ABSTRACT

Title of Document: FROM SOCIAL CHOICE TO SYSTEM CHOICE: A PROBLEM FOR LEWIS'S BEST SYSTEM ANALYSIS

Sungwon Woo, Doctor of Philosophy, 2015

Directed By: Assistant Professor Aidan Lyon
Department of Philosophy

One of the most important results in social choice theory is Kenneth Arrow's impossibility theorem (1951/1963), according to which there cannot exist any rational procedure of aggregating individual preferences into a social preference. In this dissertation, I argue that the analogue of Arrow's theorem threatens David Lewis's Best System Account (BSA) of laws of nature, as the BSA invokes the procedure of aggregating different system-choice criteria into a resultant choice of the best system. First, I examine the formal conditions of Arrow's impossibility theorem and its theory-choice variant. In the domain of theory choice, statistical model selection methods make different theory-choice standards commensurable. This inter-standard comparability may open up an escape route from the Arrovian impossibility for theory choice. Conducting a rigorous examination of those statistical methods, in particular, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), I show that these methods assume the existence of true status of nature, and that their inter-standard comparability serves as an epistemic constraint. I then argue that there is a formal analogy between social choice and system choice for the BSA and the Arrovian impossibility threatens the BSA. After rejecting various possible attempts to escape from the Arrovian impossibility for the BSA, I propose the variants of the BSA implemented with AIC and BIC as an attempt to make a case for inter-criterial comparability in system choice. I argue that, however, the proposed variants will inevitably fail to pick the best system. The failure is explained by the results in my investigation of the statistical methods. Finally, I suggest different ways in which the BSA might be able to escape from the Arrovian impossibility: a non-harmful dictatorship, a threshold-prior criterion, and the statistical method called Minimum Description Length Principle. I close the dissertation by suggesting that the BSA might have to give up the notion of 'balancing' in its analysis of laws of nature in order to avoid the Arrovian result in a way that is consistent with the Humean perspective on laws of nature.

FROM SOCIAL CHOICE TO SYSTEM CHOICE: A PROBLEM FOR LEWIS'S BEST SYSTEM ANALYSIS

By

Sungwon Woo

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Professor Aidan Lyon, Chair
Professor Allen Stairs
Assistant Professor Eric Pacuit
Professor Mathias Frisch
Professor Michael Morreau

To my parents

Table of Contents

Chapter 1. Introduction

Introduction

This chapter serves as an introduction to the main questions I will investigate throughout this dissertation. In §1.2, I will give a general introduction to aggregation problem in various domains of collective choice. In particular, I will introduce Arrow's impossibility theorem in social choice and its analogues in scientific theory choice and in system choice for Lewis's Best System Analysis of laws of nature. In §1.3, I will give an overview of some development in social choice theory, in particular Sen's information-enriched framework which allows further investigation of a wide range of measurability and comparability of individual utilities. In this framework, a number of possible escape routes from the Arrovian impossibility can open up. I will also discuss a possible escape in the domain of theory choice. This will lead us to §1.4, where I give a sketch of an escape route in theory choice, namely, inter-criterial comparability found in the literature on statistical model selection. In §1.5, I will lay out the outline of this dissertation.

1.1 Aggregation Problem in Social Choice, Theory Choice, and System Choice

There are aggregation problems in the domain of social choice. Social choice theory is a formal study of procedures of collective decision making, e.g., aggregating individual preferences, votes, or welfare into one overall social preference, vote, or welfare. One of the most important results in social choice theory is Kenneth Arrow's impossibility theorem (1951/1963). One might investigate particular social choice procedures case by case. For example, one might study various winner-choosing rules, e.g., majority rule, ranked voting rule, tournament rule, etc., and investigate when they do and do not work case by case. Arrow took a fundamentally different approach. He pioneered the axiomatic approach in which he can analyze all social choice procedures by imposing a set of reasonable axioms and mathematically deducing a theorem from them. The startling result of his theorem was that there cannot exist any reasonable procedure of aggregating individual preferences into a social preference, hence the name Arrow's impossibility theorem.

One notable feature of such an axiomatic approach of Arrow is that an analogue of his theorem obtains in other domains of aggregation problem, as long as it can be shown that the conditions for the theorem apply to the domain in question. It is at this point we can observe an interesting kind of aggregation problem in scientific theory choice. It is commonly supposed that the procedure for choosing a better theory is a multi-dimensional procedure in that there are multiple merits that good theories should display, such as accuracy, simplicity, fruitfulness, consistency, and so on. For example, Karl Popper (1959, 1963) argued that one theory can be better than another by being closer to the truth. Larry Laudan (1977), in contrast, argued that theories equipped with better problem-solving capacities are better. Thomas Kuhn (1977a) says that there are five standard criteria for evaluating the adequacy of a theory: accuracy, consistency, scope, simplicity, and fruitfulness. Carl Hempel (1983) speaks of the similar desiderata for a good hypothesis: accuracy of prediction, consistency with neighboring fields, broad scope, simplicity, and fruitfulness. Apparently, a good theory would be the one that maximize different theoretical merits.

If the above picture of theory choice is correct, then the analogy between social choice and theory choice seems to hold. In voting, all the different individual preferences of the candidates in consideration is to be aggregated into a resultant collective choice of a candidate. Analogously, then, all the different theoretical merits of the theories in consideration is to be aggregated into an overall choice of a theory. For example, consider we are comparing three rival theories X, Y, and Z, and we evaluate these theories in terms of three theoretical merits, namely, simplicity, fruitfulness, and accuracy. Suppose we evaluate them as follows, where the higher in a column the better:

Simplicity	Fruitfulness	Accuracy
X	Z	Y
Y	X	Z
Z	Y	X

Which theory should we choose? X is better than Y in terms of two merits (simplicity and fruitfulness); Y is better than Z in terms of two merits (simplicity and accuracy.) So, assuming transitivity of '… is better than…' relation, X is better than Z. Does this mean X should be our choice? Unfortunately the answer is no. Z is better than X with respect to two merits (fruitfulness and accuracy). As a result, we have a cyclic relation of 'betterness' – X is better than Z and Z is better than X. The similar can be said about the other pairwise competitions of the pair of Y and Z and that of X and Y. Cyclic ranking patterns like this is called the paradox of voting, or Condorcet paradox. This is a particular case where we have a clearly problematic case. But if it can be shown that theory choice is sufficiently analogous to social choice, then we may draw more general conclusions by applying Arrow's axiomatic analysis of social choice procedures to theory choice procedures.

Now let us make a brief jump to a popular philosophical view about laws of nature, the Best System Analysis (BSA) of laws of nature (Lewis 1973, 1983, 1994). Broadly speaking, there are three philosophical positions about laws of nature. One might take an eliminativist position, according to which laws of nature are simply non-existent, therefore they are to be eliminated from the philosophical discourse. Or, one might take a primitivist position that laws are fundamental, non-reducible, primitive element of the reality, i.e., they are weaved into the fabric of the reality. Philosophers who hold this position support the governing-laws conception of laws of nature. Or, one might take a reductionist position that there exist laws but they can be reduced to other things, without remainder; laws are just regularities; not some 'mysterious', metaphysically fundamental entities. Philosophers holding this position support the non-governing conception of laws of nature. The BSA belongs to the third category. It is a modified regularity theory. It says that regularities are laws if and only if they appear as theorems or axioms in an appropriately axiomatized collection of true propositions about the world – where 'appropriately' means that they are as simple, strong, and accurate as possible. That is, laws are what the best systemization of facts says they are.

What does this account of lawhood have to do with the aggregation problems in social choice and theory choice? Let me quote Davie Lewis:

> *…I take a suitable system to be one that has the virtues we aspire to in our own theory building, and that has them to the greatest extent possible given the way the world is. (Lewis 1983; 367)*

More specifically, the best system is defined as follows:

> *The virtues of simplicity, strength, and fit trade off. The best system is the system that gets the best balance of all three. The best system is the system that gets the best balance of all three. As before, the laws are those regularities that are theorems of the best system. (Lewis 1994; 480)*

The BSA regards the best system as ideal scientific theory. Lewis says: "Suppose there is an ideal theory of everything ... [o]n the best system account, it follows that the rules of this ideal best theory are the true laws of nature." (Lewis 1994: 231f). What he thinks of is something not too different from present-day physics, just a "presumably somewhat improved" (Lewis 1983; 364) version of physics. We may even think of his 'ideal theory of everything' as what fundamental physics is aiming for, for example, Weinberg (1992)'s "Final Theory" or Penrose (2004)'s devoted "Theory of Everything". The successes of physics to date provide reason to think that our world is susceptible to very good systematizations in fundamental terms, so it seems like a reasonable hope that laws can be find in the systemization of facts which is systemized in the same way as our fundamental science is theorized. This motivates the BSA to take standards from our practice of scientific theory choice and use them as system choice standards in its analysis of laws, as we can see in the above quote. Laws are what the ideal theory says they are.

So, the system choice procedure for the BSA can be viewed as a procedure of aggregating system choice standards (which are imported from scientific theory choice) into an overall choice of the 'best' system. Suppose the theory-choice analogue of Arrow's impossibility theorem obtains. Then we may naturally suspect that the BSA too will be susceptible to the analogue of the impossibility theorem, to the extent it has scientific theory-choice procedure as integral to the analysis of lawhood. Whether there is a sufficient formal analogy between social choice and system choice, and, if there is one, whether one or more conditions of the analogue can be relaxed, are the main questions I will investigate in this dissertation.

1.2 Cardinality, Sen's "St. George the Dragon Slayer", and Comparability

A number of solutions have been suggested to the Arrovian impossibility in the literature of social choice theory. Probably one of the most commonly discussed solutions is to adopt Sen (1970)'s 'information enriched' approach by implementing different measurement scales for individual preference. As we will see in Chapter 2, Arrow's original characterization of social choice is informationally impoverished - it only allows information about ordinal rankings of alternatives. Sen extended Arrow's framework so that cardinal information about individual preferences can be used in social choice. Following Sen, one may suggest that we should view theory choice in such a informationally rich framework.

One might respond to the Arrovian impossibility that we should utilize cardinal measures of preferences. Here is a possible line of response: In the theory choice example above, very limited information was used. The only information admitted to use in the aggregation procedure was the orderings of the theories by theoretical merits, that is, the *ordinal* rankings of the alternatives. But can we not use richer information about individual preferences? For instance, couldn't it be the case that the difference in the degrees of simplicity of *X* and *Y* is far greater than the difference of accuracy between *Y* to *X*? If it is the case that, for whatever reason, simplicity is more important than accuracy, then we may justifiably give a more weighting to the simplicity rankings of theories than the accuracy rankings. Furthermore, each theoretical merit might be represented with some numerical values. For example, simplicity of a theory might be measured by the number of theoretical entities it presupposes, the number of parameters, or the number of axioms. Each of these measure can generate a numeric representation of simplicity. Likewise, one might continue, accuracy may be measured on a cardinal scale where the degree of accuracy is represented numerically. Provided that such measures of theoretical merits are available, then, it might seem to be a perfectly reasonable procedure in which the winner is the theory that has the greatest sum of merit-scores across individuals. This might serve as an escape route from the Arrovian impossibility, one might think.

In social choice, the same idea has been long voiced. The above line of thought is formally analogous to the following:

$$W(u_1, \dots, u_n) = \sum_{i=1}^{n} \lambda_i u_i$$

This is the same form as the so-called classical utilitarianism, where u_i are utilities that individuals get from the alternative in question and λ_i are weightings to individuals. If we are to give equal weight to everyone, then $\lambda_i = 1$ for every i, we get 'Benthamite' utilitarianism. If we are to give unequal weights to individuals, then it means λ_i will get different values for different individuals ('weighted' utilitarianism). One important thing to note is that both specifications of utilitarianism presented above presuppose something more than just numeric representation of utilities. Individual values λ_i are assumed to be individually measurable and comparable across individuals. However, in order to have a metric for individual values $\lambda_{i,}$ we need interpersonal comparability of utilities. For example, the claim that person A's utility is comparable to person B's utility implies that we can assign certain weights to individual utilities. In the context of cardinal utilities, we have to be able to say something like “A certain loss of utility in person A can be compensated (i.e., can be traded for) by an equal gain in utility by person B.” This statement expresses comparisons of utility intervals in different alternatives between individuals. In short, we cardinal comparability, not just cardinality, if we were to escape from the Arrovian impossibility. We will have further discussion on interpersonal comparability in Chapter 2, but for now let us bring our attention to Sen’s characterization of the issues surrounding interpersonal comparability. The following is a snapshot of what will be discussed in that chapter.

Sen (1970) generalized Arrow's model to incorporate information richer than just orderings of alternatives. In Sen’s framework, the preferences of individuals are presented not simply as orderings (‘rankings’, so to say) but as utility functions that map the alternatives onto real numbers. In terms of utility information, it is usual to view utilities as being *ordinal* or *cardinal*. So, if preferences are to be measured as ordinal utility, the delivered information is essentially same as the ordinal ranking of the alternatives. But if they are measured as cardinal utility, then the measurement delivers more information than just the ordering of alternatives, say, ‘intensity’ of preference.

One important finding in social choice was that having cardinal utilities is not by itself enough to avoid an impossibility result. In addition, utilities have to be interpersonally comparable (Sen 1970; Ch8). Sen (1970) and Kalai and Schmeidler (1977) show that if no interpersonal comparisons of preference are permitted, then the impossibility conclusion of Arrow’s theorem remains true, even if Arrow’s ordinal interpretation of individual utility is replaced by a cardinal interpretation. Cardinality alone cannot open up an escape route, as Sen says about St. George the dragon slayer:

> *Given non-comparability, the relative preference intensities of individuals over any pair can be varied in any way we like except for reversing the sign, i.e., without reversing the ordering, so that cardinality is not much of an advance over individual orderings when combined with non-comparability. To give some bite to cardinality we have to relax one of the other conditions ... Cardinality alone seems to kill no dragons, and our little St. George must be sought elsewhere. (Sen 1970; 124 -5)*

What is needed in addition to cardinality is interpersonal comparability of utilities or preferences. In the context of cardinal utilities, in order to avoid the Arrovian impossibility, we have to be able to say something like "A certain loss of utility in person A can be compensated (i.e., can be traded for) by an equal gain in utility by person B." This statement expresses comparisons of utility intervals in different alternatives between individuals. For example, classical utilitarianism as we saw earlier requires an interpersonal comparison in which the individual cardinal preferences are summed into an overall social preference. In the context of ordinal utilities, the relevant comparability would be the utility level comparability across individuals. For example, Rawlsian utilitarianism requires comparison of utility levels of the worst-off individual in each alternative state.

Now let us turn to the matter of theory choice. Assuming that social choice and theory choice are formally analogous, it is natural to seek the same type of escape route for theory choice as social choice. Theories are ranked by each dimension of scientific merits just as alternatives are ranked by each individual in the society. Just as we ultimately need interpersonal comparability of cardinal preferences in social choice, we likewise need inter-dimensional comparability of scientific virtues in order to save theory choice from the Arrovian impossibility.

What would it be like to have such inter-criterion cardinal comparability in theory choice? That there is cardinal inter-criterion comparability implies that theories are measured on a cardinal scale for each criterion and that there is an exchange rate between criteria. Cardinal comparability in this context would mean then that we can justifiably make judgments like "*This* amount of loss in simplicity can be compensated for with *that* amount of gain in accuracy." To put it different way, this would mean something like "The metric for trade-off, i.e., the exchange ratio, between simplicity and accuracy is such-and-such." With this kind of 'recipe' for the trade-off between standards, theories may be compared in a consistent way.

Turning to system choice now, as we will see in Chapter 4, the BSA imports our actual scientific theory-choice procedure for system-choice procedure. Assuming there is formal analogy between theory choice and system choice (and there should be, given that the BSA takes the scientific theory choice standards and elevates them to the constituents of laws of nature; see Chapter 4 and Chapter 5 for further discussion), if we were to search for the same kind of escape routes for system choice as in theory choice, we need to identify: (a) measurability of each system-choice criterion, (b) inter-criterion comparability of each system's 'score' with respect to system-choice criteria, and (c) the form of a function which will specify an overall system ranking given measurability and comparability assumed. The second half of this dissertation will be devoted to these tasks.

1.3 Comparability in Statistical Model Selection Methods

There has been growing interest among philosophers of science in statistical model selection methods. Some think that statistical model selection methods can provide the comparability required to escape from the Arrovian impossibility (for example, Okasha 2011, Morreau 2015).

Let us briefly discuss what statistical model selection is about and how they relate to some important philosophical questions. One of the most important questions in model selection is how to minimize the risk of *overfitting*.

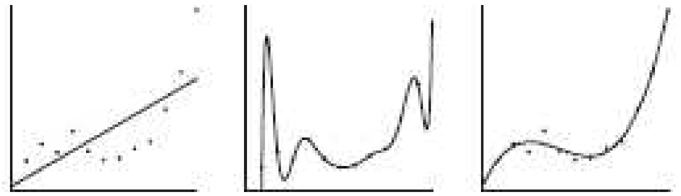

Fig 1. Curves with varying complexity (from Grünwald 2005)

Suppose the values of variables X and Y have been observed and the result is plotted in the figure above, where the dots representing the observed data points. We would like to learn the relationship between X and Y. We may fit various curves to the data points as shown above. The straight line is simple, but maybe it's 'too' simple. It seems to fail to capture the apparent pattern or trend in the data. If we want to choose a curve that perfectly fits the data, then we will choose a curve like the one in the second picture. This curve seems 'too' complex. It seems to 'overfit' the noise part of the data, i.e., the random fluctuations in the data rather than the true pattern underlying it. Probably a reasonable choice would be to choose the curve shown in the rightmost picture, which seems to capture regularity in the data without fitting too much the noise in the data. So we need a principled method which allows us to choose a right curve among all the curves that are logically compatible with the observed data. Even in a simple example like this, we may agree that the desirable model selection methods are probably the ones that give us principled ways to make trade-offs between goodness-of-fit and simplicity. But exactly how to make the trade-off in real situations is not so obvious. Suppose model A fits the data a little better than model B, but has one more parameter. How does one trade off goodness-of-fit against number of parameters?

As we will investigate further in Chapter 3 and Chapter 6, a great deal of statistical model selection methods attempt to provide such specific trade-off 'recipes'. For example, Akaike Information Criterion (Akaike 1974) and Bayesian Information Criterion (Schwarz 1978) provide the following rules and model indices:

AIC score of model *M* = [MLE of *M*] **-** [the number of parameters of *M*].

AIC rule: Choose the model that maximizes AIC score.

BIC score of model *M* = [MLE of *M*] **-** [log *n* × the number of parameters of *M*].

BIC rule: Choose the model that maximizes BIC score.

In statistics, the term [MLE: Maximum Likelihood Estimate] is one of the most widely used measure of goodness-of-fit. The number of parameters seems to give us a natural measure of simplicity of models.[1] We will conduct a careful examination of these methods in Chapter 3, but for now we can notice that they express specific trade-off forms between simplicity and fit.

[1] Of course, this is one specific sense of simplicity, among many. We will have further discussion on this in Chapter 3 and Chapter 5.

Different model selectin criteria use different definitions of simplicity and fit. For example, the model selection principle called Minimum Description Length Principle (MDL) (Rissanen, 1978; Grünwald *et al.*, 2005) take a fundamentally different approach to model selection problems. In MDL, the goal of statistical inference is to find regularity in the data, and regularity is identified with "ability to compress." The underlying idea is that there are different ways to 'summarize' the observed regularities in the data sets, and the shorter the summary is, the better. For a toy example, suppose we have a sequence 01010101010101010101010101. We could summarize this sequence as "0 appears 13 times, 1 appears 13 times, 0 appears first, then 1 appears, then alternate", or more efficiently, "'01' is repeated 13 times".[2] MDL says the latter is the better theory to explain the given sequence. More formally,

> **MDL principle** (Rissanen 1978): the best theory to explain given data x is the one that minimizes the sum of:
>
> (1) The length of the description of the theory itself, plus
>
> (2) The length of the description of the data *x* when the data is described with the help of the theory.[3]

The first term can be understood as complexity of the theory, and the second term as goodness-of-fit. Intuitively speaking, the second term tells us goodness-of-fit of the theory because, the better it fits the data, the fewer bits we would need to describe the data *given* the theory. For example, in order to 'describe the data' as in (2), we would need to describe the discrepancies between the values predicted by the theory and the

[2] Formally, in MDL sequences are described and summarized in a universal program languages. We will have further discussion on this in Chapter 6.

[3] The lengths are measured in bits. See Chapter 6.

actually observed value; but we would not describe what the theory predicts about the data, which is the job of (1) (Grünwald 2005). We can see that simplicity and fit in MDL are very different from those in AIC or BIC. Each of these statistical model selectin methods provides a certain form of trade-off in their own.

Philosophers of science who attempt to deal with the Arrovian impossibility in theory choice have been paying attention to various kinds of statistical model selection methods like above because they appear to give specific trade-off recipe, i.e., inter-criterial comparability required to escape the impossibility. For example, Okasha (2011) thinks that AIC and BIC do provide the right kind of comparability to avoid the impossibility in theory choice. We now have a question whether it provides the same kind of escape from the Arrovian impossibility in system choice for the BSA. We will have further discussion on this in Chapter 5 and Chapter 6.

1.4 Thesis outline

This dissertation will proceed as follows. In Chapter 2, I will examine aggregation problems in the domain of social choice and theory choice. First, in §2.1, as a grounding work, I will review some common understanding of rational scientific progress and skepticism of rational theory choice. In §2.2., I will discuss Arrow's impossibility theorem (1951/1963) which says there cannot exist any reasonable procedure of aggregating individual preferences into a social preference. I will carefully examine formal statements of the theorem and its conditions. In §2.3, following Okasha (2011)'s lead, I will examine if the theory-choice analogues of Arrow's condition are motivated. I will explore if the analogue of Arrow's impossibility theorem obtain in the domain of theory choice. The tentative conclusion will be that the analogue seems to obtain in theory choice, with a caveat that condition **U** does not apply to theory choice but probably its weaker counterpart **R** does. In §2.4, I will discuss possible escape routes from the impossibility in theory choice. Weakening **U** to **R** does not necessarily open up an escape route from the Arrovian impossibility, if condition **I** can be strengthened to **SN**. It will be noted that it is unclear if **SN** applies to theory choice but we saw some motivation for thinking it does. We will also have some discussion on the single-profile variants of Arrow's impossibility theorem. It will turn out that probably the clearest and most promising escape route from the Arrovian impossibility for theory choice would be to make a case for inter-criterial comparability, in Sen (1970)'s extended framework. In particular, we will be interested in some form of inter-criterial comparability expressed in some statistical model selection methods, for example Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). I will ask where the comparability in these methods comes from. This question will lead us to Chapter 3.

