mathfrak{D}_3^3$, there was only one profile of preferences which majority voting mapped to an intransitive ranking. In such a scenario, scientists would succeed in making a rational choice using the function in 215 out of 216 of the possible cases.

In the above example, there is a sense in which pairwise majority would be *less* 'rational' (in an intuitive sense) if it generated a transitive ranking from only 1 out of 216 profiles, than it would be if it did so from 215. If a scientist used the function in the former case she would be acting 'irrationally.' But she wouldn't if she did so in the latter. In fact, not using pairwise majority in such a scenario would be 'irrationally' cautious.[4] Perhaps the scientist wants to know whether or not she should go through the rigmarole of generating rankings by virtues in order to choose between a given set of alternatives. If there were little hope that her preferred function map the resulting profile to a preference ranking, then this would be a waste of her time. But if there is a high chance that her function will deliver such a ranking and she wants to make a choice between the alternatives, then she should proceed.

3 Assuming all 216 possible ways are equally likely to obtain, the chances of succeeding in rationally choosing the best theory are very low. See a discussion of this assumption further below.

4 Some level of risk aversion is undoubtedly rational. Consider the following scenario: you are offered two bets. Bet 1 gives you the chance to win 100$ with probability 1. Bet 2 gives you a chance to win 200$ with probability .5 and 0$ otherwise. Choosing Bet 1 in this instance does not seem irrational, and in fact, many people will do so. However, as the probability of winning 200$ in Bet 2 increases, Bet 2 becomes more appealing, and fewer people will avoid it. There seems to be a point at which choosing Bet 1 over Bet 2 becomes irrationally cautious (if this still doesn't appeal to your intuition, consider Bet 3 with probability .9 of winning 1,000,000$ and 99$ otherwise). In this sense, a scientist refraining from using pairwise majority, as this rule fails in 1 out of 216 cases appears irrationally cautious.

So, whether or not a function is 'rational' seems sensitive to how likely it is to deliver a transitive and complete preference ranking. And how likely it is to deliver such a ranking, for a domain $\mathfrak{D}_m^n$, depends on the likelihood assigned to each of the profiles within that domain. This requires introducing a probability measure $\mathtt{Pr}$ over $\mathfrak{D}_m^n$.[5] In the aforementioned discussion we assumed that $\mathtt{Pr}$ was the equiprobable distribution, with $\mathtt{Pr}$ assigning $1/(n!)^m$ to each profile in $\mathfrak{D}_m^n$.[6] But suppose, for comparison, that the probability of the single profile which mapped to an intransitive ranking in the latter example above were approaching 1. Then in that case the scientist would be 'irrational' to attempt to rank theories according to virtues.

This suggests the possibility of a degree measure of rationality. The degree to which an aggregation function f is rational (for a given $\mathfrak{D}_m^n$, $\mathtt{Pr}$) is simply the sum of the values $\mathtt{Pr}$ assigns to all profiles with respect to which f is minimally-rational. We denote this sum by μ and we call the resulting notion of rationality, rationality by degrees.

RATIONALITY BY DEGREES A theory choice function f, which meets UD, WP, IIA and ND (with respect to a domain, $\mathfrak{D}_m^n$) is μ-rational (or rational to degree μ), with respect to $\mathtt{Pr}$ if and only if

$$\mathtt{Pr}(\{\mathcal{P} \in \mathfrak{D}_m^n | f \text{ is minimally rational for } \mathcal{P}\}) = \mu$$

The shift from thinking about rationality in an all-or-nothing sense, to thinking about it in degrees is done in two steps. Firstly, we introduce a probability function $\mathtt{Pr}$ over the elements of $\mathfrak{D}_m^n$. Secondly,

5 The only restriction we place on $\mathtt{Pr}$ is that it assigns a non-zero probability to every profile in $\mathfrak{D}_m^n$. The motivation for this restriction is the same as the motivation for UD. If a theory choice function is rational only if it is defined over every profile in a domain, then all profiles are considered as 'live options'. Assigning to any profile a zero probability of occurring would undermine this.

6 In the social choice literature this is known as the 'impartial culture' assumption Gehrlein (1983). This assumption is discussed in more detail below.

we measure the rationality of a theory choice function f by the probability mass assigned to the set of all profiles in $\mathfrak{D}_m^n$ with respect to which f is minimally rational. Some correspondences between the notions emerge. If f is 1-rational, then it is rational for $\mathfrak{D}_m^n$, or equivalently, minimally rational with respect to every $\mathcal{P} \in \mathfrak{D}_m^n$. If, on the other hand, f is 0-rational, then f is minimally irrational with respect to every $\mathcal{P} \in \mathfrak{D}_m^n$.[7]

How μ-rational is pairwise majority then? We investigate the degree of rationality for certain values of m and n for a probability function, `Pr` assigning equal weight to all elements of $\mathfrak{D}_m^n$:[8]

		Theories		
		3	4	5
Virtues	3	.94444	.8298	.67573
	5	.93055	.7896	
	7	.92490		
	9	.92202		

Table 1: The μ-rationality of pairwise majority for n theories and m virtues

The values in Table 3 indicate that pairwise majority becomes less rational as one increases the numbers of theories under consideration and the number of virtues used to evaluate them.[9] The same

7 These correspondences rely on our restriction on `Pr` stated in fn.5 above.

8 The values in Tables 3 have been calculated in *Mathematica 10*. Please contact the authors if you wish to consult the notebooks used.

9 For a more sophisticated discussion (in the context of social choice) of the results in this table, as well as for a general formula for approximating the probability of a cycle given any number of voters (odd) and any number of alternatives, see DeMeyer and Plott (1970).

trend can be observed when calculating the probability of a Condorcet winner under majority. A Condorcet winner is an alternative which ranks better than all others in one-to-one comparisons with all the other alternatives. Notice that even if an all-things-considered ranking is intransitive there may be still one alternative which is better than all other, i.e. the cycle occurs lower in the ranking. The likelihood of a Condorcet winner has already been investigated in a series of papers in the social choice literature, i.e. Gehrlein and Fishburn (1976) and Gehrlein (1983). Table 4 collects some of the results of multiple papers.[10]

		Theories											
		3	4	5	6	7	8	9	10	11	12	13	14
Virtues	3	.94444	.8888	.8399	.7977	.7612	.7293	.7011	.6760	.6536	.6333	.6148	.5980
	5	.93055	.8611	.80048	.74865	.70424	.66588	.63243		.57682		.53235	
	7	.92498	.84997	.78467		.68168	.72908	.60551	.64090	.54703		.50063	
	9	.92202		.77628		.66976		.59135		.534		.486	

Table 2: The μ-Condorcet-rationality of pairwise majority for n theories and m virtues

So what can we learn from Tables 1 and 2? In cases in which scientists are choosing among a small number of theories, the values remain quite elevated. For instance in choosing between three theories based on five virtues, pairwise majority is .9306-rational. In other words, in less than 7% of cases will a scientist trying to use majority voting run into an intransitive all-things-considered ranking. So, refraining from eliciting the individual rankings of theories based on virtues on account of Arrow's result is irrationally cautious in this scenario. Nevertheless, as scientists choose between increasing numbers of alternatives, using increasing numbers of virtues, the μ-rationality of pairwise majority decreases. Finding a precise

10 The reason for this is that performing these calculations is a computational demanding task and some of the older papers did not have the technical means of obtaining all results.

threshold for when μ is high enough to warrant starting the aggregation procedure, or low enough to refrain from doing so, is not our focus here. It suffices to note that the μ-rationality of a theory choice function is sensitive to the numbers of alternatives and virtues under consideration. For relatively low numbers of both, theory choice using majority voting is rational *enough*.

3.4 DISCUSSION

In this section we address some possible objections to Okasha's framework and our analysis in this paper.

3.4.1 *Kuhn vs. Okasha*

According to Kuhn, scientists guide their choice of theories by looking at how those theories fare with respect to a series of virtues, such as simplicity, accuracy, scope, etc. Okasha interprets this claim as saying that each of these virtues induces a complete ranking over the set of theories and that each scientist aggregates (according to an algorithm set a priori) all of these rankings into an all-things-considered ranking. This then models the order in which the scientist endorses the theories under consideration. We take Okasha's challenge to be that, per Arrow's impossibility theorem, such an aggregation cannot be guaranteed a priori. That is, prior to eliciting the individual rankings, a scientist cannot be sure that the algorithm chosen to aggregate them will deliver an ordering. In consequence, it seems that a scientist wishing to decide what theory to endorse based on this 'Kuhnian' procedure is irrational. In this paper we argue that there are plausible theory choice situations in which a scientist would appear irrationally cautious not to employ this Kuhnian procedure as long as the algorithm she uses is

pairwise majority voting.

But one could question Okasha's interpretation of Kuhn's ideas on multi-criterial theory choice. Firstly, Kuhn does not say that scientific virtues induce complete rankings over the set of alternatives. Secondly, he does not construe theory choice as an algorithmic decision from a set of individual rankings into an all-things-considered ranking. Thirdly, Kuhn does not talk about scientists having a complete all-things-considered ranking over the set of theories. These considerations suggest that Okasha might be diverging from Kuhn's project more than he lets on. But the purpose of this article is not to engage in Kuhnian hermeneutics, but rather to reply to Okasha's challenge to the rationality of theory choice. and whether an algorithmic procedure for arriving at an all-things considered best theory is possible is an interesting, albeit less Kuhnian than Okasha sells it to be, question. This paper shows that the viability of such an algorithm hinges on allowing an all-or-nothing vs. a degrees view of scientific rationality and on setting a threshold for what counts as rational suitable for the theory choice situation one is facing.

3.4.2 *Impartial culture*

In articulating the notion of rationality by degrees we remarked that we require a probability distribution defined over the space of all possible profiles definable over m virtues and n theories. There we assumed that this probability distribution is the equiprobable one. In the social choice literature this kind of assumption is known as the 'impartial culture' (IC) assumption (Gehrlein, 1983). Impartial cultures are a natural starting point: they make computations much easier and they have been widely studied in the social choice

literature.[11]

But our primary motivation for assuming the equiprobable distribution is epistemic.[12] Prior to beginning the process of eliciting the rankings according to each virtue, a scientist cannot deem how likely it is for a particular profile to obtain. Therefore, from the perspective of the scientist, IC functions as a principle of indifference with respect to the different ways virtues rank the competing theories. Of course if one were to assume a different probability distribution, then one would expect the probability of majority cycles occurring to change. But as far as we can see, there is no reason to assume that a different probability distribution (for example, one where different virtues were less likely to submit the same ranking as one another than is assumed in IC) would *increase* the probability of majority cycles. Moreover, it's difficult to see how such a probability distribution could be justified from the *ex ante* perspective of the scientist, before she has elicited the rankings of the theories by virtues.

3.4.3 *How many alternatives?*

Returning to the problem of theory choice, the numbers presented in Tables 1 and 2 suggest that if a theory choice situation is placed in the upper left corner - 3 to 5 virtues and 3 to 5 theories, then the threat of Okasha's Arrovian result is quite small. And consequently,

11 Interestingly, it has been proven that as the number of voters tends to infinity, any deviation from impartial culture will reduce the probability of majority cycles, as long as the deviation isn't in favour of a distribution that assumes the Condorcet paradox from the start (Tsetlin et al., 2003). However, since we are working in a context with a finite, relatively small number of virtues (voters) we cannot rely on this result to motivate impartial culture here.

12 We are grateful to an anonymous referee for encouraging us to motivate this assumption.

moving to a degree notion of rationality would save the rationality of theory choice (at least in that choice situation). But as we are moving away from the upper left corner, and especially if we are increasing the number of theories under consideration, the likelihood of there being a Condorcet winner in a theory choice situation becomes so low that one can no longer make the claim that theory choice is rational enough.

If this model is supposed to capture the kind of theory choice Kuhn was considering, i.e. the choice between different paradigms, then it should only care about very few theories. However, most of the theory choice situations Kuhn discusses are binary, such as the move from the Ptolemaic to the Copernican model of the Solar system or the shift from Newtonian to relativistic physics. These are situations in which Arrow's theorem does not create any problems. Okasha identifies this as a challenge and suggests that, in fact, we should be interested in more realistic theory choice situations such as:

> "statistical estimation, where a researcher might want to estimate the value of a real-valued parameter in the unit interval; the alternatives that must be chosen between are uncountably many. So focusing exclusively on binary choice, as a way of trying to avoid the Arrovian predicament, is at odds with scientific practice." (p. 95)

Although we have no way of gauging how many alternatives are usually in play in theory choice situations, the scenario Okasha sketches above is not one that we consider worrisome for the results in this paper. The reason for this is simple. The choice of a real-value parameter is not a multi-criterial one. Such a choice is one in which there is a single criterion: accuracy.[13] Nothing changes

13 We are grateful to an anonymous reviewer for pointing out that if this involves Bayesian statistical model selection that the criterion might not be accuracy, but

in terms of simplicity, scope, etc. when a real-valued parameter is assigned a different value.

3.5 CONCLUSION

This paper contrasted the view that scientific rationality is an all-or-nothing notion with the view that scientific rationality comes in degrees. We showed that the choice between these two views can have significant implications to how rational we think theory choice is in the face of Okasha's Arrovian challenge.

There may be some for whom the mere possibility, irrespective of how small, of the virtues leaving them without an ordering is enough to make them doubt the possibility of using the virtues to select between competing theories. To them we can only respond that the purpose of this paper was to gauge the threat Okasha raised and evaluate what the prospects of rational theory choice remain in the aftermath of applying Arrow's impossibility theorem to theory choice.

---

rather posterior probability. However, as Okasha (2011, pp.105-110) notes such an approach to model selection is immune from the Arrovian challenge.

4

SUBJECTIVITY, AMBIGUITY AND RATIONALITY

If theory choice were purely an objective matter, then this would be a significant problem. But Okasha fails to pay attention to Kuhn's claims about the subjective elements involved in theory choice; in particular, the role scientists play in disambiguating the scientific virtues. We demonstrate how to do this in the framework Okasha proposes, and show via simulation that this blunts the threat posed significantly.

We proceed as follows. In Section 4.1 we discuss the subjective elements involved in the context of theory choice, i.e. the ambiguity of scientific virtues. In Section 4.2 we argue that Okasha's proposed way of dealing with this in the social choice framework is unsatisfactory. In Section 4.3 we propose an alternative treatment which we then feed into the definition of rationality, in Section 6. Finally, in Section 7 we demonstrate that with this in place, the impossibility result can be bypassed. Perhaps counter-intuitively, the subjectivity involved in disambiguating scientific virtues turns out to be a good thing.

In contrast to these authors, rather than attempting to reformulate the conditions, our focus is on one of Okasha's modelling assumptions. Once profiles of competing theories ranked by virtues are provided, for the purposes of this paper we can grant that Okasha's

result kicks in. What we question instead is whether or not the virtues provide such rankings. If theory choice were perfectly objective then this would be a natural assumption to make. But if Kuhn is correct, then rational theory choice involves subjective elements as well. In particular, at least for our current purposes, subjective elements are required to disambiguate between different ways the same virtue would evaluate competing theories. And if virtues are ambiguous in the sense to be outlined in Section 4.1, then particular theory choice situations do not supply a unique profile with which to augment a theory choice function. And as we demonstrate in Section 4.5, if enough virtues are assumed ambiguous to a certain extent, the threat posed by Okasha's argument dissolves.

4.1 SUBJECTIVITY AND AMBIGUITY

Kuhn (1972) goes at great lengths to explain the impact of subjective as well as objective factors in his model of theory change. According to Kuhn, the way scientists evaluate the adequacy of scientific theories is guided by the scientific virtues. These form a shared and, in Kuhn's view, objective list of adequacy conditions according to which every scientist evaluates every theory. Therefore when confronted with a list of competing theories, scientists will produce rankings of these theories according to the virtues taken into consideration.

Prima facie, each scientific virtue i supplies a unique preference ranking $\prec_i$. So for a given theory choice situation, a scientist who starts the aggregation procedure is faced with one and only one profile from which to generate an all-things-considered ranking. This is a natural assumption to make in the context of orthodox social choice theory, but it fails to capture the appropriate notion of ambiguity of scientific virtues. Kuhn claims that:

> 'Individually the [virtues] are imprecise: individuals may legitimately differ about their application to concrete cases' (1972, p. 357)

and relatedly:

> 'Individuals must then still choose and be guided by the [virtues] when they [choose]. For that purpose, however, each must first flesh [them] out ... and each will do so in a somewhat different way' (1972, p. 364).

This is the sense in which subjective elements, according to Kuhn, enter into the way scientists choose among competing theories, models or hypotheses.[1] For example, two scientists may disagree with respect to how to interpret 'simplicity'. Suppose the competitors are hypotheses in the form of mathematical equations, polynomials for instance. One scientist might believe that equation x is simpler than equation y if and only if x contains strictly fewer parameters than y. Another might use the order (the largest exponent) of the equations as their guide to simplicity. A third might use the computational labour required to generate solutions to the equations). Alternatively, suppose the competing theories consist of qualitative statements that consist of equal numbers of universally quantified conjunctions. One way of comparing them with respect to accuracy would be to simply count the number of strictly true conjuncts.

1 There is another natural way of thinking about the subjectivity involved in choosing between competing scientific theories (Kuhn, 1972, p. 358). When trying to arrive at an all-things-considered ranking of the theories presented with, different scientists may assign different weights to how much the virtues 'count' for the final ranking. For instance, radical empiricists will most likely assign a very high importance to accuracy to the detriment of all the other virtues, whereas others might be interested in a mix of accuracy and simplicity. So, although the virtues form a common template according to which theories are being evaluated, different scientists may disagree to how important some of them are. This second sense in which subjectivity appears in the context of theory choice raises interesting problems, but we will not address it in this paper.

Another would be to compare the absolute number of falsifying instances (by summing the number of falsifiers across the conjunction). These may plausibly deliver different results. Irrespective of the different reasons for taking virtues to be ambiguous, Kuhn believes that:

> 'The considerable effectiveness of [scientific virtues] does not ... depend on their being sufficiently articulated to dictate the choice of each individual who subscribes to them' (1972, p. 362).

Therefore, although the virtues according to which decisions are being made are objective and shared by every member of a particular scientific community, the way each virtue ranks theories is not a matter of fact. There are three senses in which this could be the case, mapping to three different ways of understanding the "concrete cases" Kuhn talks about. First, the same scientist is free to adopt a different interpretation in different theory choice scenarios (i.e. where choosing among different sets of competing theories, models, or hypotheses). Second, the same scientist is free to adopt a different interpretation in different theory application scenarios (i.e. where the theories in question are fixed, but applied to different target systems). In some of these scenarios she may interpret simplicity in one way (number of parameters say), and in another she may adopt a different interpretation (computational labour required to make predictions say). In other words, there is no threat of irrationality stemming from this kind of inter-context inconsistency with respect to the interpretation of a virtue.

But more significantly from our current perspective, even in a particular theory choice scenario, where the pertinent virtues, competitors, and application are fixed, different scientists may legitimately disagree about how to interpret each scientific virtue in that context. Such different interpretations can presumably lead

to different orderings of theories, and therefore different profiles with respect to which to apply an aggregation procedure. In this sense, even in a particular theory choice context, there is no 'matter of fact' with respect to how a scientific virtue orders the competing theories. Different rankings can be equally 'correct', and thus a scientific virtue can be ambiguous between them. To explain why someone ranks theories in a particular context according to simplicity, is a question for the sociologist and psychologist, thinks Kuhn, and not for the analytical philosopher of science.

So, to sum up. In particular theory choice contexts, scientific virtues can be ambiguous across multiple orderings of the competitors. Different sociological and psychological facts about the scientists involved can legitimately lead them to disambiguate a virtue in a different manner. Ambiguity thus allows for some freedom of movement between different rankings. How much freedom of movement differs from context to context and also depends on the subjective reasons guiding individual scientists towards particular rankings. In this paper we model this space of movement and investigate the impact this extra dimension of theory choice has on the notion of scientific rationality and ultimately on bypassing the impossibility Okasha discusses.

4.2 OKASHA'S TREATMENT

Okasha claims that '[d]isambiguation can always be carried out by sub-dividing an ambiguous [virtue]' (2011, p. 85) into non-ambiguous distinct preference rankings, each of which are included in the profile of rankings that a theory choice function takes as an argument. So if, for example, simplicity is ambiguous between two distinct rankings $\prec_{si}$ and $\prec_{si'}$, then both are included in the profile.

This approach is unsatisfactory as it does not capture Kuhn's claims about legitimate disagreement with respect to how virtues apply to concrete cases. Rather than modelling $\prec_{si}$ and $\prec_{si'}$ as competing disambiguations of simplicity, Okasha's approach treats them as being compatible alternatives. So multiple scientists who disagreed with respect to how to disambiguate simplicity in a concrete instance of theory choice would be treated as agreeing that each disambiguation should be used for the purpose of generating an all-things-considered ranking. Okasha's approach thus multiplies the objective element of theory choice – an additional virtue for each disambiguation of a virtue – rather than recognising Kuhn's claim that there is a subjective element involved in turning these less than fully articulated notions into ones that can guide choice.

We therefore believe that an alternative account of ambiguity in theory choice is worth pursuing. We propose such an account below and argue that it formally captures the idea that virtues are ambiguous in Kuhn's sense, thus capturing the subjective element involved in theory choice, and that it blunts the threat posed by Okasha's argument.

4.3 A KUHNIAN CONSTRUAL OF AMBIGUITY

We have argued that on Kuhn's construal of theory choice scientific virtues cannot provide an objective (or in any sense 'true') ordering of theories. Therefore, given their ambiguity scientists can have some freedom of movement in between different orderings of the same set of theories under the same virtue. How different can two rankings be so that a scientist treats them as being different disambiguations of the same virtue? This is sensitive to how many different ways there are to rank the alternatives, which is sensitive to the size of the choice set under consideration. For n alternatives,

there are $n!$ distinct (strict) rankings over them. So, with respect to our example of choosing between $\mathfrak{T} = \{x, y, z\}$, there are 6 possible rankings over $\mathfrak{T}$. A virtue that specified a precise ranking would be unambiguous. A virtue which was maximally ambiguous between all of these 6 would be uninformative. Between these two extremes, there are multiple ways of setting a sensible threshold.

One obvious way to proceed is to define a notion of 'closeness' between two rankings in terms of the number of their pairwise disagreements. Two rankings are 'close', for instance, if and only if they differ on only one pair of alternatives. To understand this notion of 'closeness' better assume $\mathfrak{T} = \{x, y, z, u, v\}$ and consider the four rankings which are close to the ranking $v \prec_i u \prec_i z \prec_i y \prec_i x$.

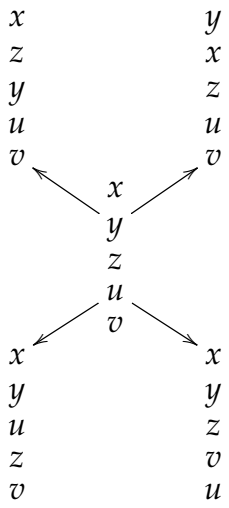


Figure 1: Example of close rankings

This notion can be extended in a natural way to profiles. To illustrate this consider the following example. Suppose a scientific community is faced with the choice out of $\mathfrak{T} = \{x, y, z, u, v\}$, by means of 3 virtues, i, j, k and that $x \prec_j v \prec_j u \prec_j z \prec_j y$ and

$y \prec_k x \prec_k v \prec_k u \prec_k z$. In the below set all profiles are close to the profile in which $v \prec_i u \prec_i z \prec_i y \prec_i x$.

$$\left\{ \left\langle \begin{matrix} x & y & z \\ y & z & u \\ z & u & v \\ u & v & x \\ v & x & y \end{matrix} \right\rangle, \left\langle \begin{matrix} x & y & z \\ z & z & u \\ y & u & v \\ u & v & x \\ v & x & y \end{matrix} \right\rangle, \left\langle \begin{matrix} y & y & z \\ x & z & u \\ z & u & v \\ u & v & x \\ v & x & y \end{matrix} \right\rangle, \left\langle \begin{matrix} x & y & z \\ y & z & u \\ u & u & v \\ z & v & x \\ v & x & y \end{matrix} \right\rangle, \left\langle \begin{matrix} x & y & z \\ y & z & u \\ z & u & v \\ v & v & x \\ u & x & y \end{matrix} \right\rangle \right\}$$

Figure 2: Example of close profiles

There are two different ways in which the definition of 'closeness' can be relaxed. Firstly, we have so far assumed two rankings are 'close' if they differ in at most the ordering of a single pair of theories. This makes more sense if the set of alternatives is small than it does for larger sets. Consequently, in general, we may consider two rankings as being close even if they differ in the ordering of β pairs of alternatives. In the example presented in Fig. 1 $\beta = 1$, which meant that for any given ranking there are four others which are close to it. But if $\beta = 2$ then this grows to include 9 more additional rankings. And if $\beta = 10$ then this corresponds to a trivial notion of closeness in the sense that any way of ranking the alternatives is close to any other. In general, for a set of n alternatives, the maximum value β can take is $\frac{n(n-1)}{2}$. As before, this generalized relation of closeness can be naturally extended to deliver a notion of closeness for profiles.

Secondly, there is no reason why closeness between profiles be judged in reference to only one virtue. That is, for a given profile there may be profiles which are close to it in the sense in which they differ from the original with respect to pairwise disagreements on multiple virtues (or one which disagrees on virtue i and another that disagrees on virtue j). We will use α to denote the number of virtues assumed in the definition of closeness. In the example in

Fig.2, $\alpha = 1$, and this delivered four profiles. If $\alpha = 2$ then there would be 20 additional profiles.

We can combine the two generalisations and obtain the notion of α-β-closeness. As α and β grow the size of the set of profiles close to a particular profile grows as well.[2] This notion of closeness can help us model Kuhn's idea of ambiguity.

If virtues really are ambiguous, as Kuhn suggests they are, then can the profiles which majority voting[3] maps to an intransitive relation be avoided by replacing them with other profiles α-β-close to them? Such a move would be permissible since, as per Kuhn, the ranking of competing theories based on scientific virtues is not set in stone, and each virtue may be disambiguated in different ways, where different disambiguations may result for reasons specific to each individual scientist. The point then, is that although a scientist may have well justified (sociological or psychological) reasons to disambiguate each virtue the way that she does, she cannot hold each disambiguation at the same time, on pain of her favoured aggregation function mapping to an intransitive/incomplete result. But, if she cannot hold each disambiguation at the same time, then this provides evidence that something has 'gone wrong' in the way that she has disambiguated the virtues, at least when considered together. And thus, there is nothing stopping her from revisiting

2 Notice that if we increase α and β too much, i.e to $\alpha = m$ and $\beta = \frac{n(n-1)}{2}$, then this set collapses into the domain that profile belongs to.

3 Pairwise majority voting can be defined thus: for a set of virtues $\mathfrak{V}$, let $\Delta^+ =_{df} \{i \in \mathfrak{V} : y \prec_i x\}$, $\Delta^- =_{df} \{i \in \mathfrak{V} : x \prec_i y\}$. Then: y $\prec$ x if and only if $|\Delta^+| \geq |\Delta^-|$. For the remainder of this paper we will focus on pairwise majority voting and assume there is an odd number of virtues. This means that the output of the aggregation, under pairwise majority, will always be a strict ordering (if an ordering, at all). We use pairwise majority vote for its simplicity in illustrating our argument. We do not thereby suggest that it be the actual aggregation function used; clearly, for example, one may want to weight different virtues differently.

them and adopting alternative close disambiguations for (at least some) of the virtues.

It bears noting here that we are not claiming that our notion of closeness, in terms of pairwise disagreements with the original disambiguations, is the only (or even the best) way of thinking about ambiguity in the formal framework under consideration. This (similar to our choice of focusing on pairwise majority voting) is a modelling assumption, and it would be an interesting question to consider different ways of capturing how a scientific virtue could be ambiguous over multiple rankings. However, it does seem plausible that if a scientist were to be forced to revisit how she disambiguated a certain scientific virtue in a given theory choice situation, then adopting an alternative disambiguation that radically disagreed with her original one would be undesirable. It could even be taken as a sign that her initial attempt at interpreting the virtue was not sufficiently well grounded. Alternatively, the sociological or psychological reasons that led to the original disambiguation might be seen as having an 'anchoring' effect. The claim is that when faced with an intransitive/incomplete all-things-considered value, alternative disambiguations that largely agree (in terms of pairwise disagreements) with the original disambiguations are to be preferred.

To illustrate the potential of such an approach to bypass Arrow's impossibility result, consider the following case: assume three virtues are used to rank three alternatives. It is trivial that majority voting maps some profile 3-3-close to any Condorcet-like profile[4] into a transitive ranking, since any element of the domain under consideration is 3-3-close to any other (so think of the profile in

4 A Condorcet profile is simply a profile that pairwise majority voting maps to an intransitive value.

which all virtues agree). This is not very useful as these particular large values of α and β do not appear plausible for such a small number of theories and virtues.

Nevertheless as the number of competing theories and virtues grow, higher values of α and β become intuitively plausible, since for more competitors and virtues, fixed values of α and β span proportionally less of the corresponding universal domain. A virtue faced with a large number of theories may be such that multiple pairwise disagreements between competing disambiguations are allowed. And if there are a larger number of virtues we may allow for more of them to be ambiguous.[5] Notice we are not suggesting replacing the troublesome profiles (those that instantiate a Condorcet-pattern of preferences) with simply any other profile in the domain. We want to restrain the possible replacements as much as possible, and we do so by only looking at profiles which are α-β-close to the troublesome ones, for plausible values of α and β.

But what precisely are plausible values? This is a difficult question to answer in the abstract for two reasons. First, as noted above, it does seem reasonable to assume that the plausibility of certain values of α and β depends on the number of alternatives (n) and virtues (m) involved in the theory choice situation. Secondly, and more importantly, even fixing the numbers n and m, whether or not particular values of α and β are plausible will still presumably depend on the particular field of research in question, and the individual scientist doing the aggregation. A scientist with deeply entrenched

5 We conjecture, that for each domain, there are minimal values of α and β such that majority voting will succeed somewhere in each neighbourhood of profiles, and that the values of α and β are low enough to make the rationality of theory choice nontrivial. Proving this conjecture is beyond the scope of this paper, but is a viable avenue for future research.

reasons for disambiguating in a way that led her favoured aggregation function delivering an intransitive/incomplete value may well resist considering alternative disambiguations which disagreed significantly with her original disambiguation (thereby restricting the value of β). And she may well resist the idea that she has to revisit a large number of the virtues she has been using to guide her choice (thereby restricting the value of α). As such, a detailed discussion of which α and β are plausible would require a detailed discussion of particular theory choice scenarios, which we cannot do in this paper. Instead, we aim to offer a proof of concept highlighting the viability of such an approach to the conceptualisation of scientific rationality and show that it blunts the threat raised by Okasha. In the following section we explain how α-β-closeness can lead to a weakened notion of rationality in the context of theory choice. And then we present some results obtained by applying this relaxed notion of rationality to several simple theory choice scenarios.

4.4 WEAK RATIONALITY

Suppose that scientific virtues, such as simplicity, accuracy, scope, etc. do deliver complete rankings of competing theories. Then, for any scientist, a theory choice situation can be represented by a profile collecting all the ranking of theories according to the individual virtues. Okasha construes the rationality of theory choice in the following way:

RATIONALITY Theory choice is rational only if there exists a function that respects UD, WP, ND, IIA and that takes every profile in UD to a transitive and complete all-things-considered ordering.

If one were to accept Okasha's principles governing theory choice discussed in Section 2, then Arrow's result indeed shows that theory

choice is irrational. In other words, there is no such function that outputs a transitive and complete all-things-considered ordering no matter what rankings the virtues supply. Using the notion of α-β-closeness introduced in the previous section, we aim to offer a double-tier weakening of Rationality.

WEAK RATIONALITY Theory choice is rational only if there exists a function that respects U, WP, ND, IIA and that for at least γ profiles, takes those profiles, or profiles α-β-close to them, into transitive and complete, all-things-considered orderings.

Rationality and *Weak Rationality* differ in two respects. Firstly, whereas *Rationality* stipulates that the aggregation function ought to deliver a transitive and complete all-things-considered ranking for any profile, *Weak Rationality* is concerned only with γ profiles. For simplicity, we will express γ as a percentage of profiles in the domain for which *Weak Rationality* holds. Secondly, *Weak Rationality* not only checks the behaviour of a function applied to a profile, but also its behaviour applied to all profiles α-β-close to it. So, in case we try to apply pairwise majority voting, say, to a Condorcet profile, *Weak Rationality* will be satisfied if pairwise majority voting can deliver, for at least one profile α-β-close to the Condorcet profile, a transitive, all-things-considered ranking. In contrast, Condorcet profiles represent the counterexamples to satisfying *Rationality* with pairwise majority voting.

The first weakening is motivated by the following idea. Arrow's theorem tells us that theory choice is not rational according to *Rationality* because for any function there will be at least one profile that function will not map to a transitive and complete all-things-considered ordering. But what if there really is a single such profile for a particular function? Let f denote your favourite aggregation function. There is a sense in which f would be *less* 'rational' (in an intuitive sense) if it generated a transitive and

complete all-things-considered ordering from only 1 out of 216 profiles (the space of all profiles formed out of three virtues ranking over three theories), than it would be if it did so from 215. If a scientist used the function in the former case she would be acting 'irrationally.' But she wouldn't if she did so in the latter. In fact, not using f in such a scenario simply because of the threat posed by the single profile f fails to map to an ordering, would be 'irrationally' cautious. Suppose a bookie offered you a choice between two bets. the first bet returns £100 with probability 1, the second £200 with probability .5 and £0 otherwise. Preferring the first bet in this instance seems rational. However, as the probability of winning £200 increases, the second bet becomes more appealing. There seems to be a point at which preferring the first to the second bet becomes irrationally cautious. Analogously, a scientist refraining from using f because of its failure in only 1 out of 216 cases appears irrationally cautious. As such, rationality can be treated as a matter of degree.

The second weakening is inspired by Kuhn's idea that the way in which virtues rank theories is subjective. We understand this as saying that in case a particular profile delivers, under an aggregation function, an intransitive (or incomplete) all-things-considered relation, moving to an α-β-close profile is permissible. We have discussed Kuhn's idea of ambiguity in Section 3 and the meaning of α-β-closeness in Section 5. The issue of what values α and β should take is still beyond the scope of this paper, but in the next section we present a proof of the viability of this proposal for saving the rationality of theory choice, if this is construed as *Weak Rationality*.

4.5 AMBIGUITY TO THE RESCUE

To illustrate the viability of moving from *Rationality* to *Weak Rationality*, consider a very simple theory choice scenario. Take α and β to be 1. That is, treat two profiles as being close if they differ in how they rank only one pair of adjacent theories according to a single virtue. Also take majority voting to be the method used to aggregate the individual rankings supplied by the virtues into an all-things-considered ranking. Then, for three virtues and three theories every Condorcet-like profile is close to a profile for which majority voting succeeds in mapping that profile to a transitive and complete all-things-considered ranking.