In Chapter 3, I will conduct a rigorous examination of the statistical model selection methods AIC and BIC, and their background assumptions. In §3.1, I will introduce statistical model selection problem and its philosophical implications. I will compare two quite different frameworks: the Best-Case strategy and Akaikean framework. It will be shown that the former framework will run a high risk of overfitting. In §3.2, I will rigorously examine the key concepts underpinning AIC: Kullbak-Leibler divergence, predictive accuracy, and estimated predictive accuracy. In §3.3, the proof for AIC will be sketched. In §3.4, I will examine BIC and its underlying assumptions.

The conclusion of the chapter will be that the examined statistical model selection methods express specific exchange ratios for the trade-off between fit and simplicity, if we understand likelihood as fit and the number of parameters as simplicity, while different methods give different weight to simplicity (in the specific sense above). This may be understood as the comparability required in order for theory choice to escape the Arrovian impossibility.

In Chapter 4, our focus will be on David Lewis's Best System Account of laws of nature. In §4.1, I will first survey three philosophical views of laws of nature: eliminativism, primitivism, and reductionism. In §4.2, I introduce Lewis's Best System Account (BSA). We will discuss the motivations for the account, the Humean Supervenience (HS) thesis about laws, and some typical objections to HS about laws. In §4.3, I will assume the task of precisifying the Best System Account of laws. We will see that the best system in the BSA should be understood as an extended, idealized version of our actual science (in particular, fundamental physics). In the process of precisifying the account, it will be revealed that the HS thesis has a specific range, and that the BSA relies heavily on what I call the *Hope thesis*. It will be suspected that the BSA's reliance on the concept of 'balance' among the system-choice criteria makes the account vulnerable to the Arrovian impossibility.

In Chapter 5, I explore if the analogue of the Arrovian impossibility in the domain of system choice holds for the BSA. In §5.2, we will see that the condition **U** (Universal Domain) does not apply to system choice but the condition **R** (Rich Domain) probably does. But we will also note that, even if **U** is weakened to **R,** a variant of Arrow's impossibility theorem obtains provided the strong neutrality condition (**SN**), a stronger version of **I**, is met. In §5.3, I will discuss **SN** in connection with the Humean Supervenience (HS) thesis. I will argue that **SN** applies to system choice. While the HS thesis seems to reject the multi-profile framework for system choice, assuming **R** is met in system choice, since **SN** applies to system choice, the single-profile variant of the Arrovian impossibility seems to obtain in system choice. In §5.4 I will suggest that **IIU** is a desirable property of system choice procedure. In §5.5, I will discuss a number of possible attempts to make a case for the cardinal measurability of fit, strength, and simplicity, the three criteria invoked by the BSA. I will conclude that most of the attempts are unsatisfactory, at least as they stand. It will be also noted that, even if they are cardinally measurable, without inter-criterial comparability one cannot escape from the Arrovian impossibility. In §5.6, I propose a variant of the BSA as an attempt to make a case for inter-criterial comparability between fit and simplicity. I will conclude that its prospect does not look good due to the context gap between epistemological and metaphysical justification of implementing inductive method to the analysis of lawhood. In doing so, I will provide a counterexample where the BSA fails to pick the best system. The BSA will still have the last resort: the Hope thesis. But this will also raise the concern that the BSA relies too much, in ad-hoc manner, on the Hope thesis every time it faces a problem. In Chapter 6, I will suggest different ways in which the BSA might be able to avoid the Arrovian impossibility: the concept of 'non-harmful dictator', and the statistical model selection called Minimum Description Length Principle (MDL). These ideas are not fully matured hence in need of more extensive research, but it will be suggested that probably they are the best possible escape routes available to the BSA, if it can be shown that they work. I close the dissertation by suggesting that the BSA might have to give up the notion of ‘balancing’ in its analysis of laws of nature; this would allow the BSA to avoid the Arrovian result in a way that is consistent with the Humean perspective on laws of nature.

Chapter 2: Maximizing Rationality in Society and in Science

Introduction

In this chapter, I will examine some aggregation problems in the domain of social choice and theory choice. Social choice theory is a formal study of procedures of collective decision making, e.g., aggregating individual preferences, votes, or welfare into one overall social preference, vote, or welfare. One of the most important results in social choice theory is Kenneth Arrow's impossibility theorem (1951/1963). Rather than investigating particular social choice procedures case by case, Arrow pioneered the axiomatic approach in social choice theory by imposing set of reasonable axioms and mathematically deducing a theorem from them. The startling result of the theorem was that there cannot exist any reasonable procedure of aggregating individual preferences into a social preference, hence the name *Arrow's impossibility theorem*.

There has been growing interest among philosophers in the findings in social choice theory, in particular Arrow's theorem. Okasha (2011) observed an analogy between social choice and theory choice and explored whether the analogue of Arrow's theorem obtains in theory choice and, if it does, whether there are escapes form the impossibility. Following his lead, in this chapter, I will examine Arrow's impossibility theorem, its theory-choice analogue, and possible escape routes from it. First, in §2.1, I will examine Kuhn's view on scientific theory choice. According to him, since there are inevitably subjective elements in theory-choice standards that scientists legitimately employ, there is no neutral theory-choice algorithm. Okasha interprets this as the claim that there are too many legitimate theory-choice procedures and later makes dramatic contrast with Arrow's impossibility theorem. Following this lead, in §2.2, I will introduce Arrow's theorem and examine formal statements of the theorem and its conditions. This task is required for our later examination of plausibility of the theory-choice analogue of the Arrovian impossibility in §2.3. The result of the examination will be that the analogue seems to obtain in theory choice, with a caveat that it is not perfectly clear if conditions about the domain of theory choice procedure are met. In §2.4, we will explore possible escape routes, focusing on a different types of measurability and comparability in Sen (1970)'s extended framework. In particular, we will focus on cardinally measurable inter-criterial comparability expressed in the literature of statistical model selection methods. I will conclude that searching for comparability in statistical model selection is probably the clearest and most promising escape route from the Arrovian impossibility in theory choice. This will lead us to the next chapter.

2.1 Theory Choice and Kuhn: Maximizing Overall Theoretical Merit

Science does not simply develop or change, but rather progresses. Making scientific progress comes down to choosing better theories. It is commonly supposed that the procedure for choosing a better theory is a multi-dimensional process in that there are multiple merits that good theories should display, for instance: accuracy, simplicity, fruitfulness, and so on. Choosing a better theory, then, seems to mean maximizing the *overall* theoretical merit. Throughout this chapter, we will examine some difficulties associated with this familiar conception of scientific progress. In this section, I will first discuss Kuhn's view that there is no neutral theory-choice algorithm due to its subjective factors, and Okasha’s interpretation that there are too many legitimate theory choice procedures.

2.1.1 Scientific Virtues and Their Aggregation

Making scientific progress amounts to choosing better theories. The nature of theory choice, however, is a topic of controversy in the philosophy of science. For one thing, ‘better’ is a multi-criteria normative term, in that there are more than one criteria against which competing scientific theories can be evaluated. Philosophers have recognized several ways in which one theory can be better than another. Just to name a few examples: Kuhn offers a list of scientific standards, or virtues, that he believes can provide a common basis for theory choice: accuracy, consistency, scope, simplicity, and fruitfulness (Kuhn 1977). Popper (1959, 1963) argues that one theory can be better than another by being closer to the truth. Laudan (1977) argues that theories equipped with better problem-solving capacities are better. Carl Hempel (1983) speaks of the similar desiderata for a good hypothesis: accuracy of prediction, consistency with neighboring fields, breadth of scope, simplicity, and fruitfulness. Sober (1994) says that simplicity and predictive accuracy are important virtues of scientific theories, especially in statistical analysis of the observed data. Miller (2006) takes logical strength to be a virtue of scientific theories. Evidently, there are multiple virtues or dimensions, with respect to which one theory can be better than another. Choosing a better theory would then amount to choosing the one that maximizes these virtues. In other words, the procedure of theory choice may be seen as that of *aggregating* these individual virtues into a theory's *overall* scientific virtue. The difficulty with this kind of concept of scientific progress is that there seems to be no objective and rational aggregation procedure. This is because some standards seem to invoke subjective judgments and also because the standards in question often conflict with each other. Probably the best example that serves a clear illustration of the associated difficulties would be Kuhn’s skeptical argument about rational theory choice.

2.1.2 Kuhn’s Skepticism about Rational Theory Choice

As Kuhn (1977) sees it, it is possible that the standard of simplicity might lead us to choose one theory while the standard of accuracy might dictate that we choose another theory. Kun gives an example of comparing Ptolemaic system and the Copernican system. Ptolemaic system augmented with many epicycles describes the apparent planetary movements more accurately than the early form of Copernican system does. But the latter seems to be simpler than the former in our ordinary meaning of simplicity. Here we seem to have a conflict between the virtue of simplicity and that of accuracy. Additionally, an individual scientist or a group of scientists might envisage differently which standards are important and how to weight their relative importance. For instance, one scientist might judge that explanatory power is the most important virtue, while another might believe that the most important virtue of a theory is that it be maximally consistent with other pre-existing theories. This interpersonal discrepancy is possible even when the scientists in question are in complete agreement regarding what should be contained in the glossary of scientific virtues; they can nonetheless disagree about how much weight should be given to each virtue.

For this reason, Kuhn argues that there is no determinate procedure of theory choice within a paradigm, not to mention theory choice across different paradigms (Kuhn 1970/1977). To quote:

> *There is no neutral algorithm for theory-choice, no systematic decision procedure which, properly applied, must lead each individual in the group to the same decision. (Kuhn 1970; 200)*

Let us examine in detail how Kuhn was led to such a skeptical conclusion. Kuhn's famous five scientific criteria, so called the 'big five,' for theory choice are: accuracy, consistency (with itself and other accepted theories on relevant aspects of nature), breadth of scope, simplicity, and fruitfulness (Kuhn 1977; 322). These provide the shared, objective basis for theory choice. But Kuhn sees two sorts of difficulties with the use of these criteria for theory choice. First, each individual criterion is very imprecise. Kuhn's own examples will be helpful at this point. Consider accuracy, for example. The oxygen theory accurately account for the observed weight relation in chemical reactions, while the phlogiston theory accounts for the metals being much more alike than the ores from which they were formed. To choose on the basis of accuracy a scientist would need to decide the area in which accuracy was more significant. Secondly, the criteria jointly often conflict one with each other. Kuhn's well-known example (1977; 323) for this point is the comparison between Ptolemy's system and Copernicus's system. Ptolemy's system and Copernicus's system were equally internally consistent. But Ptolemy's was more consistent than Copernicus's with the existing scientific theories at that time. On the other hand, Copernicus's was simpler than Ptolemy's in that the former requires less mathematical equations to explain the observed planetary movements. Hence the criterion of consistency and simplicity are in conflict. Generalizing this point, Kuhn argues, in applying those criteria to theory choice, different scientists may give different weights to them - due to individual differences in their experience in and outside science, in their personal traits, and such. In this way, the procedure of theory choice contains some subjective elements. So, for Kuhn, scientific theory choice is a mixture of subjective and objective factors.

Kuhn's important observation here is that scientific theory choice admits of such subjectivity because the theory choice criteria are not rules but *values*. If they were rules, they would simply *dictate* theory choice – Kuhn thought this was hardly the case. Scientists deploy values as reasons for or against certain choices of theories. So scientists' theory choices are essentially *value judgments*, and reasons have to be given for their judgments, as for any type of value judgment. Recognized as values, then, the criteria like accuracy, simplicity, scope, and such specify a great deal as to how theory choice procedure should proceed, given that list. It would specify what a scientist must consider in reaching a decision, what he may and may not consider relevant, and what he can legitimately be required to report as the basis for the choice he has made (*ibid.*, 331). At the same time, however, values admit of individual variations in their applications. Since they are values, different scientist can legitimately assign different weights to the criteria; scientists can reasonably disagree with each other in their application of values. Then, the five criteria as values underdetermine theory choice, and the underdetermined are to be determined by fleshing out the criteria in ways that vary from one scientists to another. For these reasons, scientists who are “committed to the same list of criteria for choice may nevertheless reach different conclusions” (*ibid.*, 324).

The upshot is that there may be shared criteria, but not a uniquely shared algorithm, for scientific theory choice. Since the criteria are values, there are multiple ways of fleshing out each criterion and of giving weightings to criteria jointly in action. Accordingly, there are multiple possible choices and there are always some “*good reasons for each possible choice”* (*ibid.*, 328). The unique algorithm, which presupposes a fixed interpretation of individual criteria and a fixed weight function between them, therefore, is just an unattainable ideal.

In the light of the above, what Kuhn means when he says there is no neutral algorithm may be that there are too many sufficiently good algorithm.[4] In short, Kuhn’s skepticism about theory choice may be formulized as: Different scientists may employ very different but perfectly legitimate theory choice procedures and reach very different conclusions.

That there are too many legitimate theory choice algorithm is at one extreme. At the opposite extreme is located a nihilistic claim there is no legitimate theory choice algorithm. Recent literature in philosophy of science has paid attention to this nihilistic claim, drawing on some famous results from social choice –Arrow's

[4] This is how Okasha (2011) understands Kuhn. Recently Morreau (2015) argues that Kuhn may not have meant such ‘too many good algorithms’ thesis; what Kuhn seems to have meant is that, since the theory choice criteria are not rules but values (see the earlier quotes from Kuhn), and, for something to be called an ‘algorithm’ it should be a sufficiently specific and prescriptive to dictate choices like a rule, scientists do not operate under such thing as ‘algorithm’ in the first place (Kuhn 1977; 329-31). It is not the aim of this dissertation to settle the proper way to interpret Kuhn. But as my discussion of the Best System Account will involve the issue of subjective and relative aspects of the system choice procedures invoked by the account, for the sake of the dialectic of this dissertation, I will adopt the interpretation of Kuhn as saying ‘there are too many good algorithms’.

theorem, in particular– and their implication to theory choice. In the next section, let us examine Arrow's theorem in social choice. Then we will discuss connection between the theorem and theory choice in Chapter 3.

2.2 Democratic Social Choice and Arrow's Theorem: Getting at Society's 'Will'

The meaning of democracy is "rule by the people". The people, however, can disagree. So the following question naturally arises: How are we to extract the will of society as a whole from individual wills? Social choice theory is the formal study of that question. Its main subject of study are the collective choice procedures by which *individual preferences* are rationally aggregated into a *social preference*. Arguably, the most important breakthrough in social choice theory is Kenneth Arrow's *impossibility theorem* (Arrow, 1951/1963). Arrow's theorem states that there cannot exist any rational aggregation procedure that satisfies certain reasonable conditions. This theorem has shaped the contemporary form of social choice theory.[5] It has also had a significant influence on economics, political science, and philosophy. In this section, I will introduce social choice theory, and one of its most important finding, Arrow's impossibility theorem, and formal statement of the theorem and its conditions.

2.2.1 Individual Preferences and the Difficulty with the Preference Aggregation

One of the most important questions about democracy is the question of how to aggregate individual preferences over alternatives into an overall social ordering of the alternatives. A variety of voting methods serve as examples of procedures that

[5] Suzumura 2002 gives a nice historical overview of the impact of Arrow's theorem on social choice theory.

map individual preferences over candidates into an overall ranking of the candidates in question. A serious difficulty with such a preference aggregation procedure arises when individual preferences are in conflict. A particularly interesting example of such a difficulty is the so-called *paradox of voting*.

Suppose Alf, Betty, and Charlie are trying to decide where to eat dinner from among three restaurants: *x*, *y*, and *z*. However, Alf, Betty, and Charlie disagree with each other on where to go. Alf prefers restaurant *x* to restaurant *y* and restaurant *y* to restaurant *z*; Betty's preference is *z* to *x* and *x* to *y*; and Charlie's preference is *y* to *z* and *z* to *x*. This information is represented below in Table 1. Now, what should be the overall choice in this case? To begin with, *y* wins over *z* in the *y* vis-à-vis *z* pairwise comparison, because *y* is preferred to *z* by two individuals (Alf and Charlie). Next, *x* wins over *y* in the *x*–*y* pairwise comparison because *x* is preferred to *y* by two individuals (Alf and Betty). In sum, *x* wins over *y* and *y* wins over *z*. Assuming preferences are *transitive*, the resulting social preference should be *x* to *z*. Does this settle the case of the restaurant problem for Alf, Betty, and Charlie? No. This is because *z* is also preferred to *x* by two individuals (Betty and Charlie), and so the social preference should be *z* to *x*. The resulting contradiction is that the social preference should be *x* to *z* (by transitivity) and also *z* to *x* (by majority). We have a preference cycle which deems their collective choice incoherent. This generates a case called the paradox of voting, also known as the *Condorcet paradox*.

Note that this paradox concerns a particular pattern of individual preferences. This is why it created a huge reaction to it when Kenneth Arrow (1951) showed that this paradox is generalized into a theorem concerning *all* possible patterns of individual preferences with three or more alternatives. This theorem is now referred to as *Arrow's impossibility theorem*. This impossibility theorem, and its applications, are the main topic of what follows.

Alf	*Betty*	*Charles*
x	*z*	*y*
y	*x*	*z*
z	*y*	*x*

Table 1: the Condorcet Paradox

2.2.2 Impossibility of Rational Social Choice

Arrow's *impossibility theorem* states that there cannot exist any rational aggregation procedure that maps individual preferences into a single social preference ranking, while at the same time satisfying certain reasonable conditions. Here, a 'rational' procedure is defined as one that satisfies a certain set of seemingly plausible conditions imposed by Arrow himself. Given its conditions' plausibility, the impossibility theorem is sometimes interpreted as a destructive blow to the possibility of democratic voting systems, since the theorem seems to imply that rational social choice is unavoidably impossible.[6]

Arrow's impossibility theorem is of particular interest to us because of its implications. As a formal theorem with a few assumptions, Arrow's theorem seems to obtain a wide applicability. The pessimistic conclusion of this theorem may be seen as being extended to any collective choice procedure so long as it satisfies Arrow's

[6] Some argue against interpreting Arrow's theorem in this way. For example, Riker (1982) takes Arrow's theorem as an illustrating example for unattainability of populism, which views voting as the procedure of translating the people's will into the actions of the elected officials.

conditions. This seems to imply that many of the collective choice problems philosophers grant as rational might be bound to be irrational. Given this, Arrow’s theorem appears to have the capacity to make real trouble for many philosophical enterprises. For example, David Lewis’s counterpart theory has been widely accepted by non-essentialists as one the most effective tools for analyzing counterfactuals. Morreau (2010) argues that the multi-dimensional notion of comparative similarity – which underlies standard counterpart theory and some part of Lewis's philosophy such as his analysis of counterfactuals – faces a variant of Arrow's impossibility theorem. As another example, see how Stegenga (2013) applies Arrow’s negative result to the case of evidence amalgamation in epistemology. Zwart and Franssen (2007) connect Arrow’s impossibility theorem with Popper’s concept of verisimilitude. Okasha (2011) suggests that scientific theory choice is formally identical to social choice and is therefore subject to the Arrovian impossibility. I will revisit Okasha’s work on the theory choice analogue of Arrow’s theorem in §2.3. For now let us examine in detail the theorem, its conditions, and some variants of the theorem.

2.2.3 Arrow’s Impossibility Theorem and Its Conditions

Arrow (1951/1963) first presents a set of conditions that he believes any rational social choice procedure must satisfy. They are: *Unrestricted domain* (**U**)*, weak Pareto* (**P**), *Independence of irrelevant alternatives* (**I**), and *Non-dictatorship* (**D**). As a simple illustration, let us see first what these conditions amount to in the restaurant choice example earlier. It doesn’t seem unreasonable to stipulate that Alf, Betty, and Charlie's group choice procedure has to meet the following requirements. First, each of them should be free to rank the alternatives in any order they like and each person's rankings are independent. That is, there should be no constraints on the order of their preferred restaurants (an analogue to **U**). Also, if everyone strictly puts one restaurant above another in their restaurant rankings, then the ranking as a whole group should put the first restaurant above the second (**P**). It also seems like a reasonable requirement that there should be *no* individual whose preference dictates the group ordering regardless of what the others prefer (**D**). And the group choice between *x* and *y* should depend only on the individual preferences over *x* and *y* and should not be affected by the presence or absence of an irrelevant alternative such as *z*. That is, the competition between *x* and *y* should be determined by Alf, Betty, and Charlie's rankings over *x* and *y*, but not *z* (**I**). All these requirements seem quite reasonable. Arrow rigorously proved that there cannot be any collective choice procedure that satisfies all of these reasonable-seeming requirements, and thus that there will always be a violation of some conditions. That is, it is impossible for a social choice procedure to satisfy all of these conditions; hence, Arrow's *impossibility* theorem.

Now it is time for some formality. I largely follow Roberts (2005), Gaertner (2009) and Morreau (2014b)'s denotation. First, let me introduce some preliminary notions.

The weak preference binary relation R. We first need to define *a weak preference binary relation R*. *R* is defined as a subset of ordered pairs in the Cartesian product of X × X, in such a way that *xRy* can be interpreted as '*x is at least as good as y*'. Let me explain. The Cartesian pair of any two sets is the set of all ordered pairs generated by taking the first element from one set and the second from the other. For example, Z ×

$\mathrm{W} = \{(z, w) | z \in \mathrm{Z} \text{ and } w \in \mathrm{W}\}$. So the Cartesian product of X with itself is a set of the all ordered pairs on $\mathrm{X} \times \mathrm{X} = \{(x, x) | x \in \mathrm{X}\}$. A binary relation is a set of ordered pair. That R is a binary relation on X, then, means that R is a subset of ordered pairs on X × X. For example, if $\mathrm{X} = \{x, y, z\}$, then the Cartesian product of X with itself is $\mathrm{X} \times \mathrm{X} = \{(x, x), (x, y), (x, z), (y, x), (y, y), (y, z), (z, x), (z, y), (z, z)\}$. Take a subset of this in such a way that it can be used as a representation of weak preference relation; in the current example, such subset would be $R = \{(x, y), (y, z), (z, x), (x, x), (y, y), (z, z)\}$.[7]

We will use xRy to mean that (x, y) is an element of R. R is assumed to have certain characteristic features as follows. R is *reflexive*, meaning that for all $x \in \mathrm{X}$, (x, x) is an element of R. This follows from the fact that R represents a weak preference including 'as good as'. R is *complete*, meaning that for any two elements $x, y \in \mathrm{X}$, either xRy or yRx or both. That is, all elements of X are connected with each other as ordered pairs in R. R is *transitive*, meaning that if xRy and yRz then xRz. We can call R a preference *ordering* on X in that it is reflexive, complete, and transitive.

Using R as a starting point for analysis, we can define the *strict preference relation P* such that xPy if and only if xRy and not-yRx. The *indifference relation I* can be defined such that xIy if and only if xRy and yRx. Defined in this way, xPy can be interpreted as 'x is strictly better than y,' and xIy as 'there is indifference between x and y.

[7] As we will see shortly, R here displays the desired characteristics to be a weak preference relation.

Preference in social choice. In the context of social choice, R (without subscript) refers to society's preference relation; R_i represents to individual i's preference relation. Following convention, the alternatives will be represented using lower case letters from the end of the alphabet as $x, y, z, \ldots$; the set of all these alternatives is denoted by X. Individual voters, assume to be finitely many, will be represented as numbers $1, \ldots, n$.

Social welfare function. Let X be the set of alternatives. Let N be the society of individual voters and each voter is represented as a number: $N = \{1, 2, ..., n\}$

Let $\mathcal{R}$ be the set of all possible weak preference relations on X. Each voter's preference R_i is drawn from $\mathcal{R}$, so $R_i \in \mathcal{R}$ for all i in N. A preference profile is a list of individual preference relations for all voters: $\langle R_1, R_2, ... , R_n\rangle = \langle R_i\rangle_{i \in N}$. We will use a shortened expression $\langle R_i\rangle$ to denote such a profile when there is no risk of confusion. To denote other profiles, we will use $\langle R'_i\rangle$, $\langle R''_i\rangle$, and so on. Then a profile $\langle R_i\rangle$ is an element of $\mathcal{R}^n$, when $\mathcal{R}^n$ is the n-times Cartesian product of $\mathcal{R}$: $\mathcal{R}^n = \mathcal{R}\times\mathcal{R}\times ... \times\mathcal{R}$.

Arrow's social welfare function (SWF) may be thought of as a procedure of aggregating individual preferences into an overall social preference. More formally, a SWF f is an aggregation procedure which generates a social ordering as a function of individual orderings: $R=f(\langle R_i\rangle)$ on X. As mentioned earlier, the convention is to use R and P for social preference, weak and strict, respectively. For example, 'xRy' means 'society weakly prefer x to y', 'zPw' means 'society strictly prefers z to w'. Now these social preferences can be viewed as being derived from $\langle R_i\rangle$, through the functional relation between individual and social preferences, as clearly seen in the notation $R=f(\langle R_i \rangle)$. The social preference derived from $\langle R'_i \rangle$ will be denoted as $R'=f(\langle R'_i \rangle)$

Arrow's theorem. Arrow's impossibility theorem states that there exists no SWF that satisfies the following four conditions:

Unrestricted Domain (**U**). The domain of f includes all logically possible profiles $\langle R_i \rangle$. In words, a SWF should be able to handle all logically possible lists of individual rankings of the alternatives.

Weak Pareto (**P**). If xP_iy for all i in N, then xPy. In words, if every individual strictly prefers x to y, then the society must prefer x to y. In words, when everyone unanimously strictly prefers one alternative to another, the social ordering generated by f should agree.

The logically stronger counterpart of the weak Pareto condition is the strong Pareto principle: For all x and y, if xR_iy for all i in N and xP_ky for some k in N, then xPy. In words, the weak Pareto condition requires that if every individual unanimously prefers x to y, so does the society. The strong Pareto condition requires that if every individual unanimously regards x as at least good as y and at least one individual strictly prefers x to y, then society must strictly prefer x to y. The strong Pareto is 'strong' in that it excludes more alternatives from being chosen than the weak Pareto. For example, suppose everyone in society initially indifferent between x and y. Under the weak Pareto, x will be socially preferred if everyone changes their mind from indifference to strict preference of x to y. But under the strong Pareto, only one person's changing mind from indifference to strict preference of x will have the effect of dropping y from the socially chosen set.

Independence of Irrelevant Alternatives (**I**). For any two preference profiles $\langle R_i \rangle$ and $\langle R'_i \rangle$, for any two alternatives x, y and for all i, if $\langle R_i \rangle$ and $\langle R'_i \rangle$ coincide over the pair x and y, then $R=f(\langle R_i \rangle)$ and $R' = f(\langle R'_i \rangle)$ should coincide over the pair x and y. In words, if every individual has exactly same preference concerning x and y in the two profiles $\langle R_i \rangle$ and $\langle R'_i \rangle$, then the social preference over x and y must exactly same for the two profiles. In other words, if two individual preference profiles agree with each other on the subset $\{x, y\}$ of X, then the society must have the same preference on that subset.

Non-Dictatorship (**D**). There is no individual i such that for all profiles in the domain of f and for all pairs of alternatives x and y in X, if xP_iy, then xPy. In words, there should be no individual who always gets his or her way regardless what the other individuals prefer.

Arrow summarizes his justification for these conditions as: "they express the doctrines of citizens' sovereignty and rationality in a very general form, with the citizens being allowed to have a wide range of values" (Arrow 1963; 31). Now we are stating Arrow's Impossibility Theorem.

Arrow's Impossibility Theorem (1951/1963): For a finite number of individuals and at least three distinct social alternatives, there is no SWF f satisfying conditions **U**, **P**, **I**, and **D**. The theorem is often described as stating that for any SWF f satisfying **U**, **P**, and **I**, there is a dictator.

Some discussion will help understand these conditions imposed by Arrow. Condition **D** seems indispensable for a democratic society; dictatorship appears to contradict the spirit of democracy. Condition **U** also seems to capture an essence of democracy in that it requires the social choice procedure to be able to take into consideration all individual preference orderings no matter how ‘odd’ they are; no one is to be disenfranchised just because his or her preferences do not sit well with other people's preferences. Condition **P** seems reasonable –a society that doesn’t prefer the unanimously strictly preferred alternative can hardly be called a society governed by the citizens.[8]

Condition **I** is more complicated than the other conditions. It requires the SWF to be “informationally parsimonious.”[9] When it comes to *x* vis-à-vis *y* comparison, for instance, the social preference should take into consideration *only* the individual preference *orderings* of *x* and *y*. Sen (1970, ch7) describes Arrow’s **I** condition as having two aspects: the “irrelevant” aspect and the “ordering” aspect. Firstly, the social preference between two alternatives should be only determined by the individuals’ preferences between *them*; for example, individuals' preferences concerning ‘irrelevant’ alternative *z* should not enter the social preference between *x* and *y*. Secondly, only the alternatives’ *rankings* matter, that is, only the information about which alternative comes first in their *ordering* is admissible. Construed this way, this condition is not as straightforward as the other conditions. As a matter of fact, there are many SWFs that are in violation of condition **I**, one of them being a social choice method called *Borda method*. In Borda’s method: for *m* alternatives, each individual voter assigns numeric point *m* to her most favorite alternative and 1 to

[8] But it doesn’t come without a problem. Sen (1970b) offered a critique of **P** in his famous *Lady Chatterley’s Lover* example, showing the weak Pareto principle can conflict with an individual’s right. In short, the example involves a cases where x is socially preferred to y by one person’s right, y is socially preferred to z by another person’s right, but by weak Pareto z is socially preferred to x, resulting in a social preference cycle.

[9] Gaertner 2006; 18.

her least favorite one. Each alternative’s points are added up and the one with the most points wins. Consider a very simple society consisting of two individuals and three alternatives. Suppose the society is concerned with the two following profile $\langle R_i \rangle = \langle R_1, R_2 \rangle$ and $\langle R'_i \rangle = \langle R'_1, R'_2 \rangle$:

Profile $\langle R_i \rangle$		Profile $\langle R'_i \rangle$	
1	2	1	2
a	b	a	b
c	a	b	c
b	c	c	a

Concerning the preference over a and b, each individual has the same preference ranking of a and b in the both profiles $\langle R_i \rangle$ and $\langle R'_i \rangle$; the first individual prefers a to b and the second prefers b to a in both of the profiles. Then condition **I** requires the social preference between a and b to be same for the both profile. However, the society using the Borda method would have the following result:

For profile $\langle R_i \rangle$: the Borda count of $a = 5 >$ the Borda count of $b = 4$, so aPb.

For profile $\langle R'_i \rangle$: the Borda count of $a = 4 <$ the Borda count of $b = 5$, so $bP'a$.

So we can see that the Borda method violates condition **I**. What happens is that the Borda method take into consideration some information about how far the relative distances are between the positions of alternatives in the profiles, while condition **I** strictly prohibits the use of such information. Also, in the Borda method, something about interpersonally comparability is assumed. Each voter gets equal weight in the Borda method in that the most favored alternative gets the count of m for any voter. So the Borda rule is a way of making interpersonal comparisons (Sen 1970, Roberts 2005). The “ordering” aspect and the “irrelevant” aspect of I jointly have the effect of excluding the irrelevant alternatives and the interpersonal comparison of ordinal preferences of individuals together.

Sen (1970) proposes the incorporation of richer information than individual preferences into Arrow's original framework. In this 'information enriched' approach to social choice, individual preferences are represented not as orderings but as utility functions that denote utility of individuals in alternative states by numerical representation. Instead of a SWF, we have a social welfare functional (SWFL) f mapping utility representations into a social ordering. In this framework, Arrow's **I** is equivalent to the conjunction of the two conditions: Independence of Irrelevant Utilities (**IIU**) and Ordinal Non-Comparability (**ONC**).[10] They can be seen as formal statements of what are excluded by the two aspects of **I**, as we saw in the Borda count example. We will have more discussion on Sen's information enriched framework and various forms of measurability and comparability in §2.4.

The proof for Arrow's theorem has been extensively explored in social choice literature and there are a number of different proofs. Arrow's original proof is in Arrow (1963). Geanakoplos (2005) provides three elegant and easily accessible proofs for the theorem. A diagrammatic proof for Arrow's theorem can be found in Blackorby, Donaldson, and Weymark (1984). Also Gaertner (2009) contains a readily accessible exposition of a number of different versions of the proof for the theorem.

Arrow's theorem is often considered as having dramatically shown the difficulty – an 'impossibility' – with aggregation procedures. Appreciating the gravity of its implication, many have been led to apply the theorem to another areas that call for

[10] See, for example, Hammond and Fleurbaey (2004).

aggregation procedures. One of such attempt is made to the problem of theory choice by Okasha (2011)'s recent work in philosophy of science. Let us turn to the question whether the Arrovian result obtain for theory choice.

2.3 Skepticism about Theory Choice

Recently Okasha (2011) explored the possibility of applying Arrow's theorem to theory choice. He argues that there is a formal analogy between social choice and theory choice, and hence that the Arrovian impossibility result in social choice also applies to theory choice. In this section, we will examine whether theory choice is sufficiently analogous to social choice and whether the theory choice analogues of Arrow's conditions are motivated.

2.3.1 Analogy between Social Choice and Theory Choice

There seems an analogy between social choice and theory choice. The analogy is that candidate theories in theory choice are like the alternatives in social choice; and that the theory choice criteria in theory choice are like the individuals in social choice. We may see that theories are ranked by each criterion just like candidates are ranked by individual voters. Just as there is a problem of aggregating individual preferences into an overall social preference of the alternatives in social choice, there seems to be a problem of aggregating theory rankings by each criterion into an overall ranking of the theories in theory choice.

As we saw in §2.1, Kuhn recognized challenges to theory choice, especially when there are multiple criteria against which competing theories are to be compared. The primary source of these difficulties was the fact that the theory choice standards in question often conflict with one another. One theory can be better than another in one dimension, (for instance, simplicity), but worse in another dimension, (for instance, fit).[11] Assuming that this can happen with regard to any standard, we can readily construct a case of the Condorcet paradox for theory choice, as shown in Table 2 below. Then, assuming that Arrow's conditions are met in the case of theory choice, we may extrapolate the Arrovian impossibility into theory choice, and are thus justified in concluding that there is no reasonable (in Arrow's sense) procedure for maximizing the different standards of scientific merits. The upshot would be that if there is no way of getting around Arrow's impossibility, then there seems to be no rational algorithm for theory choice. If this is true, then, as Okasha points out, theory choice will have to face much more serious problems than Kuhn worried about in theory choice. Let us examine whether it is indeed the case that an analogue of Arrow's theorem holds in theory choice.

Simplicity	*Fruitfulness*	*Accuracy*
X	*Z*	*Y*
Y	*X*	*Z*
Z	*Y*	*X*

Table 2: a Condorcet Paradox analogue for theory choice

2.3.2 Applying Arrow's Theorem to Theory Choice

Okasha (2011) argues that the conditions imposed by Arrow on social choice are well motivated, or at least not unreasonable, impositions in the domain of theory choice.

Non-dictatorship

[11] Baumann (2005), for example, observes this 'non-linearity' of multiple dimensions. Although he does not explicitly mention the results from social choice theory but his concept of non-linearity reflects what is called 'double-peaked' of preference profile in social choice.

Let us first consider the non-dictatorship condition (**D**). The analogue of **D** for theory choice would be that there should be no ‘dictatorial’ criterion. Violation of this condition would mean that there is a particular dictatorial criterion of theory choice such that if a theory comes out to be better than another with respect to that criterion, then that theory is just to be chosen as the overall winner regardless how it fares with respect to the other criteria. Suppose simplicity is such a dictator criterion. This would imply that a ridiculously simple theory saying “Everything true is true” is the best theory because it is the simplest, regardless of the fact that it has no merit in terms of informativeness. Clearly, this result is undesirable. Theory choice algorithms presumably must factor in all dimensions of evaluation.

At this point, however, one might wonder - for example a 'hardcore' empiricist - why accuracy (fit-to-the-data) shouldn't be a dictatorial criterion in a strong sense. She might argue that if a theory fits the observed facts perfectly then it is a decisive reason to choose that theory regardless of how well it fares with other standard. Or, a more moderate empiricist might argue that accuracy should be at least a lexicographic dictator - i.e., choose a theory that has maximum accuracy and, when theories are in tie with respect to accuracy, then move on to the other criteria to use them as tie-breakers. In either case, one who is rooted in empiricism would conclude that accuracy assumes a special status in theory choice - something like a 'benevolent dictator' - hence condition **D** is not motivated in theory choice.

Okasha provided a plausible response to this kind of question. The answer lies in the phenomenon called *overfitting*. Overfitting is a phenomenon that occurs when one fails to distinguish between noise and genuine information in the observed data. In most of the cases the observed data is noisy, so the theories that fit the data ‘too well’ have likely failed to distinguish the true information from the noise; this is likely to result in poor performance in predicting future data.[12] Then it is reasonable *not* to take accuracy as a dictator criterion, even on the grounds that are compatible with empiricism. The attempt to relax condition **D** on the basis of empiricism like above is, then, unmotivated.

Unrestricted Domain

The analogue of the unrestricted domain condition (**U**) for theory choice would be that a theory choice algorithm should handle all logically possible profiles of theory rankings with respect to the theory choice criteria. Okasha takes this condition to be well-motivated in theory choice. He argues that it would be absurd for a theory choice procedure to favor or disregard particular patterns of theory rankings.

In his recent paper, Morreau (2015) argues that **U** does not apply to theory choice, while noting that the rich domain condition (**R**), a weaker condition than **U**, may apply to theory choice. Shortly we will have discussion on **R**. For now let us focus on **U** condition.

At first blush, **U** might appear to apply to theory choice. As Okasha (2011) suggests, it might seem like a reasonable imposition that there should be no *a priori* restrictions on what profiles are admissible and what not to theory choice rule.[13] Morreau (2015) argues this is not the case. Once we fix the sense of simplicity to use to rank theories, simplicity ranking of two theories is invariant regardless of data. Consider the context of statistical model selection, for example. Suppose model LIN claims that the true

[12] See Forster and Sober 1994. We will have further discussion on this point in Chapter 3.

[13] In response to Morreau (2015), Okasha (2015) concedes that **U** does not apply to theory choice.

relation between random variables x and y can be represented in the form of "$y=ax+b$", and model PAR claims that the true relation can be represented in the form of "$y=ax^2+bx+c$". In this case, once the relevant sense of simplicity is fixed as the number of parameters, then LIN is simpler than PAR no matter what.[14] The simplicity ranking between LIN and PAR could not be reversed in any possible situation; the simplicity ranking is rigid. The strength ranking of theories also seems rigid in some cases. If strength of a theory is defined as the set of its logical consequences, in case where the set of one theory's logical consequences is a proper subset of the set of another theory's logical consequences, the second theory will come out be stronger than the former no matter what.[15] Hence, Morreau argues, the admissible profiles for system choice are fairly restricted, so **U** is not applicable to theory choice.

Rich Domain

It is now agreed that Unrestricted domain is not required for the Arrovian impossibility; something strong enough is required, that is, the domain should be diverse enough (Kelly 1978; ch7, Pollak 1979; 76-7, Campbell and Kelly 2002; 64-5, for example). For example, the so-called Pollak diversity condition requires that, for any logically possible profile over three 'hypothetical' alternatives (x, y, z), then there exist three alternatives (a, b, c) such that the profile restricted to that triple coincide with the profile over the hypothetical triple. Morreau (2014a, 2015) elegantly subsume Pollak's pioneer work and similar studies on diverse domain conditions under *Rich domain* (**R**) condition. Morreau defines a *pattern* as a list of weak orderings of some

[14] We also need to assume that the coordinate system for the two models is fixed. See Priest 1976 for an argument that curve-fitting is susceptible to language dependence problem.

[15] However, there are problems with measuring strength in this way. I will discuss it in §4.3 and §5.5.

set of logical variables (not actual alternatives). A profile is said to *realize* a pattern if there is a matching between a set of variables in the pattern and a set of alternatives in the profile. Morreau's Rich domain condition is: A domain is *rich* if for every suitable pattern P of three variables, there is some profile in this domain that realizes P. In words, a domain is rich if orderings of three alternatives are showing patterns which coincide in one way or another with all possible orderings of three hypothetical, not actual, variables. According to **R**, what should be unrestricted in theory choice is the patterns of variables. This rich domain, along with the other suitably modified analogue[16] of Arrow's condition, is enough to give a rise to a variant of the Arrovian impossibility.

As we saw, **U** does not apply to theory choice. Rich domain (**R**) may or may not apply to theory choice.[17] But literature in social choice and theory choice (Parks 1976, Pollak 1979, Hammond 1976, Kemp and Ng 1976, Roberts 1980, Rubinstein 1984, Feldman and Serrano 2008; Morreau 2014a, 2015) shows that even if **U** is replaced by **R** (or some similar conditions in early literature)**,** a variant of Arrow's impossibility theorem obtains provided the strong neutrality condition (**SN**), a stronger version of **I**, is met. That is, simply weakening **U** to **R** in theory choice does not open up an escape route from the Arrovian result. Shortly we will have discussion on **SN** and possible escape routes from the Arrovian impossibility. For now, let us continue to examine the theory choice analogues of the conditions for Arrow's theorem.

[16] It is Strong neutrality condition (**SN**). Shortly we will have discussion on it.

[17] Morreau (2014a) provides an illustrating example for **R** in theory choice. See §5.2 for further discussion.

Weak Pareto

The theory choice analogue of the weak Pareto condition (**P**) would be that if a theory fares better than another under every criterion, then that theory is the winner. This seems a well-motivated analogue. As briefly noted, there has been inquiries about the adequacy of **P** in social choice concerning a potential conflict between individual rights and **P** (Sen 1970b, for example). In the domain of theory choice, no such concerns seem to arise because the theory choice standards are not like agents with power to exercise their rights as in Sen's critical discussion on **P**. It seems plausible that theory choice procedures have to respect unanimity among the theory choice criteria.

Independence of Irrelevant Alternative

The theory choice analogue of Independence of irrelevant alternatives condition (**I**) would be that if two theory merits (standards) profiles agree with each other on the pair of two theories, then the theory choice procedure must have the same preference on that pair. For example, if two profiles agree on the orderings of T_1 and T_2 with respect to the theory choice criteria, than the resultant overall ordering of theories should not depend on the presence or absence of a third, irrelevant alternative theory T_3. Suppose T_1 comes out better regarding accuracy and informativeness and worse regarding simplicity and fruitfulness than T_2, and a theory choice algorithm ranks T_1 above T_2 in their overall theoretical merit ranking. Then **I** in this example would mean that the algorithm should produce the same overall ranking whenever any two competing theories exhibit the same ranking patterns as above. This condition seems reasonable because presumably it would not make sense to allow the overall ranking of T_1 and T_2 to be affected by something other than their own rankings with respect to the theory choice standards.

At first blush, the theory choice analogues of the four social choice conditions seem reasonable. From this, Okasha suggests, we can infer that the Arrovian impossibility result will arise for theory choice as well.

2.3.3 The Nihilistic Result Regarding Rational Theory Choice

If the Arrovian impossibility obtains for theory choice, how serious problem would it be for theory choice? If an analogue of Arrow's impossibility results holds in theory choice, it would mean that theory choice cannot be rational, at least in its ideal sense. Assuming that the conditions imposed by Arrow are all reasonable, and that the theory choice analogues of them are motivated in theory choice, one would face a gloomy conclusion that there is no coherent theory choice algorithm.

It is at this point that Okasha makes an important observation about Kuhn’s ‘no unique algorithm’ thesis about theory choice and the implication of Arrow’s theorem regarding theory choice. As we saw in §2.1.2, Kuhn says that there is no determinate and unique theory choice algorithm. Even if the scientists in question share the same criteria for theory choice, they may very well give different weights to different criteria. Each of them is justified in using their own weight metric. So, in a sense, there are ‘too many’ legitimate algorithms, and we have no objective and 'transcendent' ground to determine which algorithm is more appropriate. Upon applying Arrow’s theorem, we now have the result that there is ‘no’ consistent algorithm at all. As Okasha describes, Kuhn makes theory choice very difficult, and Arrow makes it impossible.

When Okasha's observation is coupled with Arrow's theorem, an even more radical irrationalism about theory choice can be drawn than Kuhn's view about science. Kuhn at least admitted the possibility of consistent theory choice relative to a given paradigm during the "normal science" phase.[18] The Arrovian result for theory choice, however, implies that even during the normal science phase within a paradigm, there is no rational and consistent procedure for theory choice. This appears to deal an even more devastating blow to the rationality of science than that of Kuhn.

2.4 An 'Escape Route': Cardinalism with Comparability

Solutions have been suggested to the Arrovian impossibility in the literature of social choice theory. Probably one of the most commonly discussed solutions is to adopt Sen (1970)'s 'information enriched' approach by implementing different measurement scales for individual preference. Recall that Arrow's original characterization of social choice is informationally impoverished - it only allows information about ordinal rankings of alternatives. Okasha (2011) proposes, with some caveats, that this notion of information broadening can be an escape route from the Arroviam impossibility in the context of theory choice - it could potentially salvage the rationality of science. Following his lead, I will examine if the said route is indeed open to theory choice as Okasha proposes in this section. First, I will introduce Sen's extended framework in which we can utilize more information than Arrow's framework. Then I will have a brief discussion on the possibility of relaxing Universal domain condition of Arrow and the Arrovian impossibility in the single-profile framework. The rest of the chapter will focus on the possibility of cardinally

[18] Kuhn (1970)

measurable comparability between theory choice criteria, drawing on the literature in statistical model selection. This will lead us to Chapter 3.

2.4.1 Sen's Information Enriched Framework: Measurability and Comparability

An immediate response to the threat of the Arrovian impossibility would involve the use of the cardinal measure of preferences. Here is a possible line of response: In the restaurant example in §2.2, very limited information was used. The only information admitted to use in the aggregation procedure was the orderings of the restaurants by individuals, that is, the *ordinal* rankings of the alternatives. But can we not use richer information about individual preferences? For instance, couldn't it be the case that the intensity of Alf's preference of *x* to *y* is far greater than Charlie's preference of *y* to *x*? If it is the case that Alf is a very sensitive gourmet whereas Charlie cares little about what he eats as long as he gets fed, then we may justifiably give more weighting to Alf's preference than Charlie's. Furthermore, the *intensity* of their preferences may be represented with some numerical values. Provided that such a measure of individual preferences is available, then, it might seem to be a perfectly reasonable procedure in which the winner is the alternative that has the greatest sum of utilities across individuals. This might serve as an escape route from the Arrovian impossibility, one might think.