What this means is that scientists who are willing to revise the way one ambiguous virtue ranks a pair of theories then the intransitive all-thing-considered ranking can be avoided. Given that the values of α and β are minimal and γ=1 in this case, the theory choice among 3 theories based on 3 virtues appears to be *Weakly Rational*. And pairwise majority voting is the witness of this result. The below table documents some more values as a function of the number of theories and virtues under consideration (note that we have not included the, plausibly common, instance of choosing between only two competitors. As Okasha (2011, pp. 94-95) notes, the Arrovian result does not hold in this case).[6]

6 These numbers were computed using *Mathematica 10*. Contact the authors for the notebooks used.

		n		
		3	4	5
	3	1	.9826	.8836
m	5	.9907	.9375	

Table 3: Proportion of profiles which are 1-1-close to a successful profile under majority voting for m virtues and n theories

The columns in the above table denote the number of theories which need to be ranked, whereas the rows denote the number of virtues according to which these theories are assessed. Every cell contains the proportion of 1-1-close to at least one profile that majority votes maps to a transitive ordering.

Notice that for other cases than three theories ranked according to three virtues (so cases of more virtues or more theories) the behaviour of majority voting is not as 'nice'. Only .88 of the profiles comprised of 3 virtues ranking 5 theories will be such that they are at most 1-1-close to a profile that maps to a transitive all-things-considered ranking under majority voting. Is this value of γ good enough for claiming that theory choice using 3 virtues to rank 5 theories is *Weakly Rational*? This seems like a subjective decision inherent in theory choice and we do not wish to argue either way. Instead, it is more fruitful to observe that the situation of majority voting is improved if we relax the definition of closeness. The below tables document these improvements:

3 virtues		α		
4 theories		1	2	3
	1	.9826	1	1
β	2	1	1	1
	3	1	1	1

Table 4: Proportion of profiles which are α-β-close to a successful profile under majority voting for 3 virtues and 4 theories

3 virtues		α		
5 theories		1	2	3
	1	.8836	.9385	.9539
β	2	.9748	1	1
	3	.9972	1	1

Table 5: Proportion of profiles which are α-β-close to a successful profile under majority voting for 3 virtues and 5 theories

5 virtues		α		
3 theories		1	2	3
	1	.9907	1	1
β	2	1	1	1
	3	1	1	1

Table 6: Proportion of profiles which are α-β-close to a successful profile under majority voting for 5 virtues and 3 theories

5 virtues		α		
4 theories		1	2	3
	1	.9375	.9831	.9935
β	2	.9805	.9995	1
	3	.9982	1	1

Table 7: Proportion of profiles which are α-β-close to a successful profile under majority voting for 5 virtues and 4 theories

Each table above corresponds to a cell in Table 3 except for the top left cell. So, Table 4 corresponds to the case of three virtues ranking four theories, Table 5 three virtues ranking five theories, and so on. The columns in these tables denote the number of ambiguous virtues considered, whereas the rows denote the number of pairwise disagreements allowed. The numbers in the cells denote the proportion of profiles in the appropriate domain which have at least one profile, α-β close to them for which majority voting delivers a transitive all-things-considered ranking. For the case of

1 ambiguous virtue and 1 pairwise disagreement the numbers are identical to those in Table 3. However, as we move away from that cell, γ increases and also reaches 1 in each instance.

For a more interesting example, consider Table 5. Here we record the results for the case in which we need to rank 5 competing theories by means of 3 scientific virtues. Now, narrow in on the intersection of row 2 and column 1 in this table. This cell represents the situation in which we construe two profiles as being close if they differ in the rankings of at most 1 virtue ($\alpha = 1$) and on this virtue the difference between the two profiles can be in the ordering of at most two pairs of adjacent theories ($\beta = 2$). The value in this cell is .9748. This tells us that only 2% of all profiles of 5 theories ranked by 3 virtues do not have a 1-2-close profile for which majority voting delivers an overall transitive ranking. Allow two profiles to be close even if they differ with respect to the ordering of two pairs of adjacent theories, i.e. move a column to the right, and all profiles in the space have a profile 2-2-close to them for which majority voting delivers an overall transitive ranking.

In other words, assuming closeness is lax enough, all of the simple cases of theory choice are such that all profiles (even the problematic ones) will have at least one profile close to them that is mapped to a transitive all-things-considered ranking under majority voting. And notice that this possibility result has been achieved without trivialising the definition of closeness, i.e. for small values of α and β. For instance, in the case of 4 theories and 3 virtues (Table 2) as well as in the case of 3 theories and 5 virtues (Table 4), this is the case for any value of α and β greater than 1. In the case of 5 theories and 3 virtues (Table 3), an $\alpha > 1$ and a $\beta > 1$ are sufficient. The case of 4 theories and 5 virtues (Table 5) is more demanding, but since there are also more virtues and theories, slightly higher values

for α and β are still non-trivial. Finally, even in cases in which there are still profiles which are not close to any profiles for which pairwise majority voting delivers an overall transitive ranking (such as Table 5, line 1), the probability of ending up with a non-transitive profile is much lower (under any construal of closeness, i.e. any values of α and β) than in the case in which we take virtues to be non-ambiguous.

The interpretation we assign to this result is that as long as communities of scientists include enough subjective disagreement about how to disambiguate the shared and objective virtues, then at least some of them will be able to rationally aggregate the competing alternatives. And theory choice ends up being rational, albeit weakly so.

4.6 CONCLUSION

Okasha (2011) aims to prove that the situation of theory choice, if we are to construe it in Kuhnian terms, is even worse than Kuhn anticipated. It is not the case the objective elements of theory choice alone do not supply an unique algorithm for arriving at an all-things-considered ranking of theories, but rather they supply no such algorithm at all. This poses a threat to the objectivity of theory choice. In this paper we argue that Kuhn's ideas regarding the subjectivity of the rankings generated by the ambiguous virtues offer a solution to Okasha's challenge. Taking the ambiguity of scientific virtues seriously, what Okasha shows is that across the universal domain, regardless of the function used, it is not the case that every disambiguation will yield a rational choice. But he doesn't show that no disambigutation will do this. We investigate some simple cases of theory choice and prove that rational aggregation is possible as long as the definition of ambiguity is sufficiently relaxed and

more virtues are treated as being ambiguous. It therefore seems that the considerable effectiveness of scientific virtues depends on them being sufficiently unarticulated.

5

SCIENTIFIC CONSENSUS WITHOUT INCONSISTENCY

5.1 INTRODUCTION

Academic disciplines are increasingly fragmented, and this naturally leads to diverse areas of expertise within them. But if the state of a discipline as a whole is supposed to depend on the beliefs of its experts in a certain way, then we arrive at an impossibility result similar to Sen's Liberal Paradox (1970). Just as Sen's Lewd and Prude were forced into inconsistency by their decisiveness over their own personal spheres, a discipline may be forced into inconsistency if experts are taken to be decisive over their respective areas of expertise. At least this will be the case if we require that the scientific consensus in the discipline – i.e what the discipline as a whole tell us about the world – be generated by a suitably-constrained function on the beliefs of scientists. We show this by importing Dietrich and List (2008)'s result from the context of aggregating individual judgements to that of aggregating scientific expert judgements. If this is the correct way of thinking about what fragmented scientific disciplines tell us about the world, it seems that science cannot protect itself from inconsistencies. Insofar as we think that Nature is free of inconsistency, and we want our scientific consensus to reflect this, something has to give. Building on the ideas of Gibbard (1974) we suggest that the best way of avoiding

inconsistencies is for experts to waive their expertise and contribute beliefs on a par with everyone else's. As such we argue against the hegemony of experts.

The structure of this paper is as follows. We begin by outlining the notions of fragmentation and expertise we have in mind. We discuss how scientists from different fields may try to come to a consensus about what science as a whole tells us about the world. We then present Dietrich and List's model and impossibility result, along with an interpretation in terms of aggregating the beliefs of scientists. We incorporate the idea of waiving expertise into the model, prove a possibility result, and argue that that the principles that the result requires are highly plausible in the contexts under consideration.

5.2 FRAGMENTATION OF KNOWLEDGE

Most, if not all, of modern academic disciplines exhibit a huge degree of fragmentation and specialisation. By this we mean that individuals working within them work on diverse topics of investigation, and that individuals working on a topic X are experts on topic X. We take it for granted that this is the case in philosophy.[1] But our primary interest here is the fragmentation and specialisation of science. For it is here where the stakes are at their highest.

Modern science is a massively broad field. The classical categorisation of physics, chemistry, biology (and possibly social science)

1 We doubt that anyone would dispute this. But if they do, we point them to Rescher's recent editorial for the *American Philosophical Quarterly*. Rescher documents the growth in philosophical output, including the tendency for increased specialisation, and notes that the 'discipline's topical fragmentation [has] far outpaced its population growth' (2014).

provides one level of fragmentation. A physicist working on Quantum Field Theory may have little to say to a doctor conducting randomized control trials to test the efficacy of a newly developed medicine. This also provides our first level of specialisation. Very few scientists make valuable contributions to different disciplines. Different scientists are experts over different areas.

The same observations apply when we 'zoom in'. Within scientific disciplines, at the level associated with a university faculty for example, there is further fragmentation. A chemist working on the the nature of covalent bonds is far removed from one synthesising artificial molecules capable of transmitting genetic information. Again, this fragmentation leads to individuals being considered as experts over their respective areas of research. And we needn't stop there. The same considerations apply when we 'zoom in' further, to specific university departments, or further still, to particular areas of research of the sort an international conference might be dedicated to. At each level of grain we find fragmentation accompanied by expertise. Obviously, this will not hold all the way down. Our point is that it holds for many areas of research which are still considered to be the same discipline.

For the most part, the fragmentation of any field into sub-disciplines is seen as a desirable, and even natural, phenomenon. Kitcher suggests that communities of scientists who diversify their cognitive labour are far more likely to find the truth (Kitcher (1990), recently discussed in Weisberg and Muldoon (2009) and McKenzie Alexander et al. (2015)). And the sheer breadth of science demands that individuals focus on narrower areas of research.

But this invites questions concerning the state of the discipline as a whole. If we want to know what 'science' tells us about the

natural world, then we need a way of aggregating the beliefs of experts coherently. The same applies if we want to know what 'physics' tells us, or what 'molecular chemistry' tells us, and so on. For any suitably fragmented discipline X, if we want to know what X tells us about the world, we need to aggregate the beliefs of experts working on sub-fields of X. One may question whether what science tells us about the natural world relies only on what individual scientists believe at a time. In the philosophy of science literature, it has been proposed that the way science forms consensus is influenced by institutions, history, or power relations (see Okruhlik (1994) for some intriguing examples). We grant these influences on what science tells us about the world. However, we believe these considerations do not feed into the project we are interested in. Rather, they arise further upstream and concern how individual scientists form their beliefs. Once this is fixed (regardless how), the question we are interested in arises. How *should* we aggregate these beliefs to deliver what the discipline as a whole tells us?

It seems innocuous to suggest the following requirements: the state of a discipline should depend on what the disparate specialists within it believe; specialists should determine what the discipline believes about their area of expertise; and whenever everyone agrees on something, the discipline as a whole should agree on that too. However, results from Dietrich and List (2008) building on the liberal paradox introduced by Sen (1970), show that if this is indeed the case, then we cannot guarantee that the resulting consensus will be free of inconsistency. For any suitably fragmented discipline, there will be cases where individually consistent experts will, collectively, deliver inconsistent claims. And, insofar as inconsistencies do not arise in the world, we take it for granted that an inconsistent science

would require correction. The question is, can this be done whilst still respecting scientific expertise?

5.3 THE IMPOSSIBILITY OF CONSISTENT EXPERT CONSENSUS

In this paper we adopt a formal model for aggregating judgements over sentences, supplied by Dietrich and List (2008) (but also List and Pettit (2002); Dietrich (2006); Dietrich and List (2007, 2013) amongst others). In this section we outline the model and provide an interpretation in terms of aggregating the beliefs of scientists working in a fragmented discipline. Begin by considering a group of *individuals*, $N = \{1, \ldots, n\}$ with $n \geq 2$, and a set of *sentences*, $\mathfrak{L}$, closed under negation: $\varphi \in \mathfrak{L}$ if and only if $\neg\varphi \in \mathfrak{L}$. A set of sentences from $\mathfrak{L}$ is *consistent* if and only if it has a model in the specified logic.[2] For instance, $\{p, p \rightarrow q, \neg q\}$ is inconsistent in classical propositional logic, and $\{p, q\}$ is not. With respect to the intended use of this (idealised) model, we will take the set of individuals to supply the set of scientists working in a fragmented discipline and the set $\mathfrak{L}$ to consist of all the meaningful claims from that discipline.

Individuals in our model formulate judgements over sentences in an agenda. An *agenda* is a finite nonempty set $X \subseteq \mathfrak{L}$ closed under negation.[3] Define a *position* on $\varphi \in X$ as either φ or $\neg\varphi$. Then call an agenda X *connected* if and only if for any two sentences φ and ψ, there is a set of sentences $Y \subseteq X$ such that some position on φ and some position on ψ are each individually consistent with

2 We restrict our focus to monotonic logics, i.e. ones in which any subset of a consistent set of sentences in the logic is consistent.

3 More accurately, we understand an agenda as consisting not of sentences, but rather equivalence classes of sentences modulo logical equivalence. As a consequence φ and $\neg\neg\varphi$ are treated as the same element of X. For simplicity, we restrict our focus to finite agendas throughout but see fn.6.

Y, but jointly inconsistent with Y. An agenda is a set of claims a group of scientists are attempting to reach a consensus on, and we assume that it is connected. Although scientific fields are fragmented, the fragments are not logically disconnected from one another. We think that this holds even at the most general level. Interdisciplinary work shows that the classical categories overlap, and the literature on inter-theoretic reduction supports the thesis that an overarching agenda consisting of sentences from each of these categories would remain connected (Dizadji-Bahmani et al., 2010). Those uncomfortable with the general claim should note that the connectedness of agendas becomes even more plausible when we 'zoom into' more narrow fields, and the results discussed in this paper hold there as well.

The *belief set* of an individual i is a consistent set $A_i \subseteq X$. A profile of belief sets is an n-tuple of belief sets $(A_1, \ldots, A_n)$. When it comes to forming a consensus, we need a *judgement aggregation function* F that takes profiles to belief sets. The input of the function are the sentences of the agenda individually believed by scientists. The value is the scientific consensus: the set of sentences agreed upon by the discipline as a whole.

To account for the idea of expertise, we introduce the notion of an expert rights system. An *expert rights system* is an n-tuple $(R_1, ..., R_n)$, where each R_i is a (possibly empty) subset of X closed under negation. For each individual i, R_i contains the sentences concerning her area of expertise. Call an individual i *decisive* on a set of sentences $Z \subseteq X$ if and only if $F(A_1, \ldots, A_n) \cap Z = A_i \cap Z$. Fragmented disciplines contain scientists who are experts over certain areas of research. The sentences (and their negations) from those areas go into the scientists' expert rights sets. Since they are the experts over these sentences, they should be decisive over

whether or not science as a whole should adopt them. We assume throughout that all rights sets in a rights system are disjoint. So for any sentence under consideration there corresponds at most one expert (think of her as the representative of her sub-discipline if you like). If the same sentence appears in multiple rights sets, the impossibility result to come would be trivial.[4]

The idea that the beliefs of a discipline should depend on the beliefs of the scientists within that discipline is captured by the requirement of a judgement aggregation function. The following condition on the function demands that regardless of what the individual scientists believe (as long as they are individually consistent), they should be able to attempt to reach consensus following the agreed aggregation method:

UNIVERSAL DOMAIN The domain of F is the set of all possible profiles of consistent judgement sets.

The definition of Universal Domain differs slightly from the definition of Dietrich and List (2008, p. 63). They require the function F to be defined over the set of all possible profiles of consistent and complete judgement sets. Requiring completeness does not seem intuitive for scientists trying to arrive at a consensus on what the discipline tells us about nature. We needn't restrict ourselves to cases where every scientist involved in the aggregation procedure has an opinion on every sentence in the agenda. As we explain later in this section, this variation makes no difference to the impossibility result to come.

4 There are ways of generalising the below to allow for expert rights of groups. We suppress them here for brevity, but see Dietrich and List (2008, pp. 64-65) for details.

The below is a formally precise formulation of the suggestion that experts should determine what the discipline believes regarding their respective areas of expertise:

MINIMAL EXPERT RIGHTS There are at least two individuals i, j who are respectively decisive over non-empty R_i and R_j.

While the fact that whenever everyone agrees on something it should be believed by the discipline as a whole is given by:

UNANIMITY PRINCIPLE For any profile $(A_1, ..., A_n)$ in the domain of F and any sentence $\varphi \in X$, if $\varphi \in A_i$ for all individuals i, then $\varphi \in F(A_1, ..., A_n)$.

With this technical apparatus at hand, we can now present Dietrich and List's result establishing that a fragmented science that arrives at consensus via an aggregation function that respects the above conditions cannot protect itself against an inconsistent view of nature:

Theorem 1. *If the agenda is connected, there exists no aggregation function (generating consistent collective judgement sets) that satisfies universal domain, minimal expert rights, and the unanimity principle.*

Proof. See Dietrich and List (2008, pp. 72-73). Notice that they restrict their focus to a proper subset of the universal domain as we define it, i.e. the set of consistent and complete profiles. But if there is no function that satisfies minimal expert rights and the unanimity principle over their domain, then, *a fortiori* there is no function that satisfies them over a superset of that domain (our universal domain). □

The following example should make this clear (Dietrich and List, 2008, p. 60). Let the agenda under consideration consist of the following sentences (including their negations):

p: Carbon dioxide emissions are above a critical threshold.

$p \to q$: If carbon dioxide emissions are above a critical threshold, then global warming is occurring.

q: Global warming is occurring.

Let individual m be the expert on carbon dioxide emissions and n the expert on global warming. Thus $R_m = \{p, \neg p\}$ and $R_n = \{q, \neg q\}$. Suppose that m believes that carbon dioxide emissions are above the threshold, and that if they are, then global warming is occurring. Suppose that n believes the conditional, but denies that global warming is occurring. Thus $A_m = \{p, p \to q, q\}$ and $A_n = \{\neg q, p \to q, \neg p\}$. By universal domain, F must be defined over the profile (A_m, A_n). By unanimity $p \to q \in F(A_m, A_n)$. By minimal expert rights $p, \neg q \in F(A_m, A_n)$. Thus, $F(A_m, A_n)$, a very small, but very important, part of the scientific consensus is inconsistent.

5.4 SCIENTIFIC CONSENSUS WITHOUT INCONSISTENCY

Dietrich and List's result shows that allowing individuals to exercise their expertise may lead to inconsistent collective beliefs. The problem arises when the beliefs of experts come into conflict with the beliefs of other experts via unanimity. The question becomes what to do in such situations? We suggest that the experts over the claims that generated the conflict should waive their expert status with regards to those claims. This attitude in front of a collective inconsistent set of beliefs is one of intellectual modesty. Experts should take the conflict arising between their beliefs as evidence that a mistake has entered into one of their belief sets. Further, since all of them are peers, and there is no reason to suspect anyone (including oneself) over the others, they should all agree to waive their expert rights over the sentences that led to the inconsistency, and agree to perform the aggregation as if none of them were experts over these sentences (this does not

preclude retaining expertise over some elements in the agenda, though). Finding a consistent belief set does not provide certainty that the result of the aggregation is the ultimate truth on the issues on the agenda. But finding an inconsistent collective belief set is evidence that at least one of them entertains a falsehood.

To capture this move towards intellectual modesty when confronted with an inconsistency we introduce an alternative notion of expertise in the spirit of Gibbard (1974). We define an individual i's *waiver set* $W_i \subseteq R_i$ as follows:

1. Define a set Ψ (relative to a given profile and rights system):

$$\Psi =_{df} \{\psi \in X : [\forall k \in \{1, \ldots, n\} : \psi \in A_k] \vee [\exists k \in \{1, \ldots, n\} : \psi \in (A_k \cap R_k)]\}$$

2. For any sentence $\varphi \in R_i$:

$$\varphi \in W_i \Leftrightarrow \exists \Psi' \subseteq \Psi : [\Psi' \text{ is consistent }] \wedge [\Psi' \cup \{\varphi\} \text{ is inconsistent}]$$

Ψ is the set of sentences endorsed by the community in the sense that they would be guaranteed to go into the collective judgement set by unanimity or minimal expert rights, should they hold. The waiver set of an individual i, W_i, is the set of claims that if i were to exercise her right over them, they would generate collective inconsistency when combined with a consistent subset of Ψ (if everyone else also exercised their rights over their fields of expertise).

This provides the following alternative in place of minimal expert rights:

ALIENABLE EXPERT RIGHTS For every profile, and every individual i, if i is an expert over φ, i accepts φ, and φ is not

waived, then φ is included in the scientific consensus, i.e. if $\varphi \in (A_i \cap R_i \setminus W_i)$, then $\varphi \in F(A_1, \ldots, A_n)$.

Notice that this condition is analogous to what Dietrich and List (2008, 64) call *positive decisiveness* which requires that an individual i is positively decisive over their rights set if and only if $(A_i \cap R_i) \subseteq F(A_1, \ldots, A_n)$, rather than *negative decisiveness*, which requires $F(A_1, \ldots, A_n) \subseteq (A_i \cap R_i)$[5] In the context under consideration we take the positive, rather than negative, aspect of this notion to capture the relevant sense of expertise: if an expert believes a sentence then, assuming it is not waived, it should be part of the scientific consensus. To briefly motivate this, consider the case in which an expert avoids taking a position on a sentence from their area of expertise (i.e. neither believes it nor its negation) – the sentence could correspond to an 'open problem' in their field – then through negative decisiveness that sentence could not become part of the scientific consensus. But it seems unintuitive to prevent scientists working in related fields from settling the question. However, the below result continues to hold when the condition is strengthened to demand that if the group accepts φ, and φ is in R_k for some k, then A_k (the analogue of negative decisiveness).

With this in place, we provide the following possibility result:

Theorem 2. *For any connected agenda and any rights system, there exists an aggregation function (generating consistent collective judgement sets) that satisfies universal domain, alienable expert rights, and the unanimity principle.*

5 We say 'analogous to' as the introduction of the waiver condition entails that we cannot use decisiveness in the sense of Dietrich and List (2008, 63-64), since in alienable expert rights, what experts are 'decisive over' varies from profile to profile. Also notice that Theorem 1 holds when minimal expert rights is weakened to positively minimal expert rights (Dietrich and List, 2008, 64).

Proof. The following function, F, respects universal domain, unanimity, and alienable expert rights by construction:

$$F(A_1, \ldots A_n) = \bigcap_i^n A_i \cup \bigcup_i^n (A_i \cap R_i \setminus W_i)$$

It remains to demonstrate that it is guaranteed to generate a consistent collective judgement set. We show this by induction on its subsets, for some arbitrary profile. To begin with, $\bigcap_i^n A_i$ is consistent by the consistency of every individual judgement set. Suppose then Ω is a subset of $\bigcup_i^n (A_i \cap R_i \setminus W_i)$ such that $\bigcap_i^n A_i \cup \Omega$ is consistent. The inductive step is that if $\bigcap_i^n A_i \cup \Omega$ is consistent then so is $\bigcap_i^n A_i \cup \Omega \cup \{\varphi\}$, for any $\varphi \in \bigcup_i^n (A_i \cap R_i \setminus W_i)$. Assume towards a contradiction that this is not the case, i.e. $\bigcap_i^n A_i \cup \Omega \cup \{\varphi\}$ is inconsistent, for some φ (1). Notice that $\bigcap_i^n A_i \cup \Omega$ is a consistent subset of Ψ (2). Since $\varphi \in \bigcup_i^n (A_i \cap R_i \setminus W_i)$, then there must exist a unique k (by the disjointness of the rights system), such that $\varphi \in (A_k \cap R_k \setminus W_k)$. But $(A_k \cap R_k \setminus W_k) \subseteq (A_k \cap R_k)$ and hence $\varphi \in (A_k \cap R_k)$ (3). From (1), (2), and (3): $\varphi \in W_k$ and $\varphi \notin \bigcup_i^n (A_i \cap R_i \setminus W_i)$. This contradicts our assumption and since our choice of φ was arbitrary, $\bigcap_i^n A_i \cup \Omega \cup \{\varphi\}$ is consistent for any $\varphi \in \bigcup_i^n (A_i \cap R_i \setminus W_i)$.[6] This concludes the proof that F always generates a consistent collective judgment set. □

In a fragmented discipline where the expert rights system and scientists' beliefs are such that respecting these rights (and unanimity) will yield an inconsistent result, the discipline has gone wrong somewhere. Even if the scientists' expert rights sets are disjoint, if there are enough logical connections between the sentences under consideration, then an inconsistency may be generated. Alienable expert rights suggests that whenever a discipline finds itself in

6 Recall our initial assumption that the agenda is finite. In consequence, the set $\bigcup_i^n (A_i \cap R_i \setminus W_i)$ will also be finite. If we were to relax our assumption, the proof would follow along the same lines, but we would have to make an additional assumption that the underlying logic is compact.

such a situation, every expert should waive their expertise over claims which contribute to the inconsistency (which needn't be all of them). And in doing so they contribute to what goes into the consensus on an equal footing. And this is how it should be. The discipline has made a mistake. Assuming unanimity, the expertise of some individuals cannot be respected whilst respecting the expertise of others. So, if scientists were to be intellectually modest, the extent of this expertise should be reconsidered.

Notice that the notion of alienable expert rights is quite demanding: it requires that *every* scientist who is an expert over *any* sentence contributing to the inconsistency waive their expertise over those sentences. Some may consider this to be too strong, after all, some scientists may not consider others to be their epistemic peers, and the former may balk at the idea of having to waive their expertise on account of the latter. One way of justifying this attitude would be to note that the evidential basis in some sub-disciplines may be objectively more secure than in others (e.g. should someone working in thermodynamics waive their expertise just because it conflicts with the beliefs of a string theorist?). What this highlights is that different sub-disciplines may place different demands on whether an individual should include a sentence in their belief set, and this may be salient when it comes to deciding who should waive their expertise (i.e. perhaps individuals with less secure evidence for their beliefs should waive before individuals with more secure evidence). Since our model does not include a parameter measuring strength of evidence, it is insensitive to such concerns; we simply assume that all scientists are epistemic peers. However, this is not to say that the model could not be further developed to take this into account, and moreover, we think there is an important distinction between the questions of how to justify the beliefs of an individual, and how to aggregate beliefs, once they have already

been taken as justified by everyone in the community. Our model only addresses the latter question.

Moreover, we want to point out here that we are not suggesting that F, as defined above, be the method scientists should use when forming scientific consensus. It is plausible that the collective belief sets it delivers will be rather limited in many cases. For any sentence φ that is not universally agreed upon, or in someone's area of expertise, φ won't be in the value of F. And we can imagine cases involving a group of extraordinarily timid experts who do not take any position on the sentences they are experts over, and moreover do not unanimously agree on any sentence in the agenda. In such a case, F will deliver the empty set. And this would indeed be too restrictive. Our claim is that any acceptable aggregation function should be such that its value for any profile be a superset of the value of F on that profile.[7] Another way of putting it is that we take universal domain, alienable expert rights, and unanimity to be necessary conditions on any acceptable aggregation function, but they needn't be sufficient. Exploring additional conditions strikes us as an potentially fruitful avenue for future research, but in this paper we are only interested in providing a possibility result.

5.5 AGAINST THE HEGEMONY OF EXPERTS

The upshot of requiring alienable expert rights of a judgement aggregation function is that the beliefs of all scientists will be respected. Every scientist can still contribute their beliefs to the aggregation function and none of them have to reconsider anything in order to successfully form an inconsistency-free consensus. This

7 Notice that if alienable expert rights is strengthened to include the analogue of negative decisiveness as discussed above, then this will place an additional constraint on any expansion of F.

is achieved by asking scientists to waive their expert status over those sentences which would otherwise lead to an inconsistent collective judgement set.

It is in this respect that Theorem 2 provides a more plausible possibility result than others suggested in the literature, mainly by Dietrich and List. They show that by restricting the domain to profiles that contain at least one *deferring* individual (where i is deferring in a profile $(A_1, ..., A_n)$ if and only if $A_i \cap R_j = A_j \cap R_j$ for every $j \neq i$), or at least one *agnostic* individual (where i is agnostic in a profile $(A_1, ..., A_n)$ if and only if A_i is consistent with every set of the form $B_1 \cup ... \cup B_{i-1} \cup B_{i+1} \cup ...B_n$ for each $j \neq i, B_j \subseteq R_j$, then a possibility result can be generated by setting the value of the aggregation function to the belief set of the deferring or agnostic individual (in the latter case combined with the other experts' beliefs regarding their areas of expertise).[8]

Comparing our approach to Dietrich and List's highlights what we take to be the take home message of our possibility result. The correct attitude when arriving at an inconsistent judgement set should not be venerating expertise at the expense of the beliefs of individual experts, but relinquishing it. Dietrich and List's results require fragmented disciplines where at least one scientist simply will not disagree with what the experts say about their respective areas of expertise. A problem with this is that if there does not exist such a scientist (which seems likely, just because a scientist is an expert over area X does not preclude them from having beliefs about areas related to X), we are not told how to shift into a profile where there is one. Furthermore, the most plausible way of doing

8 They also investigate a possibility result regarding severely restricted agendas and expert rights systems, but assuming the scientists are aiming to achieve a consensus which could contain logically connected sentences, the type of result List and Dietrich have in mind is implausible in this context.

this requires that at least one scientist change their mind in a fairly radical way. Such a scientist has to either defer to every other expert, or withhold judgement over any sentence in someone else's expert rights set, and there is no guarantee that such a scientist will be found. A further problem is that the way the possibility results are generated does suggest that scientists might be motivated to defer, or remain agnostic, since in doing so their belief set determines the value of the aggregation function (Dietrich and List, 2008, 74). Although they do not act as a dictator in the Arrovian sense (the same named individual determines the value of the function for any profile), they do determine whether or not any sentence not in anyone's rights set goes into the collective judgement set. And this is not a desirable result.

The possibility results considered in this paper suggests a trade off. On the one hand, we can respect the notion of expertise, which comes with the cost of demanding belief revision and generating a quasi-dictator. On the other, we can respect the beliefs of all scientists involved, but this comes with the cost that experts must be modest in the sense they they relinquish their expert status over troublesome sentences. Alienable rights provides a way of doing the latter, whilst still retaining the idea of expertise for all sentences which do not contribute to the inconsistency. No individual is forced to change her mind, and she may still contribute her entire belief set to the aggregation function, albeit without expert status in certain cases. It is for this reason that we take it to be the most plausible way of ridding scientific consensus of any inconsistencies. And it is for this reason that we think our result shows that the hegemony of the experts should not be saved at any cost.

5.6 CONCLUSION

Despite its apparent desirability, the fragmentation and specialisation of many, if not all, academic disciplines poses a problem. How are we to determine the state of the discipline as a whole? When the discipline in question is science, or a suitably fragmented subfield thereof, this concerns what our best scientists believe about the natural world. Under certain plausible assumptions, Dietrich and List's impossibility result shows that this cannot be answered without the threat of inconsistency. We proved that this can be avoided by replacing expert rights with alienable expert rights, when an expert right is alienable when it conflicts with expert beliefs and the beliefs unanimously accepted. We argued that this provides a more plausible way of avoiding inconsistency in the scientific consensus about the natural world than the idea of venerating expertise at the expense of respecting the beliefs of the individuals involved.

Part III

INDIVIDUAL RATIONALITY

6

INTRODUCTION

In this part of the thesis I turn my attention to several proposals for enriching the standard view on rationality beyond the mere satisfaction of the Kolmogorovian axioms.

In Chapter 7 I focus on Elga's (2004) restricted principle of indifference (RPI) for self-locating belief and in Chapter 8 on Titelbaum's (2013) defense of RPI. I show that both Elga's and Titelbaum's arguments in support of RPI fail for the same reason the argument for 'staying' in the *Monty Hall* problem fails. To wit, both Elga and Titelbaum fail to appropriately model the probabilistic scenarios their arguments rely on. In the appropriate sophisticated models of their scenarios Elga's and Titelbaum's conclusions no longer follow.

Elga's and Titelbaum's mistake in constructing the correct probability models for their scenarios is traced back to an old and often overlooked discussion by Glenn Shafer (1985) who argues that in order for Bayesian conditionalisation to be a good formal counterpart to learning, one ought to specify a partition of events which an agent can conditionalise upon. Luc Bovens and Jose Luis Ferreira (2010) interpret Shafer's insight as "[w]hen we are informed of some proposition, we do not only learn the proposition in question, but also that we have learned the proposition as one of the many propositions that we might have learned. The

information is generated by a protocol, which determines the various propositions that we might learn." This point can be found in the formal literature on bayesian epistemology such as in the works of Pearl (1988) and Halpern (2003), but is often ignored by philosophers. When we model the scenarios of Elga and Titelbaum in light of Shafer's insight, their arguments become significantly weaker.

Before proceeding it is important to clarify the status of RPI according to Elga (2004). Here is a statement of the principle:

RPI A rational agent ought to assign equal credence to worlds that agree on all uncentred propositions and are centred on agents whose experiences are indistinguishable.[1]

Initially, RPI was introduced by Elga (2000) as part of his argument for the Thirder answer to the Sleeping Beauty problem.

> SLEEPING BEAUTY Some researchers are going to put you to sleep. During the two days that your sleep will last, they will briefly wake you up either once or twice, depending on the toss of a fair coin (Heads: once; Tails: twice). After each waking, they will put you to back to sleep with a drug that makes you forget that waking. When you are first awakened, to what degree ought you believe that the outcome of the coin toss is Heads? (Elga, 2000, p. 143)

Trying to answer this question Elga reasons:

1 This is an elaboration of Elga (2004, p. 387)'s *Indifference*. The way I formulate this principle here: 1) brings to the fore that it is imposing a constraint on the credal state of any rational agent; and 2) elucidates the scope of the principle. See Weatherson (2005, p. 614) for a detailed discussion.

> "If (upon first awakening) you were to learn that the toss outcome is Tails, that would amount to your learning that you are in either T1 ["The coin came out Tails and it's Monday"] or T2 ["The coin came out Tails and it's Tuesday"]. Since being in T1 is subjectively just like being in T2, and since exactly the same propositions are true whether you are in T1 or T2, even *a highly restricted principle of indifference* yields that you ought then to have equal credence in each." (Elga, 2000, p. 144, my emphasis)

Elga only needs a further premise to the one in the above quote to be able to conclude that upon awakening, the probability of the coin having landed heads is $\frac{1}{3}$. Assume Beauty were to learn upon awakening that it is Monday. Then, she should divide her credence between a 'Monday and Heads' (call this M1) world and a 'Monday and Tails' equally (T1). After all, the coin was fair. But if M1 and T1 should receive equal probability and T1 and T2 (upon the above argument) should also receive equal probability, it means that all three worlds are equally likely. Since these jointly form the entire sample space, then Beauty should assign $\frac{1}{3}$ to each.