This line of thought is on the right track but is still missing a very important element: interpersonal comparability. Let me illustrate the point with a simple example. The procedure of getting the sum of individuals' cardinal utilities may be represented as follows:

$$W(u_1, \dots, u_n) = \sum_{i=1}^{n} \lambda_i u_i$$

where u_i are utilities that individuals get from the alternative in question and λ_i are weightings to individuals. If we are to give equal weight to everyone, then $\lambda_i = 1$ for every *i* (‘Benthamite’ utilitarianism). If we are to give unequal weights to individuals, as in the example of one person being a very sensitive gourmet and another being a dull person, then of course λ_i will get different values for different individuals (weighted utilitarianism). Note that both specifications of utilitarianism presuppose that individual values λ_i are individually measurable and comparable across individuals.[19]

At this point, it would be convenient to introduce a framework in which we can utilize the formal characterizations of various types of utility measures and comparability. Sen’s ‘information enriched’ framework provides such a framework. Sen (1970) generalized Arrow's model to incorporate information richer than just orderings of alternatives. Morreau (2014b) and List (2013) provide a thorough overview of the framework and recent developments. The following presentation is largely taken from them.

In Sen’s framework, the preferences of individuals are presented not simply as orderings R_i but as utility functions U_i that map the alternatives onto real numbers: $U_i(x)$ is the utility that *i* obtains from *x*. A utility function U_i contains at least as much information as an individual preference ordering because we can reduce it to an ordering by putting xR_iy if $U_i(x) \geq U_i(y)$. Given different utility functions can

[19] Gaertner (2009) gives a readily accessible overview of the topic.

be reduced to the same ordering, using utility functions generally delivers more information than orderings. In Sen's framework, a preference profile is a list of utility functions: $\langle U_i, \dots, U_n \rangle$. Instead of a SWF, now we have a social welfare *functional* (SWFL) *f* mapping each profile onto a social weak ordering of the alternatives.

In terms of utility information, it is usual to view utilities as being *ordinal* or *cardinal*. In the literature of social choice, the idea of the *invariance transformation* (Roberts 1980, 2005) is generally used to represent various types of utility measures formally. An invariance transformation ϕ_i has the property that $\phi_i(U_i)$ is informationally equivalent to U_i. As we saw above, ordinal utility means that utility information defines an ordering of alternatives - the levels of utility can be ordered - but no more, in a way that the ordering is given a numerical utility representation by assigning higher utility to more preferred alternative. So, if preferences are to be measured as ordinal utility, invariance transformation ϕ_i would be strictly monotonic transformation. Cardinal utility delivers more information than just the ordering of alternatives, say, 'intensity' of preference. More specifically, if preferences are measured on a cardinal scale, then the relevant invariance transformation would be positive affine transformation: $\phi_i(\underline{U_i}) = \alpha_i U_i + \beta_i$, where $\alpha_i > 0$.[20]

One important finding in social choice was that having cardinal utilities is not by itself enough to avoid an impossibility result. In addition, utilities have to be interpersonally comparable. (Sen 1970; ch8). Sen (1970) and Kalai and Schmeidler

[20] In social science, an interval scale is the frequently used type of cardinal scale. A nice analogy is the concept of temperature. Suppose the difference in temperature between *a* and *b* is two times as much as the difference between *c* and *d.* This information remains same independently of whether we measure their temperature in Celsius or Fahrenheit. This is because moving from Celsious to Fahrenheit is simply applying a positive affine transformation to the Celsius values.

(1977) show that if no interpersonal comparisons of preference are permitted, then the impossibility conclusion of Arrow's theorem remains true, even if Arrow's ordinal interpretation of individual utility is replaced by a cardinal interpretation.

The claim that Alf's utility is comparable to Betty's utility implies that we can assign certain weights to individual utilities. In the context of cardinal utilities, we have to be able to say something like "A certain loss of utility in Alf can be compensated (i.e., can be traded for) by an equal gain in utility by Betty." This statement expresses comparisons of utility intervals in different alternatives between individuals. For example, only under circumstances such as these can classical utilitarianism be achieved through an interpersonal comparison in which the individual cardinal preferences are summed into an overall social preference. In the context of ordinal utilities, the relevant comparability would be the utility level comparability across individuals. For example, Rawlsian utilitarianism requires comparison of utility levels of the worst-off individual in each alternative state. The upshot is that the Arrovian impossibility still obtains if there is no interpersonal comparability (Sen 1970, 1986, Hammond 1976, Roberts 1980, d'Aspremont and Gevers 1987). Formally,[21]

Ordinal measurability with no interpersonal comparability (**ONC**):

Two profiles $\langle U_1, U_2, \ldots, U_n \rangle$ and $\langle U'_1, U'_2, \ldots, U'_n \rangle$ contain the same information whenever, for each $i \in N$, $U^*_i = \varphi_i(U_i)$, where φ_i is some positive monotonic transformation, different for different individuals.

Cardinal measurability with no interpersonal comparability (**CNC**):

[21] See Roberts 1980 for a clear exposition of different types of measurability and comparability. In this dissertation I follow List (2013)'s characterization. There is essentially no difference between minor differences in notations.

Two profiles $\langle U_1, U_2, \ldots, U_n\rangle$ and $\langle U'_1, U'_2, \ldots, U'_n\rangle$ contain the same information whenever, for each $i \in N$, $U^*_i = a_i U_i + b_i$, where the a_is and b_is are real numbers (with $a_i > 0$), different for different individuals.

The finding in social choice theory is that, even under **CNC** (and **ONC** as well), if there are three or more alternatives in X, there exists no SWFL satisfying **U**, **P**, **I**, and **D**. So, cardinality alone cannot open up an escape route from the Arrovian impossibility.

As mentioned earlier, Arrow's **I** condition is equivalent to the conjunction of **ONC** and **IIU** (Hammond and Fleurbaey 2004). Here is a statement of **IIU**:[22]

Independence of irrelevant utilities (**IIU**): For all alternatives x and y in X, and all utility profiles $\langle U_i\rangle$ and $\langle U'_i\rangle$, if $\langle U_i\rangle$ and $\langle U'_i\rangle$ coincide restricted to the pair x and y, then $f\langle U_i\rangle$ and $f\langle U'_i\rangle$ should coincide restricted to that pair.

In words, if **IIU** requires social preferences over a subset of a pair of social alternatives to depend only on utility levels on this subset, and not at all on utilities at any other alternatives of X. So this captures the "irrelevant" aspect of **I** in Sen's framework.

One cannot simply claim that there exists interpersonal comparability. She needs a theoretical justification and/or empirical demonstration of existence of such comparability. Does such a justification or demonstration exist? This is the question we will investigate in this chapter and also in chapter 4 in the domain of system choice. For now, let us briefly examine differences in Arrow's and Sen's perspectives

[22] From Hammond & Fleurbaey 2004, with some changes on notations to fit the notation in this dissertation.

on measurability and comparability. This examination will shed a light on what we should seek for if we were to escape from the Arrovian impossibility.

Arrow himself was a firm believer in *ordinalism,* the idea that the only meaningful form of preference is ordinal preference because cardinal utilities are not observable. In the same vein, he was also a denier of interpersonal comparability. What he says about cardinal interpersonal comparability is worth noting:

> *The oldest critique of social choice theory ... is that it disregards intensity of preference. Even with two alternatives, it would be argued that a majority with weak preferences should not necessarily prevail against a minority with strong feelings. The problem in accepting this criticism is that of making it operational.* **Theoretically, is there any meaning to the interpersonal comparison of preference intensities? Practically, is there any way of measuring them, that is, is there any form of individual behavior from which the interpersonal comparisons can be inferred?** *(Arrow 1984; 172, my emphasis)*

The emphasized part is of the most importance. Arrow himself is aware that allowing interpersonal comparison provides an escape route from his impossibility theorem; the problem is, for him, there is no theoretical or practical ground for the interpersonal comparison for it to be used in aggregating personal preferences. So we are left with the question of what the interpersonal comparison amounts to.

According to Sen, there is at least some form of interpersonal comparability we can make sense out of. For example, it is sensible to say that Emperor Nero's gain from burning Rome was outweighed by the loss on the part of all the other Romans, and that the grounds for saying something like this result from an (somewhat loose form of) interpersonal comparability of utilities. An ethical observer would say the state where Nero fiddles is socially less preferable in comparison to the state where Nero does not.[23] We cannot precisely pinpoint how these comparisons are to be made, Sen argues, but it is at least meaningful to state that the latter is better than former.[24] This Sen–Arrow debate on interpersonal comparability deserves a separate discussion, but for the purposes of this dissertation it is sufficient to note that it is very important for us to ask what this comparability consists in and where it comes from, if we are to avoid Arrovian impossibility. More importantly, if we want to escape from the Arrovian impossibility in the domain of theory choice, it is imperative for us to seek out and justify a similar type of comparability between theory choice standards. This is what Okasha (2011) explored. Inter-criterial comparability in theory choice is probably the most interesting and promising escape route, but there seem other possible routes from the Arrovian impossibility too. So, let us explore some of them first, and then move onto the question of comparability.

2.4.2 A Possible Response: Relaxing Condition U?

As mentioned in §2.3.2, Morreau (2015) observes that **U** is not motivated in theory choice. The simplicity and strength rankings of theories, at least in some cases, are rigid; their rankings *could not* be different than they actually are. So, one might hope, dropping **U** in theory choice might open up an escape route. But this is not necessarily the case. Let us see why.

[23] Sen (1999)

[24] This example expresses *partial* comparability. Roberts (2005) suggests an aggregation procedure in the case of partial interpersonal comparability. If social ranking were to be made by just sum of utilities, then the case of partial comparability will generate an incomplete ordering. Alternatively, if the expected sum of utility differences between alternatives is used, based on a probability distribution over utility functions capturing degrees of belief of individuals, then complete social ranking of alternatives can be made (requiring some conjectures, of course, about degree of beliefs).

Rich domain (**R**) may or may not apply to theory choice (Morreau 2014a gives a toy example that **R** applies; but it remains to see whether **R** applies to the realistic context of science). But the literature in social choice (Parks 1976, Pollak 1979, Hammond 1976, Kemp and Ng 1976, Roberts 1980, Rubinstein 1984, Feldman and Serrano 2008) show that even if **U** is replaced by **R,** a variant of Arrow's impossibility theorem obtains provided the strong neutrality condition (**SN**), a stronger version of **I**, is met. That is, simply weakening **U** to **R** in theory choice does not open up an escape route from the Arrovian result. Here is a statement of Strong neutrality condition[25]:

Strong Neutrality (**SN**): For all w, x, y, z in the set of alternative X, and for all profiles $\langle R_i \rangle$ and $\langle R'_i \rangle$,

If, for every i in N, [xR_iy iff zR'_iw] and [yR_ix iff wR'_iz], then [xRy iff $zR'w$] and [yRx iff $wR'z$].

As we can see, **SN** is more stringent than **I**. **I** requires consistency for each pair of alternatives separately. Figuratively speaking, **I** means that when the social welfare function aggregates individual orderings, it should take each pair of alternatives separately, paying no attention to preferences for alternatives other than the pair in question.[26] I requires consistency between two profiles over a pair each time; it leaves possibility that different pairs might be treated differently. For example, when $\langle R_i \rangle$ and $\langle R'_i \rangle$ coincide on x and y, and the individuals exhibit the exactly same pattern of

[25] See, for example, d'Aspremont and Gevers (2002; 493–494). They provide the formal definitions of Intraprofile Neutrality (**IAN**) and Strong Neutrality (**SN**), which clearly indicate that the latter is the stronger condition than the former. This is, roughly put, because **SN** requires social choice rule to be consistent over different pairs across different profiles and **IAN** requires consistency over different pairs within a profile. Given **IAN** is a special case of SN (when two profiles are equated), SN is the stronger imposition on a social choice rule than **IAN**.

[26] See Morreau (2014b)'s illustrative examples on this point.

preference orderings on *z* and *w* as they do on *x* and *y*, what **I** requires is that the social preference ordering of two profiles should agree on the pair of *x* and *y*, and agree on the pair of *z* and *w*, separately. But **I** does not require that the social ordering of two profiles are same across the pair of *x* and *y* and the pair of *z* and *w*. As it should be clear now, it is **SN** that precisely requires such consistency across different pairs. In theory choice, is the analogue of **SN** motivated? **SN** is a fairly strong condition, and in social choice, it has the effect of forbidding social choice procedure from using non-utility information.[27] In theory choice, the analogue of **SN** would require that theory choice procedure should only use information about how well theories fare with respect to the theory choice criteria; for example, the *identity* of theories should not enter the procedure. It seems unclear whether **SN** applies to theory choice. On the one hand, we can think of some examples where other kinds of information seem to be allowed in theory choice procedure. If two theories, for example a descent of Darwinianism and a descent of Creationism, are in competition, scientists may take into consideration information about the theoretical lineage of the two theories. Or, scientists working in different branches of science may judge theories in different contexts of interest. On the other hand, it seems to desirable that theory choice procedure is as 'neutral' and consistent as possible, for theory choice to be rational in the most common and intuitive sense of the term. So, we don't seem to have theoretical or empirical ground for outright rejection or acceptance of **SN** in theory choice. Noting the matter of **SN** in theory choice remains to be seen, let us continue to explore another possible escape route from the Arrovian impossibility.

[27] The view that individual utilities should be the only basis for deriving social preferences is called *welfarism*. Sen, among many others, offered critique of the view in Sen 1979.

2.4.3 A Possible Response: the Single-Profile Framework?

As we saw in §2.2, Arrow's impossibility theorem is concerned with how social choice rule is to generate a social ordering over a set of alternatives for *every* logically possible profile of individual preferences over the alternatives.[28] However, in actual situations, there is only one profile: the actual profile of how individuals actually prefer the alternatives. So, it may seem that, in a given actual situation, social choice rule only needs to generate a social ordering for one fixed, actual profile. Based on this consideration, some claimed that Arrow's nihilistic conclusion should be rejected (Little (1952), Samuelson (1967)) because the conditions imposed by Arrow are defined in the multi-profile framework. According to these objectors of Arrow, individual preferences are given and social choice procedure only need to determine the best alternative given those individual preferences; and if individual preferences change then we just have "a new world and a new order" (Little 1952; 423-424). Requiring social choice rule to be sensitive to all logically possible profiles like Arrow did is just "an infant discipline of mathematical politics" rather than that of appropriate welfare economics, hence we should "export Arrow from economics to politics" (Samuelson 1967; 42).

In response to objections like above, literature in social choice theory in the late 1970's and early 1980's showed the single-profile variants of Arrow's theorem obtains for a fixed preference profile if the profile is diverse enough and the intra-profile counterparts of Arrow's inter-profile conditions are met (Fishburn 1973, Parks 1976, Hammond 1976, Kemp and Ng 1976, Pollak 1979, Roberts 1980, and

[28] See, for example, Geneakalpolos (2005)'s the extensive use of **I** condition in his proof for the theorem, which requires social choice procedure to be consistent across the many non-actual profiles. See also Gaertner (2009) on the same point, for another example

Rubinstein 1984; See Suzumura 2002 and Feldman and Serrano 2008 for historical overview). It is now agreed that there are single profile analogues of all the results given in the multi-profile framework, provided suitably constructed single-profile conditions are met (Pollak 1979; 86, Sen 1977; 1564, Rubinstein 1984; 726). Precisely the conjunction of **SN** and **R**, with other conditions being met, is what is needed to derive the single-profile analogue of Arrow's impossibility. A statement of the single-profile variant of **SN**, sometimes called Intra-profile neutrality (**IAN**) is obtained by, in the original statement of **SN**:

Strong Neutrality (**SN**): For all w, x, y, z in the set of alternative X, and for all profiles $\langle R_i \rangle$ and $\langle R'_i \rangle$: If, for every i in N, [xR_iy iff zR'_iw] and [yR_ix iff wR'_iz], then [xRy iff $zR'w$] and [yRx iff $wR'z$]

equating the two profiles $\langle R_i \rangle = \langle R'_i \rangle$:

Intra-profile Neutrality (**IAN**): For all w, x, y, z in the set of alternative X, and for any profile $\langle R_i \rangle$: If, for every i in N, [xR_iy iff zR_iw] and [yR_ix iff wR_iz], then [xRy iff zRw] and [yRx iff wRz].

The following single-profile impossibility theorem has been proven in the literature of social choice:

If there are more than two alternatives, there is no SWFL *f* that satisfies **R**, **P**, **IAN**, and **D**.

In Chapter 5, we will revisit the concept of the multi-profile framework. In that chapter, it will be shown that some possible solution to the aggregation problem I raise for the domain of system choice does not work because the solution fails to recognize that there is the single-profile variant of Arrow's impossibility. Now let us move on to the more important possible escape route: inter-criterial comparability.

2.4.4 A Possible Escape Route in Theory Choice: Inter-Criterial Comparability

One possible escape route from Arrovian impossibility in social choice discussed earlier involves the use of a cardinal function together with well-justified interpersonal comparability. Assuming that social choice and theory choice are formally analogous, it is natural to seek the same type of escape route in theory choice. Theories are ranked by each dimension of scientific virtues just as alternatives are ranked by each individual in the society. Just as we ultimately need interpersonal comparability of cardinal preferences in social choice, we likewise need inter-dimensional comparability of scientific virtues in order to save theory choice from Arrovian impossibility. Given this, it is important to inquire as to what it is that makes different scientific virtues comparable or commensurable, if they are.

What would it be like to have such inter-criterion comparability in theory choice? That there is cardinal inter-criterion comparability implies that theories are measured on a cardinal scale for each criterion and that there is an exchange rate between criteria. Again, for example, one theory can come out to be better than another with respect to accuracy but worse with respect to simplicity; informativeness of a theory can be increased by sacrificing simplicity; and so on. Cardinal comparability in this context would mean then that we can justifiably make judgments like "*This* amount of loss in simplicity can be compensated for with *that* amount of gain in accuracy." To put it different way, this would mean something like "The metric for trade-off, i.e., the exchange ratio, between simplicity and accuracy is such-and-such." With this kind of "recipe" for the trade-off between standards, theories can be compared in a consistent way. Inter-theory comparison requires that we can *compare* theories using certain weights assigned to each standard. This is the same kind of comparability we discussed in the case of interpersonal comparability in social choice. In what follows we will inquire about (cardinal) inter-criterion comparability.

A Case for Comparability between Simplicity and Accuracy

It may seem reasonable to assume that theories can be measured on a cardinal scale for at least some of the criteria. Simplicity of a theory could be measured by how many axioms it has, how many ontological basic kinds it assumes, or the number of freely adjustable parameters if they are defined on parameter space, etc. Accuracy is often measured in terms of statistical fit. Informativeness of a theory could be measured by the range of its scope in the context of quantified study. This seems to suggest that the cardinal measurability of the criteria is, at least in some cases, plausible. We will have further discussion on this in Chapter 4, but for now let us assume that it is a plausible idea.

But this idea is less straightforward than it seems, especially when it comes to the matter of rates of exchange between the criteria. Take simplicity and fit, for instance. It is difficult to see how to make sense of the inter-criterion trade-off between them. First of all, there seems to be no single determinate use of each criterion, as Kuhn told us. Also, different criteria are implemented for different reasons. Simplicity has been considered to be a theoretical virtue, for a variety of reasons. Simplicity is often valued for epistemic reasons, e.g., cognitive costs, computational limitations, and so on. Occam's razor is commonly mentioned as a justification for using simplicity as a theory choice criterion. Accuracy is often required by a spirit of empiricism – given that the most important source of our knowledge is experience, a good theory must fit the data well. Now, the trade-off between simplicity and accuracy presupposes the exchange rate, which in turn requires a common scale on which each of them can be measured. How can there be such a common scale for such vastly different criteria in its nature and justifications?

Okasha (2011) explores a case for inter-criterial comparability in the statistical model selection literature. For example, consider we are comparing different models making different claims about the relationship between two random variables *x* and *y*. Suppose some models fit the data better than others; some models are simpler than others, and so on. We would need a principled method to select the best model among them. The literature on statistical model selection provides a wide variety of model selection methods, some of which express some form of comparability between different criteria, namely accuracy and simplicity. One typical example of such methods is Akaike Information Criteria (AIC). AIC tells us to choose the model that has maximum AIC score:

AIC score of model *M* = LogLikelihood of model *M* – the number of parameters in *M*.

The first term may be understood as accuracy and the second simplicity (or complexity). Then, assuming log-likelihood and the number of parameters can be measured on a cardinal scale, this may be viewed as a case for cardinal measurability with unit comparability between simplicity and fit. Drawing on this, Okasha suggests that at least in the domain of statistical model selection like above we have a case for cardinal comparability.

At first blush, Okasha seems right. But we need careful analysis of the said measurability and comparability. We need to deal with the important questions like: What can ground a rate of exchange for trade-off between criteria expressed in some statistical model selection methods? What exactly do different terms mean in the methods? What underlies the exchange ratio displayed by them? These are the questions I will discuss in the next chapter.

Conclusion

In this chapter, we have examined aggregation problems in the domain of social choice and theory choice. First we discussed Arrow's impossibility theorem (1951/1963) which says there cannot exist any reasonable procedure of aggregating individual preferences into a social preference. Following Okasha (2011)'s lead, we explored if the analogue of Arrow's impossibility theorem obtain in the domain of theory choice. For this, we have carefully examined formal statements of the theorem and its conditions. Then we examined if the theory-choice analogues of Arrow's condition are motivated. The result was that the analogues seem to obtain in theory choice, with a caveat that condition **U** does not apply to theory choice but probably its weaker counterpart **R** does. Weakening **U** to **R** does not necessarily open up an escape route from the Arrovian impossibility, if condition **I** can be strengthened to **SN**. It is not perfectly clear if **SN** applies to theory choice but we saw some motivation for thinking it does. Also, we saw that there are single-profile variants of Arrow's impossibility theorem. It seems that probably the clearest and most promising escape route from the Arrovian impossibility for theory choice would be to make a case for inter-criterial comparability, in Sen (1970)'s extended framework. In particular, we discussed some form of inter-criterial comparability expressed in some statistical model selection methods, for example Akaike Information Criterion. We asked where the comparability in AIC comes from. We will answer this question in the next chapter, by examining technical details and assumptions of AIC.

Chapter 3: An Answer to the Question about the Trade-Off

Introduction

As we saw in the previous chapter, the Arrovian impossibility in social choice theory appears to pose a threat to a common notion of scientific progress. The problem at hand is that we want the procedure of theory choice to be that of maximizing overall theoretical virtues. The required procedure of aggregating theoretical virtues, however, appears to be subject to the Arrovian impossibility. Just as cardinal interpersonal comparability provides a solution to Arrow's impossibility in social choice, it was suggested that inter-criterion comparability can serve as an escape route from the Arrovian impossibility in theory choice. The question now before us is how to make the different theoretical virtues commensurable. To provide a solution to this problem, we must answer the following questions: What is the exchange ratio for inter-criterial trade-off, and where does it come from? For example, is there a way to measure the amount of simplicity that would need to be increased in a theory to compensate for a certain amount of decrease in accuracy? Can these two virtues be measured by the same metric so that trade-offs are possible between them?

In this chapter, I will discuss an exemplar answer to these questions, appealing to the model selection problem in statistics as Okasha did. I will provide explication of statistical model selection problem and its implications. There is an abundance of literature about statistical model selection that covers a wide range of domains. In this thesis, I will focus on a particular case of the model selection problem: the curve-fitting problem. The curve-fitting problem is a contemporary version of the long-standing problem of induction. This will serve as an exemplar case for the question at hand because it involves conflicting virtues and attempts to resolve these conflicts by making trade-offs between those virtues. By examining how the trade-offs in question are derived and used in statistical model selection, we may discover an escape route from Arrovian impossibility for theory choice.

3.1 Statistical Model Selection and Its Philosophical Implications

It is often said that all scientific theories are underdetermined by the evidence because the finite amount of observed data (i.e., evidence) is compatible with infinitely many logically possible theories.[29] This *problem of underdetermination* has been one of the most important questions in philosophy of science because it poses a threat to the rationality of science, especially regarding theory choice. Hoping to resolve this problem, philosophers have shown significant interest in recent developments in statistics, and in particular statistical model selection criteria, by which statisticians choose models against the observed data. If we can discover well-founded theoretical justifications for statistical model selection criteria, and if these justifications are applicable to theory choice, then we would have an appealing solution to the problem of finding a case for inter-criterial comparability discussed in the previous chapter. Statistical model selection criteria have drawn a number of philosophers' attention, including, but not limited to: Sober and Forster (1994), Forster (2002), Mulaik (2001), and Kieseppä (1997, 2001a, 2001b), discussing statistical model selection criteria and their philosophical implications. This is where Okasha (2011) spots a promising solution to Arrovian impossibility. In this chapter, I will first explain

[29] The famous Duhem-Quine thesis. (Duhem [1914] 1954;187, Quine 1951;42-3)

model selection problems, making use of examples from the curve-fitting problem. I will then discuss theoretical justifications for statistical model selection criteria. In particular, I will focus on a particular criterion called the *Akaike Information Criterion* (AIC) and its theoretical justifications.

3.1.1 Statistical Model Selection Problem

A statistical model is a family of probabilistic functions or hypotheses with the same number of parameters. For example, suppose a biologist studies the size of fish (*y*) in a lake. She considers various hypotheses that purport to explain *y*: from the very simple hypothesis that *y* is not affected by anything, to the very complex one that *y* is affected by everything in the lake. These infinitely many hypotheses may be grouped into families depending on how many factors they claim affect *y*. Hypotheses that claim *y* is determined by one factor, say, the lake's average oxygen level (*O*), may be grouped together. Those that claim that *y* is determined by two factors, the average oxygen level and the average water temperature (*T*), may be also grouped together as another family. And so on. Individual hypotheses within a group will agree on which factors affect *y,* though they may disagree on *how significantly* the assumed factors affect *y.* Let us suppose that the researcher considers the following models, which differ based on the number of factors considered by each: Model 1 holds that the size of the fish is determined by the lake's average oxygen level; Model 2 holds that the size of the fish is determined by the average oxygen level and the average water temperate; and Model 3, in addition to the two factors included in Model 2, holds that the average water velocity (*V*) also affects the size of the fish. In addition to setting up models, there is something else that the researcher will have to consider. In almost all cases, the observed data involves some noise, or *measurement errors*. Assuming that these measurement errors are normally distributed, each model needs to introduce the presence of random error ε, meaning that each model corresponds to a certain probabilistic function. Then each of the above models claims that the true relation between y and the factors it involves is in the following form, and that one of its members specifies true values for the parameters (β_i) of the factors:

Model 1: $y = \beta_1 O + \varepsilon$
Model 2: $y = \beta_1 O + \beta_2 T + \varepsilon$
Model 3: $y = \beta_1 O + \beta_2 T + \beta_3 P + \varepsilon$

Note that these three models differ with each other in different aspects. First of all, their degrees of *simplicity* differ. There are different ways of measuring simplicity in statistics, but one common way is to have simplicity be measured by the number of adjustable parameters. This is because it represents how many factors are needed for specifying a member in the model.[30] In the above example, Model 1 is simpler than Model 2 because the former model has fewer free parameters than the latter. Models can also differ in their degrees of *fit*. A model's fit is generally measured by its *likelihood*, the joint probability of the observed data according to an element of the model in question. In general, the best-fitting element within the model is used to measure the model's likelihood.[31]

Characterized in this way, model selection process is a two-step process.[32]

Step 1: A statistical model ***M*** is chosen;
Step 2: A member of the model ***M*** is chosen.

[30] See, for example, Kieseppä (2001b) for a clear exposition on this point.

[31] The set of parameter values of the best-fitting element in a model is called *Maximum Likelihood Estimate* (MLE).

[32] The following discussion on model selection problem and curve-fitting problem are largely drawn from Sober and Forster (1994), Kieseppä (2001b)'s characterization of them.

Step 2 is not so controversial – the rational choice in Step 2 would be, without much question, the member of ***M*** that 'fits' to the observed data best after ***M*** has been chosen in Step 1. The difficult question is how to choose ***M*** in Step 1, especially given the possibility that models fare differently when equipped with different virtues.[33] This is where the model selection criteria come into play. A variety of model selection criteria have been suggested. One criterion that has drawn philosophical attention is the *Akaike Information Criterion* (AIC). This criterion will be discussed in §3.2. It will be useful to examine model selection methods in a sufficiently specific context. The context of curve-fitting problem provides such a useful framework. So, let us examine the problem of curve-fitting in detail.

3.1.2 The Curve-Fitting Problem

The problem of curve-fitting is a special case of the model selection problem. Suppose an experiment on the relationship between *X* and *Y* has been conducted[34] and its result is plotted in the figure below, with the dots representing observed data points.

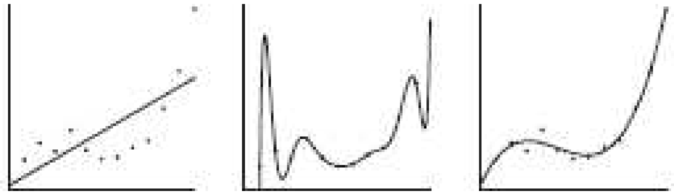

Fig 2. Curves with different complexity (from Grünwald 2005)

[33] One might wonder, probably in the spirit of empiricism, why not just choose the model displaying the maximum model likelihood in Step 1. As discussed in the previous chapter, doing so entails accuracy is a dictatorial criterion. As will be discussed shortly, accuracy as a dictator is harmful particularly due to the phenomenon of the over-fitting.

[34] In the case of experimental science. In the case of observational science, an observation would have been made.

What can be said about the relation between *X* and *Y*? Here we aim to make generalized statements concerning the observed data as well as unobserved data. Such generalization may be graphically represented as a curve. In the above figure, we can see different curves imposed on the same data. They can be understood as different claims about the relationship between the variables in question. We may name a linear curve LIN, which can be seen as the claim that all *X* values have a certain linear relationship with all *Y* values; PAR that all *X*s are in a parabolic relation with *Y*; CUB a cubic relation, and POLY-9 a polynomial of the degree of nine; NONE says there is no relation whatsoever between them. In this way, a curve can be viewed as a generalized statement about *X* and *Y*. As a toy example, we might consider a curve as a theory.[35]

The problem of underdetermination is vividly underscored in the context of a curve-fitting problem such as this. The observed data alone cannot decisively tell us which curve to choose. Most scientists would avoid POLY-9; they would pursue simpler theories. At the same time, no scientist of sound mind would pursue NONE, although it is the simplest theory in the sense that it has zero adjustable parameters. So here we can observe a conflict between the two scientific virtues: simplicity and accuracy. To recap, in the context of the curve-fitting problem, a curve can be seen as a specific scientific claim about the relation between variables; simplicity plays a particular role in the curve-fitting process, so does accuracy, and the observed data alone doesn't appear to determine which curve to choose.

[35] Sometimes a pure inductive generalization becomes a significant law or theory (e.g., Chargaff's rules about DNA). Other interesting examples would be the law of definite and multiple proportion and the law of Mendel, which were discovered by observation and induction and explained by theories (Dalton's atomic theory, and chromosomal theory, respectively).

As in the model selection problem, we can utilize the concept of models in the context of the curve-fitting problem. To reiterate, a model is a family of the curves that possess the same number of adjustable parameters. In the context of the curve-fitting problem, for instance, all linear curves have the same number of free parameters (in the form of $y=ax+b$, so that there is one adjustable parameter, a.[36]) Therefore, they can be said to belong to the same family, namely, LIN. Likewise, all parabolic curves belong to PAR family (in the form of $y = ax^2+bx +c$), cubic ones to CUB, and so on. Characterized this way, as in the model selection problem, the curve-fitting problem is a two-step process: one must choose a model, then choose a particular member of that model.

It is worth discussing why models are to be used in the curve-fitting problem. One might wonder whether Step 1 is necessary. The answer is that using models provides us with a natural measure of simplicity. Note that particular curves with its parameters being filled with specific values do not have free parameters because their values are already specified (or 'saturated'). Therefore, without characterizing them in terms of models, there is no objective way of ranking the simplicity of different curves (for example, "Because this straight line looks simpler to me than that curve does" wouldn't work.).[37] Once we use the language of models, then there is a natural measure of *how much* simpler one model is than another. The interval of simplicity between LIN and PAR is one unit of simplicity, as PAR has one more adjustable parameter than LIN; a member from LIN is two units simpler than that from CUB, and so on.

[36] Usually the Y-intercept b is not considered as a parameter because b can be readily factored out by rescaling.

[37] This point is taken from Sober and Forster (1994).

3.1.3 Model Selection Criteria

Now we need to examine model selection criteria. There is a great variety in statistical model selection methods, but here we are going to discuss two broadly-construed conceptual frameworks. This categorization is drawn from Sober and Forster (1994). The first is the Best Case strategy, and the second is the Akaikean framework.

3.1.3.1 Best Case Strategy

The Best Case strategy operates as follows. First, find the best case in each family (LIN, PAR, etc.,) with respect to the observed data. Then compare those best cases with each other, in terms of how well they match the observed data. The one which fits the data best among them is the best case, and hence the model that contains it is the best model.[38]

The accuracy of a curve can be measured in different ways: it may be measured by the squared sum of the discrepancy between the data and the curve [*Sum of Square*: SOS]; or by the conditional probability of the observed data, given the curve [*Likelihood*], if the curve represents a particular probability density function. Which measure we choose to adopt will have little effect on our discussion. Mostly I will use likelihood as a measure of accuracy. One thing to be noted is that referring to a case as "best" means that it is the best with respect to the observed data. Note that, in most cases, bumpier curves fit the observed data better than simpler ones. In the earlier example, let's say H_1, H_2, H_3, and H_4 are the best cases of the model LIN, PAR, CUB, and POLY-9, respectively. H_4 falls exactly on all the observed data points; H_1

[38] In statistics, Maximum Likelihood Estimate method is a typical example of the Best Case strategy described here.

through H_3 do not (and they cannot, precisely because they are not bumpy enough). In this case, the Best Case strategy yields the model POLY-9 as the best model. The Best Case strategy will almost always pick a very complex model as the best one. The Best Case strategy has long history in statistics, and it is useful in certain contexts of statistical estimation, but it may not be an appropriate solution to the curve-fitting problem.

3.1.3.2 The Overfitting Problem

The Best Case strategy runs the risk of *overfitting*. Overfitting is a phenomenon that occurs when we fail to distinguish between noise and genuine information in the observed data. Models which are too simple are likely to fit too little of the data. If we opt for more complex models, then we have more 'degrees of freedom' to fit the observed data. So, it is generally a wise move to select more complex models. However, as mentioned earlier, the process of collecting data almost always involves some noise. It could be due to observation errors, or unobserved latent factors, or the stochastic nature of measurement due to indeterministic characteristics of nature, (if there are any). It doesn't matter where errors come from; what matters is that the existence of such errors should be taken into consideration in the curve-fitting. However, the models that fit the data 'too well' have likely failed to distinguish the true information from the noise.[39] This is the risk of overfitting. The Best Case strategy almost always selects the most complex model. However, given that data points involve noise or measurement errors, the most complex curves might not be the best curves, for the reasons put forward above.

[39] This tension between 'fit too little' and 'fit too well' is exactly what underlies the phenomenon called *bias-variance tradeoff* in statistics. See Forster (2001).

Overfitting is particularly problematic when it comes to prediction. A curve which is extremely bumpy but that exactly matches the data would perform badly in *predicting* the future data from the same population because it is likely to have committed overfitting. Assuming that noise is normally distributed, if we repeat the same experiment, then it is likely that a slightly different set of data will be observed; the result can be somewhat different every time. This means that, when a curve “overfits” the present data, it will most likely not fit the future data. Therefore, a bumpy curve that exactly matches the data will likely possess poor *predictive accuracy*. To recap, since the Best Case strategy only concerns accuracy as its model selection criteria, it will most likely end up with the most complex model among the models under consideration. Given this, it runs the risk of overfitting, which results in poor predictive accuracy.

An Analogy. Consider the following analogy. Bruce, a speedster whose sole criterion for car choice is average top speed, wants to buy a new car. He has an unlimited budget. Doing some market research, he realizes that it is not a good idea to make comparisons between every single individual car, since there are simply too many to compare. Instead, Bruce considers *classes* of cars: a class of $10K and below, a class of $10K - $20K, a class of $20K - $30K, and so on. Which class would be the best choice for Bruce? One might think that Bruce should choose the most expensive class and then pick the best car within that class. After all, he has an unlimited budget. However, this would probably be a very bad idea. The key to this is the phenomenon of overfitting. Suppose that Bruce decides to select the most expensive class of car. He drops by a dealer that carries the most expensive cars like luxury sports cars, test drives them on the test track prepared by the shop, and then picks the one that marks the best speed record on those test runs. Unfortunately, this car will most likely not mark a good top speed on the usual roads, which have pot holes, hills, curves, and so on. Bruce's choice is 'overfitted' to those particular test runs. By opting for the most expensive class of car, Bruce was given *too large a degree of freedom* to fit the particular test runs and consequently he ended up with a car (say, a Formula One car) that runs *too well* on the particular occasions he observed. Had Bruce chosen a somewhat moderate class of car and made a choice within that class, he would have had less room to tailor-fit his preferences to the particular test runs, but instead the chosen car would likely mark a much better overall top speed than the Formula One car. An important lesson for Bruce in this analogy is that, even when Bruce has an unlimited budget, the best choice for him may be to *restrict* himself to a somewhat lower class of car. This is fundamentally due to the fact that the only resources available to Bruce are particular test runs on a certain track, from which he has to make an inference about how well a given car will perform under all possible road conditions. In other words, he is with epistemic limitations with respect to how well a given car will run on the roads that he has not experienced.

3.1.3.3 Akaikean Framework and Akaike Information Criterion: an answer to the question of the trade-off

The Akaikean framework is a model selection framework with a predictive point of view.[40] This framework is based on a consideration of what the purpose of modeling is and what models are to be used for. In the Akaikean Framework, (Forster and Sober 1994, Kieseppä 1997) the primary purpose of statistical modeling is not to

[40] The term 'Akaikean Framework' is drawn from Kieseppa 1997, Sober and Forster 1994.

accurately describe current data.[41] Rather, in the framework, the purpose of statistical modeling is to predict future data as accurately as possible. Akaike's model selection criterion, the Akaike Information Criterion (AIC), is designed to maximize such predictive accuracy. AIC has drawn philosophers' attention for its interesting features.[42] The most important feature of AIC for our present purpose is that it seems to provide a specific exchange ratio between the different virtues of models. It tells us to choose a statistical model *M* that has the minimum in the following AIC formula:

> **AIC formula**: –2[Maximum Log-Likelihood Estimate of *M*] + 2[number of parameters of *M*].

Rephrased to give a simpler reading by reversing the signs, this formula gives us a score of the predictive accuracy of the model in question:

> **AIC score of model *M*** = [MLE of *M*] - [the number of parameters of *M*].

The AIC rule can be characterized as follows:

> **AIC rule**: Choose the model that maximizes AIC score.

What is noteworthy about AIC is that it seems to express a specific exchange ratio for the trade-off between fit and simplicity, if we understand likelihood as fit and the number of parameters as simplicity. On this understanding, what AIC effectively says is "sacrificing *one* unit of simplicity can be compensated only when doing so will

[41] In the Akaikean framework, estimating the "true distribution" is given lower priority. In most of the cases, there is no significant difference between the point of view of inferring the true structure and that of making a prediction if the size of data is large enough. However, in modeling based on a finite quantity of real data, there is a significant gap between these two points of view, because an optimal model for prediction purposes may differ from one obtained by estimating the "true model." See Konish & Kitagawa (2007; ch1) for further discussion.

[42] The original formulation of AIC can be found in Akaike 1974. For an accessible, technical discussion of AIC, see Anderson and Burnham 2002, and Konish & Kitagawa 2007. Forster 1995; 353-4, Kieseppä 1997; 23. For philosophical implications of AIC, see Sober and Forster 1994, Forster 1995, and Kieseppä 1997, 2001a, 2001b.

increase fit by *one* unit or more." This seems to be the exact kind of comparability required in order for theory choice to escape the Arrovian impossibility, as we saw in §2.4. In this way, AIC seems to answer the questions we asked at the end of the previous chapter: What is the exchange ratio between the virtues? The answer is: it is a one-to-one ratio (in AIC's specific sense of simplicity and fit). We also have a partial answer to the second question, regarding *where* the trade-off comes from, that is, what drives such a specific form of trade-off. The phenomenon of *overfitting* is what drives the form of trade-off specified in AIC. Simplicity and fit (in AIC's specific sense) are made *commensurable* in terms of their degree of contribution to the predictive power of a theory, and it is *how much* they contribute that determines the specific ratio expressed by AIC. To discover exactly how much simplicity and fit contribute to the predictive power of a theory, we have to look at what justifies the exchange ratio displayed in AIC. We will go over the proof for the AIC formula and its related assumptions in what follows.

3.2 Akaike Information Criterion

Two things must be noted before we proceed. First, in the Akaikean framework, one is not concerned with accuracy with respect to the observed data; rather, one is concerned with the distance to the true curve from the fitted curve. The true curve is the curve that is assumed to have 'generated' the observed data.[43] Understanding this difference is important. Second, the actual estimation process in the Akaikean framework involves some correction for the risk of overfitting. For now, it suffices to

[43] Statisticians who don't wish to commit themselves to the existence of such 'true curve' often uses the term 'quasi-true curve' meaning the curve that is speculated to have generated the observed data. In §5.6, I will discuss the concept of true curve in the context of the Best System Analysis.

note that the Akaikean framework involves the process of such a correction. The details of this will come later.

Let us now examine how the curve-fitting problem is approached in the Akaikean frame work step-by-step. The following is drawn from Sober and Forster 1994, and Kieseppä 1997.

The derivation of AIC is largely composed of three parts. The first part is to define the distance between two given curves or probability distributions. Of particular interest to us will be the distance of a given curve to the true curve. The second part is to define a *model's predictive accuracy* in terms of the *average* distance between the true curve and the model's best curves with respect to *every possible data set* generated from the true curve. The third part is to derive a statistical estimation of the model's predictive accuracy.

3.2.1 Part I: Kullback-Leibler Divergence as Distance Between Curves

In what follows let us adopt a God's eye view. That is, in the Akaikean frame work, models are not evaluated merely with respect to the particular observed data, but rather with respect to *all possible data* along with the curve that generates the observed data.

First of all, we are going to assume that each curve is composed of a deterministic part and a random error part. For example, a curve in the form of $y = ax + b$ is to be understood as [$y = ax + b + Error$]. In this case we can say that each curve, including the true curve, defines a certain probability distribution function [p.d.f]. Let θ be a hypothesized curve and $θ_0$ be the true curve. The p.d.f defined by each of them can be represented as: $f(\cdot|θ)$ and $f(\cdot|θ_0)$.

One natural way of understanding θ and θ_0 would be that θ is a working hypothesis that we want to test against the observed data, and θ_0 is the true curve under which that data has been generated. Crudely speaking, θ is a specific theory tested in relation to the data, and θ_0 is the truth.

Now we are going to define the distance, or *divergence*, between a given curve and the true curve. A curve defines a certain probability distribution, and the same holds for the true curve because it is also a curve. Given this, in order to evaluate how good our hypothesis θ is, we will need to measure the divergence of θ's probability distribution from θ_0's probability distribution. There are different ways of measuring this, and here we are going to use a distance metric called the *Kullback-Leiber divergence*.

Kullback-Leibler divergence. In the present example of curve-fitting, we have data points. For each data point, we can see how probable the given point is under hypothesis θ. Also, and importantly, we from a god's eye view can see how probable the same data point is under the true curve θ_0. Note that the true curve is not deterministic – i.e., it will generate probabilistically different data every time we make an observation.

Then, for each of the observed data points, we can obtain the ratio of its probability under θ and θ_0. For example, take one observed data point, y_1. Suppose it is observed to have a certain value, a_1. Now, suppose that θ says the probability of obtaining $y_1=a_1$ is 0.8; and θ_0 says that the probability is 0.9. In this case the ratio of the two probabilities is 0.8/0.9 = 8/9. Given this, we may say that θ is quite close to θ_0, with respect to y_1. Now, take another data point, y_2, and obtain the ratio of θ and θ_0, with respect to y_2. Repeat this procedure over all observed data, and obtain the final average. As a result of this procedure, we can obtain the weighted average ratio of the probability distributions defined by θ and θ_0.[44] This ratio can be a natural index of how far our hypothesis θ is from the truth θ_0, i.e., the distance between θ and θ_0. (Note again that this distance is defined with respect to a set of data.) This way of measuring the divergence between the two probability distribution has been suggested by Kullback and Leibler (1951):

K-L divergence D_{KL} between the probability distribution P andQ:

$D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) log \frac{P(x)}{Q(x)}$

Also note that the law of large numbers will apply here. That is, the more data points we have, the more information we have about the accurate ratio between θ and θ_0.[45] Each of the observed data points contains some reflection of the true curve. However, since the true curve is itself probabilistic, the reflection cannot be perfect, that is, some level of noise is involved. Our theory θ says how much each data point reflects the true curve. Given this, there can be a discrepancy between what θ says and what θ_0 says. Such a discrepancy can be considered to be the distance between θ and θ_0. When $\theta = \theta_0$, the discrepancy will be minimum (i.e., zero).[46] In other words, when $\theta = \theta_0$, θ comes to have the maximum reflection (information) of the true curve. If θ is not equal to θ_0, then there is a loss of information in proportion to the distance from θ

[44] Strictly speaking, we don't just take the average ratio over the observed data points. We take logarithm of each ratio and then obtain the average of them, so that the result behaves like distance.