Therefore RPI appears to be essential to Elga's argument for the Thirder answer to the Sleeping Beauty problem. Moreover, one of the early defenders of Elga's Thirder answer, i.e. Dorr (2002, p. 294), also makes RPI an integral part of his argument. This suggests that if RPI is shown to fail, this can affect at least some of the best known Thirder arguments.

But what is the status of RPI, according to Elga? Elga might take RPI to merely be rationally permissible. At least in certain cases (and perhaps Sleeping Beauty is among them), an agent can align her subjective credences according to RPI and assign equal credence

to worlds that agree on all uncentred propositions and are centred on agents whose experiences are indistinguishable (e.g. to being Monday or Tuesday). Alternatively, Elga might endorse the view that RPI is rationally mandated and that any rational agent ought to align her credences according to RPI. I contend Elga (2004) offers an *argument* for RPI and that Elga (at least in that paper) endorses the latter view: RPI is rationally mandated. Here is what Elga says:

> My defense of [RPI] has two parts. I'll describe a basic case involving an agent who divides his credence between a pair of similar centered worlds, and argue that he ought to assign those worlds equal credence. And I'll say how that argument can be generalized. But since the controversial parts of the argument arise even in the basic case, I relegate the generalization to the Appendix. (Elga, 2004, p. 388)

Moreover, Titelbaum writes:

> Elga defends [RPI] by defending this result for *Duplication*.[2] It's fairly easy to generalize his arguments about *Duplication* to analogous situations with arbitrary (finite) numbers of duplicates, and Elga does so in an appendix. The generalization is uncontroversial, so we'll confine our attention to Elga's argument concerning the original *Duplication* case (Titelbaum, 2013, p. 253) . . . I have three objections to Elga's argument for [RPI] (Titelbaum, 2013, p. 254)

In other words, Elga (2004) isn't an illustration of the consequences of adopting RPI but an argument in defense of it. Therefore, RPI is taken by Elga to be a proper enrichment of our theory of rationality.

2 See the next chapter for a discussion of this puzzle.

Finally, in Chapters 9 and 10 extend the above discussion by showing how the the formal mistake Elga and Titelbaum make when constructing the probability models for their scenarios is more widespread and I highlight it in Christensen's (2010) argument against an intuitive reflection principle and in Mahtani's (forthcoming) discussion of the opaque proposition principle. These last two chapters also engage with two proposals for extending our understanding of rationality, however, they are primarily meant as a further illustration of the importance of constructing a precise probabilistic model when analysing probabilistic scenarios. To sum up, in appropriate sophisticated models, the conclusions of both Christensen and Mahtani no longer go through.

7

ELGA'S RESTRICTED PRINCIPLE OF INDIFFERENCE

The Restricted Principle of Indifference (RPI) states that "similar centred worlds deserve equal credence" (Elga, 2004, p. 387). I understand this principle as making the following claim:

RPI A rational agent ought to assign equal credence to worlds that agree on all uncentred propositions and are centred on agents whose experiences are indistinguishable.

RPI is restricted in two ways. Firstly, it is more restricted than usual principles of indifference from the philosophy of probability literature as it only applies to centred worlds. Secondly, it only applies to centred worlds that agree on all uncentred propositions. In other words, RPI wouldn't apply in the following case: suppose W is the actual world centred on you and W' is a Matrix-like world in which one of the people connected to machines has the exact same subjective experiences you have in W. RPI does not recommend assigning equal credence to W and W'. Elga labels a principle of indifference that would apply in such a situation the "absurd claim that I don't endorse." (Elga, 2004, p. 387) Elga's argument for why he wouldn't endorse such a claim is that if RPI would cover cases in which the uncertainty spans over individuals located in different worlds, then RPI would collapse into a full-blown principle of indifference and would no longer be resricted:

> let AT be the actual world, centered on you, now. Let VAT be a world centered on a brain in a vat who is in a state subjectively indistinguishable from yours. ABSURD-CLAIM-THATI-DON'T-ENDORSE entails that you ought to assign AT and VAT equal credence. (Elga, 2004, p. 387-8)

As the above discussion shows, there are meaningful senses of self-locating indifference that don't seem to fall under RPI and one can find even more in Weatherson (2005). This may be a problem as even if we establish RPI, we are still falling short of accomplishing the goal of offering a comprehensive principle of indifference for self-locating belief. Nevertheless, the focus of this chapter is on RPI alone and not on any possible variations of it.

In this chapter I show that Elga's argument for RPI is not valid. In Section 1, I begin by gauging the relevance RPI has in the literature on self-locating beliefs and I argue that RPI is not only a common place, but that moreover it plays an integral part in many argumentative strategies. I then introduce Elga's argument in favour of RPI (Section 2). In Section 3 I present the *Monty Hall* problem and explain what lessons we can draw from it for one of the scenarios instrumental to Elga's argument (Section 4). I conclude that Elga's argument is not valid (Section 5).

rmation accrues toThe following chapter will prove that the only other justification of RPI, viz. Titelbaum's (2013), relies on the same fallacy. I end by remarking that despite its widespread use, to date we have no reason for accepting RPI.

7.1 THE RELEVANCE OF RPI

In this section I attempt to gauge the relevance RPI has in the literature on self-locating beliefs. I show that RPI plays an integral part in many argumentative strategies even if all papers mentioned in this section (and many others I leave out due to brevity considerations) do not attempt to justify the principle and instead direct the reader to Elga (2004).

Firstly, it isn't just Thirders that rely on RPI. Some Halfers also rely on RPI in their arguments. For instance, Lewis (2001) writes:

> Elga writes, "Since being in T1 is subjectively just like being in T2, and since exactly the same propositions are true whether you are in T1 or T2, even a highly restricted principle of indifference yields that you ought then to have equal credence in each" (144). By 'proposition' he means an uncentred possibility. The reason the same propositions are true whether Beauty is in T1 or T2 is that the centred worlds that are members of T1 are collocated with the corresponding members of T2. I accept Elga's 'highly restricted principle of indifference'. So we have a further point of agreement. (p. 172)

This is not a trivial point of agreement. Lewis acknowledges that in order to derive his solution to the *Sleeping Beauty*, viz. what Lewis calls proposition L2 (Lewis, 2001, p. 174), he requires the following to hold (I preserve Lewis's numbering and his notation where $P_{\_}(\cdot)$ is Beauty's credence before being put to sleep and $P(\cdot)$ her credence upon waking up): (1) $P(T_1) = P(T_2)$, (5) $P_{\_}(HEADS) = 1/2 = P_{\_}(TAILS)$ and (6) Beauty gains no new uncentred evidence relevant to HEADS vs. TAILS between the time she has credence function $P_{\_}(\cdot)$ and the time when she has credence function $P(\cdot)$. (Lewis, 2001, p. 173) Proposition (1) is

Elga's RPI and without it Lewis's conclusion wouldn't follow.

Secondly, Ross (2010) argues Thirders in the Sleeping Beauty problem are committed to what he calls a 'rational dilemma'. An agent is faced with a rational dilemma if they find two (or more) principles that conflict with one another equally plausible and hence are forced to continue accepting them both. The rational dilemma thirders are faced with is between a Generalized Thirder Principle and the principle of Countable Additivity. For Ross's argument to go through, however, he has to argue that dropping any of these two principles is not a plausible way of defusing the tension between them. The part of the argument we are interested in for the purposes of this chapter is why we cannot drop the Generalized Thirder Principle, and more specifically, why the fact that the principle relies on an indifference principle is not enough to warrant its rejection. Ross writes

> Since several of the arguments for the [Generalized Thirder Principle] appeal, implicitly or explicitly, to a finitistic principle of indifference, one could reject these arguments so long as one denies that such indifference principles apply even in finitistic cases. One might claim instead that when Beauty awakens, she could rationally have more credence in *Tails and Monday* than in *Tails and Tuesday* , or viceversa, or else one might claim that she should not have precise credences in these possibilities at all. (...) These moves, however, involve considerable costs. (Ross, 2010, p. 443)

To explain the costs of rejecting the 'finististic principle of indifference' Ross simply points the reader to Elga (2004) (see fn. 28). This is an implicit acknowledgement of the validity of Elga's argument.

Finally, Leitgeb and Bradley (2006) argue that there are cases in which betting odds and rational beliefs come apart. They present the following scenario modelled after the Sleeping Beauty problem:

> Suppose that if the coin lands Tails, you will be offered two real bets on Heads (of the same flip), one after the other. There is no funny business here. But if the coin lands Heads, you will be offered a real bet on Heads and you will also hallucinate being offered a bet on Heads. You won't know whether the hallucination occurs at the first stage or the second stage. You do know that one of the bets will be real and one will be a hallucination. Whether or not you accept the hallucinatory bet, you will later wake up and find your wallet untouched. (Leitgeb and Bradley, 2006, pp. 123-4)

In this example, the agent thinks she is offered four bets (out of which only three are real). Let's assume the stake is s. If the coin comes up Tails the outcome of the first bet (a) is $-s$, as the coin has actually landed Tails and not Heads. By the same reasoning, the outcome of the second bet (b) is also $-s$. If, however, the coin lands Heads, then the real bet (c) will have an outcome of s, whereas the hallucinatory one (d) will have an outcome of 0 (nothing is gambled, nothing is won). To get more precision in the calculation, Leitgeb and Bradley invoke Elga's principle:

> How do we divide these probabilities up further between (a) and (b) (and (c) and (d))? Using Elga's (2004) Restricted Principle of Indifference, accepted by all concerned including Hitchcock (and implicitly used earlier), each of these 4 possibilities gets a probability of 25% (p. 124)

This is not a totally innocous use of RPI. If instead of being $\frac{1}{2}$, the probability of 'Stage 1 and Heads' plus the probability of 'Stage 1

and Tails' is in fact 1, the expected utility of the bet becomes 0, and Beauty would no longer be Dutch-bookable. A natural response to this would be that any other probability assignment except for 1 would support Leitgeb and Bradley's conclusion, and there seems to be something irrational about such a probability assignment. This may be so, but without RIP, Leitgeb and Bradley requires a further premise to their argument which is not easy to motivate (even if it feels very intuitive). Moreover, anticipating our argument later in the next chapter, if one were to claim that in such cases as those invoked by Leitgeb and Bradley, the events 'Heads and Stage 1' and 'Heads and Stage 2' (and likewise for the events 'Tails and Stage 1' and 'Tails and Stage 2') are actually non-measurable, the argument put forward by Leitgeb and Bradley would have to be more substantially revised.

What this section purports to show is that Elga's argument for RPI plays an integral role in many arguments in the literature on self-locating belief. If it is shown that the argument is fallacious this, in effect, affects 1) Elga's and Dorr's arguments for the Thirder answer to *Sleeping Beauty*, 2) Lewis's Halfer answer to the same problem, 3) Ross's argument that there are such things as rational dilemmas, and finally 4) Leitgeb and Bradley's argument that sometimes betting odds and credences come apart. Can these arguments be saved even if RPI fails? Perhaps. But I shall not attempt to do so in this thesis. My focus, instead, will be on the validity of Elga's and Titelbaum's arguments to which I dedicate the following two chapters.

7.2 DR. EVIL AND RPI

Elga's defense of RPI proceeds from the story of *Dr. Evil*. Weatherson summarizes Elga's argumentative strategy as follows: "Elga argues for [RPI] by arguing it holds in a special case, and then ar-

guing that the special case is effectively arbitrary, so if it holds there it holds everywhere." (Weatherson, 2005, p. 627). In this section I present Elga's argument as it unfolds in his paper; later in this chapter I offer a more structured reconstruction of Elga's argument.

> DR. EVIL Safe in an impregnable battlestation on the moon, Dr. Evil had planned to launch a bomb that would destroy the Earth. In response, the Philosophy Defense Force (PDF) sent Dr. Evil the following message: 'Dear Sir, (...) We have just created a duplicate of Dr. Evil. The duplicate - call him "Dup" - is inhabiting a replica of Dr. Evil's battlestation that we have installed in our skepticism lab. At each moment Dup has experiences indistinguishable from those of Dr. Evil. For example, at this moment both Dr. Evil and Dup are reading this message. We are in control of Dup's environment. If in the next ten minutes Dup performs actions that correspond to deactivating the battlestation and surrendering, we will treat him well. Otherwise we will torture him. Best regards, The PDF' (Elga, 2004, p. 383)

Elga argues that upon receiving this message, Dr. Evil should assign equal credence to being himself and to being Dup. In this section I rationally reconstruct Elga's argument, while in the remaining sections in this chapter I show that it relies on the kind of mistaken reasoning that recommends the 'staying' strategy in the *Monty Hall* problem.

Consider the following variation of *Dr. Evil*:

> COMATOSE DR. EVIL Just like *Dr. Evil*, only that the scientists tell Dr. Evil they will flip a coin with bias .9 towards Tails and that the laws of nature do not allow

> for two subjectively indistinguishable consciousnesses to exist at the same time. If the coin lands Heads, only Evil wakes up. If the coin lands Tails, only Dup wakes up.[1]

If the coin lands Heads, Dr. Evil is reading the message from PDF. If the coin lands Tails, Dr. Evil is in a coma on the Moon and Dup is reading the message back in the skepticism lab. In *Comatose Dr. Evil*, Elga argues Dr. Evil ought to align his credence that he is Dr. Evil to the bias of the coin. In other words, upon reading the message, Dr. Evil's degree of belief in being Dr. Evil ought to be .1.[2]

Consider now another variation of *Dr. Evil*:

> COIN TOSS DR. EVIL Just like *Dr. Evil*, only that the scientists tell Dr. Evil that while they were duplicating him they flipped a coin with bias .9 towards Tails. But they assure him the coin toss had no impact on the duplication process.

1 This is a variation of *Coma* in Elga (2004, pp. 390-1). Elga in fact moves away from *Dr. Evil* and develops his entire argument for RPI based on a completely analogous set of scenarios involving Al and his Duplicate. Nevertheless, there is no need to do that, and I will present his reasoning as it applies to *Dr. Evil*.

2 Those familiar with Elga (2000)'s discussion of the *Sleeping Beauty* problem may find surprising what he says about Dr. Evil's degrees of belief in *Comatose Dr. Evil*. Such a view goes against the Thirder answer to the *Sleeping Beauty* problem. Titelbaum (2013) has already noticed this tension:

> it was Elga himself who originally argued for the 1/3 answer to the Sleeping Beauty Problem, an answer that is incompatible with the Relevance-Limiting Thesis's position on the irrelevance of centered evidence to uncentered propositions. A thirder about Sleeping Beauty can't just assume that [Dr. Evil] should assign a degree of belief of 0.10 to heads when he awakens in [Comatose Dr. Evil]! (Titelbaum, 2013, 353)

In *Coin Toss Dr. Evil*, upon receiving the message from PDF, Dr. Evil should assign probability .1 to the coin having landed Heads (H, and T for Tails). Secondly, since the coin toss is independent from the duplication process, he should assign the same probability conditional on him being Dr. Evil (E, and D for Dup). That is:

$$P(H) = .1 \tag{1}$$

$$P(H|E) = .1 \tag{2}$$

Suppose in *Coin Toss Dr. Evil*, PDF were to send Dr. Evil a second message saying that if the coin landed Heads then Dup fell in a coma and Dr. Evil is now reading the message and if the coin landed Tails, Dup is reading the message and Dr. Evil is in a coma on the Moon, that is $(H\&E) \vee (T\&D)$. Elga argues that in such case, Dr. Evil's credal state in *Coin Toss Dr. Evil* upon reading the second message should align with his credal state in *Comatose Dr. Evil* upon reading the message of that scenario.[3] In other words,

$$P(H|(H\&E) \vee (T\&D)) = .1 \tag{3}$$

(1)-(3) are enough to derive Dr. Evil's degree of belief in being Dr. Evil in *Coin Toss Dr. Evil* after being told about the coin toss but before being announced that $(H\&E) \vee (T\&D)$:

$$\begin{aligned}
\text{From (2) and (3)} &: P(H|(H\&E) \vee (T\&D)) = P(H|E) \\
\text{By def. of cond. prob.} &: \frac{P(H\&E)}{P(H\&E) + P(T\&D)} = \frac{P(H\&E)}{P(H\&E) + P(T\&E)} \\
\text{By simplification} &: P(T\&D) = P(T\&E) \\
\text{By independence} &: P(T)P(D) = P(T)P(E) \\
\text{By simplification} &: P(D) = P(E)
\end{aligned}$$

3 "So when [Evil] wakes up in the [*Comatose Dr. Evil*] case, he has just the evidence about the coin toss as he would have if he had been awakened in [*Coin Toss Dr. Evil*] and then been told [$(H\&E) \vee (T\&D)$]." (Elga, 2004, p. 391)

Therefore, Dr. Evil should assign equal credence to being Dr. Evil and to being Dup in *Coin Toss Dr. Evil*, after being told about the duplication, but before being told that $(H\&E) \vee (T\&D)$. Since the coin toss in *Coin Toss Dr. Evil* has no causal impact on the duplication process, Dr. Evil's credal state after being told about the duplication and the coin toss (but before receiving the second message) is the same as his credal state in *Dr. Evil* upon simply being told he had been duplicated. It is true that in *Coin Toss Dr. Evil* the scientists tell Dr. Evil more than in *Dr. Evil*, but that additional information has no bearing on whether he is Dup or Dr. Evil. Therefore, in *Dr. Evil* he should divide his credence equally between being Dr. Evil and being Dup upon receiving the message about duplication from the scientists.

Finally, *Dr. Evil* is taken by Elga to be a prototypical example of a rational agent contemplating worlds that agree on all uncentred propositions and are centred on agents whose experiences are indistinguishable. Consequently the move from 'a rational agent' to *Dr. Evil* is done without loss of generality. That means that whatever rational requirements bind Dr. Evil's credal state, they ought to bind, on pain of irrationality, any agent. In particular, if Dr. Evil is rationally required to assign equal credence to the centred worlds he is contemplating, so should any agent. Since the above argument establishes, according to Elga, that Dr. Evil should indeed be indifferent between the world centred on Dr. Evil and the one centred on Dup, so should any other rational agent, and RPI follows.

In the remainder of this chapter I show that the problem with Elga's argument for RIP lies in the formal modelling of *Dr. Evil* and its accompanying variations. In particular, Elga's reasoning makes an implicit assumption about the probability with which information is passed on to Dr. Evil and I prove that if that information is explicitly

introduced in the formal model, Elga's conclusion no longer follows. This is in effect the lesson we learnt via Shafer (1985); Bovens and Ferreira (2010); Halpern (2004) from *Monty Hall*. The following section explains why this is so.

7.3 THE MONTY HALL PROBLEM

At the same time PDF is trying to thwart Dr. Evil's plans, on some TV set Monty Hall attempts to trick a contestant into making the losing choice in a game show:

> MONTY HALL Monty presents a game contestant with three doors. Behind two of these doors there is a goat. One of the doors, however, hides a brand new car. The contestant is asked to pick a door. Monty then opens one of the other two doors such that he doesn't give the prize away. Afterwards he asks the contestant which door she wants to open - the one she initially chose, or the other remaining closed door.

Suppose the door behind which the car is hidden is chosen at random. Suppose further that the contestant first picks Door 1. Monty hopes the contestant will reason in the following way:

> 'initially, there was a $\frac{1}{3}$ chance the car was behind Door 1. Now that Monty opened one door hiding a goat, there are only two possible locations the car could be in, i.e. behind Door 1 or behind the door Monty left unopened. Therefore the probability the car is behind the Door 1 selected increased to $\frac{1}{2}$ and the probability the car is behind

> the other unopened door is also $\frac{1}{2}$. So there is no reason for me to switch.'[4]

As it is well known, however, this reasoning is incorrect. Bovens and Ferreira (2010, pp. 474-6), following Sneed (1985)'s discussion of Shafer (1985), explain the mistake in terms of the fact that when we are informed of some proposition "we do not only learn the proposition in question, but also that we have learned the proposition as one of the many propositions that we might have learned." (Bovens and Ferreira, 2010, p. 474) The difference between updating on some proposition rather than updating on learning that proposition is nicely highlighted by Halpern, following an example first introduced by (Howson, 1995, p. 9). I discuss this example in detail in a later chapter.

> If I think my wife is much more clever than I, then I might be convinced that I will never learn of her infidelity should she be unfaithful. So, my conditional probability for Y, 'I will learn that my wife is cheating on me', given X, 'She will cheat on me', is very low. Yet, the probability of Y if I actually learn X is clearly 1. (Halpern, 2004, pp. 128-9)

Applying this insight to *Monty Hall* Bovens and Ferreira explain the contestant's mistaken reasoning by the fact that she updated only on the content of the information Monty gave her when he opened Door 2 and revealed a goat. If she instead were to consider how the information Monty can pass on to her is constrained she would notice that the probability Monty would open a particular door is not the same irrespective of the state of world.

4 This last step is usually assumed in the discussions of this puzzle, but notice it relies on a type of conservativism: a rational agent should not revise her strategy, unless she has a positive reason to do so.

This is easy to see: assume the car is behind Door 3, then the goats are behind Doors 1 and 2. Monty cannot open the former, as this is the one the contestant chose at the beginning of the round. Therefore Monty is forced to open Door 2. An analogous reasoning applies if the car is behind Door 2. But if the car is behind Door 1, then Monty can open either Door 2 or Door 3. So the probability with which he would open Door 2, say, in this case can be lower than 1. This asymmetry in how Monty can communicate with the contestant is made clear by considering the protocol under which information can accrue to the contestant. A conditional probability table, such as Table 8 can be used to specify a protocol.

	Car 1	Car 2	Car 3
"Goat 2"	$\frac{1}{2}$	0	1
"Goat 3"	$\frac{1}{2}$	1	0

Table 8: Protocol 1 for Monty Hall

In this table, each row corresponds to possible item of information the contestant could receive. Each cell corresponds to the probability the contestant will learn that item of information at each possible world. This table can be used to construct a sophisticated event space in which we take into consideration every piece of information the agent could receive. Such a space would contain four atomic events with non-zero probability: (Car 1, "Goat 2"), (Car 1, "Goat 3"), (Car 2, "Goat 3"), (Car 3, "Goat 2"). We can now calculate how the contestant should change her degrees of belief upon Monty opening Door 2, say, and revealing a goat.

$$
\begin{aligned}
&P(\text{Car 1}|\text{"Goat 2"}) \\
&= \frac{P(\text{"Goat 2"}|\text{Car 1})P(\text{Car 1})}{P(\text{"Goat 2"}|\text{Car 1})P(\text{Car 1}) + P(\text{"Goat 2"}|\text{Car 2})P(\text{Car 2}) + P(\text{"Goat 2"}|\text{Car 3})P(\text{Car 3})} \\
&= \frac{\frac{1}{2} \times \frac{1}{3}}{(\frac{1}{2} \times \frac{1}{3}) + (0 \times \frac{1}{3}) + (1 \times \frac{1}{3})} \\
&= \frac{1}{3}
\end{aligned}
$$

Therefore, taking into account the asymmetry of the way in which information may accrue to her, the contestant learns something new about where the car may be. Is this the only protocol that would make sense in Monty Hall? Although the puzzle is quite detailed with respect to how information is being delivered to the contestant, the scenario does not say Monty flips a fair coin in order to choose which door to open when the car is behind Door 1. Another protocol compatible with the story would be:

	Car 1	Car 2	Car 3
"Goat 2"	$\frac{3}{4}$	0	1
"Goat 3"	$\frac{1}{4}$	1	0

Table 9: Protocol 2 for Monty Hall

This represents a situation in which Monty would have a preference for opening Door 2 when the car is behind Door 1 and the contestant chooses Door 1 at the beginning of the game. In this case, upon learning that Door 2 hides a goat, the contestant's credence in the car being behind Door 1 should go from $\frac{1}{3}$ to $\frac{3}{7}$.

So what is the correct answer then: $\frac{1}{3}$ or $\frac{3}{7}$? The analysis of *Monty Hall* in terms of protocols shows that the answer to this question is sensitive to what the structure of the interaction between Monty and the contestant is. The puzzle is silent on some of the details and hence we cannot talk of a 'right' answer to this question.

Are we rationally required, though, to take protocols into account? Bovens and Ferreira (2010, p. 480) and Shafer (1985, p. 264) claim it is implicit in the Principle of Total Evidence that an agent's probability model should give probabilities for all the different ways her learning may turn out. The motivation for this goes back to the observation that when you receive some information Y, you don't only

learn the propositional content of Y, but also that you have received Y instead of Y′. So, insofar as Y represents your evidence, so does the fact that you learned Y instead of Y′. Therefore if you take the Principle of Total Evidence as a requirement for forming rational beliefs, then protocols should be taken into account.

7.4 THE PROTOCOL OF COIN TOSS DR. EVIL

So what would a protocol for *Coin Toss Dr. Evil* look like? There are four possible states of the world: either the person reading the message from PDF is Dr. Evil or he is Dup; and either the coin landed Heads or it landed Tails. Then Elga tells Dr. Evil the PDF could send a second message saying "$(H\&E) \vee (T\&D)$". We don't know anything else about what other information the scientists could include in that second message. Consider the following protocol, where a and b are arbitrary parameters:

	$H\&E$	$T\&E$	$H\&D$	$T\&D$
"$(H\&E) \vee (T\&D)$"	a	o	o	d
⋮	⋮	⋮	⋮	⋮

Table 10: Protocol 1 for Coin Toss Dr. Evil

Given this protocol, the probability of the coin having landed Heads given the scientists' message is:

$$
\begin{aligned}
&P(H|''(H\&E) \vee (T\&D)'') = \\
&= \frac{P(''(H\&E) \vee (T\&D)''|H)P(H)}{P(''(H\&E) \vee (T\&D)''|(H\&E))P(H\&E) + P(''(H\&E) \vee (T\&D)''|(T\&E))P(T\&E) + \\
\qquad &(''(H\&E) \vee (T\&D)''|(H\\&D))P(H\&D) + P(''(H\&E) \vee (T\&D)''|(T\&D))P(T\&D)} \\
&= \frac{P(''(H\&E) \vee (T\&D)''|(H\&E) \vee (H\&D))P((H\&E) \vee (H\&D))}{aP(H\&E) + bP(T\&D)} \\
&= \frac{\frac{P(''(H\&E) \vee (T\&D)'' \& ((H\&E) \vee (H\&D)))P((H\&E) \vee (H\&D))}{P((H\&E) \vee (H\&D))}}{aP(H\&E) + bP(T\&D)} \\
&= \frac{P((''(H\&E) \vee (T\&D)'' \& (H\&E)) \vee (''(H\&E) \vee (T\&D)'' \& (H\&D)))}{aP(H\&E) + bP(T\&D)} \\
&= \frac{P((''(H\&E) \vee (T\&D)'' \& (H\&E))) + P((''(H\&E) \vee (T\&D)'' \& (H\&D)))}{aP(H\&E) + bP(T\&D)} \\
&= \frac{P((''(H\&E) \vee (T\&D)'' \& (H\&E))}{aP(H\&E) + bP(T\&D)} \\
&= \frac{P(''(H\&E) \vee (T\&D)''|(H\&E))P(H\&E)}{aP(H\&E) + bP(T\&D)} \\
&= \frac{aP(H)P(E)}{aP(H)P(E) + bP(T)P(D)} \\
&= \frac{aP(E)}{aP(E) + 9bP(D)}
\end{aligned}
$$

Elga claims that $P(H|''(H\&E) \vee (T\&D)'')$ should be equal to the probability of Heads, that is .1. Solving the equation

$$\frac{aP(E)}{aP(E) + 9bP(D)} = .1,$$

we obtain that

$$aP(E) = bP(D).$$

Therefore (assuming there are no extreme values) the probability of being Dr. Evil is equal to the probability of being Dup if and only if a=b. In other words, the agent should consider it equally likely to be told $''(H\&E) \vee (T\&D)''$ in a Heads world in which he is Dr. Evil as in a Tails world in which he is Dup. This is by no means certain. One could easily assume something like the protocol in Table 11

	$H\&E$	$T\&E$	$H\&D$	$T\&D$
"$(H\&E) \vee (T\&D)$"	$\frac{1}{2}$	0	0	1
"$(H\&D) \vee (T\&E)$"	0	$\frac{1}{2}$	$\frac{1}{2}$	0
"$(H\&E) \vee (T\&E) \vee (H\&D)$"	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0

Table 11: Protocol 2 for Coin Toss Dr. Evil

underwrites the exchange in *Coin Toss Dr. Evil*.

In this case, PDF can send three messages to Dr. Evil and they have different likelihoods based on the possible world that obtains. The relevant asymmetry is that the scientists will definitely announce "$(H\&E) \vee (T\&D)$" whenever (T&D) obtains but will only announce it with probability $\frac{1}{2}$ when (H&E) obtains. In this case, even if one were to accept Elga's claim that $P(H|"(H\&E) \vee (T\&D)") = .1$, then $P(E) = 2P(D)$, which means that the agent would consider it twice more likely to be Dr. Evil than Dup.

Nevertheless, the assumption that $a = b$ is not incompatible with Elga's *Coin Toss Dr. Evil* (the scenario underdetermines the different messages PDF could send to Dr. Evil). So prima facie it may seem that Elga's argument simply requires an additional innocuous assumption about the protocol underlying *Coin Toss Dr. Evil* for the conclusion that $p(E) = p(D)$ to go through.

7.5 ELGA'S ARGUMENT, CAREFULLY

The fact that the conditional probability of the scientists' announcement in (H&E) has to be equal to the conditional probability in (T&D) spells trouble for Elga's argument for RPI. Recall, Elga's argumentative strategy:

CLAIM A Dr. Evil's credal state after receiving the message from PDF in *Comatose Dr. Evil* is identical to his credal state in *Coin Toss Dr. Evil* after being told he has been duplicated and learning "$(H\&E) \vee (T\&D)$".

CLAIM B Therefore Dr. Evil should assign equal credences to being Dr. Evil and being Dup upon being told he has been duplicated in *Coin Toss Dr. Evil* (and before receiving the second message).[5]

CLAIM C But upon learning he has been duplicated in *Coin Toss Dr. Evil* (and before the receiving the second message), his credal state is identical to his credal state in *Dr. Evil* (modulo the irrelevant difference that he now knows a coin toss independent of his duplication has been flipped).

CLAIM D Therefore, in *Dr. Evil*, he should assign equal credences to being Dr. Evil and being Dup.

CLAIM E Given *Dr. Evil* is a prototypical scenario for the restricted principle of indifference for self-locating beliefs, RPI holds.

Consider Claim A. The argument in the previous section establishes that Claim A only holds if a particular restriction is placed on the protocol under which information is passed to Dr.Evil/Dup by the scientists. Not all possible learning scenarios will support Claim A. So if Claim A were to hold, then the scientists should follow a protocol compatible with $a = d$, and both Dr. Evil and Dup should be aware of this protocol.

One could reply to this that there is nothing preventing us from tweaking *Coin Toss Dr. Evil* in order to account for this protocol. Assume we come up with a story that makes the receipt

5 This follows from (1)-(3), above.

of the second message from PDF equally likely in (H&E) as in (T&D). Let's call this new scenario *Coin Toss Dr Evil*$^+$. Claim A only holds if we replace *Coin Toss Dr. Evil* with *Coin Toss Dr. Evil*$^+$.

Now, consider Claim C. If Claim C were to hold, then the same knowledge of the protocol which is now embedded into *Coin Toss Dr. Evil*$^+$ should obtain in *Dr. Evil*. Dr. Evil should be aware that the scientists could flip a fair coin independently of the duplication process, and if they do flip it, they could announce that the coin came up heads to Dr. Evil or that the coin came up tails to Dup. Finally, Dr. Evil should also be aware that it is as likely for them to announce this if he indeed is Dr. Evil and the coin came up heads as it is if he is in fact Dup and the coin came up tails. So, for Claim C to hold *Dr. Evil* has to be replaced with *Dr. Evil*$^+$.

However, Dr. Evil's credal state in *Dr. Evil*$^+$ contains this protocol and hence his credal state is no longer a prototypical credal state of an agent faced with worlds that agree on all uncentred propositions and are centred on agents whose experiences are indistinguishable. Hence, *Dr. Evil* can no longer serve as the instantiation of an arbitrary rational agent as the move from a general instance of RPI to *Dr. Evil*$^+$ cannot be done without loss of generality. To wit, Dr. Evil assigns a credence of $\frac{1}{2}$ to being Dr. Evil not in a prototypical case of RPI, but in a case in which information accrues to him according to a particular protocol. In consequence, the step back from *Dr. Evil* to establishing RPI is no longer warranted.

To sum up, either Elga's argument fails at the very outset when credences from *Comatose Dr. Evil* are imported to *Coin Toss Dr. Evil*, or at the last step when Dr. Evil's credences cannot be attributed to an arbitrary rational agent dealing with worlds agreeing on all uncentred propositions and centred on agents whose experiences

are indistinguishable.

Before concluding, here is another way of making the same argument as in the above pages. Suppose Elga's argument is correct and hence:

> In [*Coin Toss Dr. Evil*], the coin toss is irrelevant to whether and how the duplication occurs. So [Evil]'s state of opinion (when he awakens) as to whether he is [Evil] or the duplicate ought to be the same in [*Coin Toss Dr. Evil*] as it is in [*Dr. Evil*] (Elga, 2004, p. 388)

Consider now a variation of *Coin Toss Dr. Evil* in which it is made clear that Protocol 2 underwrites the informational exchange between PDF and Evil and Evil knows this. In such a scenario the toss of the coin would also be "irrelevant to whether and how the duplication occurs". Therefore, by Elga's reasoning, Evil's credence function in *Dr. Evil* ought to match his credence function in this modified scenario, too. But as we saw above, with Protocol 2 in place, $P(E) = 2P(D)$. Consequently in *Dr. Evil*, Evil ought to believe both that the probability of being himself is equal to that of being Dup and equal to $1/2$, and that it is twice the probability of being Dup. This would make Evil probabilistically incoherent.

7.6 CONCLUSION

In this chapter I show that Elga's argument for RPI fails. This failure is interesting for two reasons. Firstly, the restricted principle of indifference is part of both the halfer and thirder answers to the Sleeping Beauty problem as well as part and parcel of several arguments in the literature on self-location. This is so despite the fact that Weatherson (2005) already provides a criticism to Elga. His main argument, though, attacks the way Dr. Evil's degree of

belief in Heads in *Comatose Dr. Evil* is determined. This amounts to challenging the soundness of Elga's argument – is the first premise that Dr. Evil should assign the same probability to Heads before and after awakening in *Comatose Dr. Evil* true? This is a substantial question and one that is reminiscent of *Sleeping Beauty*. As Titelbaum already noticed (see fn. 3), answering this question in the positive assumes a halfer strategy. Weatherson disagrees with this, but "think[s] the rest of Elga's argument is right" (Weatherson, 2005, p. 628). In this paper I show that irrespective of how this question is answered, Elga's argument is not valid and thus fails no matter what one's intuitions about the halfer/thirder debate may be. So in this respect this paper shows that there is a need for a new grounding of a principle of indifference for self-locating beliefs if we are to have one at all. The next chapter will look at Titelbaum's attempt to save Elga's argument and can be construed as an attempt to offer a solid grounding to RPI (although Titelbaum stops short of claiming this).