[45] But note that the term "distance" should not be taken literally. The Kullback-Leibler distance is not symmetric because it is a measure of the expected value weighted by the presumed probability distribution; so, when reversed, we have different weightings by a different distribution. Therefore, $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$.

[46] Konish and Kitagawa (2007).

to θ_0.[47] In this way, we have defined the distance between any given curve and the true curve.

3.2.2 Part II: Predictive Accuracy of a Model

The curve-fitting problem is a two-step process. The choice of model should be made before a particular member of a model is chosen. The basic idea of measuring the worth of a model is as follows: First, pick a representative curve of a given model and then evaluate how well it represents that model. Use the Best Case strategy for picking a representative curve. This means that one should pick the representative curve based on the observed data. Let us name the best curve chosen based on the observed data Best $\text{Case}_{\text{w.r.t. the observed data}}$. Now, measure the distance between Best $\text{Case}_{\text{w.r.t.the observed data}}$ and the true curve *with respect to every possible data set*, and obtain the average of them. That is, obtain the distance averaged over all possible data.

We have selected our representative curve solely based on the observed data. However, when we repeat the same experiment again, we will most likely observe somewhat different data. This suggests that Best $\text{Case}_{\text{w.r.t. the observed data}}$ may not fare well when it comes to predicting future data, because the future data will be somewhat different from the extant data.[48] And this was precisely why the Best Case strategy runs the risk of overfitting. Now, the degree to which future data will differ

[47] This idea of information loss in substituting θ for θ_0 is expressed in the exactly same form as *Shannon Information Entropy*, which is one of the most important formulae in information theory. This is why Akaike's method has *Information* in its name, AIC – Akaike *Information* Criteria.

[48] One might also give a counterfactual analysis of the situation. That is, the same true curve that has generated the observed data *could have* generated different data. Had it been the case, then we *would have* chosen another curve as Best Case than the one we chose in the actual world. Although this reading makes a good sense, this counterfactual reading requires much heavier assumption that the normality assumption made by AIC. This point will be discussed in more detail in Chapter 4 on David Lewis's best system account of lawhood.

from our current data is of course governed by the probability distribution of θ_0 – the true curve. This means that when we go through the averaging process over all possible data, we need to obtain the *weighted* average according to the probability distribution of θ_0. This average – the weighted average distance between Best Case$_{\textbf{w.r.t each data set}}$ and θ_0 over all possible data sets – is called the *predictive accuracy of a model*, in the sense that it can serve as an index of the average performance of a given model in predicting future data *independently of which data set happens to be observed*. That is to say, this average can serve as an index of how good a given model is from a predictive point of view.

3.2.3 Part III: Estimating Predictive Accuracy of a Model

Our discussion so far has taken place from a god's eye view. The concepts discussed so far, i.e., distance to the true curve and the predictive accuracy of a model, require the probability distribution defined by the true curve, $\theta_{0.}$ In reality, however, θ_0 is unknown because the only available resource is the observed data and the best curve associated with that data. In light of this, we need a way to acquire a reliable estimate of the predictive accuracy of a model from the best curve, with respect to the observed data.

Akaike (1974) proposed a formula for acquiring such an estimate, which is as follows (rephrased using the terms I have been using in earlier sections):

AIC value of the Model ***M***

= [Log-likelihood of the Best Case$_{\text{with respect to the observed data}}$ of ***M*** – the number of adjustable parameters in ***M***].

Akaike proved that, based on certain normality assumptions, the predictive accuracy of ***M*** can be reliably estimated by AIC value of ***M***. (See §3.2.4 for discussion on assumptions; see §3.3 for a proof) So, the above AIC value can serve as a good estimate of the predictive accuracy of ***M***. Naturally, Akaike's model selection criterion would then require us to choose the model that has the highest AIC value. The two terms that appear in the above equation needs some explanation. The first term amounts to the statistical fit (i.e., accuracy) of the best member within ***M***. This reflects the fact that, in statistical model selection, our resources are limited to the observed data, (which is a sample of the population). Although it is imperfect, the observed data is a reflection of θ_0, at least in some aspects.[49] Given this, we have no better alternative than to use the best available tool relative to the observed data, i.e., Best Case$_{\text{w.r.t. the observed data}}$. This explains the appearance of the Best Case in the first term of AIC.

Regarding the second term (= the number of adjustable parameters of ***M***): This term serves as a *penalty for the complexity* of ***M***. We now know that the Best Case strategy faces the problem of over-fitting. Given this, overcomplexity is a vice. But, at the same time, to oversimplify would also be a vice. This means that we need some metric for the trade-off between simplicity and fit. Now, AIC provides us with a recipe for making such a trade-off. Increasing complexity will only be allowed if it significantly increases goodness-of-fit. More specifically, an increase in complexity is

[49] Also note that θ_0 is just one aspect of the truth; θ_0 is assumed to define a certain type of probability distribution. So, at least we are working under assumption that the truth can be somehow manifested in a certain type of distribution. Under this assumption, we are replacing truth with θ_0.

allowed only if choosing a model with one more parameters results in more than one unit increase in Log-likelihood of Best Case$_{\text{w.r.t. the observed data}}$.

In sum, under an Akaikean frame work, models are to be evaluated with regard to the truth (by the distance from the true curve averaged over all possible data generated by the true curve). However, the true curve is not known. So, in reality, we need to evaluate models in regard to the observed data, imposing a penalty on the complexity of models.

3.2.4 Assumptions

Now let us check the assumptions for AIC. Let θ^* denote the vector of the true parameter values and $\theta(k)$ denote the best estimate of the parameter values in the model with k parameters. A key assumption of AIC is that the squared distance $|\theta^* - \theta(k)|^2$ is chi-square distributed with k degrees of freedom. This in turn relies on the assumption that $\theta(k)$ is normally distributed around θ^* on the vector space representing parameter values (Akaike 1974; 718, 1977; 31, Forster 1995; 353-4, Kieseppä 1997; 23, Konish & Kitagawa 1996; 888). More specifically, suppose the model M is true in that in this model there is a particular set of the parameter values $\alpha_1^*,\ \alpha_2^*, \dots, \alpha_k^*$, which corresponds to the true curve. The above assumption says that, if we repeatedly measure y values and each time compute an estimate of the true parameter values based on the measurement, the estimate will be centered around the true values $\alpha_1^*,\ \alpha_2^*, \dots, \alpha_k^*$ of the parameters. This assumption is supported by the so-called Central Limit Theorem (CLT) (Forster 1995; 354, Konish & Kitagawa 2007; 51).[50]

In addition to the above, if we assume the error distribution is normal (that is, p($y|\theta$*) follows normal distribution), then AIC is valid in most of the cases (Kieseppä 1997). Note that the assumption of normal error distribution (e.g., the assumption that the data itself is normal in that the observed *y* values will form a bell curve centered on the true value *y**) is from the earlier assumption of normal distribution of parameter values.

In short, the essential assumption of AIC is the CLT. Additional, the assumption of normal error distribution is needed for wide applicability of AIC. Let us grant both assumptions for now. I will revisit the CLT assumption of AIC in Chapter 5 where I discuss the AIC-implemented-BSA.

3.3 Proof for AIC

Defining Predictive Accuracy. Let's suppose we made *n* number of observations of Zs controlling for Xs. We are trying find a mathematical relationship between X and Z based on the outcome of the experiment by fitting a curve to the data plotted on the X-Z plane. (I'm reserving Y for all possible data).

Let Z be the set of the observed data = {(x_1, z_1) … (x_n, z_n)}.

Let $θ_k$ be a specific function (curve) within a family *k* such that $θ_k$ specifies values for parameters of $θ_k$ = {$θ_1$... $θ_k$}. If I don't use subscript *k*, it means I am discussing a curve in general. Assuming random errors, each θ tells us the probability of obtaining each observation to be z_1 … z_n, given θ. That is, each θ defines its own probability

[50] The conditions for the (classic) CLT is that samples are independent, samples are sufficiently large, and sample variance is finite.

distribution function f(Z|θ).[51] Let θ_0 be the true curve, and f(·|θ_0) be the true probability distribution.

We are going to use *likelihood* to measure *fit-to-the-observed-data*. Likelihood of θ is the probability of obtaining the observed value given θ. We have multiple data points, so we need to compute the joint probability of obtaining the observed values for each of $z_1 \ldots z_n$, controlling for the control variable X. Then, we can define likelihood as follows:

Likelihood of θ, with respect to Z = $\prod p(z_i|\theta, x_i)$

Take logarithm of it, we'll get log-likelihood of θ:

$L(\theta) = \sum \log p(z_i|\theta, x_i)$

Within a given family *k*, we may find a curve that has the maximum log-likelihood with respect to the observed data Z. Call it $\hat{\theta}_{Z;k}$. In other words, for all $\theta_k$in *k*, there is no such $\theta_k$that has greater L(θ_k) then L($\hat{\theta}_{Z;k}$). This maximum member is called *Maximum Likelihood Estimate* [MLE]. Note that MLE is defined in regard to a certain family *k* and the observed data Z.

There are certain conditions to be noted: The log-likelihood function should be *k*-differentiable, single-peaked, and meeting some usual conditions for the task of finding maximum value. Then, the log-likelihood of MLE of a family k, with respect to the data, is

$L(\hat{\theta}_{Z;k}) = \sum \log p\left(z_i \middle| \hat{\theta}_{Z;k}, x_i\right)$.

[51] In what follows I sometimes use p(Z) instead of f(Z). They mean the same thing, except that the former is used for the discrete cases, the latter for the continuous cases. This doesn't affect our discussion.

How to get the distance between MLE and the true curve, with respect to the data, in a family *k:*

In order to evaluate how good the particular MLE that we have chosen with respect to Z is, we need to consider the probability of obtaining the observed data Z with respect to the probability distribution of *all possible data*, i.e., the probability distribution for Y. In other words, we need to consider the *weighted average of log-likelihood* of $\hat{\theta}_{Z;k}$ - weighted by the probability distribution of Z, which is defined by θ_0.Call such value the *expected* log-likelihood of $\hat{\theta}_{Z;k}$ and denote it by $l_N^*(\hat{\theta}_{Z;k})$. And this is obtained by the following equation:

$$l_N^*(\hat{\theta}_{Z;k}) = E_Y\ [L(\hat{\theta}_{Z;k})] = \sum \int \log p(y = z_i|\hat{\theta}_{Z;k}, x_i) \log p(y = z_i|\theta_0, x_i)dy.$$

The above expected log-likelihood is maximum when $\theta = \theta_0$. Then,

$$l_N^*(\theta_0) - l_N^*(\hat{\theta}_{Z;k})$$

is the distance between $\hat{\theta}_{Z;k}$ and θ_0.

Note that the first term will be a fixed value, and that it is zero when $\theta = \theta_0$. So the resultant value of this equation behaves like a distance metric. So, this can be considered as the distance between the MLE and the true curve.

Predictive accuracy of a given family k, as the distance averaged over all possible data.

Now our task is to choose the *family* that is expected to, on average, yield the largest value for $l_N^*(\hat{\theta}_{Z;k})$. It can be obtained by averaging $l_N^*(\hat{\theta}_{Z;k})$over all possible data, weighted by the probability distribution of Y, defined by $f(Y|\theta_0)$.

So what we need is Expectation$_Z$ [$l^*_N(\hat{\theta}_{Z;k})$], which can be expressed with the following equation

$$E_Z\,[l^*_N(\hat{\theta}_{Z;k})] = \iint \dots \int l^*_N(\hat{\theta}_{Z;k})\,p(z_1|\theta_0)\,p(z_2|\theta_0)\dots p(z_n|\theta_0)\,dz_1\dots dz_n$$

This value is called *predictive accuracy* of a given family *k*.

In Akaikean frame work, the criteria for model section is that we have to choose a family *k* that yields highest value of predictive accuracy as defined above. (Again, recall that in the current context of curve-fitting, we don't just choose a single curve – we choose a model, or family.)

So far we have seen somewhat technical definition for the predictive accuracy of family K. A scientist's goal who is utilizing AIC is to choose a K that has best predictive accuracy as defined. Back to Bruce the speedster example, he now knows, at least in theory, how to 'measure' how much good a chosen *class* of cars will do him. But the measurement is theoretically based on the assumption that he a priori knows how often he will face which type of road, of all the possible types. Of course, he doesn't have such knowledge in reality. Then, he would better have some rule-utilitarian perspective: that is, adopt a systematic rule for class choice that is expected to yield maximum utility for him, on average; a class chosen by that rule might or might not be the ideal one depending on the case, but it is the highest expected utility.

Deriving an Estimate for Predictive Accuracy. In most of the cases θ_0 is unknown. Therefore, we need a way to obtain a reliable estimate of the predictive accuracy of a given family. AIC holds that the following equation holds:

$\textit{Predictive accuracy of the family } k \approx L(\hat{\theta}_{Z;k}) - k$.

What this equation implies is that, if we repeat the same experiment over and over, picking L($\hat{\theta}_{Z;k}$) every time, on average, the value [L($\hat{\theta}_{Z;k}$) – *k*] will tend to be equal to the predictive accuracy of the family k. In other words, *expectation* value of [L($\hat{\theta}_{Z;k}$) – *k*] is equal to the predictive accuracy of the family *k*.

The proof for this claim is based on the empirical assumption that the numerical difference between $\hat{\theta}_{Z;k}$ and θ_0 [the distance defined in earlier subsection] on the parameter space follows a chi-square distribution. Note that this is the 'empirical' assumption. The behaviors of certain parameters on the parameter space are expected to show certain patterns, in this case statistical pattern called chi-square distribution – but there is no further a priori justification for why they follow a chi-square distribution. And this assumption is, as we have seen in §3.2.4, essentially supported by the Central Limit Theorem (CLT). It seems safe to assume that other statistical methods at their root have assumptions of this kind. We will come back to this aspect of statistical model selection in Chapter 5.

Before we move onto the next chapter, let me provide a diagrammatical illustration of the behavior of the distance between θ and θ_0 on the different dimensions of parameter space. That is, how the distance behaves in different cases of different number of parameters. For simplicity, I am going to illustrate a case where the 'truth' can be defined as a point on three-dimensional space, θ_0 on a second-dimensional space, and θ on one-dimensional line. (Recall that θ_0 is just one aspect of the truth; it is only an aspect that is responsible what has been manifested before our eyes. It represents an aspect of the truth that is responsible for the relevant domain of data we observe.)

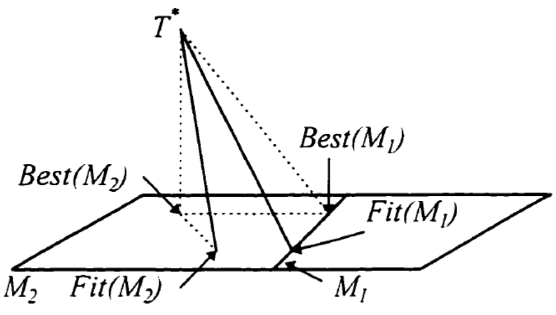

Fig 3. Graphical Illustration of AIC (From Kruze 1996)

T* is truth. Fit(M_1) and Fit(M_2) are the maximally-fitting-to-the-observed-data hypotheses (i.e., MLE) in each family. These will be the hypotheses we choose within each family. So the distance between Fit(M_i) and T* is what we are interested in minimizing. If we decide to go with M_2, for example, then Fit(M_2) is the hypothesis that we end up with; if M_1, then Fit(M_1). Best(M_i) is the best hypothesis in the family *i* in that it is the closest to the T* within that family. In general, the distance of Best(M_i) to T* will be shorter (hence better) as *i* increases. This distance is *model error*, in that such distance from T* is unique for and inherent in each particular model (family). Choosing higher *i* will result in lesser model error.

The problem for us, being agents with epistemic limitations, i.e., the only available resource being the observed data, is that we don't know how to obtain the Best(M_i) in a given model. In other words, due to measurement error, the distance between Best(M_i) and Fit(M_i) varies and we don't know how far it is for a given data set. Each time a set of data pulled from the truth, the location of Fit (M_i) varies. And here comes the normality assumption, saying that Fit(M_i) will be normally distributed around Best(M_i). In the above example, every time a set of data pulled from T*, Fit(M_2) will fall on a certain point on the plane; Fit(M_1) on a certain point on the line.

That point will be normally distributed around Best(M_2) and Best(M_1), respectively. Then, as we can see, the distance between Best(M_1) and Fit(M_1) is always equal to or shorter than the distance between Best(M_2) and Fit(M_2), no matter which data set we pull.

In short, there is tension between the task of minimizing model error and that of minimizing measurement error, so over-complexity will need to be penalized, and that is exactly what AIC tells us to do.

3.4 Another Model Selection Criterion: Bayesian Information Criteria

Let us examine another model selection criteria, Bayesian Information Criteria (Schwarz 1978). This section will be just focusing on what the criteria is and what its assumptions are.

A typical Bayesian approach to model selection problems is that we should choose the model which has the largest *posterior* probability. The posterior probability of a model *M* is the probability after the observation has been made. The posterior probability of *M* is determined by 1) how well the observation fits *M* and by 2) the *prior* probability of *M*, i.e., the probability assigned to M *before* the observation has been made. The following discussion is drawn from Kieseppä (2001a, 2003).

Formally, denote the probability distribution of Y (the observed *y* values) given X (the observed *x* values) and *M* by *prob*(Y given X and *M*). The Bayes theorem is:

$$\mathbf{P}(H|E) = [\mathbf{P}(H)\times\mathbf{P}(E|H)]/\mathbf{P}(E)$$

So, according to the Bayes theorem,

probability(*M* given X and Y) = [*prior*(*M*)×*prob*(Y given X and *M*)]/*prob*(Y given X).

We can ignore the term *prob*(Y given X) in comparing models, so the useful fact in the Bayesian framework is that *probability*(*M* given X and Y) is proportional to *prior*(*M*)×*prob*(Y given X and *M*). Therefore, the model that has the largest posterior probability of *M* is the one that has the largest value of *prior*(*M*)×*prob*(Y given X and *M*). Then *prior*(*M*) is to be fixed and *prob*(Y given X and *M*) is to be computed.

In words, computing *prob*(Y given X and *M*) can be done in the following way.

Within each model M, there are individual curves (in the context of curve-fitting). Each of those curve can be presented a set of specific parameters values. For example, PAR is a family of the curves in the form of "$f(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$", so each member of PAR can be presented as a list of particular values for ($\alpha_0, \alpha_1, \alpha_2$). Each of this hypothesis is assumed to have a certain probability given PAR. So, if we multiply [probability of the observed data given each hypothesis in PAR] by [probability of each hypothesis given PAR], we can obtain *prob*(Y given X and PAR).

Formally, the *prob*(Y given X, *M*, and $\alpha_0, \alpha_1, \ldots, \alpha_k$) is the probability density of individual curves in the model *M*. On the other, for model selection, we need to calculate the quantity prob(Y given X and *M*). For this we first need to introduce a *prior* probability distribution of the parameter values, for a given model M. Such distribution can be denoted by *prior*($\alpha_0, \alpha_1, \ldots, \alpha_k$ given *M*). Then this and the probability density of individual curves combined are sufficient to determine the value of *prob*(*Y* given *X* and *M*).

The real question for this Bayesian approach (like any Bayesian approaches) is what priors we should assign for models and for the parameter values within a model. That is, what probability distributions we should use for *prior*(*M*) and *prior*($\alpha_0, \alpha_1, \ldots, \alpha_k$ given *M*). Note that neither of them can be determined by consulting to the observed data. These are, after all, *priors* which we assign before we make observations. The question about the model prior *prior*(*M*) may be answered by assigning the same prior probability to all the models in consideration (recall that the model prior is assigned before observation). This is quite a common practice and not an unreasonable one. What of *prior*($\alpha_0, \alpha_1, \ldots, \alpha_k$ given *M*)?

Kieseppä (2001a) suggests that our choice of the prior distribution should be based on the use of a quantitative measure of informativeness of the prior distribution. That is, use the information from the data to contrast the models and identify the model that likely generated the data. That is, we observe some sufficiently large data first, then, under certain normality assumption, by computing variance and covariance observed in data, we can set the prior parameter distribution according to the observed data. Konish and Kitagawa (2007; 212) take the same approach. Then, we obtain BIC formula, which bears some similarity to AIC:

BIC score of model *M* = [MLE of *M*] **-** [log *n* × the number of parameters of *M*].

Accordingly, BIC rule can be characterized as follows:

BIC rule: Choose the model that maximizes BIC score.

Assumptions. As expected, BIC relies on the assumption that, as we repeatedly measure *y* values given *x* values and each time get an estimate of parameter values based on the measurement, the estimate will be centered around the true values $\alpha_1^*, \alpha_2^*, \ldots, \alpha_k^*$ of the parameters, assuming the distribution is a normal distribution

(Kieseppa 2001a; S148-50, Konish & Kitagawa 2007; 216) This assumption is supported by the CLT, as in the case of AIC.

Conclusion

In this chapter we have examined the following statistical model selection methods: AIC and BIC. They are:

AIC score of model *M* = [MLE of *M*] **-** [the number of parameters of *M*].

AIC rule: Choose the model that maximizes AIC score.

BIC score of model *M* = [MLE of *M*] **-** [log *n* × the number of parameters of *M*].

BIC rule: Choose the model that maximizes BIC score.

What was interesting about them is that they seem to express a specific exchange ratio for the trade-off between fit and simplicity, if we understand likelihood as fit and the number of parameters as simplicity. AIC and BIC give different weight to simplicity (in the current sense), and Kieseppä (2001a; s151) shows that various information criteria can be generally expressed as:

[MLE of *M*] – [*f*(*n*)× the number of parameters of *M*],

where *f*(*n*) can be understood as a kind of weight-giving function for simplicity. This seems to provide a range of comparability required in order for theory choice to escape the Arrovian impossibility, as we saw in §2.4. Another important point of this chapter was that these information criteria ultimately rely on the normality assumption, which is supported by the Central Limit Theorem. We will carefully examine their philosophical implications in the context of a new problem I will raise for the Best System Analysis in Chapter 5.