Secondly, the mistake in Elga's argument is in itself interesting, as it illustrates the need for specifying a precise sample space when applying conditionalization. In this respect, this chapter shows that *Monty Hall* still has important lessons to teach us. The following chapters in this part of the thesis will make this point over and over again.

8

TECHNICOLOR EVIL AND THE MONTY HALL PROBLEM REDUX

In the previous chapter we saw that if the informational context in which the scientists make the announcements to Dr. Evil is explicitly modelled, Elga's argument for the Restricted Principle of Indifference (RPI) fails. In a recent book, Michael Titelbaum (2013, Section 11.1.2) expresses misgivings with Elga's argument and offers a new justification for Evil's indifference with respect his identity in *Dr. Evil*. In this section I will show the same reason why Elga's argument is not valid applies against Titelbaum's argument too.

Titelbaum runs his argument for indifference in *Dr. Evil* in a quasi-bayesian framework which he calls the *Certainty-Loss Framework* (CLF). In this chapter I first show how Titelbaum's conclusions can be obtained in an orthodox Bayesian framework if we explicitly take into account the informational context. Secondly, I argue that his argument in favour of RPI fails as well and I explain why I believe CLF doesn't have the resources to overcome the criticism articulated in this chapter.

8.1 FROM DR. EVIL TO TECHNICOLOR EVIL AND BACK

Titelbaum introduces the following variation to the story of *Dr. Evil*:

> TECHNICOLOR EVIL The same as *Dr. Evil*, except Dr. Evil has a spy in the PDF. After the duplication, the spy flips a fair coin (without revealing the result to Evil). If the coin comes up heads, he will show Evil a piece of red paper and show Dup a piece of blue paper. If the coin comes up tails the colours are reversed. Both Dr. Evil and Dup know this.[1](Titelbaum, 2013, p. 255)

Let t_0 be the time before the duplication, t_1 the time after Dr. Evil was duplicated but before seeing any piece of paper, and t_2 the time after Evil sees the coloured paper. Titelbaum claims the following hold of *Technicolor Evil*: Dr. Evil's t_2 credence he is Evil should be $1/2$; his t_1 credence he is Evil conditional on seeing the red paper is also $1/2$; and his unconditional t_1 credence he is Evil is once again $1/2$. Finally, his unconditional t_1 credence he is Evil in *Technicolor Evil* should be the same as his unconditional credence that he is Evil in *Dr. Evil* after he learns of the duplication (Titelbaum, 2013, see p. 255). If Elga is right that *Dr. Evil* is a general case of self-location ignorance, RPI holds.[2]

Titelbaum argues in favour of the above claims using CLF. Below I show how they can be derived from an orthodox Bayesian model

1 Titelbaum, too, follows Elga and discusses the puzzle of *Dr. Evil* in terms of Al and his Duplicate. I follow the strategy of the previous chapter and I adapt Titelbaum's argument to the problem of Evil. Nothing is lost in this translation.

2 Note, however, that Titelbaum isn't convinced *Dr. Evil* is general enough to play this role. One of his criticism to Elga is that this scenario represents a prototypical case of a transitional story only - from self-location certainty to ignorance. A story in which certainty about one's location was never had would presumably also fall under the remit of RPI, but could not be argued for through the story of *Dr. Evil*.

which takes into account the protocol the scientists and the spy follow.[3] Let's first rationally reconstruct Titelbaum's argument. For this purpose assume the following notation: E stands for the proposition "I am Dr. Evil" (and D for "I am Dup"), Red for "I see a piece of red paper" (and Blue for "I see a piece of blue paper"), H for "the coin came up heads" (and T for tails) and let q capture Dr. Evil's credence function in *Technicolor Evil* (and let p continue to be his credence function in *Dr. Evil*).

CLAIM 1 $q_1(E|Red) = q_1(H|Red)$

CLAIM 2 $q_1(E|Red) = q_1(E)$

CLAIM 3 $q_2(H) = \frac{1}{2}$

CLAIM 4 $q_1(E) = \frac{1}{2}$

CLAIM 5 $p_1(E) = q_1(E)$

Let's now carefully model *Technicolor Evil* in a Bayesian framework that explicitly models the information that can be passed on to the agent. First, just like in the case of *Dr. Evil*, there are 4 ways the world could be like: either the coin lands head or it lands tails and either the agent performing the reasoning after the duplication is Dr. Evil or he is Dup. In addition to this, the agent believes he could receive two announcements from his spy in PDF, either Red (in head worlds if he is Evil and tail worlds if he is Dup) or Blue (in tail worlds if he is Evil and head worlds if he is Dup). This generates the following protocol (which in turn determines a sophisticated probability space, as briefly discussed in the previous chapter):
Technicolor Evil is similar to *Coin Toss Dr. Evil* but differs from the latter in two significant ways. Firstly, it makes explicit what

3 This raises some doubts regarding the need for the complicated and non-standard apparatus of CLF. Although I won't engage in a full proof of the expressive poverty of CLF, I will formulate some additional reasons why we should be wary of using this framework for *de se* conditionalisation.

	$H\&E$	$T\&E$	$H\&D$	$T\&D$
Red	1	0	0	1
Blue	0	1	1	0

Table 12: Protocol 1 for Technicolor Evil

information Evil could receive after the duplication, i.e. Red or Blue and with which probabilities. As such, Titelbaum's puzzle is a combined version of *Coin Toss Dr. Evil* and *Comatose Dr. Evil*. Secondly, as opposed to the set-up envisaged by Elga, this puzzle uses a fair (as opposed to a biased) coin.

Given Protocol 1 for *Technicolor Evil*, we can assess Titelbaum's argument. I will look at the first four claims in the remainder of this section and dedicate the next to evaluating his fifth claim.

CLAIM 1 Given Protocol 1 above, the first claim can be easily shown to hold in the case of *Technicolor Evil*:

$$
\begin{aligned}
q_1(H|Red) &= \frac{q_1(Red|H)q_1(H)}{q_1(Red)} \\
&= \frac{\frac{q_1(Red \cap H)}{q_1(H)} \times q_1(E)}{q_1(Red|H\&E)q_1(H\&E) + q_1(Red|T\&D)q_1(T\&D)} \\
&= \frac{q_1(Red \cap \{H\&E, H\&D\})}{q_1(Red|H\&E)q_1(H\&E) + q_1(Red|T\&D)q_1(T\&D)} \\
&= \frac{q_1(Red|H\&E)q_1(H\&E) + q_1(Red|H\&D)q_1(H\&D)}{q_1(H\&E) + q_1(T\&D)} \\
&= \frac{q_1(H\&E)}{q_1(H\&E) + q_1(T\&D)}
\end{aligned}
$$

And

$$
\begin{aligned}
q_1(E|Red) &= \frac{q_1(Red|E)q_1(E)}{q_1(Red)} \\
&= \frac{\frac{q_1(Red \cap E)}{q_1(E)} \times q_1(E)}{q_1(Red|H\&E)q_1(H\&E) + q_1(Red|T\&D)q_1(T\&D)} \\
&= \frac{q_1(Red \cap \{H\&E, T\&E\})}{q_1(Red|H\&E)q_1(H\&E) + q_1(Red|T\&D)q_1(T\&D)} \\
&= \frac{q_1(Red|H\&E)q_1(H\&E) + q_1(Red|T\&E)q_1(T\&E)}{q_1(H\&E) + q_1(T\&D)} \\
&= \frac{q_1(H\&E)}{q_1(H\&E) + q_1(T\&D)}
\end{aligned}
$$

This establishes that Evil's t_2 credence he is Evil should match his credence that the coin came up heads in the case in which between t_1 and t_2 he sees the red piece of paper[4]. This holds for a coin with an arbitrary bias, but relies on the probabilities in Protocol 1 being what they are. To see this, consider an alternative protocol corresponding to a variation of *Technicolor Beauty* in which the spy flips a fair coin to decide whether to show Dr. Evil the red or the blue piece of paper.

	$H\&E$	$T\&E$	$H\&D$	$T\&D$
Red	1/2	1/2	0	1
Blue	1/2	1/2	1	0

Table 13: Protocol 2 for Technicolor Evil

Under Protocol 2,

$$
\begin{aligned}
q_1(H|Red) &= \frac{\frac{1}{2} \times q_1(H\&E)}{q_1(H\&E) + q_1(T\&D)}, \text{ but} \\
q_1(E|Red) &= \frac{\frac{1}{2}(q_1(H\&E) + q_1(T\&E))}{q_1(H\&E) + q_1(T\&D)}
\end{aligned}
$$

4 I assume without loss of generality that Dr. Evil will see the red piece of paper.

These can only be equal if Evil's t_1 credence that he is himself is 0. This seems implausible. What this exercise shows is that Titelbaum's first claim relies on the actual details of the protocol underwriting *Technicolor Evil*. If we were to make a change to it (like the one in Protocol 2), the claim would no longer hold.

CLAIM 2 Given Protocol 1, we can also determine the connection between Evil's unconditional t_1 credence in E and his credence in E conditional on seeing the red piece of paper.

$$\begin{aligned} q_1(E|Red) &= \frac{q_1(H\&E)}{q_1(H\&E) + q_1(T\&D)} \\ &= \frac{q_1(H)q_1(E)}{q_1(H)q_1(E) + (1 - q_1(H))(1 - q_1(E))} \end{aligned}$$

If (and only if) $q_1(H) = \frac{1}{2}$, then:

$$\begin{aligned} q_1(E|Red) &= \frac{\frac{1}{2} \times q_1(E)}{\frac{1}{2} \times q_1(E) + \frac{1}{2} \times (1 - q_1(E))} \\ &= q_1(E) \end{aligned}$$

In other words, given Protocol 1 his unconditional t_1 credence in E matches Evil's conditional credence in E given Red if and only if the bias of the coin is 1/2. So in the presence of Protocol 1, it is only a fair coin that allows us to derive Claim 2.

CLAIM 3 Claim 3 doesn't have the same status as the first two in the sense that it isn't derivable from the scenario using the probability calculus, though it is compatible with it. As such it isn't an endogenous constraint on Evil's credence function, but it could be stipulated as an exogenous constraint on it. Making an exogenous stipulation on an agent's credence function is not forbidden by Bayesianism as long it does not violate the laws of probabilistic coherence - and the suggested stipulation in Claim 3 doesn't do

that. Moreover exogenous constraints are acceptable to some extent by all Bayesians as even the strictest form of Bayesianism requires agents to fix their priors and the event space on which to define them in an exogenous way. What is less trivial about the current stipulation is that it makes demands of how an agent ought to shift her credences given that she learns new information. And some Bayesians would like these shifts to be done fully endogenously. For our present purposes we will simply adhere to the position of the moderate Bayesians and allow any exogenous shifts that do not violate probabilistic coherence.[5]

One reason why we might endorse this particular exogenous constraint on Evil's credence function is the Relevance-Limiting Thesis: "it is never rational for an agent who learns only self-locating information to respond by altering a non-self-locating degree of belief." (Titelbaum, 2013, p. 232) This is a principle that halfers usually espouse in reasoning about the Sleeping Beauty Problem and Elga relies on when discussing Evil's rational credences in *Comatose Dr. Evil*, above. Nevertheless, irrespective of the reason for supporting Claim 3, supposing that Evil's t_2 credence in H should match his t_1 credence in the same proposition is not creating any problem for the agent's probabilistic coherence.

CLAIM 4 Given Claims 1,2 and 3, Claim 4 follows.

To wit, Titelbaum's arguments for Claims 1 to 4 can be reconstructed in a standard Bayesian model. We showed how this is done and on what aspects of the story of *Technicolor Evil*, each argument relies. Three key aspects were identified: the relevance-limiting thesis, or something providing a similar exogenous constraint on Evil's cre-

5 For further discussion of exogenous and endogenous requirements on rationality according to Bayesianism see Urbach and Howson (1993, p. 285), Miller (2016, p. 773) and Schoenfield (forthcoming)

dence function[6] (without which Claim 3 wouldn't be justified), the bias of the coin (without which Claim 2 wouldn't hold), and finally the protocol (without which Claim 1 wouldn't hold). A change in any of them and one part of the overall argument delivering that $q_1(E) = \frac{1}{2}$ would fail. So far, so good. In the next section we evaluate Titelbaum's fifth claim which purports to deliver RPI[7] based on the story of *Technicolor Evil*.

8.2 TITELBAUM'S ARGUMENT FOR CLAIM 5, CAREFULLY

Titelbaum argues for Claim 5 in the following way:

> We now want to argue that [Evil's] required t_1 degree of belief that he's [Evil] in [*Technicolor Evil*] equals his required degree of belief that he's [Evil] in the original [*Dr. Evil*] story. This seems justified on the grounds that [Evil's] relevant evidence is identical in the two circumstances (...) at t_1 [Evil] hasn't seen any colored papers yet. The only information he has at t_1 in [*Technicolor Dr. Evil*] that he doesn't have after the duplication in [*Dr. Evil*] is that his [spy] has been doing some stuff with papers and coins in another room, and that a colored paper will shortly be revealed to him. *Surely this information doesn't provide [Evil] with any evidence in either direction about his identity*. So he is left to set his degree of belief that he's [Evil] based on whatever considerations were

6 I will not mention this assumption in the next section. The reason for this will become apparent when I discuss below the merits of CLF, below.

7 Recall that Titelbaum isn't convinced it does indeed provide the general principle Elga wishes as the story of *Dr. Evil* only models a transitional story from self-location certainty to ignorance. Nevertheless, if Titelbaum's argument is correct Elga, or someone sympathetic to his position, could claim RPI to have been vindicated.

> appropriate in the original case. (Titelbaum, 2013, p. 257, my emphasis)

The discussion in the preceding section shows that in fact the extra information that is contained in *Technicolor Beauty* does provide evidence with respect to Evil's identity. Change the way information accrues to Dr. Evil between t_1 and t_2 and he will no longer be indifferent between being himself or being Dup. Change the bias of the coin his spy is flipping and the same thing would happen. In other words, Evil's t_1 credence in being himself is $1/2$ only in a model in which he knows "his spy has been doing some stuff with papers and" *a fair* "coin in another room, and that a colored paper will shortly be revealed to him" *following a particular protocol*. Evil's credal state after learning of his duplication in *Dr. Evil* contains none of this information and it cannot be identical to Evil's credal state in *Technicolor Beauty*. Suppose, nevertheless, that it is. Then we can concoct the following puzzle:

> CONTRADICTION DR. EVIL The same as *Technicolor Evil* except that 1) the spy flips a coin with bias $2/5$ towards heads; and 2) if the coin comes up heads, he will show Evil either a red piece of paper or a green piece of paper (he will decide which by flipping a second, and independent, fair coin), and he will show Dup a blue piece of paper. If however, the coin comes of tails, he will show Evil a blue piece of paper and Dup either a red or a green piece of paper (he will decide which by flipping a third and independent coin with bias $1/3$ towards red). Both Evil and Dup are aware of this.

The following protocol underwrites this new puzzle.
In this new scenario Evil's posterior credence in being himself upon seeing a red piece of paper matches his posterior credence in the

	$H\&E$	$T\&E$	$H\&D$	$T\&D$
Red	1/2	0	0	1/3
Blue	0	1	1	0
Green	1/2	0	0	2/3

Table 14: Protocol for Contradiction Dr. Evil

coin having landed heads. Let $r(\cdot)$ denote Evil's credence function in *Contradiction Dr. Evil*.

$$\begin{aligned}
r_1(H|Red) &= \frac{r_1(Red|H)r_1(H)}{r_1(Red)} \\
&= \frac{\frac{r_1(Red \cap H)}{r_1(H)} \times r_1(E)}{r_1(Red|H\&E)r_1(H\&E) + r_1(Red|T\&D)r_1(T\&D)} \\
&= \frac{r_1(Red \cap \{H\&E, H\&D\})}{r_1(Red|H\&E)r_1(H\&E) + r_1(Red|T\&D)r_1(T\&D)} \\
&= \frac{r_1(Red|H\&E)r_1(H\&E) + r_1(Red|H\&D)r_1(H\&D)}{r_1(Red|H\&E)r_1(H\&E) + r_1(Red|T\&D)r_1(T\&D)} \\
&= \frac{\frac{1}{2} \times r_1(H\&E)}{\frac{1}{2} \times r_1(H\&E) + \frac{1}{3} \times r_1(T\&D)}
\end{aligned}$$

And

$$\begin{aligned}
r_1(E|Red) &= \frac{r_1(Red|E)r_1(E)}{r_1(Red)} \\
&= \frac{\frac{r_1(Red \cap E)}{r_1(E)} \times r_1(E)}{r_1(Red|H\&E)r_1(H\&E) + r_1(Red|T\&D)r_1(T\&D)} \\
&= \frac{r_1(Red \cap \{H\&E, T\&E\})}{r_1(Red|H\&E)r_1(H\&E) + r_1(Red|T\&D)r_1(T\&D)} \\
&= \frac{r_1(Red|H\&E)r_1(H\&E) + r_1(Red|T\&E)r_1(T\&E)}{r_1(H\&E) + r_1(T\&D)} \\
&= \frac{\frac{1}{2} \times r_1(H\&E)}{\frac{1}{2} \times r_1(H\&E) + \frac{1}{3} \times r_1(T\&D)}
\end{aligned}$$

Moreover, his t_1 credence in being himself conditional on seeing the red piece of paper is equal to his unconditional t_1 credence that he is himself.

$$\begin{aligned} r_1(E|Red) &= \frac{\frac{1}{2} \times r_1(H\&E)}{\frac{1}{2} \times r_1(H\&E) + \frac{1}{3} \times r_1(T\&D)} \\ &= \frac{\frac{1}{2} \times \frac{2}{5} \times r_1(E)}{\frac{1}{2} \times \frac{2}{5} \times r_1(E) + \frac{1}{3} \times \frac{3}{5} \times r_1(D)} \\ &= \frac{r_1(E)}{r_1(E) + 1 - r_1(E)} \\ &= r_1(E) \end{aligned}$$

Furthermore we can again impose exogenously that $r_2(H) = r_1(H)$ as this doesn't create any internal tensions for the agent, and it could also be justified via an independent somewhat substantial constraint on rationality, like the relevance-limiting thesis. Consequently, $r_1(E) = \frac{2}{5}$. Therefore, Evil's unconditional t_1 credence in being himself in *Contradiction Evil* is $2/5$. But if Titelbaum's argument for Claim 5 is correct an analogous reasoning would require Evil's credence in being himself upon learning of the duplication in *Dr. Evil* to be $2/5$. Thus, by Titelbaum's lights, Evil should believe he is himself in *Dr. Evil* with probability $1/2$ and $2/5$. This proves Claim 5 doesn't hold.

Nevertheless, one could argue that this contradiction obtains precisely because of the poverty of the the standard Bayesian framework for dealing with self-location. Titelbaum's analysis is carried out in CLF and perhaps this framework has the internal resources for overcoming the criticism presented here. In the following section I outline the basics of CLF. After that I show that Titelbaum's analysis of *Technicolor Evil* in CLF is open to the same criticism.

8.3 TITELBAUM'S CERTAINTY-LOSS FRAMEWORK

In this section I will attempt to give the reader a crash course in the intricacies of CLF. In doing so I am not endorsing this framework as offering the correct account of *de se* conditionalisation, which is what Titelbaum claims of it. As I remark extensively below, I think CLF fails this task. Rather, I quickly present it and apply it to two puzzles simply as a reply to a critic who might believe the above challenge to Titelbaum's argument can be overcome when we move to CLF. In writing the below sections, I tried to remain as faithful as possible to Titelbaum's presentation and analysis, and as such they are addressed to someone (a critic) sympathetic to Titelbaum's approach and who might believe CLF offers the correct answer to cases of *de se* conditionalisation. Had I wanted to defend CLF for a non-believer, I would have spent much more time investigating some of the technical machinery that underpins CLF, which Titelbaum never clearly explains.

8.3.1 *The formal framework*

The motivation for developing a new framework for accounting for rational belief change stems from the following observation: on the standard Bayesian picture, agents rationally change their beliefs by becoming certain of more and more facts. If we were to read this normatively, we would have to contend that it is irrational to forget. The CLF is meant to offer the Bayesian answer to problems involving memory loss and, what is more, to problems exhibiting context-sensitivity.

The CLF comprises a *model* together with some *systematic* and *extrasystematic constraints* and an interpretation. A *model* consists of a set of time points (T) and a modelling language (L). Over this

set of time points and the set of sentences, Titelbaum defines a probability function capturing the agent's credences at time i (P_i).

The *extrasystematic* constraints stipulate that the model assigns an agent a credence of 1 in a claim iff the story implies that the agent is certain of the truth of that claim or if the claim is a consequence of some of the agent's certainties. The extrasystematic constraints are meant to be determined by each particular story, and the *systematic* constraints are meant to be the same in all CLF models. The first systematic constraint requires credences to be finitely additive. The second is the ratio formula (if $P_i(\neg y) < 1$, then $P_i(x|y) = \frac{P_i(x\&y)}{P_i(y)}$). The third and the fourth constraints do the most interesting work in the CLF: Generalized Conditionalization (GC) and the Proper Expansion Principle (PEP). These are supposed to offer a direct response to how learning context-sensitive facts can affect an agent's credences in context-insensitive ones. I shall discuss each in turn.

> (GC) For any $t_j, t_k \in T$ and any $x \in L$, if $P_j(\neg\langle C_k - C_j\rangle) < 1$ and $P_k(\neg\langle C_j - C_k\rangle) < 1$, then $P_j(x|\langle C_k - C_j\rangle) = P_k(x|\langle C_j - C_k\rangle)$.

(GC) is a clever modification of the usual conditionalization formula. First, C_j is the set of certainties an agent entertains at time point t_j. Second, $\langle C_k - C_j\rangle$ is the conjunction of all the certainties the agent gains between t_j and t_k ($j < k$). If the agent's certainties do not change, then $\langle C_k - C_j\rangle = \top$ (the set of tautologies). The final equality is supposed to capture the intuition that "when you lose information, your resulting doxastic state should be such that were you to regain that information you would return to the doxastic state in which you began" (Titelbaum, 2013, p. 127). For instance, assume that from t_j to t_k the agent becomes certain that a and does not lose any of the certainties he entertained at t_j. Then for any arbitrary sentence x, $P_j(x|a) = P_k(x|\top)$. If the agent does not learn

any new information between t_j and t_k, but instead forgets b, then for any arbitrary x, $P_j(x|\top) = P_k(x|b)$.

Despite its wide scope, (GC) does not generate any verdicts in models in which context-sensitivity plays a crucial role (Titelbaum, 2013, Ch. 8.1.1). The reason is that if an agent's belief in sentence x goes from 1 to 0 (or vice versa) between t_j and t_k, then $C_j \cup C_k$ is inconsistent and in consequence the antecedent of (GC) will be false (Titelbaum, 2013, Appendix 6, Theorem C.9). Therefore, Titelbaum suggests a way of obtaining a context-insensitive reduction of a context-sensitive CLF model which can offer us information about the initial context-sensitive model. A reduction of a model M, is a model M^- with the same time points modelled and with a reduced language $L^- \subseteq L$ such that the analogue of any extrasystematic constraint in M^- is an extrasystematic constraint in M. If M is context sensitive, then we cannot apply (GC), but (GC) can be used on a context-insensitive reduction of M.

(PEP) tells us we can trust that any verdicts of the reduction of a model M apply to the original model if and only if

$$(\forall y \in L)(\forall t_i \in T)(\exists x \in L^-)(P_i(x \equiv y) = 1).$$

According to (PEP), the reduction can inform the initial model if for every sentence in the original context-sensitive model and for every time point, there exists a (context-insensitive) sentence in the language of the reduction which is equivalent to it.

Lastly, Titelbaum is very precise in how CLF models should be *interpreted*:

> [i]f a model's verdicts contradict each other, that model indicates that the agent's doxastic evolution violates the requirements of ideal rationality. If contradictory ver-

> dicts cannot be derived, no violation is indicated. (Titelbaum, 2013, p.s 56)

In other words, if no contradiction can be derived from a model then the agent's credences are ideally rational. And more importantly, judgments of ideal rationality always have to be made within a particular CLF model.

Before moving to its application to *Technicolor Beauty*, I would like to highlight a difference between the CLF and standard Bayesianism. In the usual Bayesian set-up, conditionalization functions as a guide to how beliefs are to be rationally changed in response to new information. In consequence, after learning new facts, an agent is prompted to conditionalize his old beliefs on the newly gathered information and derive his new credal function. This new function will assign certainty to some facts, while other facts will receive a probability of 0, and still others will be assigned an intermediate probability. However, this is not the case with (GC). (GC) can only be applied if the agent already knows what he is certain of. In other words, (GC) is no longer a guide to how to modify beliefs in response to new information, but a way of aligning your non-extreme beliefs given that you know how your certainty set has changed. But *the certainty set itself does not change through (generalized) conditionalization*.

8.3.2 *CLF and Technicolor Evil*

In this section I apply CLF to *Technicolor Evil* and derive the verdict that Evil's t_1 credence he is himself in that scenario ought to be $1/2$. Then I show how to apply CLF to a variation of *Contradiction Dr. Evil* and prove that in this framework we can derive that Evil's t_1 credence in being himself in that scenario ought to be $1/3$. So, if Titelbaum's Claim 5 holds, Evil credence he is himself in *Dr. Evil*

ought to be both $1/2$ and $1/3$, which is contradictory. CLF offers no antidote to the problem I present above.

Titelbaum models *Technicolor Evil* in two steps. First he builds two models, call them Red A02 and Blue A02, for time points $T = \{t_0, t_2\}$. The appropriate language for describing Evil's credal state at those points is $L_1 = \{E, H, Red\}$. Table 15 shows the extrasystematic constrains placed on Evil's credence function in the case in which Evil sees the red piece of paper (Red A02), whereas Table 16 presents the case in which Evil sees the blue piece of paper (Blue A02). p_0 is Evil's credence function at t_0, whereas p_2 his credence function at t_2.

	p_0	p_2
E	1	<1
Red	1/2	1
H	1/2	<1
Red ≡ H	1	<1
E ≡ H	<1	1

Table 15: Extrasystematic constraints for Red A02

	p_0	p_2
E	1	<1
Blue	1/2	1
T	1/2	<1
Blue ≡ T	1	<1
E ≡ T	<1	1

Table 16: Extrasystematic constraints for Blue A02

Focus on Table 15 which codes Evil's loss of certainty between time points t_0 and t_2. At t_0 he is certain he is Evil, he knows that he will see the red piece of paper only when the coin comes up heads, that the probability of the coin coming up heads is $1/2$, and that he will see the red paper if and only if the coin comes up heads. At t_2, Evil loses the certainty regarding his identity. But he is now certain he has seen a red piece of paper and that his identity is tied to the outcome of the coin toss. Titelbaum then invites us to consider the reductions of Red A02 and Blue A02, call them Red A02$^-$ and Blue A02$^-$, which eliminate all context-sensitive propositions from the

language of Red/Blue A02 (i.e. propositions containing indexicals). Red A02$^-$ and Blue A02$^-$ will model the same time points but will only contain one proposition in the language, i.e. $L_2 = \{H\}$. Tables 17, 18 presents their respective extrasystematic constraints.

	p_0^-	p_2^-
H	1/2	<1

Table 17: Extrasystematic constraints for Red A02$^-$

	p_0^-	p_2^-
T	1/2	<1

Table 18: Extrasystematic constraints for Blue A02$^-$

Since no certainty is gained or lost, Evil's credence in H between t_0 and t_2 ought to remain the same in both models and hence, according to (GC):

$$p_2^-(H) = p_0^-(H)$$

We establish that Red A02$^-$ is a proper reduction of Red A02 by noticing that:

$$p_0(E \equiv \top) = 1 \text{ and } p_2(E \equiv H) = 1$$
$$p_0(Red \equiv H) = 1 \text{ and } p_2(Red \equiv \top) = 1$$

And then, by (PEP):

$$p_2(H) = p_0(H)$$

From this we can derive that $p_2(H) = 1/2$, too, and finally since $p_2(E \equiv H) = 1$:

$$p_2(E) = 1/2$$

A completely analogous reasoning in the case in which Evil sees the blue piece of paper (and based on Tables 16 and 18) delivers the same rational t_2 credence. This concludes the first step in

Titelbaum's analysis.

The second step is to build two models for time points t_1 and t_2. Titelbaum's models Red $A12$ and Blue $A12$ use the same language as Red A02 and Blue A02, viz. L_1, and place the following extrasystematic constraints on Evil's credence function:

	p_1	p_2
E	<1	<1
Red	<1	1
H	<1	<1
E ≡ H	<1	1

Table 19: Extrasystematic constraints for Red A12

	p_1	p_2
E	<1	<1
Blue	<1	1
T	<1	<1
E ≡ T	<1	1

Table 20: Extrasystematic constraints for Blue A12

At t_1 Evil has been informed of the duplication but hasn't yet seen the piece of paper. Therefore, he is no longer certain with respect to his identity, he isn't certain what piece of paper he will see and whether the probability of seeing the red piece of paper, say, is connected to the coin coming up heads (this would only be so if he were indeed Evil, but he isn't sure of this fact). Finally, the probability he assigns to the coin coming up heads is no longer $1/2$, as the fact of hearing about the duplication could presumably alter this belief, but he doesn't know in which direction he should revise, if at all. At t_2, he knows he has seen the red (or the blue) piece of paper and that he is evil if and only if the coin came up heads (or tails, in the case in which he sees the blue piece of paper).

According to Red A12, between t_1 and t_2 Evil gains certainty in two propositions: Red and E≡T, therefore $p_2(E|\top) = p_1(E|Red\&(E \equiv H))$. But, at t_2, after seeing the red piece of paper Evil knows that

he is himself if the coin came up heads. So Red implies that E≡H. With this simplification we get that

$$p_2(E) = p_1(E|Red)$$

Importing the verdict of model Red A02 regarding $p_2(E)$:

$$1/2 = p_1(E|Red)$$

An analogous reasoning carried out in the case in which Evil sees the blue piece of paper, that is in model Blue A12 (and based on Table 20) would deliver:

$$1/2 = p_1(E|Blue)$$

Importing this last result into Red A12 as a further extrasystematic constraint, we obtain that whatever information Dr. Evil receives (either Red or Blue), he ought to shift his credences being himself at t_1 to $\frac{1}{2}$. Therefore, by CLF systematic constraints, $p_1(E) = 1/2$. Given Claim 5, then Dr. Evil's rational credence upon learning of his duplication in *Dr. Evil* ought to be $\frac{1}{2}$. This concludes Titelbaum's proof.

8.3.3 *A quick appraisal of CLF*

The previous section presented the CLF solution to *Technicolor Evil*. In this section I wish to raise some methodological problems with CLF before going further to explaining how one can derive contradictory verdicts about Evil's rational credences in *Dr. Evil* in the next section. These comments are intended to provide a motivation for being wary of applying CLF as your modelling strategy for accounting for the doxastic evolution of an agent in a scenario involving self-location.

Firstly, note CLF fails to offer any useful insights if applied to the entirety of scenario. In order to be able to successfully apply it to the story of *Technicolor Evil* we had to make a very careful choice with regards to which time points to model (t_1 and t_2, and t_0 and t_2).

Titelbaum's answer to this criticism (Titelbaum, 2013, Ch. 8, fn. 23) is that the doxastic evolution of an agent is captured by (GC) and that this rule is spelled out only in terms of the difference of certainties in-between time points. But, no matter how we model the different time points of a story, the agent's certainties at those time points are always the same and hence we will never obtain clashing verdicts. This is because, in the CLF, an agent's certainties at a time point are not obtained from conditioning on the information the agent had at the previous time point and the information received in between. They are directly inferred from the story as extrasystematic constraints.

I accept Titelbaum's point, but the situation is, nevertheless, surprising. On the one hand, Bayesians have argued convincingly that in order to obtain the correct results from their formal models we need to follow the principle of Total Evidence and plug into the model all the aspects of the story we seek to understand. Titelbaum, on the other hand, tells us that his quasi-Bayesian framework functions by *restricting* the evidence we take into consideration. I am willing to follow Titelbaum's cue, but, in contrast with the methodological precision of the rest of the book, he is not forthcoming with indications of when and how to restrict the evidence we should take into consideration. Given his analysis of the *Technicolor Evil*, his proposed restrictions seem *ad hoc*.

Secondly, there is the question of how we derive the verdict that Evil's t_1 credence in being himself *should* be $\frac{1}{2}$. Titelbaum says

that we simply patch together the verdicts of Red $A02$, namely $p_2(H) = 1/2$ and $p_2(E) = 1/2$, Blue $A02$, namely $p_2(T) = 1/2$ and $p_2(E) = 1/2$, $A12$, namely $p_2(E) = p_1(E|Red)$ and Blue $A12$, namely $p_2(E) = p_1(E|Blue)$. But in order for something to be a requirement of ideal rationality in his framework, it has to be a verdict of a model. Now, $p_1(E) = \frac{1}{2}$ cannot be a verdict of an A02 model, since it doesn't contain the relevant time point. Therefore it has to be a verdict of one of the A12 ones. However, it cannot be a verdict of Red A12 or Blue A12, since none of them contain enough information to derive it and furthermore their respective languages aren't rich enough to express seeing the color the other model is intended to model (there is no proposition Blue in the language of Red A12 for instance). Titelbaum strategy is to use the verict of Red A02 as an extrasystematic constraint in Red A12. Then the verdict of Blue A02 as an extrasystematic constraint in Blue A12. And moreover use the verdict of Blue A12 as an extrasystematic constraint in Red A12. Finally, he also presumably imposes extrasystematically that Red and Blue are exhaustive, so as to obtain systematically that if conditioned on any of them the probability of being Evil is $1/2$ then it ought to be that unconditionally. Figure 3 presents the interdependencies between the various models in Titelbaum's analysis.

$$\begin{array}{ccccc} RedA02 & \longrightarrow & RedA12 & \longrightarrow & p_1(E) = \frac{1}{2} \\ & & \uparrow & & \\ BlueA02 & \longrightarrow & BlueA12 & & \end{array}$$

Figure 3: Dependencies between Titelbaum's various models for Technicolor Evil

This modelling strategy demands great care from the modeller not only in terms of how she groups the different time points in a story when applying the CLF to it, but also in terms of the order in which she analyses the resulting CLF models of the different time points

of a story is relevant. Without Red $A02$, Red $A12$ cannot yield the desired verdict. Without Blue A12, Red A12 cannot yield that $p_1(E) = \frac{1}{2}$.

These two problems show that there is a lot of work to be done at an intuitive level when applying the CLF, and this gives the impression of ad hocness. However, what Titelbaum's framework achieves quite naturally is a justification for why Evil ought to assign 1/2 to the coin having landed heads at t_2. And it indeed manages to do so without invoking anything like the relevance-limiting thesis. Moreover, this is an endogenous contraint on Evil's credence in CLF.

8.3.4 *CLF and Contradiction Dr. Evil: Part II*

In this section I show that CLF doesn't overcome the criticism of section 7.2. To do so, I will follow Titelbaum's solution to *Technicolor Beauty* closely, making the appropriate changes along the way. Unfortunately, CLF doesn't generate any verdicts in *Contradiction Dr. Evil*[8], so we will need to investigate a close variation of it, *Contradiction Dr. Evil: Part II*:

> CONTRADICTION DR. EVIL: PART II The same as *Contradiction Dr. Evil* except that the bias of the coin flipped by the spy is 2/3 towards heads and 1/3 towards tails (instead of 2/5 and 3/5).

The first thing to notice is that we now require three models, one for each piece of coloured paper the spy can show Evil, let's call

8 I will not explain here in detail neither why that is the case, nor why this isn't a problem with CLF. I will accept that it isn't for the sake of argument. The purpose of this chapter isn't to engage in the exegesis of Titelbaum's CLF, but to show that it doesn't provide a way out of the challenge I mount against Claim 5 in section 7.2.

them Red B02, Blue B02 and Green B02. Each of these three models will contain two time points, t_0 and t_2, and assign credences (I will use p again to represent them) over the following set of propositions $L' = \{E, Red, H, h_1, h_2\}$. The only difference from Titelbaum's language he used to model *Technicolor Beauty* is the inclusion of two new propositions, h_1: "The second coin the spy flips comes up heads", and h_2: "The third coin the spy flips comes up heads". Recall, in *Contradiction Dr. Evil*, if the spy is trying to signal to Evil in a heads-world he will use a fair coin to decide whether to show him the red or green pieces of paper (this is captured by h_1/t_1), and if he is signalling to Dup in a tails-world he will use a coin with bias $1/3$ towards choosing the red piece of paper (this is captured by h_2/t_2); this carries over in *Contradiction Dr. Evil: Part II*. The below tables encode the extrasystematic constraints at play in the three models.

	p_0	p_2
E	1	<1
Red	<1	1
h_1	1/2	<1
h_2	1/3	<1
H	2/3	<1
Red $\equiv$ H $\wedge h_1$	1	<1
E $\equiv$ H $\wedge h_1$	<1	1

Table 21: Extrasystematic constraints for Red B02

	p_0	p_2
E	1	<1
Blue	<1	1
h_1	1/2	<1
h_2	1/3	<1
T	1/3	<1
Blue $\equiv$ T	1	<1
E $\equiv$ T	<1	1

Table 22: Extrasystematic constraints for Blue B02

	p_0	p_2
E	1	<1
Green	<1	1
t_1	1/2	<1
h_2	1/3	<1
H	2/3	<1
Green $\equiv$ H $\wedge t_1$	1	<1
E $\equiv$ H $\wedge t_1$	<1	1

Table 23: Extrasystematic constraints for Green B02

Let me explain the above probability assignments in more detail. Firstly, at time t_0 Evil knows his identity and so he knows the Red paper will be shown if the first coin comes up heads AND if the second coin turns up heads. If the second coin comes up tails he will see the green paper. The outcome of the third coin is not relevant for his credence in Red in this situation. At the same time point, he knows he will see the Green piece of paper if the outcome of the first coin is heads AND the outcome of the second coin is tails. The second coin is again irrelevant to him at t_0 when he is certain he is Evil. Secondly, assume that at time t_2 he sees the red piece of paper. He would thus know the proposition that he is Evil is now equivalent to the first coin having landed heads AND the second coin having landed heads. If he sees a blue piece of paper, the outcomes of the second and third coins are irrelevant and the proposition he is Evil is equivalent just to the proposition the first coin landed tails, whereas if he sees a green piece of paper he knows that the proposition he is Evil is equivalent to the first coin having landed heads AND the second having landed tails.

With these in place, the following equations establish that models Red $B02^{-}$, Blue $B02^{-}$, and Green $B02^{-}$ built over the language $L'^{-} = \{H, h_1, h_2\}$ are proper reductions of models Red $B02$, Blue $B02$, and Green $B02$, respectively.

$$p_0(E \equiv \top) = 1 \text{ and } p_2(E \equiv H \wedge h_1) = 1$$
$$p_0(Red \equiv H \wedge h_1) = 1 \text{ and } p_2(Red \equiv \top) = 1, \text{ and}$$
$$p_0(E \equiv \top) = 1 \text{ and } p_2(E \equiv T) = 1$$
$$p_0(Blue \equiv T) = 1 \text{ and } p_2(Blue \equiv \top) = 1, \text{ and}$$
$$p_0(E \equiv \top) = 1 \text{ and } p_2(E \equiv H \wedge t_1) = 1$$
$$p_0(Green \equiv H \wedge t_1) = 1 \text{ and } p_2(Green \equiv \top) = 1.$$

	p_0^-	p_2^-
H	2/3	<1
h_1	1/2	<1
h_2	1/3	<1

Table 24: Extrasystematic constraints for Red $B02^{-}$

	p_0^-	p_2^-
T	1/3	<1
h_1	1/2	<1
h_2	1/3	<1

Table 25: Extrasystematic constraints for Blue $B02^{-}$

	p_0^-	p_2^-
H	2/3	<1
t_1	1/2	<1
h_2	1/3	<1

Table 26: Extrasystematic constraints for Green $B02^{-}$

Given that the propositions H, and h_1 (and h_2 for that matter, but that proposition is irrelevant when thinking about the probability of being Evil) neither gain nor lose certainty between t_0 and t_2, then by (GC) we can derive that in Red $B02^{-}$, $p_2^-(H) = p_0^-(H) = \frac{2}{3}$ and $p_2^-(h_1) = p_0^-(h_1) = \frac{1}{2}$. By (PEP) $p_2(H) = p_0(H) = \frac{2}{3}$ and $p_2(h_1) = p_0(h_1) = \frac{1}{2}$. Since in Red Bo2, at t_2 the proposition E is equivalent to the conjunction $H \wedge h_1$ then $p_2(E) = \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}$.

Another application of (GC) in Green $B02^{-}$ yields that in that model, too, $p_2^{-}(H) = p_0^{-}(H) = \frac{2}{3}$ and that $p_2^{-}(t_1) = p_0^{-}(t_1) = \frac{1}{2}$. By (PEP) $p_2(H) = p_0(H) = \frac{2}{3}$, and $p_2(t_1) = p_0(t_1) = \frac{1}{2}$. Since in Green B02, at t_2 the proposition E is equivalent to the conjunction $H \wedge t_1$ then $p_2(E) = \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}$.

Finally, applying (GC) in Blue $B02^{-}$ we get that $p_2^{-}(T) = 1/3$ and $p_2^{-}(h_1) = \frac{1}{2}$. By (PEP) this becomes a verdict of Blue B02 and hence $p_2(T) = 1/3$. Since at t_2 the proposition E is equivalent to the proposition T, then it's also the case $p_2(E) = \frac{1}{3}$.

The second step is to investigate models Red $B12$, Blue $B12$ and Green $B12$. Tables 27, 28, and 29 encode their respective extrasystematic constraints in *Contradiction Dr. Evil: Part II*.

	p_1	p_2
E	<1	<1
Red	<1	1
H	<1	<1
h_1	<1	<1
h_2	<1	<1
E ≡ H ∧h_1	<1	1

Table 27: Extrasystematic constraints for Red B12

	p_1	p_2
E	<1	<1
Blue	<1	1
T	<1	<1
h_1	<1	<1
h_2	<1	<1
E ≡ T	<1	1

Table 28: Extrasystematic constraints for Blue B12

	p_1	p_2
E	<1	<1
Green	<1	1
H	<1	<1
t_1	<1	<1
h_2	<1	<1
E ≡ H ∧t_1	<1	1

Table 29: Extrasystematic constraints for Green B12

Focus on Red B12. In this model, Evil doesn't lose any certainties between t_1 and t_2, but he gains certainty in two propositions, namely in Red and E≡ H∧h_1. So, by applying (GC) to Red B12 we get that:

$$p_2(E) = p_1(E|Red \wedge (E \equiv H \wedge h_1))$$

But at t_1, Evil knows that if he sees the red piece of paper then the proposition E becomes equivalent to H$\wedge h_1$ so we can simplify the above equation to $p_2(E) = p_1(E|Red)$. Importing the verdict that $p_2(E) = \frac{1}{3}$ from Red B02 into Red B12, we obtain that:

$$\frac{1}{3} = p_1(E|Red)$$

An analogous reasoning run on Blue B12 and Green B12, will result in

$$\frac{1}{3} = p_1(E|Blue), \text{ in model Blue B12}$$
$$\frac{1}{3} = p_1(E|Green), \text{ in model Green B12}$$

Importing the latter two results into Red B12 (and presumably fixing extrasystematically that Evil can only see these three pieces of paper), we obtain that whatever information Dr. Evil receives (Red, Blue, or Green), he ought to shift his credence in being himself at t_1 to $\frac{1}{3}$. Therefore, by CLF's systematic constraints, $p_1(E) = 1/3$. Given Claim 5, then Dr. Evil's rational credence upon learning of his duplication in *Dr. Evil* ought to be $\frac{1}{3}$. But together with Titelbaum's analysis of *Technicolor Evil* this shows that by CLF and Claim 5's lights, Dr. Evil ought to assign a credence in being himself in *Dr. Evil* after he learns of the duplication of both $1/2$ and $1/3$. This shows that Claim 5 cannot hold if we believe the correct modelling strategy for these scenarios is CLF. Alternatively we could infer from this that CLF's strategy for *de se* conditionalisation is not correct. In either case, Titelbaum's argument fails to offer a new justification for Elga's verdict in *Dr. Evil* and *eo ipso* to RPI.

8.4 CONCLUSION

To sum up, despite the omnipresence of RPI in the literature on self-location and its crucial role in both the Halfer and Thirder answers

to the *Sleeping Beauty Problem*, none of the arguments purporting to justify it work. What is more, the reason why they fail is the same reason that undermines the 'staying' strategy in the *Monty Hall Problem*.

9

THE UNMARKED CLOCK AND THOMASON CASES REDUX

In the previous two chapters we have seen how thinking carefully about the probability with which evidence accrues to an agent depending on the state of the world she is in, can be used to clarify some of the puzzles discussed in the literature engaging with Elga's Restricted Principle of Indifference. In this chapter and the next, I show that ignoring protocols is not a malady of the literature on self-location but a problem which is more widespread. To this purpose I discuss two puzzles from the literature on Reflection Principles and defeating higher-order evidence, viz. Christensen's (2010) *Unmarked Clock*, and Mahtani's (forthcoming) *World's Smallest Lottery*.

Christensen (2010) formulates a quasi-formal argument against an intuitive principle purporting to bridge our first-order beliefs about the world to our higher-order beliefs about our own abilities as reasoners that he dubs RatRef. According to RatRef if we were to learn what the maximally rational credence in a proposition ought to be in our current situation, we would be irrational not to adopt that credence. Nevertheless, Christensen argues, a rational agent ought to violate RatRef on pain of probabilistic incoherence in simple scenarios such as the *Unmarked Clock*. Therefore, RatRef doesn't appear to be a viable rational requirement on an agent's credal state.

In this chapter I show that the argument Christensen formulates involves a common mistake in applying conditionalisation (Shafer, 1985; Sneed, 1985; Pearl, 1988; Halpern, 2004; Bovens and Ferreira, 2010; Halpern, 2015) and that a sophisticated probabilistic model of the *Unmarked Clock* could, if we were to make certain assumptions about the underlying informational protocol in the scenario, satisfy RatRef.

9.1 RATIONAL REFLECTION AND THE UNMARKED CLOCK

Christensen contends that intuitively there should be some connection between "what one is rational to believe, and what one is rational to believe one is rational to believe" (Christensen, 2010, p. 121). And a natural first contender for such a bridge principle is Rational Reflection (RatRef from now on)

RATREF $Cr(A|Pr(A) = n) = n$, where $Cr(\cdot)$ is an agent's credence function, and $Pr(\cdot)$ is her maximally rational credence.

In other words, when one learns that her rational credence in proposition A is n she would be irrational not to change her current credence in A into n. Note that the formula Christensen uses to formally describe RatRef is not well defined. In Kolmogorovian probability theory, probability functions are defined over subsets of the sample space and not over functions. Therefore we cannot take "Pr(a)=n" to be a function but the name of a subset of the sample space. But which one? Christensen doesn't offer any precise formal definition. See Rédei and Gyenis (2016); Gyenis and Rédei (2017) for a discussion of how principles such as RatRef, though intuitive, require more careful development.

That being said, Christensen believes RatRef runs into (a different set of) problems and invites us to consider the following scenario:

UNMARKED CLOCK Chloe walks into a room and notices an unmarked clock on the wall. The minute hand of the clock seems to her to be in the lower right quadrant, a bit under the place where the hour 4 should be marked. But she cannot be sure of this, however, as, alas, the clock is unmarked.[1] What should Chloe's rational credal state be upon seeing the clock?

Chloe cannot be certain of any hand position, but suppose it looks to her as if the hand shows minute 21. Let Pi stand for the proposition "the hand is showing minute i". Christensen claims her rational posterior credence function might be such that P21 receives probability .3, say, P20 and P22 probability .2 and P19 and P23 probability .15. Assume these credences are indeed the maximally rational ones and that moreover, this probability distribution would be the maximally rational one irrespective on the state of the world Pi, except that for each Pi the distribution would be centred around it. That is, if the hand of the clock shows minute 15, then Chloe's maximally rational credence in P15 would be .3, in P16 and P14 would be .2 and in P13 and P17 .15. This generates a table of all of Chloe's maximally rational credence functions, which Christensen calls Chloe's Chart.

1 We assume throughout the minute hand moves in one-minute increments so it never rests between two minute positions.

	P0	⋯	P19	P20	P21	P22	P23	⋯	P59
P0	.3	⋯	0	0	0	0	0	⋯	.2
⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋮	⋯	⋮
P19	0	⋯	.3	.2	.15	0	0	⋯	0
P20	0	⋯	.2	.3	.2	.15	0	⋯	0
P21	0	⋯	.15	.2	.3	.2	.15	⋯	0
P22	0	⋯	0	.15	.2	.3	.2	⋯	0
P19	0	⋯	0	0	.15	.2	.3	⋯	0
⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋮	⋯	⋮
P59	.2	⋯	0	0	0	0	0	⋯	.3

Table 30: Chloe's Chart 1

Here is how to read Chloe's Chart. Each column represents a possible world. Each row represents the worlds Chloe entertains as possible at each possible column-world. Every cell contains the probability Chloe assigns to the worlds she thinks are possible at each possible world. Now, according to Christensen, if she were to be shown her Chart she would thereby learn her posterior credences are the maximally rational ones and contradiction would ensue if RatRef were to hold. Here is Christensen's argument.[2]

> given certainty about the Chart, Chloe will be certain that that

$$a) Pr(P21) = .3 \text{ iff } P21$$

2 Note that Christensen doesn't offer the sample space on which the below probability functions are defined nor the definition of "Pr(P21)=.3" (see above why it cannot formally be a function as the notation would suggest). Therefore the below calculations are a presentation of Christensen's argument as he envisages it. My criticism of Christensen, developed later in this chapter, will be that he doesn't properly define his probability space. I do so later and show how his conclusion no longer follows.

But certainty of this guarantees (modulo coherence) that:

$$b) Cr(P21 / Pr(P21) = .3) = Cr(P21 / P21)$$

> Clearly, the right-hand side of (b) must be 1. But RatRef says that the left hand side should be .3. (Christensen, 2010, p. 124)

Christensen's argument can be reconstructed thus. Let $Chart_1$ be the proposition $P21 \leftrightarrow Pr(P21) = .3$ and assume for simplicity that by seeing her chart, Chloe only learns proposition $Chart_1$. If $Cr(\cdot)$ is Chloe's credence function before seeing her chart, then Christensen claims

$$Cr(P21 | Pr(P21) = .3 \cap Chart_1) = 1.$$

It is easy to check this is the case:[3]

$$
\begin{aligned}
& Cr(P21 | Pr(P21) = .3 \cap Chart_1) = \\
&= \frac{Cr(P21 \cap (Pr(P21) = .3 \cap Chart_1))}{Cr(Pr(P21) = .3 \cap Chart_1)} \\
&= \frac{Cr(P21 \cap (Pr(P21) = .3 \cap (P21 \leftrightarrow Pr(P21) = .3)))}{Cr(Pr(P21) = .3 \cap (P21 \leftrightarrow Pr(P21) = .3))} \\
&= \frac{Cr(P21 \cap (Pr(P21) = .3 \cap ((P21 \cap Pr(P21) = .3) \cup (\neg P21 \cap \neg Pr(P21) = .3))))}{Cr(Pr(P21) = .3 \cap ((P21 \cap Pr(P21) = .3) \cup (\neg P21 \cap \neg Pr(P21) = .3)))} \\
&= \frac{Cr(P21 \cap ((Pr(P21) = .3 \cap (P21 \cap Pr(P21) = .3)) \cup (Pr(P21) = .3 \cap (\neg P21 \cap \neg Pr(P21) = .3))))}{Cr((Pr(P21) = .3 \cap (P21 \cap Pr(P21) = .3)) \cup (Pr(P21) = .3 \cap (\neg P21 \cap \neg Pr(P21) = .3)))} \\
&= \frac{Cr(P21 \cap (P21 \cap Pr(P21) = .3))}{Cr(Pr(P21) = .3 \cap (P21 \cap Pr(P21) = .3))} \\
&= \frac{Cr(P21 \cap Pr(P21) = .3)}{Cr(P21 \cap Pr(P21) = .3)} \\
&= 1
\end{aligned}
$$

3 The only way to make sense of Christensen's argument in the absence of a precise probability model is to assume "Pr(P21)=.3" is a set of possible worlds, namely all the worlds in which the probability of P21 is .3.

This seems to establish that RatRef cannot be taken as a rational constraint on an agent's credences.[4] Nevertheless, I believe Christensen's argument against RatRef deserves a new discussion. The reason is that the argument commits the same fallacy as the one I noted in chapters 6 and 7 and will note again in the next chapter. It is also the mistake Sneed (1985) noted in the reasoning recommending the 'staying' in the *Monty Hall* problem, Pearl (1988) noted in the mistaken reasoning in the *3 Prisoners* and Bovens and Ferreira (2010) noted in Bovens (2010). In other words, it is a very common mistake in probabilistic modelling and it is worth pointing it once again in this context. It may not be the only reason why Christensen's argument fails, but it is one we should also draw a lesson from. To understand the fallacy I first present it in a different context. In the next section I will introduce an example due to Howson (1995) and meant to challenge the use of conditionalisation. I explain carefully how to overcome the challenge and what that teaches us about the *Unmarked Clock*. I conclude with a discussion of whether RatRef holds.

4 A reader familiar with the discussion on Lewis's Principal Principle might object here that Christensen's argument relies on an assumption regarding the admissibility of Chloe's chart. Christensen considers this possible objection and replies that:

> in the case of RatRef, the expert function Pr seems by definition to take into account all of the agent's evidence that bears on the matter in question, and to do so in the maximally rational manner. So motivating restrictions will be more tricky [than in the Principal Principle case]. That said, it's certainly possible that some more sophisticated relative of RatRef might allow us to avoid puzzlement in Chloe's case. (Christensen, 2010, p. 135)

I will not pursue this line of argumentation in the present chapter as I take the flaw in Christensen's argument to reside somewhere else and hence applies even if a strong argument for the admissibility of Chloe's chart can be constructed.

9.2 THOMASON CASES

Howson (1995) argues that conditionalisation shouldn't always be the way through which we update our beliefs in the face of new evidence. He makes his case based on an example he attributes to Richmond Thomason.

> THOMASON CASE A husband announces 'if my wife is unfaithful, I shall never know'; – the wife being known to be an expert in deception. The corresponding conditional probability he ascribes to his not knowing that his wife is unfaithful, given his wife's infidelity, is presumably 1 or near 1. Yet learning that his wife was unfaithful he could scarcely consistently assign probability close to 1 to not knowing what he has just learnt (Howson, 1995, p. 9)

Take $p_i(\cdot)$ to represent the husband's credences over the two time points the example refers to, viz. t_1 before the husband learns anything about the wife's unfaithfulness and t_2 after the husband learns about the wife's unfaithfulness. Let U stand for "the wife is unfaithful". Then the *Thomason Case* purports to show that $p_2(U)$ differs from $p_1(U|U)$: it seems intuitive the former ought to be very low, whereas the latter ought to be 1.

The example is intriguing but I contend all it shows is that conditionalisation can only deliver the correct result in a particular scenario if we model the scenario appropriately. This is a lesson that many have drawn in relation to other problematic cases such as *Monty Hall* (Shafer, 1985; Sneed, 1985; Halpern, 2003; Bovens and Ferreira, 2010), *Sleeping Beauty* (Halpern, 2004; Bovens and Ferreira, 2010; Halpern, 2015), *3 Prisoners* (Pearl, 1988) and the *Doomsday Argument* (Halpern, 2015). Halpern (2004) explicitly mentions the

Thomason Case as a further example of this but doesn't offer a formal analysis to it.

What is the appropriate model for the *Thomason Case*? Firstly, it should account for the fact that the wife could be unfaithful or not. But the model should also account for the husband receiving evidence that seems to suggest she is unfaithful. Let X stand for "evidence that suggests the wife is unfaithful".The following conditional probability table provides a simple visualisation of a possible probability model for this scenario.

	U	¬ U
X	.1	.2
¬ X	.9	.8

Table 31: Protocol 1 for the Thomason Case

The table should be read in the following way. Each column represents a possible state of the world: either the wife is unfaithful or she is faithful. Each row represents a possible evidential state the husband could be in: either he receives evidence that she is unfaithful or he doesn't. Each cell represents the probability the husband is in the row-evidential state given the column-possible world. Following the details of the scenario, the wife is a master deceiver so assuming that she is unfaithful it is very unlikely for the husband to obtain any evidence that she is cheating (.1). However, let's assume for the sake of the argument (nothing of theoretic importance hinges on this) that if she is faithful she will have no reason to try to deceive and so it is a bit more likely for her to do so something that makes her husband suspect she may in fact be unfaithful (.2). Let's assume that the husband believes prior to receiving any evidence about her faithfulness that she is equally likely to be faithful as it is for her not to be faithful. I will relax

this assumption below, but let's reason based on it for now. This table induces the following sophisticated space (Halpern, 2003) for the *Thomason Case*, $S = \{(U, X), (\neg U, X), (\neg U, X), (\neg U, \neg X)\}$. We can take the algebra over which the husband's credences are defined to be 2^S, and we can define his credence function as follows: $p_1(\{(U, X)\}) = \frac{1}{20}$, $p_1(\{(\neg U, X)\}) = \frac{9}{20}$, $p_1(\{(\neg U, X)\}) = \frac{1}{10}$, $p_1(\{(\neg U, \neg X)\}) = \frac{2}{5}$.

With this model we can calculate the husband's conditional credences:

$$p_1(X|U) = \frac{p_1(X \cap U)}{p_1(U)} = \frac{\frac{1}{20}}{\frac{1}{2}} = \frac{1}{10}$$

Whereas,

$$p_1(U|X) = \frac{p_1(X \cap U)}{p_1(X)} = \frac{\frac{1}{20}}{\frac{3}{20}} = \frac{1}{3}$$

That is, the likelihood that the husband will receive evidence that his wife is unfaithful, given she is, is 1/10. And conditional on receiving evidence she is unfaithful the husband's posterior credence in his wife's infidelity goes down from 1/2 to 1/3. So, if we take $p_2(\cdot) = p_1(\cdot|U)$ then $p_2(U)$=1 which is counterintuitive. But if we take $p_2(\cdot) = p_1(\cdot|X)$ then $p_2(U) = 1/3 \neq 1$.

However, if instead of being initially indifferent between U and ¬U, the husband starts off by assuming the wife is faithful, and hence $p_1'(U) = .1$ then:

$$p_1'(U|X) = \frac{p_1'(X \cap U)}{p_1'(X)} = \frac{\frac{1}{100}}{\frac{19}{100}} = \frac{1}{19} << 1$$

In other words, the husband believes that it is very unlikely he will be in an evidential state that would suggest that his wife is being unfaithful if she really were unfaithful, i.e. $\frac{1}{10}$. Nevertheless,

he believes his credence in his wife being unfaithful if he were to receive evidence she were is even lower now, i.e. $\frac{1}{19}$. So if we were to take $p_2'(\cdot) = p_1'(\cdot|X)$ then $p_2'(U) = 1/19$ (very low).

Finally, assume the protocol underwriting the *Thomason Case* is:

	U	¬ U
X	.1	0
¬ X	.9	1

Table 32: Protocol 2 for the Thomason Case

In this case, irrespective of the husband's prior probability distribution:

$$p_1''(U|X) = \frac{p_1''(X \cap U)}{p_1''(X)} = 1 = p_1''(X|X)$$

So, if we were to take $p_2''(\cdot) = p_1''(\cdot|X)$ then $p_2''(X) = 1$. What is then the rational credence the husband ought to adopt: p_i, p_i' or p_i''? I contend there is no answer to the question. The *Thomason Case* is missing relevant information for making this determination. The first credence function is rational given Protocol 1 and equiprobability over U and ¬ U. The second is rational given the same protocol but a non-equiprobable prior distribution. The last one is rational given Protocol 2. Since the scenario doesn't fix either the protocol or the husband's prior over the wife's faithfulness, it is impossible to tell what he ought to believe. Nevertheless, the above shows that there are plausible ways of filling in the story and modelling it in a Bayesian framework so that we get Howson's desired result, e.g. $p2'(U)$ is very low, whereas $p_1'(U|U) = 1$.

What was the problem with Howson's formalisation of the *Thomason Case*? Howson modeled the event of the husband learning that his wife is unfaithful as conditionalising on proposition U, i.e.

$p_2(\cdot) = p_1(\cdot|U)$. However, this is known to lead us into trouble. See elsewhere in the thesis how this strategy lead to problems both in *Monty Hall* and the *3 Prisoners*. Therefore, in order to overcome this difficulty, I introduced another proposition in the model for the *Thomason Case* to stand for the husband's learning of his wife (un)faithfulness. With this additional proposition I showed that conditionalisation no longer leads us astray. The lesson to be drawn from this example is that whenever applying conditionalisation to model learning, we need to pay attention to the likelihood with which an agent believes a particular piece of information can accrue to them and to what other information he thinks he might learn (in the present case the information that his wife is faithful). If this is left out of the formal model, counterintuitive results ensue. As in the *Monty Hall* problem in the *Thomason Case*.

9.3 CHRISTENSEN'S ARGUMENT, CAREFULLY

How are we to import this insight gained from looking at the *Thomason Case* to the *Unmarked Clock*? First, here is a formally precise way of reconstructing Christensen's argument. Let $Cr_0(\cdot)$ be Chloe's prior credence function before learning anything about her chart or any rational credence she ought to hold in the *Unmarked Clock* (but after seeing the clock). Then let $Cr_1(\cdot) = Cr_0(\cdot|Pr(P21) = .3)$. Then $Cr_1(P21|Chart_1) = 1$. The discussion above, however, suggests we shouldn't simply model Chloe's learning of her chart as her conditionalising on the proposition $Chart_1$ but rather on her learning/receiving evidence that $Chart_1$. To emphasise the importance of doing this, consider an alternative to Chart_1, call it Chart_2.

	P0	⋯	P19	P20	P21	P22	P23	⋯	P59
P0	.3	⋯	0	0	0	0	0	⋯	.3
⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋮	⋯	⋮
P19	0	⋯	.3	.3	.05	0	0	⋯	0
P20	0	⋯	.3	.3	.3	.05	0	⋯	0
P21	0	⋯	.05	.3	.3	.3	.05	⋯	0
P22	0	⋯	0	.05	.3	.3	.3	⋯	0
P19	0	⋯	0	0	.05	.3	.3	⋯	0
⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋮	⋯	⋮
P59	.3	⋯	0	0	0	0	0	⋯	.3

Table 33: Chloe's Chart 2

In Chart_2, Chloe's rational credence upon seeing the clock is to assign a .9 credence to the hand showing a position within one minute of the actual time. How could we have two charts? How could Chloe have a probability distribution over the maximally rational credences in her situation? One simple story to motivate this would be that her visual acuity is influenced by environmental conditions: light, distance to the clock, tiredness, etc. Chloe cannot quantify any of these contextual factors. But she knows that for certain constellations of factors a maximally rational agent would assign credences in a particular way, whereas for another she would assign them differently. We simplify things and assume there are only two such ways. The first encoded in Chart_1, the second in Chart_2. Now, with this second chart in play we could wonder what chart Chloe considers receiving. The below represents a possible protocol underwriting the *Unmarked Clock*.

	P0	⋯	P19	P20	P21	P22	P23	⋯	P59
"$Chart_1$"	1/2	⋯	1/2	1/2	1/2	1/2	1/2	⋯	1/2
"$Chart_2$"	1/2	⋯	1/2	1/2	1/2	1/2	1/2	⋯	1/2

Table 34: Protocol 1 for the Unmarked Clock

What this protocol says is that Chloe considers it equally likely to be informed between t_1 and t_2 that $Chart_1$ holds or that $Chart_2$ holds. So what should Chloe's credence in P21 be given the "information that $Chart_1$ holds" (let's designate this by "$Chart_1$")?

$$\begin{aligned} Cr_1(P21|"Chart_1") &= \frac{Cr_1("Chart_1"|P21)Cr_1(P21)}{Cr_1("Chart_1")} \\ &= \frac{\frac{1}{2} \times Cr_1(P21)}{\frac{1}{2}} \\ &= Cr_1(P21) \neq 1 \end{aligned}$$

Recall that $Cr_1(\cdot) = Cr_0(\cdot|Pr(P21) = .3)$, so $Cr_1(P21) = Cr_0(P21|Pr(P21) = .3)$. If RatRef holds, then $Cr_1(P21) = .3$ and then $Cr_2(P21) = .3$. In this construal, the threat of the *Unmarked Clock* to RatRef has been blunted and the scenario no longer represents a counterexample to RatRef. Are there other protocols that make sense in this context? The answer is YES. Consider the following protocol that could be underwriting the informational exchange Christensen has in mind.

	P0	⋯	P19	P20	P21	P22	P23	⋯	P59
"$Chart_1$"	0	⋯	1/3	0	1/3	1/3	1/3	⋯	1/3
"$Chart_2$"	1	⋯	2/3	1	2/3	2/3	2/3	⋯	2/3

Table 35: Protocol 2 for the Unmarked Clock

The story behind this protocol is that whenever the clock's hand is pointing towards one of the minutes which usually are marked on a

regular clock, Chloe becomes better at discriminating what time it is. When the hand falls in between locations which are usually marked, Chloe's rational distribution has fat tails. With this new protocol:

$$\begin{aligned} Cr_1(P21|"Chart_1") &= \frac{Cr_1("Chart_1"|P21)Cr_1(P21)}{Cr_1("Chart_1")} \\ &= \frac{\frac{1}{3} \times Cr_1(P21)}{\frac{1}{3}(Cr_1(P1) + Cr_1(P2) + Cr_1(P3) + Cr_1(P4) + Cr_1(P6) + \ldots)} \\ &= \frac{Cr_1(P21)}{(Cr_1(P1) + Cr_1(P2) + Cr_1(P3) + Cr_1(P4) + Cr_1(P6) + \ldots)} \end{aligned}$$

We cannot say anything more about $Cr_2(P21)$ in this context without making further stipulations about what $Cr_1(P1), Cr_1(P2), \ldots$ are. Recall that $Cr_1(\cdot) = Cr_0(\cdot|Pr(P21) = .3)$. Now, even if we were to assume RatRef held, we could only infer $Cr_1(P21) = .3$. But we still wouldn't know anything about $Cr_1(P1), etc.$ as RatRef remains silent on $Cr_0(Pi|Pr(P21) = .3)$, where $i \neq 21$.

Where does this leave us with respect to the *Unmarked Clock*? We assumed that Chloe is considering two possible charts before being told anything about what a maximally rational agent ought to believe in her situation. We then looked at two possible protocols Chloe could be entertaining in her situation and remarked that both are compatible with the scenario given that Christensen remains silent on how information about the chart accrues to Chloe. Just as in the *Thomason Case* what this shows is that in order to fully analyse the *Unmarked Clock* we require more information than it is offered. And there are ways of complementing the scenario according to which the challenge it is meant to mount disappears: assuming Protocol 1, the *Unmarked Clock* no longer violates RatRef. At the same time there are ways of complementing the scenario, viz. Protocol 2, for which cannot draw any conclusion. This shows the *Unmarked Clock* doesn't deliver on its promise. Just like the other scenarios I looked at in previous sections, it offers too little informa-

tion to allow a precise probabilistic analysis. This isn't to say that an informationally richer version of the same scenario couldn't be turned into a counterexample of RatRef, but the conclusion Christense wants to derive from the scenario is, as I have tried to show in this section, too quick.

10

THE OPAQUE PROPOSITION PRINCIPLE AND INFORMATION-GATHERING PROCESSES

In a recent paper, Anna Mahtani (forthcoming) introduces a simple probabilistic puzzle which she claims carries a lesson for a generalisation of the Expert Deference Principle (Elga, 2007), viz. *The Opaque Proposition Principle* (OP). In this paper we show that the puzzle she introduces is equivalent to a generalisation of the *3 Prisoners Problem* and explain how her analysis maps onto its intuitive, but nevertheless, mistaken solution. In other words, when employing the correct probabilistic model for understanding Mahtani's puzzle, the alleged challenge to OP disappears. We conclude by drawing once again (Shafer, 1985; Sneed, 1985; Pearl, 1988; Halpern, 2003; Bovens and Ferreira, 2010) the lesson that the process through which information is gathered is essential for correctly taking conditionalisation on X as a guide to learning X.

10.1 BASIC-KNOW, SUPER-KNOW AND THE OPAQUE PROPOSITION PRINCIPLE

Mahtani introduces a distinction between two kinds of knowledge, viz. basic-knowledge and super-knowledge. Consider the First Amendment to the United States Constitution:

> Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or

> abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances.

When asking whether someone knows the First Amendment, I might be asking two different things. One the one hand, I might be asking if they knew that Congress cannot restrict several key freedoms (of religion, of speech and of the press). Someone could presumably know this without knowing it's the First Amendment, or even in the Constitution at all. In this case, they would basic-know it. On the other hand, someone could believe Congress is actively and lawfully (although regrettably) restricting the freedom of religion in the US, but nevertheless know that the First Amendment of the Constitution says that it doesn't have the right to do that. Such a person would super-know the First Amendment, but not basic-know it. Of course, a third individual could both basic-know it and super-know it.[1]

Now Mahtani introduces a type of reflection principle and claims that if we were to interpret the highlighted occurrence of the verb 'know' in this principle as basic-know the principle fails. If we were to interpret it as super-know, it holds.