Chapter 4: The Best System Analysis of Laws of Nature

Introduction

In Chapter 2, I examined whether Arrow's impossibility theorem in social choice applies to theory choice. In Chapter 3, I explored a possible escape route from the Arrovian impossibility in theory choice. In particular, I investigated the inter-criterial comparability shown in statistical model selection methods. In this chapter, I will deal with a philosophical theory of laws of nature: the Best System Account of laws of nature. The account seems to invoke an aggregation procedure of different system choice standards and it requires an exchange ratio between the standards being aggregated. So, the BSA might be susceptible to the Arrovian impossibility. In order to find it out, we first need to make the account precise, which is the main task of the current chapter. In §4.1, I will examine three different views about laws of nature: eliminativism (§4.1.1), primitivism (§4.1.2), and simple reductionism (§4.1.3). The limitations of each view will lead us to a more sophisticated reductionism, the Best System Account. In §4.2, I will examine the philosophical motivations for the BSA (§4.2.1), David Lewis's characterization of the BSA (§4.2.2), and the underlying principle for the BSA: the Humean Supervenience thesis (§4.2.3). I will examine a few typical objections to the Humean view about laws: Earman, Tooley, and Carroll's counterexamples (§4.2.4). Examining the objections and responses will help us clarify what conception of laws is underpinning the BSA. Then I will assume the task of precisifying the key elements of the BSA in §4.3. First I will investigate why the BSA defenders do little work on precisifications (§4.3.1 and §4.3.2). Then, by examining the objections about the BSA's reliance on actual standards of science, I will refine what I call the Hope thesis, which turns out to play a very important role in the BSA (§4.3.3). In §4.3.4, I will discuss the questions about the BSA's reliance on epistemic standards; namely, what justifies the use of epistemic standards in metaphysics. This task will help us clarify the conception of laws the BSA assumes. I will then precisify each system choice standard and the balance between them (§4.3.5 through §4.3.8). I will conclude this chapter suggesting the Arrovian impossibility result may apply to system choice for the BSA.

4.1 Conceptions of Laws of Nature

The notion of laws of nature is important to both science and philosophy. They are important for scientists. Arguably, science aims at discovering laws of nature and the concept of laws plays a central role in science. For example, in statistical mechanics, the concept of a law plays a key role in differentiating dynamical and logical possibilities with respect to a given state space; in astronomy, distinguishing what are laws and what not allows to assess competing hypotheses about the origin of the universe.[52] It is also an important task for philosophers of science to develop an ontology of laws and to explicate the role of these laws in science. The notion of laws of nature is also essential to many other philosophical issues. For example, laws of nature are invoked in the analysis of modal concepts, counterfactuals, causation, explanations, the connection between the mental and the physical, and so on. Clearly, we need a plausible philosophical account of what laws of nature are.

But the current philosophical discourse on laws of nature is very complicated. Earman (2004) calls it a 'scandal', saying

[52] See, for example, Roberts (2008) for more examples like this.

> *It is hard to imagine how there could be more disagreement about the fundamentals of the concept of laws of nature—or any other concept so basic to the philosophy of science—than currently exists in philosophy. A cursory survey of the recent literature reveals the following oppositions (among others): there are no laws versus there are/must be laws; laws express relations among universals versus laws do not express such relations; laws are not/cannot be Humean supervenient versus laws are/must be Humean supervenient; laws do not/cannot contain ceteris paribus clauses versus laws do/must contain ceteris paribus clauses.* (Earman 2004; 1228)

Taking a position in the philosophical debate about laws of nature is closely connected to one's position about other issues in metaphysics and philosophy of science: about the fundamental metaphysical structure of the reality, the relation between epistemology and metaphysics, the ontological status of theoretical entities in science, and so on. More specifically, these debates include debates over the metaphysical and epistemological nature of time, space, causation, explanation, chance, and so on. It is not easy to neatly categorize different views about these issues, but it is agreed that, broadly speaking, there are three possible positions one may take.[53] One might take an *eliminativist* position, according to which the entities in question are simply non-existent hence can be eliminated from the philosophical discourse. Or, one might take a *primitivist* position that the entities in question are fundamental, non-reducible, primitive element of the reality, i.e., they are weaved into the fabric of the reality. Philosophers who hold this position support the *governing*-laws conception of laws of nature. Or, one might take a *reductionist* position that there are the entities in question but they can be reduced to other things, without remainder. Philosophers holding this position support the *non-governing* conception of laws of nature. Let us examine each of these positions in turn.

[53] See, for example, Mumford (2005) for categorization of the three positions.

4.1.1 Eliminativism

Eliminativism about laws of nature is the view that there are no such things as laws of nature; laws are neither reducible to other things and nor are they a distinct category in their own right. Some philosophers from this camp argue that laws of science, for instance, are at best the approximation of truth and are only instrumentally useful fictions for a limited set of scientific inquiries. For example, Cartwright (1983) has argued that natural processes are not governed by laws and that propositions of laws of nature are not true at all, while they sometimes are instrumentally useful as descriptions of causal powers. Explanatory laws in physics fail due to their inability to correspond to complexity of reality, so the fundamental laws of physics do not describe true facts about reality (Cartwright 1980; 865). Some posit that there is no plausible account that can provide coherent truth conditions for statements concerning laws and that therefore the whole concept of laws of nature should be abandoned. For example, van Fraassen (1980) argues that laws as commonly characterized in philosophical discourse do not have place in science and none of the existing philosophical theories about laws provides an adequate account. He says that the aim of science lies in empirical adequacy, not truth. A scientific theory is empirically adequate if it truthfully says about the *observable* features of the world, that is, if it "saves the phenomena" (1980; 12). Therefore, he argues, scientific realists are mistaken as they claim that what (successful) scientific theories say about laws of nature do reflect the objective reality of nature. Some hold that we may understand certain aspects of nature by abstract models but that these don't deserve to be called laws (Giere 1999, for example). Some hold a sort of hybrid form of eliminativism. For example, Mumford (2004) claims that in the essence of natural properties are there genuine metaphysical necessity properties and relations and we happened to call them 'laws of nature'. But the concept 'laws of nature' should be abandoned as it is too much infected by the jurisprudential metaphor from the seventeenth century (Mumford 2004; 201-5). Mumford calls views like his as *lawless realism*, in the sense that the view endorses certain nomic entities as fundamental, primitive elements of the reality but denies that 'laws of nature' genuinely refer to them.

In general, eliminativists about laws of nature share that the statements of laws of nature are not genuine claims about objective features of the world, but are just descriptions of certain aspects of mathematical, abstract models; there are no objective laws of nature corresponding to law-statements. Therefore, they tend to claim, the traditional conceptions of laws of nature do no substantial work either in science or in philosophy.

Considering the significant roles that laws of nature play in much of scientific and philosophical projects, the eliminativist view strike many unappealing. At least, it seems prudent that we wait until we thoroughly investigate the plausibility of the alternative views before we resort to the eliminativist view; it seems too hasty to conclude that laws of nature are nothing more than an outdated philosophical illusion. Also, the concept of laws of nature sometimes plays indispensable roles in science. Roberts (2008) present such an example. Statistical mechanics requires a distinction between dynamically possible trajectories (trajectories consistent with the underlying laws) and merely logically possible trajectories. This distinction enters the definition of the statistical measure: that measure must be such that it is invariant under all of the dynamically possible trajectories, but it need not be invariant under all of the logically possible trajectories. While examples like this does not decisively show us how we should understand the distinction between laws and no laws, Roberts claims, it suggests that it is a bit too hasty to merely claim that the idea of a law of nature is a just metaphysical leftover from a traditional theological worldview.[54] It is a concept that does at least some real work in at least some real science, and the philosophical problem of explicating it is a legitimate problem for the philosophy of science (Roberts 2008; 16). Now let us turn to the second position about laws of nature: primitivism about laws of nature.

4.1.2 Primitivism

Primitivism about laws of nature holds that laws are genuine, fundamental, non-reducible elements of the reality; they are weaved into the fabric of the reality. Primitivists generally endorse the *Governing-Laws* conception. The Governing-Laws view posits that there are genuine laws of nature and that these laws do govern the universe. The leading figures in this camp were Dretske (1977), Tooley (1977), and Armstrong (1983), hence the view is often labeled as the *DTA* view. Philosophers in this camp tend to argue that there must be something that binds or 'glues' instances of properties; that laws of nature are exactly what do that. The instances of law-like regularities in the universe are governed by laws of nature, hence we live in a 'law-governed' universe.

This position carries the burden of metaphysical and epistemological explanation. The conception of law-governing needs to be fleshed out in order for it to be more

[54] *Cf.* See Jane Ruby (1986). She claims that a common perception of the origin of the concept of laws is mistaken; the origin of the concept does not lie in the notion of a divine legislator as commonly thought, but in the analogy to mathematical and logical laws.

than just a metaphor; if laws do govern the universe, exactly in what does the governing consist? It seems that any promising answer to the question needs to appeal in one way or another to some necessity relationship between universals. It is indeed the approach that philosophers in this camp tend to take. For example, Armstrong characterizes lawhood as follows:

> *It is a law that Fs are Gs if and only if F-ness necessitates G-ness; the necessitation relation between the property of being F and that of being G is what binds an instance of F to that of G. (Armstrong 1983, p. 85).*

According to this view, for example, Gallileo's law of gravity is the necessitation relation between two properties: the property of falling freely and the property of having an acceleration of 9.8 m/s^2. Such a necessitation relation between the two is what governs regularities between instances of first and second properties.

What primarily motivated this governing law conception is that it can readily explain the difference between accidental generalizations and law-like generalizations. For example, consider the following true generalization statements:

All gold spheres are less than a mile in diameter.

All uranium spheres are less than a mile in diameter.

Both are true. However, the former is merely an accidental generalization, while the latter is a statement of a law (according to our current understanding of nature). This is because uranium's critical mass is such as to guarantee that sphere of that size will never exist.[55] Necessarily,[56] then, it holds that all uranium spheres are less than a mile in diameter. In contrast, the truth of the former statement is not a matter of necessity.

[55] This example is from van Fraassen, *ibid.*

[56] Physical, not logical, necessity.

So, according to the necessitarian conception of laws, what makes the latter a law but not the former is whether the necessitation relationship holds between the properties in question. Primitivists argue that mere universal truths cannot play the role of holding the properties in question. It is for this reason that Dretske, for example, thinks we have to make 'ontological ascent' (Dretske 1977; 263) from talking about the objects or events instantiating certain properties to the properties themselves and their relations. That is, we need to make ascent from the universal truths like "All Fs are Gs" to the property relations like "F-ness → G-ness"; the latter entails the former, but not vice versa, so the latter is a law, not the former. In general we are not in position of knowing whether a given universal statement expresses accidental generalizations or necessary relations. Despite the possibility of such epistemic limitations, the necessitarians about laws hold that only the governing conception can provide support for counterfactual statements and explain what must, rather than what will, happen. (Dretske, 1977; 263)

Other notable, more recent primitivists include Carroll (1987, 1994). Expressing his sympathy to the view he calls 'nomic platonism' (1994; 161), he claims that laws are 'primitive and irreducible' (1987, 267). The concept of laws of nature is deeply embedded in the commonsensical, scientific, and philosophical discourse and we are ontologically committed to laws of nature as the primitive cement of the universe (1994; 160). Maudlin (2007) is another notable primitivist. He suggests that we should accept laws as fundamental entities in our ontology (2007; 18); the notion of a law cannot be reduced to other more primitive notions. He believes that our conceptions of laws serve roles as building blocks in our beliefs in other domains like counterfactuals, scientific explanations, physical possibilities, and so on.

But many think that Drestke's ascent, nomic Platonism, and other similar approaches, have a critical disadvantage: it renders laws unintelligible. Exactly what is the necessitation relation? It is not something we can experience. Even if we could experience all occurrences of everything throughout time and space, we would still not experience any necessitation.[57] Lewis illustrates this problem:

> *The mystery is somewhat hidden by Armstrong's terminology. He uses 'necessitates' as a name for the lawmaking universal N; and who would be surprised to hear that*
> *if F 'necessitates' G and* a *has F, then* a *must have G? But I say that N deserves the name of 'necessitation' only if, somehow, it really can enter into the requisite necessary connections. It can't enter into them just by bearing a name, any more than one can have mighty biceps just by being called 'Armstrong'"(Lewis 1983, 366)*

Bas van Fraassen (1989) also made a similar criticism, which he calls the identification problem. When observing universal relations in nature, how could creatures like us identify if it is a law or not? In a similar spirit, Roberts (2008) makes a criticism the governing-law conception allows no epistemic access. If laws are things that govern the universe, rather than simply pervasive regularities in the course of events, then how can we have any epistemic access to them? We cannot empirically detect whether it happened because it was necessitated by a law, or whether it happened just as a brute fact. This also creates a semantic problem for the primitivists because it remains indeterminate how the terms referring to such necessitation relations have determinate extensions (Roberts 2008).

[57] For a survey of criticisms in this line see Bird 1998, Lange 2009.

In short, a common objection to the primitivism about laws is that laws remains metaphysically and epistemologically mysterious; the view has to demystify the concept of ‘governing’ by precisifying in what does this 'governing' consist and how we could have epistemic access to it.

Now let us turn to a reductionist view about laws of nature. First we will examine the naïve regularity view in the next section. Then we will move on to the more sophisticated regularity view in §4.2.

4.1.3 Simple Reductionism: Naïve Regularity View

In this section, we will examine a branch of the reductionist view about laws of nature: naïve regularity view. According to the naïve regularity account of laws, it is a law of nature that P just in case P is a true universal generalization. This account might have the virtue of being the simplest possible philosophical account of laws. Either naïve or sophisticated, philosophers in the camp of the regularity view of laws usually hold the non-governing conception of laws of nature. This conception of laws may be seen as an attempt to strike a balance between the two extremes of the No-Laws and the Governing-Laws. The non-governing view states that there indeed are such things as laws of nature, but these laws do not govern the universe, because they are just regularities found in the universe. An example (from Bird 1998) will illustrate this idea. If it is a law that all free falling bodies accelerate at 9.8 m/s^2, then a particular object falling and accelerating at 9.8 m/s^2 is an *instance* of this law. This law is essentially the collection of all of such instances of free-falling objects accelerating at this rate. The relevant law amounts to the universal generalization of these instances. In sum, it is a law that *Fs* are *Gs* if and only if all instances of F are

G. This view is often called the simple (naïve) regularity (SR) theory of lawhood because it claims that laws are nothing but true regularities expressed in universally generalized statements.

One apparent advantage of this conception of laws is that it requires no heavy metaphysical assumptions like the necessitarian conception of laws. The SR is an empiricism-friendly view as it makes laws epistemologically accessible, at least in principle. On this view the concept of a law can be explicated by our experiences of instances of the target regularities. This idea may be related well to empiricism in that our concepts are explicable in terms that relate to our experiences.

One critical problem for the SR view is that it cannot distinguish accidental generalizations from laws. Consider again the gold sphere example. The regularity concerning the gold sphere is not a law; it just happens to be the case that there is no gold sphere that is greater than a mile in diameter. Suppose that, as a matter of fact, all the coins in my pocket are silver. However, we could be hardly warranted in saying that the generalized statement that all the coins in my pocket are silver should be referred to as a law[58]. For a similar example, consider laws in Newton's *Principia* and the movements of the planets. It is true but not a law that all of the planets go around the sun in the same direction. According to the SR account, this cannot be: if it is true that all planets go around the sun in the same direction according to Newton's theory, then this regularity must be among the laws of that theory.[59] Similar counterexamples may be readily constructed. The point of these counterexamples is that there exist law-like regularities and also *regularities that are not laws*. However,

[58] This example is from Carroll (1994).

[59] Roberts 2008;129.

according to the SR, any regularity should be counted as a law. This is troublingly counterintuitive.

The problems with the SR are well known and there are many.[60] To name a few: the SR cannot relate laws and counterfactuals; it has troubles for no-case laws (vacuous generalization); single-case laws (not a generalization); and probabilistic laws. Let me give a brief overview of these problems. Obviously, the SR cannot take into consideration *unrealized* physical possibilities. But we usually invoke unrealized but lawful physical possibilities in our counterfactual reasoning in science. For example, while it happens to be a contingent truth that all gold lumps are of a volume less than a cubic mile and all uranium lumps are of a volume less than a cubic mile, we may ask whether it *would* be physically possible to make a lump with a volume greater than a mile in gold and in uranium, and answers to questions like this have to invoke laws of nature. But this is not possible in the picture of the SR (Armstrong 1983). The trouble for the SR with respect to no-case laws is this. As we saw, the SR takes as laws true universal statements like “All Fs are Gs”. But, if there are no Fs and the universal generalization that all Fs are Gs is contingent and unrestricted, then, according to the SR, that generalization is a law (Carroll 1994). According to Armstrong, single-case regularities are ubiquitous in the sense that every object is different from each other in at least one or more microscopic properties, so every object makes true a universal statement in the form of “All Fs are Gs”, where ‘all’ only picks out the single object in question. The SR will have to count all of such single-case regularities as laws, but it is clearly counterintuitive (Armstrong 1983).

[60] Armstrong (1983), Carroll (1994), and Mumford (2004) contain a critical survey of the problems with the SR view.

The SR cannot account for probabilistic laws. Presumably, many fundamental laws of nature are probabilistic laws. For example, there seems to be a law about the probability distribution of a uranium atom decay over certain time intervals. But since the instances of the decay even does not always obtains (hence probabilistic), they cannot be formulated in a universal statement. Therefore the SR will end ep with rejecting such probabilistic laws; this is undesirable for a theory of laws of nature. So far I have discussed only a subset of the problems with the SR but they seem to be sufficient to reject the SR as an adequate theory of laws of nature.

Each view surveyed so far seems to have its own shortcomings. Of particular interest to us is the more refined version of reductionist view about laws, namely the *systems approach* to lawhood, which gains more and more popularity among philosophers of science. We will bring our attention to the view in the next section.

4.2 Best System Account of Laws of Nature

In this section I discuss the Best System Account (BSA) of laws of nature, which is a sophisticated version of the regularity theory about laws and probably is the best regularity view up to date. In particular, I will focus on David Lewis's BSA (Lewis 1973, 1983, 1994). On Lewis's account, law-statements are still statements of regularity and nothing more. Not all statements of regularities count as law-statements, however. The law statements are the axioms or theorems in the *best* deductive system in a world. The best deductive system is the one which achieves an optimal balance between simplicity and strength (and fit, when systems talk about chancy events). This account is sometimes called the *web of laws theory* (Psillos 2002; 148-54, cited in Mumford 2004; 40). The phrase 'web of laws' brings out one important feature of the BSA: Whether something is a law is not a purely intrinsic feature of it. Rather, something is a law when it is part of a systematic account of the world. I think this characterization nicely captures the essence of the system approach to laws.

In what follows, I will discuss philosophical motivations for the BSA (§4.2.1) and David Lewis's characterization of the BSA (§4.2.2). Then I will discuss the Humean supervenience thesis about laws (§4.2.3). Then I will examine some typical objections to the Humean conception of laws (§4.2.4). The examination of them will help clarify the range of the Humean supervenience thesis.

4.2.1 Motivation for the Best System Account

Facing the problems with the naïve regularity view, those rooted in empiricism may have to put some reasonable constraints on what kind of regularity is entitled to be a law. One such constraint can be supplied by selectively characterizing propositions expressing laws in terms of their relation to other propositions in an idealized theoretical *system*.[61] It could be the case that the difference between the propositions about the movements of all the planets gold spheres and uranium spheres lies in their relation within a system of true physics - say, the former is something that is merely derivative or peripheral to the true physics theory, while the latter is some basic law-like axioms or theorems of it. In light of this, we ought to develop a sophisticated way of understanding what an ideal system amounts to and exactly what role law-expressing propositions have within it. Arguably the most popular attempt in line with this aim is the Best System account (BSA).

[61] Carroll 1994; 47.

The BSA is a modified regularity theory. It says that regularities are laws if and only if they appear as theorems or axioms in an appropriately axiomatized collection of true propositions about the world – where ‘appropriately’ means that they are as *simple* and *strong* as possible. This view is often called the Mill-Ramsey-Lewis (MRL) theory of lawhood; Mill (1843) hinted at the idea of obtaining the fewest general propositions from all the regularities in the universe, while Ramsey (1929) characterized laws as axioms of the collection of knowledge of everything, organized as simply as possible. David Lewis later fleshed out this idea in full detail.

In *System of Logic* (1843), Mill says

> *According to one mode of expression, the question, What are the laws of nature? may be stated thus: What are the fewest and simplest assumptions, which being granted, the whole existing order of nature would result? Another mode of stating it would be thus: What are the fewest general propositions from which all the uniformities which exist in the universe might be deductively inferred?*

Mill here is hinting at the idea of systemizing facts in a simple but deductively powerful way. Here is an example from Mill (1843). Kepler expressed the regularity which exists in the observed motions of the heavenly bodies by the three general propositions. They were called laws in that these three simple suppositions which would suffice to construct the whole scheme of the heavenly motions. But these three “laws” of Kepler’s are not the best deservers of the title of laws, Mill says, because there is even simpler and more general law; that is, the phrase ‘laws of nature’ should be reserved for the simpler and more general laws, like Newton’s law (ibid.; Book III Ch IV sec1).

In *Universals of Law and of Fact* (1929), Ramsey says

> *laws are consequences of those [general] propositions which we should take as axioms if we knew everything and organized it as simply as possible in a deductive system' (Ramsey 1929; 150)*

According to him, even "if we knew everything, we should still want to systematize our knowledge as a deductive system, and the general axioms in that system would be the fundamental laws of nature" (1928; 143). Ramsey himself had abandoned this idea on the ground that it is impossible that we know everything and systemize it as a deductive system (1929; 150-1). But the core idea of laws as axioms or theorems in best systemization was revived by contemporary regularity theorists, and the most important figure is David Lewis.

In *Counterfactuals*, Lewis recasts the definition as "a contingent generalization is a law of nature if and only if it appears as a theorem (or axiom) in each of the true deductive systems that achieves a best combination of simplicity and strength" (1973; 73). On this view, laws are not just any regularities as the SR says. Rather, laws are some 'special' regularities which allow many facts about a world to be derived from them. And this can be done when they are the axioms or theorems of the best systematization of the world. The 'best' here means the right balance between being as simple as possible by having fewest possible axioms and being as strong as possible by allowing enough of the facts about a world can be derived from them.

Lewis says the BSA has a number advantages, which in effect are what motivate the BSA. They are (Lewis 1973; 74), with rephrasing:

1) The account explains why lawhood is not just a matter of the generality of a single sentence. The generality earns lawhood if it fits with other truths in the best system.

2) The account explains why lawhood is a contingent matter. In different worlds, different generalizations might earn lawhood by being a part of the best system at the world.
3) The account explains how we can know, by exhausting the instances, that a generalization is true while not yet know whether it is a law.
4) The account allows laws of which we have no knowledge. *Being* a law is not the same as being *regarded* as a law. On the account, there can be laws we don't know of.
5) The account explains why we have reason to take the theorems of well-established theories provisionally to be laws. Our actual scientific theorizing is an attempt to approximate the true deductive systems which strikes the best balance between simplicity and strength.
6) It explains why lawhood has seemed a rather vague and difficult concept: our standards of simplicity and strength, and our standard of the proper balance between them, are only roughly fixed standards.