> OP If: for some claim H and value v, an agent knows that there is a true proposition E such that if (s)he were to come to *know* E (and not to learn or forget anything

1 Mahtani believes this distinction plays an important role in the understanding of a probabilistic principle, but despite that, she doesn't formally distinguish between basic-know and super-know. In particular, it is not clear from Mahtani's paper how a formal model engendering basic-knowledge would differ from one engendering super-knowledge. Nevertheless, I believe one wouldn't need to explore this distinction further, as the correct probabilistic model for Mahtani's puzzle will offer a solution to the problem she raises without requiring any special construal of 'knowledge'.

> else), then his or her credence in H should be v
> Then: that agent ought to have a credence of v in H.
> (Mahtani, forthcoming, p. 3)

We take the principle in its basic-know reading to mean the following: For a credence function p over an algebra S, for $E \subseteq S$ such that p(E)=1 and some proposition $H \subseteq S$ and value $v \in [0,1]$: if for any $e \in E$ were an agent to come to know it (and not to learn or forget anything else), then his or her credence in H would be v, then that agent ought to have a credence of v in H.[2] In order to show that this principle fails Mahtani provides a counterexample, the *World's smallest lottery* (WSL from now on):

> WSL Suppose that you have bought a ticket in the World's Smallest Lottery. There are four tickets (tickets 1, 2, 3 and 4) and four players, each of whom has bought a single ticket. There are two prizes, and so two of the four tickets have been selected randomly as the winners. Your ticket is number 1. Assuming that you are rational, what is your credence that (WIN1) your ticket number 1 is a winner? (Mahtani, forthcoming, p. 2)

To clarify the details of WSL, she explains the procedure by which the winning tickets are selected as follows:

> First (s)he picks out one ticket at random from amongst the four; then (s)he picks out a further winning ticket from the remaining three tickets. Thus there are two ways that ticket 1 might be selected: one way is by being selected on the first draw, and the other is by not

2 Mahtani distinguishes between two senses of basic-know. In this paper we focus on what she calls the variable reading of OP under basic-know. We gloss over the distinction between this reading and the other as it doesn't impact the discussion below.

> being selected on the first draw but being selected on the subsequent draw. (Mahtani, forthcoming, fn. 1)

First, Mahtani claims that the probability of WIN1 is $\frac{1}{2}$ (forthcoming, fn. 1). This is correct, but to make it precise let's construct the probability space carefully so that we can actually get this result as the verdict of a probabilistic model. Consider the following model of WSL, $\langle X, S, p \rangle$: X={(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)}; S= 2^X; p({(1,2)}) = p({(1,3)}) = p({(1,4)}) = p({(2,3)}) = p({(2,4)}) = p({(3,4)}) = $\frac{1}{6}$.

Each number i stands for "Ticket i wins a prize". Each element of the sample space represents a pair of winning tickets (assuming the order in which they are declared winning tickets is irrelevant to the set-up). The algebra over which we define our measure is simply the power set of the sample space. The probability function works as the scenario indicates: the probability of WIN1, i.e. {(1,2),(1,3),(1,4)}, is the same as the probability of WIN2, i.e. {(1,2),(2,3),(2,4)}, etc. and is equal to $\frac{1}{2}$. The same holds of WIN3 and WIN4. There is nothing puzzling about this fact as WIN1, WIN2, WIN3 and WIN4 are not disjoint, so it is not surprising that they don't sum up to 1.

Second, Mahtani (forthcoming, p.2) calculates the probability of WIN1 conditional on learning that WIN2:

$$p(WIN1|WIN2) = \frac{p(WIN1 \cap WIN2)}{p(WIN2)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3},$$

and notices that the probability of WIN1 given WIN2 is the same as the probability of WIN1 given WIN3, and as the probability of WIN1 given WIN4. From this Mahtani raises the following challenge:

> You know then that there is some true proposition (either WIN2, WIN3 or WIN4) such that if you were to come

> to know it, you would – and rationally should – have a credence of 1/3 in WIN1. Doesn't it follow that your actual current credence in WIN1 should be 1/3? After all, you know that there is a true proposition that would rightly drive your credence in WIN1 down to 1/3 if you but knew it. You do not know this proposition – it is opaque to you – but isn't knowing that there is such a proposition enough? (Mahtani, forthcoming, pp. 2-3)

In other words, in WSL, it seems there exists some true proposition (either WIN2, WIN3, or WIN4) such that if you were to come to know it then your credence in WIN1 should decrease to $\frac{1}{3}$. If OP were true, then your credence in WIN1 should already be $\frac{1}{3}$ even before hearing which of WIN2, WIN3 and WIN4 actually holds. This reading of OP is the basic-know one, as you do not know which proposition E actually is. You know that E can be one of three distinct propositions that jointly exhaust the ways in which the world could be (that is, $p(WIN1) + p(WIN2) + p(WIN3) = 1$), but you do not know which one it is, as it is a different one depending of the state of the world and that is unknown to you. You would have super-known E if you knew that E was WIN2, say.[3]

But, Mahtani argues, OP cannot be true (at least not under this reading). We could easily conjure up propositions LOSE2, LOSE3 and LOSE4 ("Ticket i is a losing ticket") and then, on the one hand, one of these propositions must be true, and on the other p(WIN1 | LOSE2) = p(WIN1 | LOSE3) = p(WIN1 | LOSE4) = $\frac{2}{3}$ (Mahtani, forthcoming, p. 5). Therefore, if OP were to hold p(WIN1) should at the same time be 1/3 and 2/3.

3 In this paper I focus solely on Mahtani's rejection of OP under its basic-know reading, and won't discuss super-knowledge any further.

To sum up, 'to know' can be disambiguated into 'to basic-know' or to 'super-know' and a reflection-like principle, viz. OP, fails if the relevant instance of know in its formulation is interpreted as basic-know. According to Mahtani, WSL witnesses the failure of the principle in its basic-know reading. I believe this claim in incorrect and that the mistake lies in the way Mahtani applies conditionalisation on X as formal construal of 'learning X'. As puzzles such as the *Monty Hall* and the *3 Prisoners* have taught us this has to be done very carefully and by paying attention to the context in which the agent learns the new information - we can follow Mahtani in calling this the information-gathering process (Bovens and Olsson, 2000). The failure of Mahtani's analysis of WSL is thus instructive by offering us yet another opportunity to think about what conditions are required for correctly applying conditionalisation to model every day learning.

The next section presents the *3 Prisoners* problem, the intuitive, but incorrect, solution to it, the correct solution to it and finally how this solution applies to several generalisations of the standard problem. The last generalisation discussed will be shown to be perfectly equivalent to the WSL and the exact correspondences between the two puzzles will be explained.

10.2 THE 3 PRISONERS PROBLEM

In this section we discuss the *3 Prisoners Problem* and a few generalisations and variations thereof. The purpose of this exercise is to lead the reader from the very well-known solution to the standard version of the problem to the solution of one of its generalisations which is perfectly analogous to Mahtani's WSL.

> 3 PRISONERS PROBLEM Three prisoners, A, B, and C, have been tried for murder, and their verdicts will be read and their sentences executed tomorrow morning. They know only that one of them will be declared guilty and will be hanged to death while the other two will be set free; the identity of the condemned prisoner is revealed to the very reliable prison guard, but not to the prisoners themselves. In the middle of the night, Prisoner A calls the guard over and makes the following request: "Please give this letter to one of my friends - to one who is to be released. You and I know that at least one of them will be freed." The guard takes the letter and promises to do as told. An hour later prisoner A calls the guard and asks "Can you tell me which of my friends you gave the letter to? (...)" The guard answers "I gave the letter to Prisoner B; he will be released tomorrow."(Pearl, 1988, p. 58)

What is A's chance of being executed upon learning that B is to be released. Here is a quick way of reasoning about this puzzle: initially A had one chance in three to be executed. Now that he knows B is to be released, he knows that either him or C will be executed. Therefore after hearing the guard's answer, he will increase his credence in being executed from one in three to one in two.

Here is a quick probability model to show this. Take the probability model $\langle Z, F, p \rangle$. $Z = \{Ae, Be, Ce\}$, where Ae means A will be executed, Be means B will be executed and Ce means that C will be executed. Let $F = 2^Z$ and $p(\{Ae\}) = p(\{Be\}) = p(\{Ce\}) = \frac{1}{3}$ (from now on I will omit certain curly brackets to simplify the notation). Assuming Prisoner A's credence function after hearing the guard's answer is obtained by conditionalization, i.e. his posterior credence

function will be $p(\cdot|\neg Be)$, we can calculate A's posterior credence in being executed:

$$\begin{aligned} p(Ae|\neg Be) &= \frac{p(Ae \cap \neg Be)}{p(\neg Be)} \\ &= \frac{p(Ae \cap \{Ae, Ce\})}{p(\{Ae, Ce\})} \\ &= \frac{p(\{Ae\})}{p(\{Ae\}) + p(\{Ce\})} \\ &= \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{3}} \\ &= \frac{1}{2} \end{aligned}$$

This answer is incorrect. And Judea Pearl explains the problem with it by pointing his readers towards the following subsequent argument A could make upon raising his credence in Ae to $\frac{1}{3}$:

> "...Worse yet, by sheer symmetry, my chances of dying would also have risen to 50% if the guard had named C instead of B - so my chances must have been 50% to begin with. I must be hallucinating ..." (Pearl, 1988, p. 59)

The problem with this reasoning is that in modelling the scenario we have ignored relevant information about the information-gathering process by which A comes to know that $\neg Be$. By this I mean the full range of answers that A could have obtained from the guard. To understand why the context (as Pearl calls it) is important it suffices to suppose the guard always answers "I gave the letter to Prisoner B", irrespective of what state of the world he is in. Then, A's rational response should be to ignore his answer altogether, and not change his credence function to $p(\cdot|\neg Be)$. This isn't presumably the case, but the puzzle remains silent on the actual details of the protocol the guard will be following when delivering the letter and answering

A's question. So in one sense, it is impossible to say what *the* correct answer is - the problem is under-specified. Nevertheless, there seem to be some salient assumptions we could make, and we can call them the protocol[4] that the guard will follow

1. the guard will never give A's fate away - that is it will only reveal information about one (and only one) of the other prisoners;
2. the guard will never lie - to wit, he will never answer that a prisoner will be released if they are sentenced to hang or vice versa; and finally
3. if the guard has a choice as to whom to deliver the letter, the guard uses a fair random device to make a decision.

The following conditional probability table captures these assumptions about the protocol the guard follows (since this section will generalise the problem along several dimensions I will use the following notation: N-k-m Prisoners is the variation of the standard 3 Prisoners problems in which there are N prisoners, k death sentences and m letters/answers the guard delivers/offers)

	Ae	Be	Ce
B will be released	1/2	0	1
C will be released	1/2	1	0

Table 36: Protocol 1 for 3-1-1 Prisoners

In this table, each column corresponds to a possible world - one in which A is executed, one in which B is executed or one in which C is executed - and each row corresponds to a possible answer the guard could give. Consequently, the numbers in the cells represent

4 Some authors call the information-gathering process in a scenario, the protocol of that scenario. The two concepts are assumed to be interchangeable in this paper.

the probability A will receive the row-answer in the column-world. This table is just a nice visualisation of a sophisticated sample space (Halpern, 2003) in which we take into consideration every answer the prisoner could receive. Such a space would contain four atomic events with positive probability: (Ae, B will be released), (Ae, C will be released), (Be, C will be released) and (Ce, B will be released). The probability of (Ae, B will be released) and (Ae, C will be released) will be equal to the probability of (Be, C will be released) and to the probability of (Ce, B will be released), that is $1/3$. (Ae, B will be released) and (Ae, C will be released) will both receive probability $1/6$. All the conditional probability tables below will induce in a similar fashion a sophisticated probability model, but we won't present that in detail. We assume throughout that A assigns equal priors to all states of the world. With this in place we can re-calculate A's posterior credence of being executed:

$$\begin{aligned} p(Ae|\text{B will be released}) &= \frac{p(\text{B will be released}|Ae)p(Ae)}{\sum_{i\in\{Ae,Be,Ce\}} p(\text{B will be released}|i)p(i)} \\ &= \frac{\frac{1}{2}\times\frac{1}{3}}{\frac{1}{2}\times\frac{1}{3}+0\times\frac{1}{3}+1\times\frac{1}{3}} \\ &= \frac{1}{3} \end{aligned}$$

In other words, upon learning the guard's answer, A has no reason to change his credence regarding his own fate. However, not the same is the case regarding C's fate:

$$p(Ce|\text{B will be released}) = \frac{2}{3}$$

Notice however, that we A has no reason to worry he might be hallucinating, as

$$p(Ce|\text{C will be released}) = 0$$

This result is based on the three assumptions above. Can we make different assumptions such that A's posterior credence in being executed to indeed be $\frac{1}{2}$? Yes, for instance we could adopt assumptions (1), (2), and (4).

4. if A is to be executed, the guard will always take the letter to Prisoner B.

This set of assumptions gives rise to a different protocol for the guard:

	Ae	Be	Ce
B will be released	1	0	1
C will be released	0	1	0

Table 37: Protocol 2 for 3-1-1 Prisoners

In this case, we can recreate the posterior credence arrived at by the naive reasoning above, but the threat of hallucinating has once again disappeared:

$$p(Ae|\text{B will be released}) = \frac{1}{2}$$
$$p(Ae|\text{C will be released}) = 0$$

This is a well-known puzzle in probability theory and is structurally analogous to the *Monty Hall* problem. Unsurprisingly this solution to the *3 Prisoners* has been given to the *Monty Hall*, as well (Sneed, 1985; Shafer, 1985; Bovens and Ferreira, 2010), and discussed further in relation to other puzzles such as the *Sleeping Beauty Problem* and the *Doomsday Argument* (Halpern, 2004; Bovens and Ferreira, 2010; Halpern, 2015). It can also be found in several textbooks (Pearl, 1988; Halpern, 2003) and we do not claim to bring anything new to it. Furthermore this reasoning naturally extends to finitely many prisoners, executions and answers the guard may give (Wechsler

et al., 2005). Below, I present the cases of 4-1-1 Prisoners and 4-2-1 Prisoners, respectively, and show how the latter is a structural analogue of Mahtani's WSL.

Firstly, imagine that instead of 3 prisoners expecting the execution we now have 4. Everything else stays the same, including assumptions (1)-(3) above. Then the following table captures the protocol the guard could follow in this scenario:

	Ae	Be	Ce	De
B will be released	1/3	0	1/2	1/2
C will be released	1/3	1/2	0	1/2
D will be released	1/3	1/2	1/2	0

Table 38: Protocol for 4-1-1 Prisoners

A's posterior credences upon hearing that B will be released will be:

$$\begin{aligned} p(Ae|\text{B will be released}) &= \frac{p(\text{B will be released}|Ae)p(Ae)}{\sum_{i\in\{Ae,Be,Ce,De\}} p(\text{B will be released}|i)p(i)} \\ &= \frac{\frac{1}{3}\times\frac{1}{4}}{\frac{1}{3}\times\frac{1}{4}+0\times\frac{1}{4}+\frac{1}{2}\times\frac{1}{4}+\frac{1}{2}\times\frac{1}{4}} \\ &= \frac{1}{4} \end{aligned}$$

Again, A's credence that he will be executed doesn't change upon hearing the guard's answer. It was 1/4 before and it stays the same. Just as before, A's credence regarding C's and D's fates does change, but an easy verification will establish that no threat of hallucinating is present.

We can now extend the scenario in another dimension too and assume 2 prisoners out of the 4 who will be executed. Suppose further the guard still only delivers one letter and, hence, Prisoner A only

learns about the fate of one other prisoner (never himself - assumptions 1-3 still hold). Such a scenario could be modeled in the following way. Let *AeBe* mean that prisoners A and B will be executed and suppose the order in which they are executed does not matter (the decision of whom to execute is done all at once, say). Finally, we assume once again all resulting 6 possible worlds have equal priors according to Prisoner A.

	AeBe	AeCe	AeDe	BeCe	BeDe	CeDe
B will be released	0	1/2	1/2	0	0	1
C will be released	1/2	0	1/2	0	1	0
D will be released	1/2	1/2	0	1	0	0

Table 39: Protocol 1 for 4-2-1 Prisoners

In this model we can retrieve the event corresponding to 'Prisoner A being executed', call this AE, as the union between (AeBe), (AeCe) and (AeDe). The prior of this event is unsurprisingly $\frac{1}{2}$. Then, we can calculate again Prisoner A's posterior credence that he will be executed (let $Y = \{AeBe, AeCe, AeDe, BeCe, BeDe, CeDe\}$):

$$\begin{aligned}
p(AE|\text{B will be released}) &= \frac{p(\text{B will be released}|AE)p(AE)}{\sum_{i\in Y} p(\text{B will be released}|i)p(i)} \\
&= \frac{\frac{p(\text{B will be released}\cap AE)}{p(AE)} \times p(Ae)}{\sum_{i\in Y} p(\text{B will be released}|i)p(i)} \\
&= \frac{p(\text{B will be released} \cap AE)}{\sum_{i\in Y} p(\text{B will be released}|i)p(i)} \\
&= \frac{\sum_{j\in AE} p(\text{B will be released}|j)p(j)}{\sum_{i\in Y} p(\text{B will be released}|i)p(i)} \\
&= \frac{0 \times \frac{1}{6} + \frac{1}{2} + \frac{1}{6} + \frac{1}{2} \times \frac{1}{6}}{0 \times \frac{1}{6} + \frac{1}{2} + \frac{1}{6} + \frac{1}{2} \times \frac{1}{6} + 0 \times \frac{1}{6} + 0 \times \frac{1}{6} + 1 \times \frac{1}{6}} \\
&= \frac{1}{2}
\end{aligned}$$

Now, let's depart from the formulation of the 3 Prisoner problem above and reflect on a situation in which we ask what credence Prisoner A assigns to being released upon hearing the guard's answer. To look at this, it is easier to re-model the scenario as such (let now ArBr stand for prisoners A and B will be released)

	ArBr	ArCr	ArDr	BrCr	BrDr	CrDr
B will be released	1	0	0	1/2	1/2	0
C will be released	0	1	0	1/2	0	1/2
D will be released	0	0	1	0	1/2	1/2

Table 40: Protocol 2 for 4-2-1 Prisoners

An easy calculation will establish that

$$p(Ar|\text{B will be released}) = \frac{1}{2}, \text{ but}$$
$$p(Cr|\text{B will be released}) = \frac{1}{4}$$

We can further re-imagine the standard presentation of the scenario to reflect on what would happen if the guard won't say who will be released, but rather who will be executed:

	ArBr	ArCr	ArDr	BrCr	BrDr	CrDr
B will be executed	0	1/2	1/2	0	0	1
C will be executed	1/2	0	1/2	0	1	0
D will be executed	1/2	1/2	0	1	0	0

Table 41: Protocol 3 for 4-2-1 Prisoners

Another easy calculation will establish that in this case, too:

$$p(Ar|\text{B will be executed}) = \frac{1}{2}, \text{ but}$$
$$p(Cr|\text{B will be executed}) = \frac{3}{4}$$

Finally, we can imagine the guard having a choice between taking the letter to another Prisoner who is to be released or one who is to be executed. To construct the protocol for the guard in this case, we need to make some refinements to assumption 3 above. We can tweak assumption 3 in two salient ways. One possible (and salient) protocol would be the following:

5. The guard will use a random device to decide whom to deliver the letter to.

	ArBr	ArCr	ArDr	BrCr	BrDr	CrDr
B will be released	1/3	0	0	1/3	1/3	0
C will be released	0	1/3	0	1/3	0	1/3
D will be released	0	0	1/3	0	1/3	1/3
B will be executed	0	1/3	1/3	0	0	1/3
C will be executed	1/3	0	1/3	0	1/3	0
D will be executed	1/3	1/3	0	1/3	0	0

Table 42: Protocol 4 for 4-2-1 Prisoners

With this protocol in mind, Prisoner A's posterior credence in being released after hearing what the guard tells him about prisoners B, C, or D will be released will be:

$$p(Ar|\text{B will be released}) = \frac{1}{3},$$
$$p(Ar|\text{C will be released}) = \frac{1}{3},$$
$$p(Ar|\text{D will be released}) = \frac{1}{3}.$$

But this looks very similar to what we identified as the incorrect reasoning above. Wouldn't Prisoner A in this case think that he may be hallucinating and that his prior in A should be $\frac{1}{3}$, too (see the quote from Pearl, above). No! Although Prisoner A believes the

guard could give any of these three answers (i.e. B will be released, C will be released and D will be released), they do not exhaust all the pieces of information Prisoner A believes the guard could pass on to him. The guard could also tell him that B will be executed, C will be executed or that D will be executed. His posteriors upon learning any of these will be:

$$p(Ar|\text{B will be executed}) = \frac{2}{3},$$
$$p(Ar|\text{B will be executed}) = \frac{2}{3},$$
$$p(Ar|\text{B will be executed}) = \frac{2}{3}.$$

To wit, it isn't the case that Prisoner A can be thinking that whatever the guard tells him, his credence in being executed would drop to 1/3. It could also increase to 2/3. He can be sure that the information he receives will have an impact on his credences regarding his own predicament, but he cannot be sure what that impact will be. So no danger of hallucinating.

There is a different salient way in which Assumption 3 could be tweaked and this in turn generates yet a further possible protocol.

6. The guard will use a random device to decide whether to deliver the letter to one of the prisoners awaiting execution or to one of those who will be released and then another (independent) random device to decide which one of the soon to be released/executed prisoners should receive it.

	ArBr	ArCr	ArDr	BrCr	BrDr	CrDr
B will be released	1/2	0	0	1/4	1/4	0
C will be released	0	1/2	0	1/4	0	1/4
D will be released	0	0	1/2	0	1/4	1/4
B will be executed	0	1/4	1/4	0	0	1/2
C will be executed	1/4	0	1/4	0	1/2	0
D will be executed	1/4	1/4	0	1/2	0	0

Table 43: Protocol 5 for 4-2-1 Prisoners

On this protocol, the answer the guard gives Prisoner A won't have an impact on his credence regarding his own predicament, but it will affect what he believes regarding the fate of others:

$$p(Ar|\text{B will be released}) = \frac{1}{2},$$
$$p(Ar|\text{B will be executed}) = \frac{1}{2}, \text{ but}$$
$$p(Cr|\text{B will be released}) = \frac{1}{3},$$
$$p(Cr|\text{B will be executed}) = \frac{2}{3},$$

10.3 THE OP, CAREFULLY

Let's recall the principle: For a credence function p over an algebra S, for $E \subseteq S$ such that p(E)=1 and some proposition $H \subseteq S$ and value $v \in [0,1]$: if for any $e \in E$ were an agent to come to know it (and not to learn or forget anything else), then his or her credence in H would be v, then that agent ought to have a credence of v in H.

Recall also that the model $\langle X, S, p \rangle$ presented in section 1 appeared to show the principle didn't hold. In that model, there were two

possible sets E, viz. {WIN2, WIN3, WIN4} and {LOSE2, LOSE3, LOSE4}, but the agent's posterior credence in WIN1 given she came to learn any of the elements of the former was 1/3, whereas learning any of the elements of the former would increase her posterior to 2/3.

However, what the *3 Prisoners* (and its variations) teaches us is that one has to be very careful when modelling the information an agent comes to know. In particular, one has to take into account the explicit or implicit information-gathering process underlying the act of learning described in a scenario. With the information-gathering process clearly articulated, it is not longer the case that any proposition expressed in the algebra can come to be known. Only those which are stipulated by the information-gathering process can come the agent's way. What this shows is that OP is indeed ambiguous, but the ambiguity relates to how we construe 'coming to know': in the naive way which leads Prisoner A to worrying he may be hallucinating, or in a more sophisticated way, which delivers the correct posterior credences. The discussion of the *3 Prisoners* above should stand as evidence that we ought to follow the latter modelling strategy.

Now, carefully accounting for the information-gathering process turns the set of propositions on which the agent can conditionalise into a partition. This is easy to observe in relation to all the examples discussed in the previous section. Each element of that set is one of the possible items of information the agent could learn. This is implicit in Shafer's (1985, Appendix 1) formal model of protocols as trees and is discussed at length in Grünwald (2013). The latter also formulates a rule of thumb:

> Briefly, for general spaces $\mathcal{Y}$, if the set of events $\mathcal{X}$ on which you can condition is not a partition of $\mathcal{Y}$, then

> conditioning on any of these events is unsafe. (Grünwald, 2013, p. 243)

In consequence, when 'coming to know' is restricted to only refer to the propositions that accrue to the agent given the information-gathering process, OP becomes a theorem of the probability calculus. The proof is immediate and is also included in Mahtani (forthcoming, fn. 21). Mahtani, however, seems to believe the set of propositions you can learn can only form a partition under the super-know reading of OP and doesn't recognise its relevance for the basic-know reading under the information-gathering process.

With these conceptual clarifications let's see why the 4-2-1 Prisoners under Protocol 4 (the reason we focus on this example will become apparent in the next section) doesn't represent a counterexample to OP. According to Protocol 4, there are only 6 answers the guard could give Prisoner A, so only 6 viable propositions the agent could come to know and consequently, 6 possible candidates for elements for the set E stipulated in the formulation of OP. As we saw above, although the first three possible answers would all change A's beliefs with respect to his own fate in the same way, the latter three answers the guard could give him have a different impact on his credences. An easy verification will show that this holds of all possible propositions H, in the event space generated by that protocol. Therefore, the antecedent of OP is false for the case of 4-2-1 Prisoners so the principle continues to hold.

10.4 MAHTANI'S ARGUMENT, CAREFULLY

It is easy to observe the case of *4-2-1 Prisoners* is completely analogous to Mahtani's WSL: let the 4 prisoners stand for the four lottery tickets; to win the lottery is to be released, to lose the lottery is to be executed; and the guard's answer represents being informed

that a particular ticket is a winning or a losing one. Why then did Mahtani reach a different conclusion?

Mahtani makes two passes at analysing WSL. The first one is presented in section 1, above. And its mistake is the same as the intuitive analysis of the *3 Prisoners*: it ignores the information-gathering process and assumes learning that 'Ticket 2 is a winning ticket' is the same as conditionalising on WIN2.

Mahtani makes a second pass at analysing WSL, though, this time thinking explicitly about an information-gathering process analogous to Protocol 4 above: "the organizer selected a ticket from amongst tickets 2, 3 and 4 at random, and told you the outcome for that ticket regardless of whether it won or lost." (Mahtani, forthcoming, p. 21) Nevertheless she quickly argues that this wouldn't impact the result of her original analysis of WSL in a significant way. That casts a doubt over the previous section. Her argument is as follows:

> suppose that E is WIN3: similar reasoning applies if E is instead either WIN2 or WIN4. Then when WIN3 is revealed to you, you come to know both WIN3 (E), and (F) the fact that ticket 3 has been randomly selected from amongst tickets 2, 3 and 4 to be revealed to you. Thus your new credence in WIN1 should equal your prior credence conditional on this new evidence, i.e. $Cr(WIN1 | WIN3 \& F) = Cr(WIN1 \& WIN3 \& F)/Cr(WIN3 \& F)$. $Cr(WIN1 \& WIN3 \& F) = Cr(WIN1 \& WIN3) \times Cr(F)$,because F is independent of (WIN1&WIN3): which ticket was randomly selected to be revealed to you does not depend on which tickets have been selected to win the lottery. Thus $Cr(WIN1 \& WIN3 \& F) = Cr(WIN1 \& WIN3) \times Cr(F) = 1/6 \times$

> 1/3 = 1/18. Similarly Cr(WIN3&F) = Cr(WIN3)×Cr(F) = 1/2 × 1/3 = 1/6. Thus Cr(WIN1 | WIN3&F) = 1/18/1/6 = 1/3. (Mahtani, forthcoming, fn. 28)

If OP were to hold, the prior credence in WIN1, Mahtani argues, should drop to 1/3. The above reasoning is, however, incorrect. To see this begin by noticing that Mahtani's modelling language is not clear: she re-uses the names WIN1, WIN2, WIN3 and WIN4 without explaining what propositions they denote. One natural assumption would be to suppose they denote the same propositions as before (see the explanation of the model underlying her initial analysis of WSL in section 1). However in this passage Mahtani isn't analysing WSL, but a puzzle like WSL which also specifies a particular information-gathering process. The first glimpse we are in a different set-up is when Mahtani introduces the proposition F. But again she doesn't explain what its denotation is. In the next paragraph I show that there are two possible denotations for the names WIN1, WIN2, WIN3 and WIN4 Mahtani could be assuming in the above passage, but for both of them Mahtani's conclusion doesn't follow.

Let the probabilistic model $\langle X', S', p\rangle$ represent WSL with the information-gathering process presented above. The sample space should contain information about both the outcome of the lottery X={(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)} and of the ticket to be revealed, that is Y={F,F',F''}, where F is the proposition that ticket 3 is to be revealed, F' that ticket 2 will be revealed and finally, F'', that ticket 4 will be revealed. Notice the elements in Y cannot be expressed in terms of the elements in X - as Mahtani correctly points out, a good model of the scenario would make the choice of the ticket to be revealed independent of the winning tickets. Therefore, $X' \neq X$ (this is the reason why the denotation of WIN3, say, cannot remain the same as in the initial modelling of WSL without the

information-gathering process). One natural way of defining the sample space is $X' = X \times Y$ and we can then take $S' = 2^{X'}$. In this model "Ticket 3 is a winning ticket" is the proposition W3, that is {(1,3,F),(1,3,F′),(1,3,F″),(2,3,F),(2,3,F′),(2,3,F″),(3,4,F),(3,4,F′),(3,4,F″)}, whereas "Ticket 3 will be revealed" is $X \times F$. The probability function is defined in the intuitive way: $p(X \times F) = p(X \times F') = p(X \times F'') = \frac{1}{3}$ and $p(W1) = p(W2) = p(W3) = p(W4) = \frac{1}{2}$.

What is WIN3, say, in Mahtani's argument? The first possibility is that what she calls WIN3 is in fact W3, that is the proposition that ticket 3 is a winning ticket. If that is the case, however, upon learning it you don't "come to know both WIN3 (E), and (F) the fact that ticket 3 has been randomly selected from amongst tickets 2, 3 and 4 to be revealed to you". It is easy to check W3 is compatible with both F′ and F″ and, hence, on this interpretation Mahtani's argument doesn't follow. The second possibility is that WIN3 is $W3\&F$. In this case learning it will reveal that "ticket 3 has been randomly selected from amongst tickets 2, 3 and 4 to be revealed to you". However, E cannot simply be the set {WIN2, WIN3, WIN4} because in this model, the propositions $W2\&F$, $W3\&F$,$W4\&F$ do not exhaust the entire space. The world can be such that both W3 and F′: that ticket 3 is a winning ticket but nevertheless ticket 2 is randomly selected to be revealed. In other words, given the information-gathering process Mahtani specifies, the organiser could also announce that ticket 2 has lost. Nevertheless, the posterior credence in the proposition that ticket 1 is a winning ticket after learning that ticket 2 has lost, say, is $\frac{2}{3}$. So the antecedent of OP is still not satisfied.

10.5 CONCLUSION

To sum up, for conditionalization to be justified, one first needs to know what information could come one's way. In other words,

conditionalization cannot be the method used for incorporating new evidence unless you know before the new information comes what you can (and cannot) learn. Then, when you learn X, you also learn 'in some way' that you have learnt it (what Mahtani would probably call 'coming to super-know' X), but this means that you learn that out of the different pieces of information that could have been delivered, X was the one that actually was sent to you.

Without formally capturing (in the probability model) the information-gathering process, conditionalisation is not guaranteed to deliver the correct answer to what it means to learn a proposition. If a scenario doesn't contain any explicit reference about the information-gathering process - the scenario is under-specified, and if there isn't any salient way of filling in that gap or if we don't want to commit to any way of doing so, then you should be very wary of applying conditionalisation to model learning in that scenario. This is the lesson of the *Monty Hall*, the *3 Prisoners*, etc.

Part IV

RATIONALITY IN PRACTICE

11

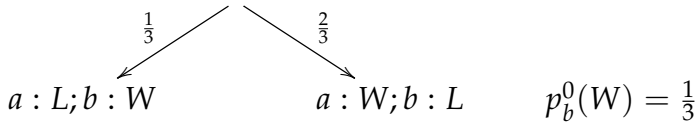
ON A DILEMMA OF REDISTRIBUTION

McKenzie Alexander (2013) presents a dilemma for a social planner who wants to correct an unfair distribution of an indivisible good between two equally worthy individuals or groups:

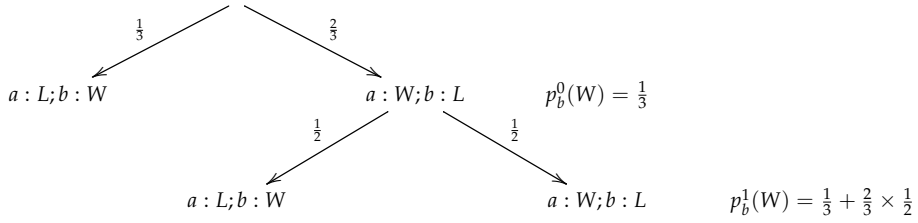
DILEMMA *Either* social planner guarantees a fair outcome, *or* she follows a fair procedure (but not both).

The argument is disconcertingly simple. Suppose the initial distribution is biased against b. If b nevertheless receives the good against all odds, as it were, it would seem unfair to take it away from her. However, if a receives the good then the social planner would want to intervene and redistribute. There are two strategies the social planner could follow when redistributing: the redistribution could be fair, offering equal chances to a and b of winning the good redistributed, or it could be unfair. McKenzie Alexander proves that if the social planner follows the former strategy, then *ex ante*, a and b have unequal chances of receiving the good. The procedure for redistributing is fair, but the outcome is that b is favoured (overall). On the other hand, if the social planner follows the latter strategy, then equal chances can be guaranteed *ex ante*, assuming the social planner chooses the appropriate biased lottery, but the redistribution would be biased against b. To wit, the social planner can either employ a fair redistribution procedure or guarantee a fair distributive mechanism *ex ante*. But not both!

In this paper I show that *Dilemma* only holds if the social planner can redistribute the good in question at most once. Consider the scenario that McKenzie Alexander discusses:

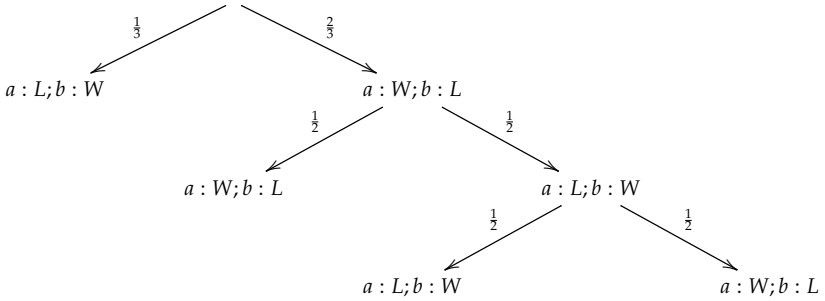


There are two ways the scenario can play out. Either a wins (W, and b loses, L) or b wins. The chance of a winning is $\frac{2}{3}$ and the chance of b winning is the complement, $\frac{1}{3}$. We assumed a and b are equally worthy and thus, following, they have an equal claim to the good (Broome, 1990). Since b's chance of receiving the good is less than $\frac{1}{2}$ as a result of this distribution, it means she is aggrieved. If the b nevertheless wins, Mckenzie Alexander contends the social planner should refrain from interfering. Taking the good away from b would be like punishing her for making it despite the odds which were stacked against her. So the social planner should only interfere when a wins this distribution. Assume then she follows the first strategy outlined above.



At the level of the redistribution both a and b are given an equal chance of winning the good by the social planner. This is in line with their (equal) claim. However, if this is how the social planner interferes, the redistributive mechanism he thus creates awards b an *ex ante* higher chance of winning the good than her claim, $p_b^1(W) = \frac{2}{3}$. If we assume that a had no doing in the initial bias in his favour, we have a strong intuition this set-up is unfair. This is the first horn of

Dilemma. By redistributing fairly, i.e. according to the claims of the two individuals involved, the social planner generated an *ex ante* unfair mechanism. Can the social planner do anything to correct this *ex ante* unfairness? The answer is YES. He can redistribute once again if the individual aggrieved by the last redistribution performed does not win the good. In this case, after the first redistribution, a is left with a chance of winning less than his claim. So whenever a loses the good, the social planner seems entitled to offer him another, fair chance.



$$p_b^0(W) = \frac{1}{3} < \frac{1}{2}$$
$$p_b^1(W) = \frac{1}{3} + \frac{2}{3} \times \frac{1}{2} = \frac{2}{3} > \frac{1}{2}$$
$$p_b^2(W) = \frac{1}{3} + \frac{2}{3} \times \frac{1}{2^2} = \frac{1}{2} = \frac{1}{2}$$

Evaluate the situation after the first redistribution: a is now aggrieved since a's *ex ante* chance of winning the good, $p_a^1(W) = \frac{1}{3} < \frac{1}{2}$. Therefore the social planner can redistribute again whenever a loses the redistribution. After the second redistribution, both individuals a and b now have equal *ex ante* chances of winning the good which is being distributed and hence there is no need for the social planner to correct when one of them loses. This is good news, but notice that the analysis was dependent on the initial bias. It worked for $p = \frac{1}{3}$. Does the solution work for all initial biases (for all unfair

distributions)? The answer is again YES. Figure 4 tracks how the *ex ante* chances of winning the good evolve over 10 redistributions for values of the initial bias between 0 and $\frac{1}{2}$ in 0.01 increments. A formal proof is provided at the end.

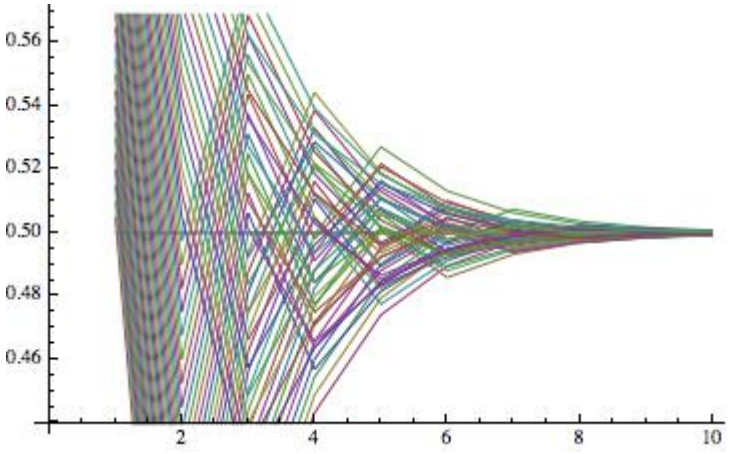


Figure 4: The evolution of the *ex ante* chances of winning the good over 10 redistributions for values of the initial bias between 0 and 1/2 in 0.01 increments

McKenzie Alexander writes that

> [s]ometimes the correct response to an injustice generated by an unfair decision procedure is to use another unfair decision procedure, which appears to disadvantage (in some sense) the same person again. In these cases, *two wrongs do make a right*.(McKenzie Alexander, 2013, p. 230, my emphasis)

The result of this paper then is that the bias against an aggrieved individual always washes out when we allow for sufficiently many redistributions. In other words, we do not have to make a second wrong in order to make right by the aggrieved: at most infinitely many rights will do. This is encouraging. But even if it is always the

case that a social planner can correct an initial unfair distribution by behaving fairly towards both the aggrieved and the favoured, no social planner has infinite time and resources. Can anything better be done for real social planners? The answer is one last time YES.

I contend it is unproblematic to assume people are not sensitive to minute differences in probabilities. Then let the sensitivity of the most sensitive member of the two person society we are concerned about in this paper be δ. I investigated two possible values for δ: $\delta_1 = 0.001$ and $\delta_2 = 0.01$. Under δ_1 the individuals in the society cannot tell a .500 chance of winning the good apart from a .5001 chance. Under δ_2 they cannot tell apart a .50 from a .51 chance of winning the good. It turns out that for δ_1 it takes *at most* nine redistributions for the probability of winning for b to reach the interval $[.499, .501]$ and hence become identical to $\frac{1}{2}$. For δ_2, the probability of winning for b reaches a value in $[.49, .51]$ in *at most* six redistributions.[1] The result is interesting as it tells us that no matter what the bias of an initial distribution is, it is always possible for a social planner to offer the two participants to the distribution equal *ex ante* chances of winning the good in at most six fair redistributions (assuming that the most sensitive of the aggrieved and the favoured of the original distribution has sensitivity δ_2). This is something that a real social planner can actually make use of.

In conclusion, contrary to McKenzie Alexander's point, there is no tension between procedural and outcome fairness as long as the social planner is given the opportunity to redistribute sufficiently many times. It may be the case that "sometimes... two wrongs make a right" but so do *a wrong and infinitely many rights*. And in

1 This resulted by testing all values of the initial bias, p, between 0 and $\frac{1}{2}$ in 0.01 increments in Mathematica 8. Notebooks used are available upon request.

fact, a wrong and *sufficiently* many rights (depending on p and δ) are right *enough*.

PROOF OF MAIN RESULT

Let x_i stand for the probability of the aggrieved of the initial distribution winning after the i^{th} redistribution,

$$\begin{aligned} x_1 &= p + (1-p)\frac{1}{2} \\ x_n &= x_{n-1} + (1-p)\frac{1}{2^n} \text{ iff } x_{n-1} < \frac{1}{2} \\ &= x_{n-1} - (1-p)\frac{1}{2^n} \text{ iff } x_{n-1} > \frac{1}{2} \\ &= \frac{1}{2} \text{ iff } x_{n-1} = \frac{1}{2} \end{aligned}$$

We can now formally state the main result of this paper: $lim_{n\to\infty}(x_n) = \frac{1}{2}$. Let the following sequences stand for the elements in (x_n) less than $\frac{1}{2}$ and greater than $\frac{1}{2}$, respectively:

$$\begin{aligned} (a_n) &= \{a \in (x_n) : a < \frac{1}{2}\} \\ (b_n) &= \{b \in (x_n) : b > \frac{1}{2}\} \end{aligned}$$

In order to prove Theorem 1 it is enough to prove that the limit of both (a_n) and (b_n) is $\frac{1}{2}$. The proofs are symmetrical and we will only show the proof for (b_n). We will first show that (b_n) is decreasing and then (by the Squeezing Theorem) that its limit is indeed $\frac{1}{2}$.
Take $b_m \in (b_n)$. By construction, $b_m \in (x_n)$. Suppose it corresponds to element $x_n \in (x_n)$. Remark that m may differ from n. Since $b_m > \frac{1}{2}, x_n > \frac{1}{2}$. Therefore $x_{n+1} = x_n - (1-p)\frac{1}{2^{n+1}}$. If $x_{n+1} > \frac{1}{2}$ then $x_{n+1} = b_{m+1}$ if not, $x_{n+2} = x_n - (1-p)\frac{1}{2^{n+1}} + (1-p)\frac{1}{2^{n+2}}$ and so

on. Therefore, depending on the value of p, $b_{m+1} = x_{n+k_n}$, for some natural number k_n.[2]

$$\begin{aligned} b_{m+1} &= x_n - (1-p)\frac{1}{2^{n+1}} + (1-p)\frac{1}{2^{n+2}} + \cdots + (1-p)\frac{1}{2^{n+k_n}} \\ &= x_n - (1-p)(\frac{1}{2^{n+1}} - \frac{1}{2^{n+2}} - \cdots - \frac{1}{2^{n+k_n}}) \\ &= x_n - (1-p)\frac{1}{2^{n+k_n}} \end{aligned}$$

In consequence,

$$\begin{aligned} b_m - b_{m+1} &= x_n - x_n + (1-p)\frac{1}{2^{n+k_n}} \\ &= (1-p)\frac{1}{2^{n+k_n}} > 0 \end{aligned}$$

This concludes the proof that (b_m) is a decreasing sequence. In order to show that the limit of all elements in (x_n) greater than $\frac{1}{2}$ when $n \to \infty$ is $\frac{1}{2}$ it is enough to show that

$$\frac{1}{2} - \frac{1}{2^n} \leq (x_n)_{x_n \geq \frac{1}{2}} \leq \frac{1}{2} + \frac{1}{2^{n+1+k_{n+1}}}$$

If this is the case, by the Squeeze Theorem

$$lim_{n\to\infty}(x_n)_{x_n \geq \frac{1}{2}} = lim_{n\to\infty}(\frac{1}{2} - \frac{1}{2^n}) = lim_{n\to\infty}(\frac{1}{2} + \frac{1}{2^{n+1+k_n}}) = \frac{1}{2}$$

The first inequality obviously holds since all elements of the sequence $(\frac{1}{2} - \frac{1}{2^n})$ are at most $\frac{1}{2}$. And all elements of $(x_n)_{x_n \geq \frac{1}{2}} \geq \frac{1}{2}$, by construction. Then we only need to check the second inequality. We do this by induction:

$$x_1 = p + (1-p)\frac{1}{2} \leq \frac{1}{2} + \frac{1}{2^{1+1+k_1}}$$

2 k_n has to be at least 1, in which case both x_n and x_{n+1} are greater than $\frac{1}{2}$; and $k_{n+1} \geq k_n$

Since k_1 is 0 (as $x_1 \geq \frac{1}{2}$ for all values of p), the right hand side of the inequality will equal $\frac{3}{4}$ which is the highest value (x_n) reaches:

$$\begin{aligned} x_n &\leq \frac{1}{2} + \frac{1}{2^{n+1+k_n}} \\ x_n - (1-p)\frac{1}{2^{n+k_n}} &\leq \frac{1}{2} + \frac{1}{2^{n+1+k_n}} - (1-p)\frac{1}{2^{n+k_n}} \\ x_{n+k} &\leq \frac{1}{2} + \frac{1}{2^{n+1+k_n}} - (1-p)\frac{1}{2^{n+k_n}} \end{aligned}$$

What the induction aims to establish is that $x_{n+k_n} \leq \frac{1}{2} + \frac{1}{2^{n+2+k_{n+1}}}$. So we need to show that (the following reasoning steps are all equivalent):

$$\begin{aligned} \frac{1}{2} + \frac{1}{2^{n+1+k_n}} - (1-p)\frac{1}{2^{n+k_n}} &\leq \frac{1}{2} + \frac{1}{2^{n+2+k_{n+1}}} \\ \frac{1}{2^{n+1+k_n}} - \frac{1}{2^{n+2+k_{n+1}}} &\leq (1-p)\frac{1}{2^{n+k_n}} \\ 2^{1+k_{n+1}-k_n} - 1 &\leq (1-p)2^{2+k_{n+1}-k_n} \\ 2^{1+k_{n+1}-k_n} - (1-p)2^{2+k_{n+1}-k_n} &\leq 1 \\ 2^{1+k_{n+1}-k_n}(1-2+p) &\leq 1 \\ 2^{1+k_{n+1}-k_n}(p-1) &\leq 1 \\ \text{But } p-1 < 0 \text{ for all values of } p & \end{aligned}$$

Therefore, for all values of p, all n: $x_n \leq \frac{1}{2} + \frac{1}{2^{n+1+k_n}}$. This concludes the proof.

12

AN EFFICIENCY ARGUMENT FOR UNISEX RESTROOMS

In this paper we construct a model to evaluate the waiting times for unisex vs. gender-segregated restrooms and show by means of simulations that, given certain plausible assumptions, the unisex set-up provides drastic reductions in the total waiting times. This translates into greater productivity at lower overhead costs in terms of estates for firms. The move to unisex restrooms will indirectly benefit members of the trans*[1] community, carers of people with disabilities, as well as parents and children of different genders. Moreover it will contribute towards increased potty parity.

In the past year we have witnessed a heated public debate regarding whether restroom access should be linked to one's sex at birth. But the debate goes at least back to the 70s: according to Mary Anne Case, part of the opposition to the Equal Rights Amendment was (at least) publicly justified by a warning that "passage of the ERA would mean a mandatory end to restrooms segregated by sex. Leaflets urging voters to reject the [Equal Rights Act] even claimed it was 'also known as the Common Toilet Law.'" (Case, 2010, p. 1) The debate has now been rekindled and focused on how our society should treat its members who identify as trans*.

1 We follow Seelman (2016) and use "*trans** to denote inclusion of a broader range of gender non-conforming identities, including those who may not use the term *transgender* for themselves." (Seelman, 2016, fn. 1)

House Bill 2 (HB2) in North Carolina[2] and *G.G. v. Gloucester County School Board*[3] have provided the crucible for debating this question. Discussions surrounding these two cases have largely been carried out in terms of justice and safety: As to justice, there is an issue of minority rights: Should the needs of a small minority be protected against the preferences of a large majority for sex-segregated bathrooms? As to safety, the question is whether accommodating the trans* minority would bring about a decrease in security of other vulnerable groups, viz. girls and women.

Both ways of framing the debate tend to engage people's moral and political intuitions from widely divergent sides of the political and socioeconomic spectrum and lead to a very polarized arena with participants becoming more and more entrenched in their positions. Moreover, each party feels that the other is disrespectful of their deeply held beliefs, making dialogue difficult. To make some headway in this debate, we propose a different way of framing the issue, one we hope will provide some common ground among all groups.

Our idea is simple: de-segregate restrooms and transform all restrooms into unisex facilities. The move to unisex restrooms increases economic efficiency in two ways. Firstly, preserving the current architecture but opening all stalls to everyone, it leads to a drastic reduction in waiting times. Secondly, calculating the maximum waiting time allowed by current legislation for the segregated set-up we can calculate how many toilets we can eliminate in a unisex set-up so that the waiting times are still under the current maximum threshold. Since everyone cares about economic efficiency,

2 http://www.ncleg.net/Sessions/2015E2/Bills/House/PDF/H2v4.pdf

3 http://www.scotusblog.com/case-files/cases/gloucester-county-school-board-v-g-g/

there is at least one argument in favour of unisex restrooms that all can agree on.

12.1 BACKGROUND

In the past year we have seen a surge in legislation, public debate and scandals surrounding the way we use public restrooms. In this section we very briefly try to take stock of what has happened in the US, as well as in Canada and the UK.

12.1.1 *United States*

In the 2017 legislative session, sixteen states "have considered legislation that would restrict access to multiuser restrooms, locker rooms, and other sex-segregated facilities on the basis of a definition of sex or gender consistent with sex assigned at birth or 'biological sex.'"[4] Six states have "considered legislation that would preempt municipal and county-level anti-discrimination laws,"[5] and fourteen states "have considered legislation that would limit transgender students' rights at school."[6] Nevertheless, North Carolina is the only state so far to pass a bathroom bill, i.e. the now repealed HB2.

The backlash against HB2 has been impressive. Associated Press tracked businesses that have changed their plans of investing in the State as a consequence of the bill, such as a PayPal, and entertainment events that were cancelled or postponed, such as a Ringo Star concert, the NCAA games, etc. and concluded that it would cost North Carolina an estimated $3.76 billion in lost business over

4 `http://www.ncsl.org/research/education/-bathroom-bill-legislative-tracking635951130.aspx`
5 *Idem.*
6 *Idem.*

a dozen year period if the bill were to stay in place.[7] Under the mounting pressure, the state legislature repealed and replaced HB2 by HB142[8] on 30 March 2017. Although the provisions of HB142 are themselves highly controversial (for good reasons), the bill effectively stopped the boycott of the state.

12.1.2 *United Kingdom*

In the United Kingdom, the public discourse around the issue of unisex restrooms is less heated. Firstly, more and more schools are introducing unisex facilities,[9] and with few exceptions,[10] the move seems to be unopposed.

Secondly, the Barbican Centre in London changed a pair of its restrooms into 'gender-neutral with urinals' and 'gender-neutral with cubicle' restrooms. In other words they have just converted existing men's and women's restrooms into 'gender-neutral' ones. While the move was received with some opposition,[11] the issue with the decision involved the resulting inequality between men and women in terms of access to facilities. It was not that men could enter the women's restroom that infuriated patrons, but that they would now have more access opportunities both in their old restroom (now gender-neutral with urinals) and in the women's restroom (now gender-neutral without urinals), whereas very few

7 https://apnews.com/fa4528580f3e4a01bb68bcb272f1f0f8

8 http://www.ncleg.net/Sessions/2017/Bills/House/HTML/H142v5.html

9 Several award-winning school designs include unisex facilities for the pupils. See for instance, pp. 22-23 of http://www.parliament.scot/S4_FinanceCommittee/Meeting%20Papers/2016_01_18_Public_Papers(1).pdf.

10 http://www.independent.co.uk/news/uk/home-news/unisex-school-toilets-gender-neutral-london-inclusive-bathrooms-lgbt-same-sex-a7441841.html

11 https://www.thetimes.co.uk/article/women-queue-up-to-condemn-arts-centres-unisex-lavatories-jq9mswjsp

women would have taken advantage of the possibility to enter the 'gender-neutral' with urinals restroom.

Finally, during a parliamentary debate on December 2016 on transgender equality, Caroline Flint (Labour MP for Don Valley and Shadow Secretary of State for Energy and Climate Change) stated that

> I welcome the debate, because it is vital for us to consider the issue of transgender rights, but should we not also be wary of creating gender-neutral environments that may prove more of a risk to women themselves? A recent case involving my old university, the University of East Anglia, which has gender-neutral toilets, revealed that a man had been using those facilities to harass women. He was charged and convicted. How does the right hon. Lady [Mrs. Maria Miller, the Conservative MP for Basingstoke] think we can protect women from male violence in gender-neutral environments?[12]

12.1.3 *Canada*

Finally, the discussion on restrooms in Canada has revolved around the seemingly innocuous Bill C-16. The proposal is to "[amend] the Canadian Human Rights Act to add gender identity and gender expression to the list of prohibited grounds of discrimination."[13] Activists opposing the bill have argued that adopting it would soon lead to opening women's facilities to anyone who declares to identify as such and this, in turn, will lead to an increase in violence

12 https://hansard.parliament.uk/Commons/2016-12-01/debates/D4F283FB-2C02-4C8C-8C7E-BEAB889D1425/TransgenderEquality

13 https://openparliament.ca/bills/42-1/C-16/?tab=mentions

against women.[14] Interpreting the bill as a first step towards giving people who identify as trans* the choice of which restroom to use has some support especially given that a previous attempt to pass this bill in 2015 failed due to an amendment that explicitly exempted restrooms and changing rooms.[15] So the opinion of the Canadian legislative seems to be that people who identify as trans* should have their right to choose the restroom they use protected by law. Despite the dissent, Bill C-19 was voted into law on June 19, 2017.

12.2 THREE PROBLEMS REGARDING ACCESS TO PUBLIC FACILITIES

In the past year the media has focused almost entirely on how bathroom bills impact the members of the trans* community. And justly so since the laws have been directly targeted at them. But, nevertheless, they are not the only group affected. A second group further marginalized by defining bathroom access in terms of one's sex or gender is that of people with a disability who require a carer. In many circumstances the carer and the person they are caring for have different genders and making a choice as to which facilities to use raises significant issues. Finally, this paper will also discuss a problem related to bathroom access which is not directly connected to the recent legislation, viz. the differential access to restrooms for women vs. men. I expand on all three issues below and explore some of the solutions that have been offered on both sides of the ideological spectrum.

14 http://womanmeanssomething.com/

15 https://www.theguardian.com/society/2016/sep/08/canadians-support-transgender-rights-poll

12.2.1 *Access trans**

According to the largest survey of the experiences of trans* people in the US (James et al., 2016), 59% of respondents (out of 27,715 surveyed) sometimes refrained from using a restroom outside of their home in the previous year. One of the main rationale given was fear of confrontations. The same survey also found that 24% were asked in the previous year at least once whether they were in the right restroom and 9% were deliberately denied or stopped from using a restroom over the same period of time. Finally, and more worryingly, 12% of respondents were "verbally harassed, physically attacked, and/or sexual assaulted when accessing or while using a restroom in the past year", 32% refrained from drinking or eating and 8% developed a urinary-tract infection or other kidney-related problems (due to refraining from using the restroom). Therefore, access to bathrooms represents not only an encroachment on this minority's rights and liberties, but poses a public health risk as well. A solution is required and in recent years we have witnessed both conservative and liberal attempts to tackle the problem. The former involves creating 'special' facilities for trans* individuals (which may take the form of unisex single-stall facilities). The latter, allowing all individual access to the facility matching the gender they identify with (and expanding the Civil Rights Act to protect this right against discrimination).

The best know example of the Conservative strategy to tackle the problem of bathroom access for members of the trans* community is that of Gavin Grimm. At the beginning of his sophomore year in 2014, Gavin informed his school of his intention to begin transitioning in all aspects of his life and with his principals's approval began using the men's restrooms.[16] Nevertheless, upon receiving

16 https://www.washingtonpost.com/local/education/gavin-grimm-just-wanted-to-use-the-bathroom-he-didnt-think-the-

complaints from some some of the parents the Gloucester School Board voted 6-1 on 9 December 2014 to adopt a policy limiting access to restrooms "to the corresponding biological genders"[17] and requiring "students with 'gender identity issues' to use an alternative private facility."[18] Gavin and the ACLU fought the decision up to the Supreme Court, but before the latter could rule, the case was sent back to the Fourth Circuit Court of Appeals "to be reconsidered in light of the Departments of Justice and Education rescinding of a Title IX guidance clarifying protections for transgender students."[19]

The best known liberal solution to tackling the problem of restroom access for trans* individuals is the controversial Charlotte City Council Ordinance 7056, passed on February 22, 2016, and extending the list of protected characteristics included in the City Code to cover 'sexual orientation', 'gender identity' and 'gender expression'.[20] Moreover, the Ordinance removed Section 12-59 of the City Code that was allowing for sex-based discrimination in access to restrooms, shower rooms, dormitories, etc. The Ordinance took effect on 1 April 2016 and generated ample public debate especially after the State legislature passed HB2 and in effect repealed it.

The advantage of the conservative solution is that it offers secure access to restroom facilities for members of the trans* community

---

nation-would-debate-it/2016/08/30/23fc9892-6a26-11e6-ba32-5a4bf5aad4fa_story.html?utm_term=.15eacd85a3c3

17 https://www.nbcnews.com/news/us-news/u-s-supreme-court-rejects-transgender-rights-case-n729556

18 https://www.nytimes.com/2016/04/20/us/appeals-court-favors-transgender-student-in-virginia-restroom-case.html

19 https://www.aclu.org/cases/gg-v-gloucester-county-school-board

20 http://charlottenc.gov/NonDiscrimination/Documents/NDO%20Ordinance%207056.pdf#search=ordinance%207056

by offering them separate facilities. However, at the same time it denies their claim to transition socially and be *recognised* as having the gender they wish to be identified with. On the other hand, while the liberal solution does offer them recognition and legal protection, it doesn't guarantee their safety in public restrooms (at least not by itself). This problem is even more acute as it affects several members of the trans* community more than others:

> Transgender men (75%) were far more likely to report sometimes or always avoiding using a public restroom, in contrast to transgender women (53%) and non-binary respondents (53%). Undocumented residents were also more likely to report sometimes or always avoiding using a public restroom in the past year (72%). Eighty percent (80%) of respondents who said that others could always or usually tell that they were transgender and 72% of those who said that others can sometimes tell they are transgender reported avoiding using public restroom, in contrast to 48% of those who said that others can rarely or never tell that they are transgender (James et al., 2016)

In other words, those who can 'pass' more easily are less affected by the laws of urinary segregation as they have fewer problems inside restrooms than those for whom it is more difficult to pass as the gender they identify with (if any).

12.2.2 *Access for people with disabilities*

An unintended target of the Bathroom Wars generated by the situation in North Carolina has been the group of people with disabilities, including physical and developmental disabilities, who sometimes cannot safely use the restroom without assistance. According to Sam Crane, the legal director and director of public

policy for the nonprofit Autistic Self Advocacy Network, "[o]ften, a person's assistant will be someone of a different gender."[21] Indeed, according to Paraprofessional Healthcare Institute, 89% of personal care attendants are female, while according to the U.S. Census Bureau, around 17% of men and 20% of women have a disability.[22] Therefore, assuming that an equal percentage of disabled men and women are in need of a carer, it is quite likely for men with disabilities to have female carers. And in these cases, bills such as HB2 pose a significant threat to how they access public restrooms.

Despite HB2 being silent on this issue, conservatives have recently become more aware of this concern. While their approach is to continue limiting access to restrooms based on one's 'biological sex',[23] some of their proposals provide exceptions to this requirement for the case of carers (and sometimes parents accompanying children). For instance, out of the 16 states considering legislation restricting access to multiuser restrooms, Texas's Senate Bill 6, Virginia's House Bill 1612 and Washington's House Bill 1011 explicitly made exceptions.[24] The latter stipulates:

> Nothing in this section prevents a minor child or a person with a disability from entering a facility segregated by gender when the child or person is a different gender from the gender for which the facility is segregated if: (a) A parent, guardian, supervisor, or caretaker is escorting the minor child or the person with a disability to or from

21 https://www.washingtonpost.com/news/parenting/wp/2017/05/16/why-parents-of-kids-with-special-needs-are-fighting-bathroom-bills/?utm_term=.1e9bc6cc222d

22 https://www.paraquad.org/blog/bathroom-bills-affect-people-with-disabilities/

23 Different state laws have different provisions, but many define biological sex as the sex assigned at birth and recorded in one's birth certificate.

24 http://www.ncsl.org/research/education/-bathroom-bill-legislative-tracking635951130.aspx

> the facility, (b) the child or person is under the custody, control, supervision, or care of the parent, guardian, supervisor, or caretaker, and (c) the gender of the parent, guardian, supervisor, or caretaker is the same as the gender for which the facility is segregated.[25]

As of September 2017, out of the three only Washington's HB1011 is still being considered.

12.2.3 *Access for women*

It is a well-established fact that during the intermission of any kind of event a long queue forms for the women's restrooms, whereas men walz in and out of their restrooms. This is taken by many as evidence of an inherent inequality in the legislation on (and architecture of) public venues. The claim that the situation should change by bringing about parity not in terms of number of toilets, or square meters, but waiting time, is known as the demand for potty parity.

One of the cases that galvanised the public debate in the US is the case of a woman who in 1990 was fined for causing a disturbance at a Houston country-western concert by going into a men's restroom. Her decision was motivated by the long queue for the women's restroom. The Municipal Court acquitted her, but the media dubbed the public debate that ensues 'Pottygate.'[26]

The US federal standards, viz. regulation 1910.141(c)(1)(i) of the US Department of Labor's Occupational Safety Health Admin-

25 http://lawfilesext.leg.wa.gov/biennium/2017-18/Pdf/Bills/House%20Bills/1011.pdf

26 http://www.nytimes.com/1990/11/03/us/woman-is-acquitted-in-trial-for-using-the-men-s-room.html

istration,[27] aim to ensure equality of facilities and set the same minimum number of facilities for both men and women. But what this usually means is that men receive many more facilities: although men's restrooms may have fewer stalls than women's, they manage to fit in more urinals (less wall-space and surface) and thus have more access opportunities. Moreover, research suggests that women take longer to use the restroom (Baillie et al., 2009) and they do so more often. All of these differences compound into creating waiting time inequality in public venues. We will discuss these matters in more depth below.

A natural answer to this problem is to increase the number of facilities mandated by law. For instance, in 2005 New York City Council passed the Women's Restroom Equity Bill, requiring all new establishments to observe a ratio of 2:1 on women stalls vs. men stalls *and* urinals.[28] Already existing establishments are required to comply with this when they undergo extensive renovation. Although this kind of legisaltion is welcomed, it still falls short of being a solution for two reasons: 1) its effects will take a long time to be seen given the need of current establishments to go through significant architectural change; and 2) it is not obvious that all relevant establishments will be able to comply. Many theatres and concert halls are housed in listed buildings thus making any structural renovations subject to an impenetrable barrier of administrative checks and approvals.

27 https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=STANDARDS&p_id=9790#1910.141(c)(1)(i)

28 http://www.nytimes.com/2005/05/26/nyregion/council-passes-a-bill-to-shorten-the-line-at-the-ladies-room.html

12.3 ONE SOLUTION TO THREE PROBLEMS: UNISEX RESTROOMS

In this paper we propose a solution that would contribute towards (partially) solving all three of the above problems: shifting from gender segregated restrooms to unisex restrooms. With respect to the first problem, unisex facilities eliminate the need to make a decision as to which restroom to use and also eliminate the gender-expectation of the patrons using that facility. With regards to the second problem, unisex restrooms allow caregivers (and parents) access to the facilities appropriate for the persons they are assisting. Finally, with regards to the third problem, shifting to unisex restrooms increases the number of facilities available to women. As such they ensure parity for all patrons and (for certain assumptions to be discussed below) the parity does not come at the cost of leveling-down.

This idea is not new and both academic articles and op-ed pieces have questioned the motivation of segregated facilities.[29] For instance, Terry Kogan writes that:

> Some argue that one solution is to convert all public restrooms to unisex use, thereby eliminating the need to even consider a patron's sex. This might strike some as bizarre or drastic. Many assume that separating restrooms based on a person's biological sex is the "natural" way to determine who should and should not be permitted to use these public spaces.[30]

The author of this piece in *Slate* asks:

29 For a history of the laws of urinary segregation see Kogan (2007).

30 https://www.theguardian.com/commentisfree/2016/jun/11/gender-bathrooms-transgender-men-women-restrooms

> Why is the bathroom seen as an untouchable, unchangeable safe space? Naturally, everyone wants to be comfortable when taking care of bathroom business, but how is a restroom different than other public spaces in which people want to be left alone? Is it simply a social construction? If comfort is the main concern, why is the comfort of some people privileged over that of others? And are we comfortable with that?[31]

And Kathryn Anthony and Meghan Dufresne observe that:

> Gender-segregated restrooms no longer work for a significant part of the population. Yet, family-friendly or companion-care restrooms that allow males and females to accompany each other, as well as unisex restrooms, are still all too rare. (Anthony and Dufresne, 2007, p. 268)

Finally, Case writes that

> Basic queuing theory confirms that making fully enclosed single user facilities available to either sex on demand, as airplane toilets are, would cut down on overall waiting times and promote the most efficient use of available toilet facilities. Case (2010, p. 7)

Despite all of this discussion, this paper represents as far we are aware, the first attempt at gauging the impact of such a policy in terms of the economic advantage and efficiency it would bring about. To do so, the paper sets itself the task of answering the following questions:

QUESTION 1 How much time do women waste queuing for female facilities?

31 http://www.slate.com/blogs/outward/2013/12/26/gender_neutral_bathrooms_all_bathrooms_should_be_open_to_all_users.html

QUESTION 2 How much time do men waste queuing for male facilities?

QUESTION 3 How much time would people waste queuing for facilities if they were to be open to all genders?

In the next section we present the method by which we attempted to answer these three question and in the following section we discuss the answers we came up with.

12.4 METHODOLOGY

According to US Department of Labor's Standard 29 CFR 1910.141(c)(1)(i) and the interpretation thereof employers have to allow their employees as many restroom breaks as they require.[32] Therefore the time they spend in the restroom is included in their work contract and a company needs to allow for less productivity during a day on account of restroom breaks.[33] However, the same does not hold for the time spent queuing in front of the restroom. This is something an employer does have the legal right to try to minimize. We can calculate the total waiting time incurred by waiting to use a restroom by adding the waiting time of all

32 The interpretation makes it explicit that the standard "requires employers to make toilet facilities available so that employees can use them when they need to do so. The employer may not impose unreasonable restrictions on employee use of the facilities." https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=interpretations&p_id=22932

33 See though *Zwiebel v. Plastipak Packaging* in which the court's decision stated that "[w]hile there is a clear public policy in favor of allowing employees access to workplace restrooms, it does not support the proposition that employees may leave their tasks or stations at any time without responsibly making sure that production is not jeopardized. In recognition of an employer's legitimate interest in avoiding disruptions, there is also a clear public policy in favor of allowing reasonable restrictions on employees' access to the restrooms." https://scholar.google.com/scholar_case?case=5386413555521481804

employees who require a restroom break. Less waiting time would translate into a more efficient set-up which would allow for more productivity. In this chapter we argue that the answers to Questions 1 to 3 above will show employers that it makes business sense to introduce unisex restrooms

In order to answer Questions 1,2 and 3 above, we first need to answer some preliminary questions. The answers to these questions will determine the parameters of the model:

What is the time interval over which we evaluate the impact of the different set-ups (segregated vs. unisex)? For the purposes of this model we assumed this to be 120 time points, corresponding to 120 minutes. Our interpretation is that this time period spans the period over which every employee has to use the bathroom (no more and no less than) once. Here is the motivation behind this. It is usually assumed that people make between 6 and 8 'visits' to the restroom a day. Assuming 8 hours of sleep, 8 hours of work and 8 hours of 'rest', we stipulated that during a work day an employee will visit the restroom 4 times on average. Therefore, each employee is assumed to visit the restroom once per every 2-hour block during the workday. We assume these visits happen 'at random' within each 2-hour block.

How long does a bathroom visit last? There is very little evidence on the average time people spend in the restroom, although there is ample anecdotal evidence of employers complaining about employees taking 'too many' restroom breaks and women 'taking longer'. For the purposes of this model we assume the results of Baillie et al. (2009). They tracked 120 college students using public restrooms in a library and found that women take on average 178.9 sec. while men take 118.4 sec. We round these values to 3 time

points (minutes) for women and 2 time points (minutes) for men. Note that other studies arrived at different results. For instance, (Case, 2010, p. 3) cites a study done by a Cornell student suggesting that women spend on average 79 seconds, whereas men only spend 45 seconds in the restroom. Our model is flexible enough to accommodate other values, but we will not do so in the present paper.

One thing to note about all the studies on this matter is that they only measure the time spent in the restroom from going in to coming out. There is no analysis of what aspect of the restroom experience causes the difference in timing. Usual explanations include: behaviour (women use the bathroom for more than just bodily function), clothing (women's clothing is less efficient than men's) and, finally, architectural ('navigating' a stall is more time consuming than 'navigating' a urinal). In any case, all studies suggest a slightly longer time for women than for men.

What is the number of restrooms available for each gender? In order to model the number of bathroom stalls available to each gender we followed regulation 1910.141(c)(1)(i) of the US Department of Labor's Occupational Safety and Health Administration which states that:

> Except as otherwise indicated in this paragraph (c)(1)(i), toilet facilities, in toilet rooms separate for each sex, shall be provided in all places of employment in accordance with table J-1 of this section. The number of facilities to be provided for each sex shall be based on the number of employees of that sex for whom the facilities are furnished. Where toilet rooms will be occupied by no more than one person at a time, can be locked from the inside, and contain at least one water closet, separate toilet rooms for each sex need not be provided. Where such

> single-occupancy rooms have more than one toilet facility, only one such facility in each toilet room shall be counted for the purpose of Table J-1 [see Table 44, below].

Number of employees of each sex	Minimum number of toilets per sex
1 to 15	1
16 to 35	2
36 to 55	3
56 to 80	4
81 to 110	5
111 to 150	6
Over 150	1 additional one for each 40 employees

Table 44: Table J-1

How about the use of urinals? The presence of urinals is usually assumed to be responsible for the shorter average time men spend in the restroom as opposed to women. For the purposes of this model we assumed no urinals are installed - they are not required by law, though they are permitted. In the case of fewer than 35 employees this assumption is legitimate as the above regulation contains the following clarification:

> When toilets will only be used by men, urinals may be provided instead of toilets, except that the number of toilets in such cases shall not be reduced to less than two-thirds of the minimum specified.

This means that for fewer than 35 employees, no urinals can be installed to replace some of the required stalls (as there would be two or fewer stalls). For simplicity, we also assume the absence of urinals also for larger companies.

How many stalls do employers install? We assume that employers never install more than the minimum required number of stalls.

What is the ratio of men to women? We assume there are as many women as men.

When do the employees need to use the restroom? For each employee we select an arbitrary integer between 1 and 120 under a uniform distribution.

How do we ensure robustness? For each number of employees we simulate both set-ups (segregated vs. unisex) 10,000 times. The outcome reported is the average of all the 10,000 values obtained for each set-up and each number of employees.

Finally, here is how the algorithm works. Assume we are in a situation for which the above legislation requires n facilities. The first n arrivals incur 0 waiting time. Men will occupy the facility for 2 minutes, women for 3. So for all subsequent arrivals we can track the time they have waited before being able to occupy the facility. If the (n+1)-th arrival happens very soon after the n-th, they may find all facilities occupied. Then they will wait until the employee who arrived first leaves. The number of minutes in which the employee is idle waiting to occupy the facility is tallied. We tally the waiting times for firms respecting the federal minimum requirements listed above that have between 1 and 150 male employees and 1 and 150 women employees under segregated conditions and 2 and 300 people under unisex conditions. We then simulated each situation 10,000 times to determine the expected waiting times for women (Question 1), for men (Question 2), and for an employee under the unisex set-up (Question 3).

12.5 RESULTS

In this section we present the results of our simulation.[34] Figure 5 presents the expected waiting times women and men incur in firms from 1 to 150 employees for each gender. The measure for time is minutes per 2-hour interval. Figure 6 shows the waiting times for women and men combined for firms with equal number of women and men and the minimum number of segregated bathrooms for each sex vs. the expected waiting times for the same firms were they to open facilities to all employees. We model firms with 2 employees up to firms with 300 employees.

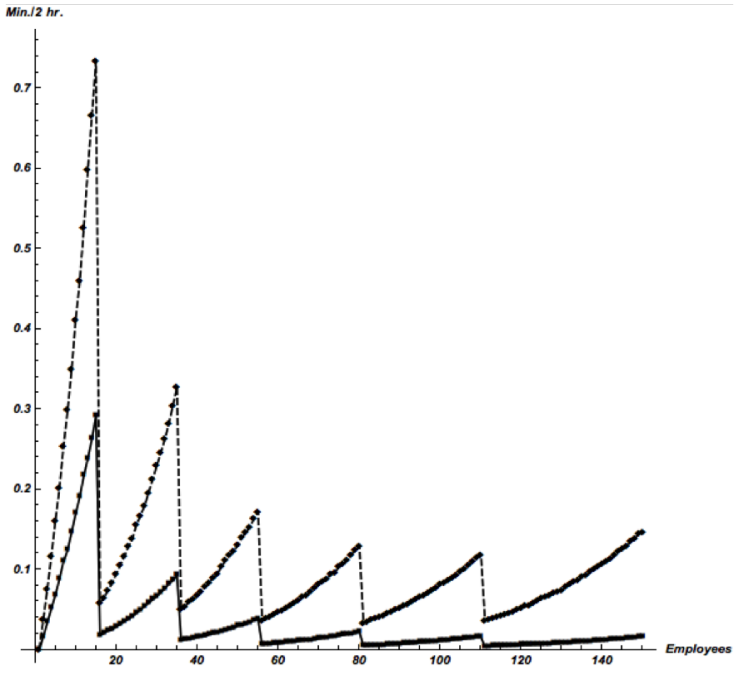


Figure 5: Expected waiting time: Women (top) vs. Men (bottom)

What we find is that there is a significant difference between the average expected waiting times for female vs. male employees. For

34 If you wish to consult the notebooks used for this chapter, please contact the author.

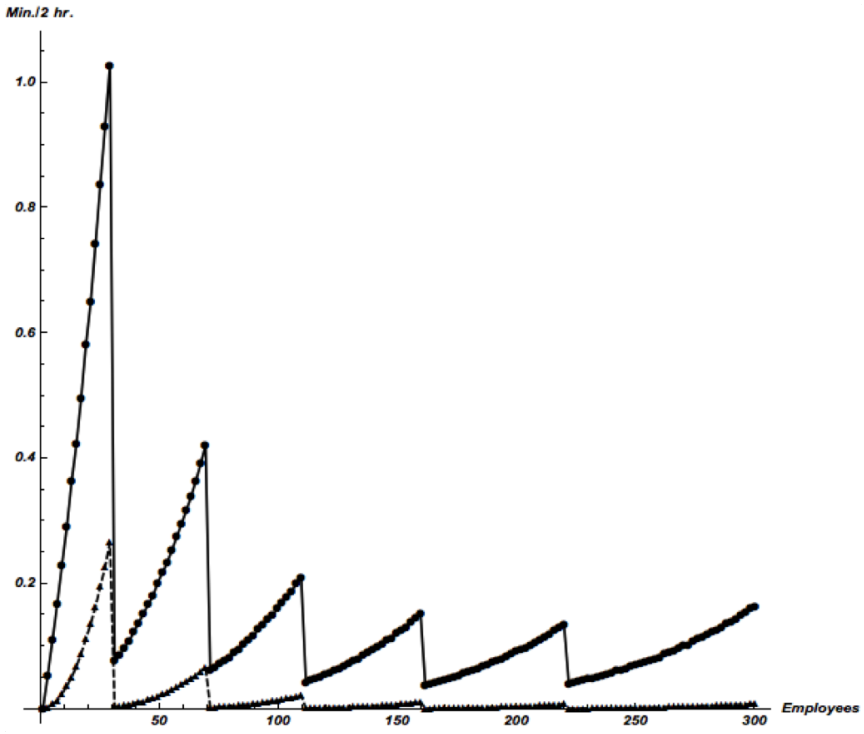


Figure 6: Expected waiting time: Segregated (top) vs. Unisex (bottom)

instance, in large firms of 300 employees, there can be a tenfold difference in the average waiting time women incur as opposed to men. Moreover, moving to a unisex set-up seems to generate a drastic reduction in the waiting time per employee. Looking again at large firms of 300 employees, the reduction between the average expected waiting time of an employee in the segregated set-up and the unisex set-up can be approximately eighteen-fold.

Figure 7 puts together the above two figures showing the comparisons between women, men and unisex. As can be easily observed from this picture, the waiting times for the unisex set-up are better than the waiting times for men employees only in the segregated set-up. This makes the introduction of unisex facilities a Pareto improvement over the current model. That being said, this is affected by the difference we take to be in terms of 'restroom time'

between women and men. The more similar we assume they are in terms of restroom usage, the likelier it is to obtain a Pareto improvement by shifting to unisex facilities. We do not explore different input values further in this paper.

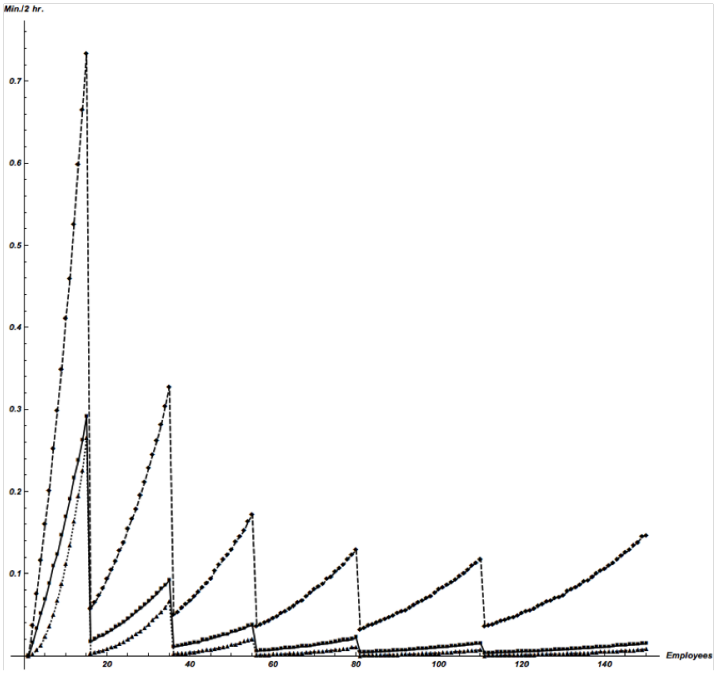


Figure 7: Expected waiting time: Women (top), Men (middle) and Unisex (bottom)

With the average expected waiting times per every 2-hour block, we can calculate how much women- and men- working-time a firm loses during a year depending on their number of employees. Figure 8 shows the results for the segregated set-up. In contrast, Figure 9 shows the expected hours of labour lost in a year for both male and female employees in the segregated set-up vs. the unisex

set-up.[35]

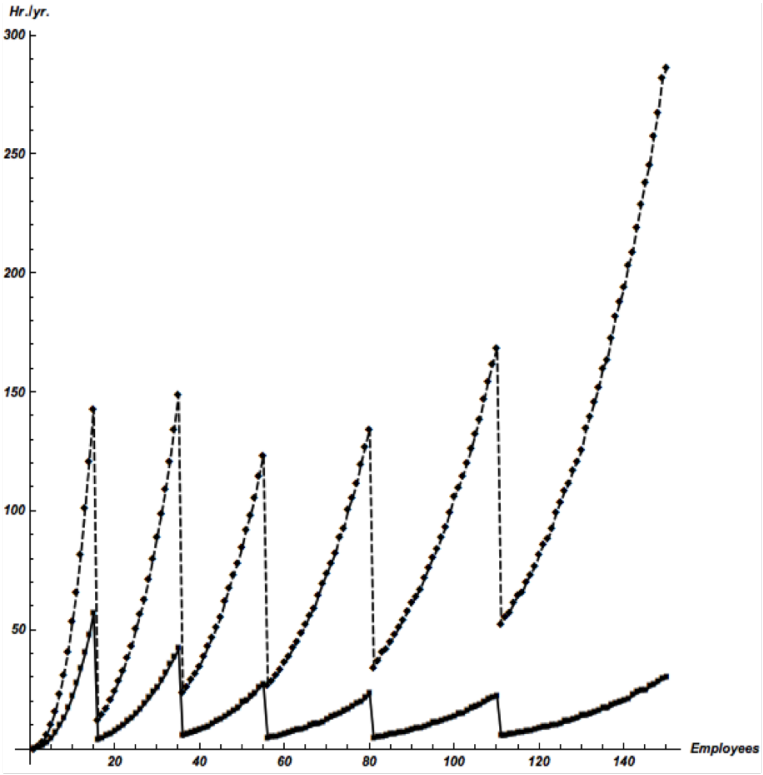


Figure 8: Expected loss in labour time: Women (top) vs. Men (bottom)

The results show a significant difference between the hours women lose as opposed to the hours men lose. Again, in certain conditions, this difference can be approximately tenfold. One way of interpreting this gender-difference is from the perspective of potty parity. Women have less access to office facilities given their needs under the current legislation than men do. This leads not only to discomfort and possible health issues, but it may also affect differences in inclusion and performance on the job If a firm were to pursue our suggestion and change to a unisex set-up it would

35 In order to perform these calculations we assume one of the visits employees making during their workday happens during their lunch break and hence should not count towards decreased efficiency.

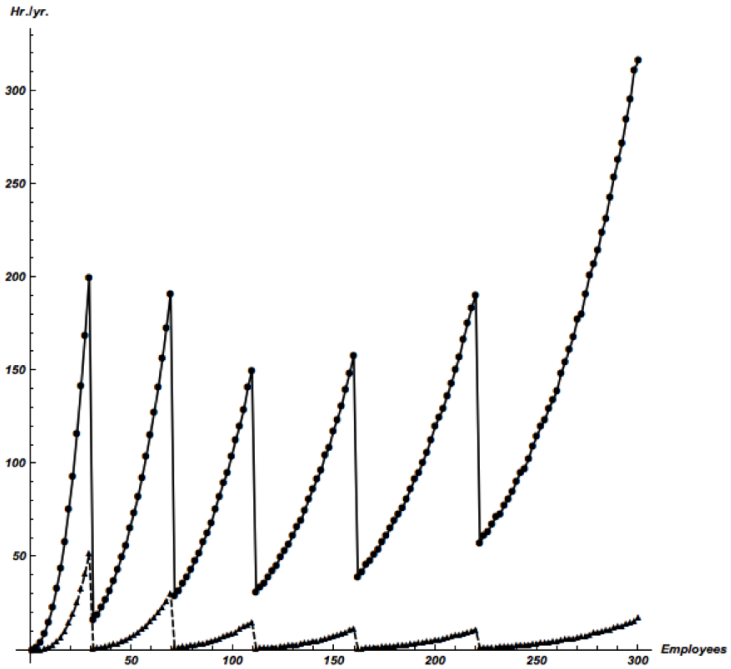


Figure 9: Expected loss in labour time: Segregated (top) vs. Unisex (bottom)

not only increase efficiency all-around, but it would also eradicate this gender imbalance.

A possible criticism at this point could be that although the relative time difference between women and men is significant, in absolute terms the situation is not that bad. The worst it can get is that in a firm with 300 employees, over a year, a female employee is expected to waste approximately 2 hours queuing to use the restroom. This is significantly worse than what her male colleague is expected to waste (12 minutes), but it does not feel like a large enough number to warrant any kind of policy change by itself. In other words, such a critic would probably say that we shouldn't pursue equality for equality's sake if no real gain is to be obtained from the change.

We are not fully convinced by this criticism,[36] but let us grant it for a moment. Then we can still approach the issue from a different angle. Consider the following question.

QUESTION 4 Assuming that the maximum wasted time for women/men according the current legislation is considered 'acceptable' how many unisex facilities would we need?

The idea is simple. Looking at the current legislation as the status quo, what is the worse situation under the status quo? What's the longest time an employee has to wait? Figure 10 repeats Figure 5 but clearly indicates the benchmark: the longest average expected waiting time for an employee (regardless of gender) is .73 minutes. If this were the acceptable standard for maximum waiting time, then how many facilities could we remove from a unisex set-up and still guarantee the average expected waiting time for an employee is under this standard of .73 minutes. Figure 11 depicts the average expected waiting time for an employee in a unisex set-up, assuming we have a combined number of 5 facilities.

Table 45 presents the maximum number of facilities for both genders in the segregated set-up (according to the US Federal regulations discussed above) and the unisex set-up for every employee bracket. The results are telling. If we were to cap the maximum average expected waiting time for an employee at the worst level permitted under the current set-up, we would be able to eliminate up to half of the facilities in a unisex set-up. This would bring about a significant reduction in overhead costs for firms.

36 And in fact this appears to be merely an artifact of the situation I am modelling in this chapter. Bovens and Marcoci (December 1, 2017) use the above algorithm to investigate a situation of two segregated restrooms with 6 stalls each, used by 150 men and 150 women, respectively over an interval of 1 hour. Under these specifications, a man's expected waiting time is 27 sec. while a woman's is 7 min and 40 sec.

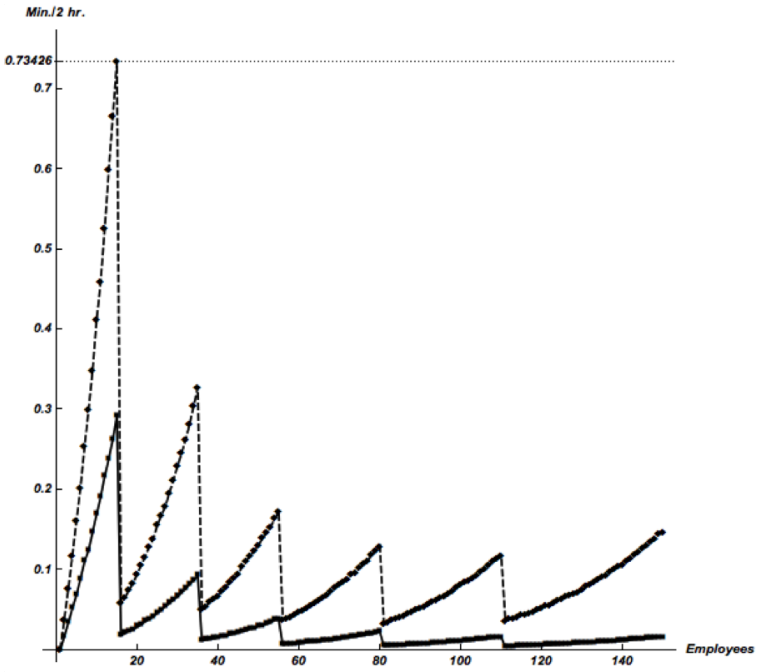


Figure 10: Expected waiting time: Women (top) vs. Men (bottom)

12.6 POSSIBLE OBJECTIONS AND REPLIES

We end by considering a few possible criticism to the recommendation of introducing unisex facilities.

THREAT OF VIOLENCE AGAINST WOMEN One of the main lines of argument against the Charlotte City Ordinance, and all other legislation opening restrooms to members of different 'biological sexes' (at birth) is that this will permit sexual predators to enter women's restrooms and attack girls and women. We were unable to find any reliable evidence that such a move would have an impact on the incidence of crimes against women.[37] Absence of evidence is

37 Most of the arguments are found on activist blogs. The best two activist arguments that violence will increase are perhaps this *Federalist* article suggesting that opening up restrooms to people who identify as trans* will bring about a wave of voyeurism-related incidents `http://thefederalist.com/`

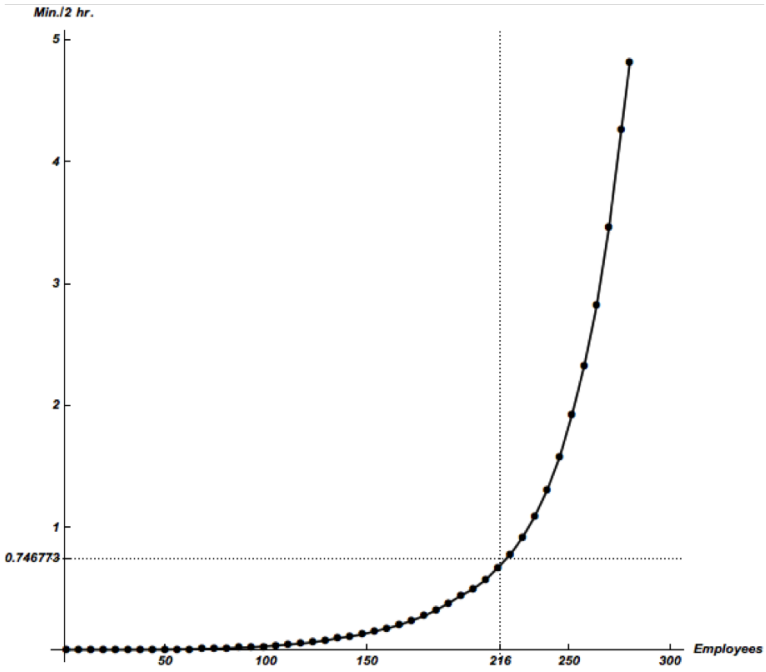


Figure 11: Maximum number of employees/5 facilities

not evidence of absence, and hence our recommendation is that before opening up all restrooms to people of different genders a pilot study should be conducted monitoring the impact this would have on safety. Furthermore, special attention should be given to micro-aggression against women.

MODESTY AND DISCOMFORT Another commonly raised issue with allowing people of different genders to use the same restroom is that the experience would be uncomfortable to them. Beyond the problem of modesty, this is particularly worrying since according to the International Paruresis Association, an estimated 21 million

2017/02/09/data-suggests-unisex-bathrooms-bonanza-male-perverts/, and this database put together by a Canadian group opposing Bill C-16, http://womanmeanssomething.com/violencedatabase/.

No. of Empl.	Min. No. Segregated	Min. No. Unisex
2 - 19	2	1
20 - 30	2	2
31 - 62	4	2
63 - 70	4	3
71 - 110	6	3
111 - 160	8	4
161 - 216	10	5
217 - 220	10	6
221 - 270	12	6
270 - 300	12	7

Table 45: Comparisons between minimum no. of facilities required to keep average expected waiting times under the current maximum level

Americans[38] suffer from paruresis or "shy bladder" syndrome.[39] It seems plausible that the symptoms would be aggravated in a unisex setting, though to date no research on this issue exists. We contend that before adopting a unisex set-up more research has to be done on the discomfort of using unisex facilities and the public health implications associated with an increase of "shy bladder" syndrome. In particular, different nudging strategies should be attempted. For instance, one could retain a few segregated facilities, but, given the smaller number, this would mean a longer walk for many employees. Their freedom to use segregated facilities is not taken away, but they are being nudged to become comfortable with unisex facilities.

LOSS OF URINALS Another argument is that moving to a unisex set-up will decrease the efficiency with which men are able to use

38 `http://paruresis.org/`

39 They also note that there may be as many as 220 million people suffering from this social anxiety disorder worldwide.

the restrooms. The reason is that a unisex restroom will typically do away with urinals (following the Swedish model), and thus the access opportunities of a unisex set-up would be less than the sum of the access opportunities of the gender-segregated set-up. Firstly, this isn't necessarily so. Several architecture firms across the US are currently developing unisex restrooms which provide space for urinals, but have a separating wall. But secondly, and more importantly, we believe the efficiency question should be asked with respect to the decrease in average waiting times per employee and not per gender. In other words, a set up that improves the efficiency per employee (or that does not increase it at least) is to be preferred even if it is detrimental to men. In our model the waiting times for both genders decreased. However, there are plausible inputs for our model that bring about a decrease in the waiting times for women and an increase in the waiting time for men. However, overall they do bring about a decrease in the waiting time per employee. We believe showing that is enough to recommend such a policy.

12.7 CONCLUSION

To sum up, this paper is making one of the first extensive arguments for changing the current way in which public restrooms are designed from a gender-segregated one to a unisex one. Our proposal goes further than recent inclusive legislation in two ways: (1) we propose the introduction of multi-stall unisex restrooms (and not simply requiring single-stall restrooms be unisex);[40] and (2) we propose opening the restrooms to everyone, and not only people who identify as trans*. We show how, based on some plausible (but still somewhat idealised) assumptions this will lead to an increase in

40 Compare to the recent decision of the New York City Council, http://www.reuters.com/article/us-new-york-lgbt/new-york-approves-unisex-bathrooms-in-nod-to-transgender-people-idUSKCN0Z72XX

economic efficiency. This would provide a shared basis for discussion for both sides of the ideological spectrum. Nevertheless, we are wary of the possible risks such a move would bring about, and we believe our proposal should first be thoroughly piloted. In particular, we think special consideration should be given to the (possible) increase in micro-aggressions against women (and not just of violent crimes) and the (possible) increase in discomfort that resists any attempts to nudge it.

BIBLIOGRAPHY

Kathryn H. Anthony and Meghan Dufresne. Potty parity in perspective: Gender and family issues in planning and designing public restrooms. *Journal of Planning Literature*, 21:267–294, 2007.

Kenneth Arrow. *Social Choice and Individual Values*. John Wiley, 1951.

Michelle Baillie, Shawndel Fraser, and Michael Brown. Do women spend more time in the restroom than men? *Psychological Reports*, 105(3):789–790, 2009.

L. Bovens and J. L. Ferreira. Monty Hall drives a wedge between Judy Benjamin and the Sleeping Beauty: a reply to Bovens. *Analysis*, 70:473–481, 2010.

L Bovens and E.J. Olsson. Coherentism, reliability and Bayesian networks. *Mind*, 109(436):685–719, 2000.

Luc Bovens. Judy benjamin is a sleeping beauty. *Analysis*, 70(1): 23–26, 2010.

Luc Bovens and Alexandru Marcoci. To those who oppose gender-neutral toilets: they're better for everybody. *The Guardian*, December 1, 2017. URL `https://www.theguardian.com/commentisfree/2017/dec/01/gender-neutral-toilets-better-everybody-rage-latrine-trans-disabled`.

Seamus Bradley. Constraints on rational theory choice. *British Journal for the Philosophy of Science*, forthcoming.

John Broome. Fairness. *Proceedings of the Aristotelian Society*, 91:87–101, 1990. URL `http://www.jstor.org/stable/4545128`.

Mary Anne Case. Why not abolish the laws of urinary segregation? In Harvey Molotch and Laura Norén, editors, *Toilet: Public Restrooms and the Politics of Sharing*, pages 1–32. New York University Press, 2010.

David Christensen. Rational reflection. *Philosophical Perspectives*, 24 (1):121–140, 2010.

Frank DeMeyer and Charles R. Plott. The probability of a cyclical majority. *Econometrica*, 38(2):345–354, 1970.

Franz Dietrich. A generalised model of judgment aggregation. *Social Choice and Welfare*, 28(4):529–565, 2006.

Franz Dietrich and Christian List. Arrow's theorem in judgment aggregation. *Social Choice and Welfare*, 29(1):19–33, 2007.

Franz Dietrich and Christian List. A liberal paradox for judgment aggregation. *Social Choice and Welfare*, 31(1):59–78, 2008.

Franz Dietrich and Christian List. Propositionwise judgment aggregation: the general case. *Social Choice and Welfare*, 40(4):1067–1095, 2013.

Foad Dizadji-Bahmani, Roman Frigg, and Stephan Hartmann. Who's afraid of nagelian reduction? *Erkenntnis*, 73(3):393–412, 2010.

C. Dorr. Sleeping Beauty: In defence of Elga. *Analysis*, 62:292–296, 2002.

A. Elga. Defeating Dr. Evil with self-locating belief. *Philosophy and Phenomenlogical Research*, 69:383–396, 2004.

Adam Elga. Self-locating belief and the Sleeping Beauty problem. *Analysis*, 60(2):143–147, 2000.

Adam Elga. Reflection and disagreement. *Noûs*, 41(3):478–502, 2007.

Wulf Gaertner and Nicolas Wuethrich. Evaluating competing theories via a common language of qualitative verdicts. *Synthese*, 2016.

William Gehrlein. Condorcet's paradox. *Theory and Decision*, 15(2): 161–197, 1983.

William Gehrlein and Peter Fishburn. The probability of the paradox of voting: A computable solution. *Journal of Economic Theory*, 13: 14–25, 1976.

Allan Gibbard. A Pareto-consistent libertarian claim. *Journal of Economic Theory*, 7(4):388–410, 1974.

Peter Grünwald. Safe probability: Restricted conditioning and extended marginalization. In Linda C. van der Gaag, editor, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 12th European Conference, ECSQARU 2013, Utrecht, The Netherlands, July 8-10, 2013. Proceedings*, pages 242–253. Springer, Berlin, Heidelberg, 2013.

Zalán Gyenis and Miklós Rédei. A principled analysis of consistency of an abstract principal principle. In Gábor Hofer-Szabó and Leszek Wroński, editors, *Making it Formally Explicit: Probability, Causality and Indeterminism*, pages 3–33. Springer International Publishing, 2017.

J. Halpern. The role of the protocol in anthropic reasoning. *Ergo*, 2: 195–206, 2015.

Joseph Halpern. Sleeping beauty reconsidered: Conditioning and reflection in asynchronous systems. In Tamar Szabo Gendler and John Hawthorne, editors, *Proceedings of the Twentieth Conference on Uncertainty in Ai*, pages 111–142. Oxford University Press, 2004.

Joseph Y. Halpern. *Reasoning about uncertainty*. MIT Press, 2003.

Colin Howson. Theories of probability. *The British Journal for the Philosophy of Science*, 46(1):1–32, 1995.

S. E. James, J. L. Herman, S. Rankin, M. Keisling, L. Mottet, and M. Anafi. *The Report of the 2015 U.S. Transgender Survey*. Washington, DC: National Center for Transgender Equality, 2016.

Philip Kitcher. The division of cognitive labor. *Journal of Philosophy*, 87(1):5–22, 1990.

Terry Kogan. Sex-separation in public restrooms: Law, architecture, and gender. *Michigan Journal of Gender and Law*, 14:1–57, 2007.

Thomas Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1970.

Thomas Kuhn. Objectivity, value judgment, and theory choice. In *The Essential Tension*. University of Chicago Press, 1972.

Imre Lakatos. Falsification and the Methodology of Scientific Research Programmes. In Imre Lakatos and Alan Musgrave, editors, *Criticism and the Growth of Knowledge*, pages 91–195. Cambridge University Press, 1970.

H. Leitgeb and D. Bradley. When betting odds and credences come apart: more worries for Dutch book arguments. *Analysis*, 66:119–127, 2006.

David Lewis. Sleeping Beauty: reply to Elga. *Analysis*, 61(3):171–176, 2001.

Christian List. Social choice theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2013 edition, 2013.

Christian List and Philip Pettit. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*, 18(1):89–110, 2002.

Anna Mahtani. Basic-know and super-know. *Philosophy and Phenomenological Research*, forthcoming.

Alexandru Marcoci. Quitting certainties: A bayesian framework modeling degrees of belief, Michael G. Titelbaum. Oxford University Press, 2013, xii 345 pages. *Economics and Philosophy*, 31(1): 194–200, 2015.

Alexandru Marcoci and James Nguyen. Scientific rationality by degrees. In Michela Massimi, Jan-Willem Romeijn, and Gerhard Schurz, editors, *EPSA15 Selected Papers: The 5th conference of the European Philosophy of Science Association in Düsseldorf*, pages 321–333. Springer International Publishing, 2017.

Alexandru Marcoci and James Nguyen. Objectivity, ambiguity, and theory choice. *Erkenntnis*, Forthcoming.

Jason McKenzie Alexander. On the redress of grievances. *Analysis*, 73(2):228–230, 2013.

Jason McKenzie Alexander, Johannes Himmelreich, and Christopher Thompson. Epistemic landscapes, optimal search and the division of cognitive labor. *Philosophy of Science*, 82(3):424–453, 2015.

Brian T. Miller. How to be a bayesian dogmatist. *Australasian Journal of Philosophy*, 94(4):766–780, 2016.

Michael Morreau. Mr. Fit, Mr. Simplicity and Mr. Scope: From social choice to theory choice. *Erkenntnis*, 79(6):1253–1268, 2014.

Michael Morreau. Theory choice and social choice: Kuhn vindicated. *Mind*, 124(493):239–262, 2015.

Samir Okasha. Theory choice and social choice: Kuhn versus Arrow. *Mind*, 120(477):83–115, 2011.

Samir Okasha. On Arrow's theorem and scientific rationality: Reply to Morreau and Stegenga. *Mind*, 124(493):279–294, 2015.

Kathleen Okruhlik. Gender and the biological sciences. *Canadian Journal of Philosophy*, 24(sup1):21–42, 1994.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

Miklós Rédei and Zalán Gyenis. Measure theoretic analysis of consistency of the principal principle. *Philosophy of Science*, 83(5):972–987, 2016.

Nicholas Rescher. Growing pains. *American Philosophical Quarterly*, 51(3):ii, 2014.

Davide Rizza. Arrow's theorem and theory choice. *Synthese*, 191(8): 1–10, 2013.

J. Ross. Sleeping Beauty, countable additivity, and rational dilemmas. *The Philosophical Review*, 119:411–447, 2010.

Miriam Schoenfield. Conditionalization does not (in general) maximize expected accuracy. *Mind*, forthcoming.

Kristie L. Seelman. Transgender adults' access to college bathrooms and housing and the relationship to suicidality. *Journal of Homosexuality*, 63:1378–1399, 2016.

Amartya Sen. The impossibility of a paretian liberal. *Journal of Political Economy*, 78(1):152–157, 1970.

G. Shafer. Conditional probability. *International Statistical Review*, 53: 261–275, 1985.

J.K. Slaney. An outline of formal logic and its applications in medicine–i. *British Medical Journal*, 295(6607):1195–1197, 1987.

T.S. Sneed. Discussion of paper by G. Shafer. *International Statistical Review*, 53:276–277, 1985.

Jacob Stegenga. Theory choice and social choice: Okasha versus Sen. *Mind*, 124(493):263–277, 2015.

Michael G. Titelbaum. *Quitting Certainties: A Bayesian Framework Modeling Degrees of Belief*. Oxford University Press, 2013.

Ilia Tsetlin, Michel Regenwetter, and Bernard Grofman. The impartial culture maximizes the probability of majority cycles. *Social Choice and Welfare*, 21(3):387–398, 2003.

Peter Urbach and Colin Howson. *Scientific Reasoning: The Bayesian Approach*. Open Court, 1993.

B. Weatherson. Should we respond to evil with indifference? *Philosophy and Phenomenlogical Research*, 70:613–635, 2005.

Sergio Wechsler, L. G. Esteves, A. Simonis, and C. Peixoto. Indifference, neutrality and informativeness: Generalizing the three prisoners paradox. *Synthese*, 143(3):255–272, 2005.

Michael Weisberg and Ryan Muldoon. Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, 76(2):225–252, 2009.