We can see 1), 3), and 4) are clear advantages over the naïve regularity view. The regularity theorists may consider 2) as a great advantage as well, while the necessitarians might deny the metaphysical assumption behind 2).[62] Overall, these motivations for the BSA from 1) through 4) are rooted in Lewis's orientation in empiricism and denial of metaphysical heaviness of the necessitarian view. What additionally motivates the system approach is, as we can see in 5), that our science has been quite successful in finding laws. These successes have been achieved by creatures like us with epistemic limitations. How fruitful would be the ideal version of our science, which has the full access to the entire history of a world? The findings of such ideal physical theories would surely deserve the title of laws of nature – so the hope goes. What seems to motivate the BSA is the hope that the theorems or axioms

[62] Mumford 2008.

in the best systemizations will at least approximately or largely coincide with generalizations we presently regard as laws because the procedure employed is just an idealization or extension of the procedure actual scientists employ in discovering laws (Woodward 2013).

For example, on the BSA of chancy laws, the BSA's motivation like 5) is that "chances can be discovered by the methods of science. Again, the best system account makes this understandable, since one can reasonably hope that the methods of science get us close to the ideal theories whose probabilities are identified with the chances" (Schwarz 2014). Lewis himself (1994) gives a typical description of such motivation: "Suppose there is an ideal theory of everything ... [o]n the best system account, it follows that the rules of this ideal best theory are the true laws of nature." (Lewis 1994: 231f). And Lewis says what he thinks of is something not too different from present-day physics, though "presumably somewhat improved" (Lewis 1983; 364). We may even think of this 'ideal theory of everything' as what fundamental physics is aiming for, for example, Weinberg (1992)'s "Final Theory" or Penrose (2004)'s devoted "Unified Theory of Everything". The successes of physics to date provide reason to think that our world is susceptible to very good systematizations in fundamental terms (Loewer 2012), so it seems like a reasonable hope that laws can be found by systemizing facts in the same way as our fundamental science is theorized. All these remarks and comments point to the main theme of the BSA: laws are what the ideal theory says laws are. By extending actual scientific practice to the "somewhat improved" case in which all the data is in, we might reach what scientists have been ultimately aiming for, given the successes of physics to date.

On 6), it might seem strange to say it is an advantage of a theory that its key concepts are vague and indeterminate.[63] Some believe that the concept of a law is indeed vague so in that sense Lewis has point a here (Mumford 2008; Chapter 8). It might seem that what 6) does is the burden-shifting from the analysis to the practice of science. Two points can be noted on this. First, it seems true that the notion of laws of nature is vague and difficult. As we saw in §4.1, the notion is 'scandalously' difficult and vague (Earman 2004) in philosophy of science. In science, different fields invoke vastly different notions of laws. Secondly, it is more difficult problem for primitivists that such vagueness and indeterminacy exist in the notion of laws. Given the BSA regards the best system as an 'ideal physics theory', it can easily explain away the vagueness and indeterminacy of the notion.

We have examined the general motivations for the BSA. It is generally agreed that the analysis does a good job overall of meeting the desiderata of a theory of laws. The theory does not answer all our concerns but perhaps the theory might be better placed than any other theory of laws, at least from the perspective of the regularity view. Now let us examine David Lewis's BSA in detail.

4.2.2 David Lewis's Best System Analysis of Laws

David Lewis's best system analysis of lawhood (Lewis 1973, 1983, 1986, 1994) is an empiricist account of laws that invokes the theoretical virtues of simplicity and strength, (and statistical fit for an account of probabilistic laws), and seeks to strike a balance between these virtues. Lewis's canonical characterization is as follows:

[63] On the concept of overall similarity of worlds, which Lewis appeals to define causation via counterfactual dependence (Lewis 1973), he makes a similar remark that the vagueness of the concept of similarity is in fact an advantage.

> *Take all deductive systems whose theorems are true. Some are simpler, better systematized than others. Some are stronger, more informative, than others. These virtues compete: an uninformative system can be very simple, an unsystematized compendium of miscellaneous information can be very informative. The best system is the one that strikes as good a balance as truth will allow between simplicity and strength. How good a balance that is will depend on how kind nature is. A regularity is a law iff it is a theorem of the best system. (1994; 478)*

In essence, this is a modified version of the regularity conception of laws. It is a regularity theory in that laws are held to be regularities of particular facts, and not some "metaphysically mysterious" necessitation relations among universals as the necessitarian view says.[64] The BSA is a modified version of the regularity theory in that laws are held to be not just any regularities, but rather those regularities that systemize facts in a certain desirable way. There can be many different ways of systemizing the facts. Some systemizations will be simpler than others, while some will be more informative. Of these true systemizations, those that achieve the *best* combination of simplicity and strength are the *best systems.* Given this, a regularity qualifies as a law if and only if it is a theorem or axiom contained in such a best systemization of particular facts. This analysis seems to be in line with the epistemology behind our acceptance of some generalizations as laws. For example, we accepted Newton's first law of motion as a law, because it was an axiom in a simple, strong, and (at the time thought to be) true theoretical system: Newtonian physics.[65]

In some cases, the best system would need to deem some events chancy (e.g., atomic decay, coin tossing, dice rolling, etc.) in order to give a simple, informative, and

[64] Lewis says the motivation for pursuing an empiricist account of lawhood is "to resist philosophical arguments that there are more things in heaven and earth than physics has dreamt of." (1994; 474)

[65] Example from Carroll (1994; 48).

accurate summarization of the “chancemaking” patterns of matters of fact in that history. This is because if those chancy events constitute a non-negligible class of events in the entire history, some laws derived from the best system will need to concern those chance events. If not, it wouldn't be the best system because being silent about the non-negligible amount of chance events would cost too much in strength (Lewis 1994; 481). The (objective) chance of an event in a world is, then, what the best system says its chance is.

In the case of probabilistic laws like above, the BSA concerning chancy events is an extension of the BSA of deterministic laws. This time it involves three virtues – simplicity, strength, and *fit*– and striking a balance between them. Lewis characterizes the BSA on chance as:

> *As before, some systems will be simpler than others. Almost as before, some will be stronger than others: some will say either what will happen or what the chances will be when situations of a certain kind arise, whereas others will fall silent both about the outcomes and about the chances. And further, some will fit the actual course of history better than others. That is, the chance of that course of history will be higher according to some systems than according to others. [. . .] The virtues of simplicity, strength, and fit trade off. The best system is the system that gets the best balance of all three. The best system is the system that gets the best balance of all three. As before, the laws are those regularities that are theorems of the best system. But now some of the laws are probabilistic. So now we can analyse chance: the chances are what the probabilistic laws of the best system say they are. (1994; 480)*

In this way, the BSA concerning deterministic laws is modified to produce the chances and the associated chancy laws in one package deal. Consider deductive systems that pertain not only to what happens in history, but also to what the chances are of various outcomes in various situations—for instance, the decay probabilities for atoms of various isotopes. As before, some systems will be simpler than others.

Similarly, some systems will be stronger than others. Some systems will predict either what will happen or what the chances will be when situations of a certain kind arise, whereas others will provide no information about the outcomes and the chances. Further, some systems will fit the actual course of history better than others. In this way, the virtues of simplicity, strength and fit trade off. Some laws in the best systems are probabilistic; the chances are what these probabilistic laws say they are. Let us consider an example. Let *w* be a coin-tossing world where a coin is tossed one thousand times. Let H_w be the full history of the outcomes of these coin flips. There will be a best way of capturing these outcomes, which will strike an ideal balance between simplicity, fit, and informativeness. Then the chance of the coin landing on heads, *Ch*(Head) is equal to the chance of it coming up heads *according to* the best system. Say, in the history of 1,000 tosses of a coin at *w*, the actual frequency of its landing heads is 498 out of 1,000. The first system S_1 simply list the outcome of each and every toss; it just contains a long sequence of Hs and Ts. S_2 perfectly matches the actual frequency, claiming *Ch*(Head)=0.498 and *Ch*(Tail)=0.502. The third system S_3 says something slightly different; it rounds off the actual frequency, thereby giving one number summary for all of the tosses: *Ch*(Head)=*Ch*(Tail)=0.5. In short, they are:

S_1: HTHHHTTHHTHT… (actual sequence of the outcomes at *w*)

S_2: *Ch*(Head)=0.498 and *Ch*(Tail)=0.502

S_3: *Ch*(Head)=*Ch*(Tail)=0.5

S_1 is very strong in the sense that it rules out all the possible sequences and only talks about the true, actual sequence at *w*.[66] But the system is very complex. S_2 is more

[66] We will have discussion in detail on the notion of strength in §3.3.5 and §4.5.2.

efficient than S_1 in that it gives a nice summary of the outcomes, while losing some strength compared to S_1. Now S_3 is simpler than S_2 but S_2 fits the facts slightly better than S_3. Assuming that the gain in simplicity outweighs the loss in fit in moving from S_2 to S_3, it may seem that S_3 achieves better balance between simplicity and fit than S_2 does. So S_3 is the best system. Let X be the proposition that the coin lands heads. Then the probabilistic law about X at w is what is entailed by the best system, which in our scenario is: $P(X)=0.5$. Then, in virtue of being a part of the best system, the chance of the coin landing heads *is* 0.5 at *w*. This is how Lewis thinks the BSA is supposed to work, when Lewis says "suppose the frequency is close to some simple value—say, 50-50. Then the system that assigns uniform chances of 50% exactly gains in simplicity at not too much cost in fit. The decisive front-runner might therefore be a system that rounds off the actual frequency" (Lewis, 1994; 481).

In short, either deterministic or probabilistic, once the best system at a world *w* is determined, whatever it asserts as laws of nature *are* the laws of nature at *w* – these laws, in virtue of being asserted by the best system, earn their lawhood at *w*.

In sum, the BSA invokes the system choice criteria of simplicity, strength, and fit and the balance between them. It holds that the laws of nature in a world are theorems of the best system, which represents the simplest, most informative, and most accurate way of systematizing the categorical facts in that world.

4.2.3 The Humean Supervenience Thesis about Laws

The underlying principle of the BSA is that lawhood supervenes on nothing but the spatiotemporal arrangement of local qualities. This idea stems from the thesis which Lewis defends through the entirety of his work: the *Humean Supervenience* (HS) thesis.[67] According to the HS thesis, truth supervenes on what there is, and what there is is just a vast distribution of local particular matters of categorical facts and the spatiotemporal relations among them (Lewis 1986b: ix). This distribution is called the *Humean mosaic*. A property is categorical if its instantiation in a region of space time do not metaphysically necessitate anything about property instantiations in wholly distinct region.[68] At each point of the mosaic lie local qualities.[69] These are supposed to be supervenience bases. They are called the Humean bases. Promising candidates for the Humean bases are *charge*, *mass*, *size* and *spin*, for example. In this dissertation, our discussion will be confined to the HS thesis about laws: laws of nature supervene on the Humean mosaic.

It is worth noting at this point that there are different versions of the HS thesis put forth by Lewis. Accordingly, different HS thesis about laws can be formulated. According to the HS thesis in Lewis (1986; x), the HS thesis about laws would contain modal character of metaphysical necessity:

> For *any* two worlds which agree with the spatio-temporal distribution of fundamental qualities, laws are the same.

[67] As generally noted by many commentators on the BSA, the term 'Humean' has no direct commitment to Hume's theory of impressions, epistemic skepticism, and so on. The name comes from the historical construal of Hume as a denier of necessary connections in nature. See Lewis 1986, Loewer 2012, for example.

[68] I draw on the definition given by Loewer (2012).

[69] In Lewis (1986; xi), he mentions the worry that quantum entanglement in quantum physics is in conflict with his characterization of the HS thesis. Loewer (1996) attempts to modify the HS thesis in a way that it won't necessarily need 'local, distinct' qualities being instantiated at points of the mosaic. While this issue deserves a separate discussion, it itself won't affect my later discussion about the Arrovian impossibility for the BSA.

According to the HS thesis in Lewis (1994; 475),[70] a weaker HS thesis about laws may be formulated as follows:

> For any two *worlds like ours* which agree with the spatio-temporal distribution of fundamental qualities, laws are the same.

If the first version is adopted by the Humeans about laws of nature, then the debate concerning the Humean view about laws becomes the same form as the general philosophical debate on primitivism versus reductionism. Defenders and critics of the Humean conception of laws do generally take the first version as the official HS thesis about laws. (Hall 2012, Beebee 2000, Roberts 2001, Earman & Roberts 2005; Armstrong 2004).

The conjunction of the BSA and the HS thesis yields an account of lawhood which have a number of distinctive characteristic features. The BSA defines lawhood as a membership of the best systematization of facts; the HS thesis confines the BSA's operation domain to the Humean facts – the Humean mosaic.[71] As a result, the BSA account of laws and chances makes no commitment to ontological primitiveness of laws, necessities, causations, dispositions, or what Lewis calls "all the primitive unHumean whatnots" (Lewis 1994; 484).[72]

Let me explain how the above result follows. If we accept the HS thesis, it follows that every matter of fact in a world supervenes on the spatiotemporal arrangement of

[70] Lewis proposed the weaker version of the HS thesis because he thought the stronger version suffers from counterexamples concerning enduring objects. Lewis (1994) explains his worries about such counterexamples. Hall (2012) argues that Lewis should not have worried about them.

[71] Loewer (2007) develops a variant of the BSA which makes no commitment to the HS thesis.

[72] Throughout this dissertation, by 'the BSA' I will refer to the account resulting from the conjunction of the BSA and the HS thesis. I will clarify if I need to specifically talk about a variant of the BSA which makes no commitment to the HS thesis.

the Humean bases in that world. This entails that no two worlds can differ with respect to what is true in them without differing with respect to the arrangement of their space-time points, or with respect to the perfectly natural properties that are instantiated at those points (Hall 2012). If we accept the regularity conception of laws in addition to the HS thesis, it follows that the lawmakers themselves cannot be above and beyond local matters of fact. In this way, the BSA can be thought of as explicating *how* lawhood supervenes on the categorical facts – it does so through the best combination of simplicity, fit, and informativeness. The most important task for the BSA is, then, to make the case that *it is matters of fact that make a system the best system*.

Let me give an analogy.[73] Suppose there is a huge display screen consisting of millions of pixels and we run a little experiment. The experiment is to find the best way to watch the screen. We take a look at that screen through different pairs of 3-D glasses, each of which come in different degrees of scope, angle, focal length, at different costs, and so on. So each pair of the 3-D glasses gives us different 3-D images. Now suppose we somehow have settled on what the best pair of glasses is –of course, it depends on what 'best' means but let us assume that we have also settled its meaning– maybe the one that achieves the best cost-benefit efficiency. From now on we are to look at the screen by the chosen 'best' pair of glasses. The images produced by the conjunction of the pixels on the screen and our chosen pair of glasses will still supervene on the pixels; we cannot get different images without changes in the patterns of the pixels on the screen. What the HS claims amounts to the claim that *the*

[73] We will revisit this analogy in §5.6.2 as we discuss the problem of circular explanation of the BSA.

images themselves *are* also facts about the screen. It is *matters of fact* that make the images in question the best images.

4.2.4. Typical Counterexamples to the HS: "Same Humean Base, Different Laws"

Typical counterexamples to the Humean view about laws have used the thought experiments about possible worlds which agree with the Humean base but disagree with laws. I will examine counterexamples discussed in Earman (1986), Tooley (1977), and Carroll (1994).

"Single-particle Worlds" (Earman 1986; 212, Roberts 2008; 357)

Suppose there are two possible worlds W_1 and W_2, in each of which there exists just a single particle and nothing else. In W_1, the particle eternally travels at a constant velocity. In this world the laws of nature are exactly like Newton's laws. In W_2, the particle eternally travels at the same constant velocity. However, in this world the laws of nature is that every particle travels at the same, fixed, constant velocity.

"Ten-Fundamental-Particles World" (Tooley 1977; 669)

Suppose a world that contains only 10 different kinds of fundamental particles. So, there are 55 types of two-particle interactions. 54 of these interactions have been studied and the laws governing them have been discovered. The 55th kind of interaction, which is supposed to be between X-particles and Y-particles, never occurs because particles are located in such a way that X-particles and Y-particles will never meet. In this world, it could a law that the interaction of X-particles with Y-particles will result in the interaction of A-particles with B-particles. But it could also be a law that the interaction of them will result in the interaction of C-particles with D-particles.

The common element in the two examples above is that we seem to have two possible worlds which agree on the Humean base but disagree on laws. So they seem to be counterexamples to the Humean conception of laws. Now let us examine a more sophisticated counterexample presented by Carroll (1994).

The Mirror Argument (Carroll 1994; 57-68)

Consider a possible world U_1 consisting of exactly five *X*-particles and five *Y*-fields, and not much else. All the particles travel in a straight line, at a constant velocity, forever. Each of the five particles enters a *Y*-field at different times, exits quickly, and never returns. All of the *X*-particles get spin up when they enter a *Y*-field. But near the path of one of the five particles –particle *b*– there is a mirror on a swivel. It is fact that the mirror is in a position (position *c*) such that it does not interfere with the flight of particle *b*: the particle just travels by it. But if the mirror had been swiveled round to position *d* (or if it had just always been in position *d*), it would have interfered with the flight of particle b, and the particle would have been deflected away from its *Y*-field. Call the generalization *L* that all *X*-particles subject to a *Y*-field get spin up. *L* is a law in U_1. Now consider possible world U_2. U_2 is just the same as U_1 except that in U_2 particle *b* does not acquire spin up when it enters the Y-field. So *L* is false, so it is not a law at U_2.

Now imagine a world U_3. U_3 is the same as U_1 except that in U_3 the mirror is in position *d*, deflecting the particle *b* away from its *Y*-field. So the particle *b* never enters the *Y*-field. Should *L* be a law in U_3? If we base our intuition on U_1, then it seems that *L* is a law in U_3. But if we base our intuition on U_2, L doesn't seem to be a law in U_3. The changes in the position of a tiny mirror, intuitively speaking, cannot affect laws about the particles. So, Carroll concludes, this example shows that the Humean conception of laws is implausible.

This is Carroll's Mirror argument. The main 'trick' of this argument is that U_3 is equally similar to U_1 and U_2. To see the thrust of the argument clearly, let me introduce another world U_4. First, U_3 is:

> We get U_3 by changing the mirror position in U_1. So, U_3 is a world which is exactly same as U_1 except that the mirror is in position *d*. Assuming the changes of the mirror position cannot affect laws about the particles, U_1 and U_3 have the same laws: *L*.

Now imagine U_4:

> We get U_4 by changing the mirror position in U_2. U_4 is a world which is exactly same as U_2 except that the mirror is in position *d*. Assuming the changes of the mirror position cannot affect laws about the particles, U_2 and U_4 are in agreement on laws about them: *L* is not a law.

So, U_3 and U_4 agree on the entire history about the particles but disagree on laws about them. We seem to have a genuine counterexample to the Humean conception of laws. The validity of this argument relies on two principles, which Carroll calls (SC*) and (SC'):

> (SC*): if *P* is physically possible and *Q* is a law, then *Q* would still be a law if *P* were the case.
>
> (SC'): if *P* is physically possible and *Q* is *not* a law, then *Q* would still not be a law if *P* were the case.

On Lewis's analysis counterfactuals (Lewis 1986; 43-45), which utilizes the concept of overall relative similarities among worlds, both (SC*) and (SC') are false.[74] While noting this, Carroll argues that there are clearly intuitive reasons we shouldn't abandon (SC*) and (SC'): laws do not counterfactually depend on events like a tiny mirror's position changes (Carroll 1994; 186-87).

Beebee (2000; 589-91) argues that Carroll's mirror argument, and the similar arguments we have seen earlier, are committing question-begging against the Humean view about laws. In particular, Beebee accuses Carroll of using "intuitions" which presupposes the governing-conception of laws, the conception the Humeans reject in the first place in their metaphysical outset. Loewer (1996; 193-4) and Schaffer (2008; 95) make similar accusations. I agree with the question-begging accusation. The Humeans have no reasons to accept (SC*) or (SC'). For them, laws are some elite regularities and nothing more; therefore, changes in the tiny mirror's position can affect laws about the particles as long as those changes result in changes in the pattern of the movements of the particles. In Carroll's scenario, the mirror position change do result in such changes. Also, the intuition Carroll appeals to doesn't seem to be the right kind. In our world, a tiny mirror does seem too miniscule to generate changes in laws. But in Carroll's mirror world, the mirror takes up a huge portion of the universe; after all, all there are in that universe are the particles and the mirror. Relative to the inhabitants of the mirror world, the right intuition would be

[74] Lewis (1986) gives the famous Nixon and the nuclear missile button example. In short: at our actual world, it is possible that Nixon had pressed the nuclear button. In a possible world where he did press the button, the laws of nature are same as ours up to the point right before he pressed the button, but his pressing the button requires some 'small miracle', i.e., violation of the laws of nature, and then from the point he had pressed the button, the history diverges from our actual world's history. So in that possible worlds, overall, laws of nature are different from ours.

that the mirror as one of the main elements in the universe can affect laws of nature at their world.

Roberts (2008; 358-361) offers a new approach, based on what he calls 'meta-theoretic conception' of laws, to the typical counterexamples like Tooley's. In this approach, the truth value of a law-statement is relative not only to the possible world at which it is to be evaluated, but also to the context from which it is to be evaluated. Let us apply Robert's approach to Carroll's mirror argument. When we are asked to consider a possible world (U_3) obtained by changing the mirror position in U_1, we evaluate the possible world in question relative to U_1 and to the context of the salient theory at U_1. We may call the salient theory at U_1 the *L*-theory. The statement "*L* is a law at U_3" is not evaluated independently; it is evaluated relative to U_1 and the context of *L*-theory, and the statement comes out to be true in that context. Similarly, the statement "*L* is not a law at U_4" is evaluated in the context of *non-L*-theory, and it is true in that context. As a result, on Roberts's approach, the troubling result for the BSA which was originally drawn by Carroll from the mirror argument:

> U_3 and U_4 agree with the history about the particles but *L* is a law in U_3 and *L* is not a law in U_4.

This should be rewritten as:

> U_3 and U_4 agree with the history about the particles and *L* is a law in U_3 relative to the *L*-context and *L* is not a law in U_4 relative to the *non-L*-context.

It is not my aim in this dissertation to draw a verdict on whose view is more plausible on the arguments like the Mirror Argument. But I think we made it clear that the Humean, hence the BSA's conception of laws of nature is the non-governing conception, and that counterarguments implicitly presupposing the governing conception of laws are likely committing question-begging. The importance of clarifying this point will be clear once we discuss a new problem I raise for the BSA in the next chapter. For now, let us first precisify the BSA to see exact nature of the system choice standards for the BSA.

4.3 Making the BSA Precise

Since the best system is the one that achieves the best balance between simplicity, informativeness, and fit, the BSA must explain what the nature of each of these standards is, and what counts as the best balance between these virtues. Lewis does not precisify them. In this section, I will discuss and assess some possible attempts to flesh out the three standards and the balance between them.

4.3.1 Why the BSA Advocates Do Little Work On Precisification: Delegating To Science

The defenders of the BSA have done relatively little work on precisifying the three standards and the balance between them. Here is Woodward (2014)'s diagnosis as to why the BSA defenders tend to do little work on precisification. As we saw in §4.2.1, the BSA takes or 'imports' from the theory choice procedure of actual science as its system choice procedure. This import is based upon the reasonable (to the BSA advocates, at least) hope that the theorems or axioms in the best system will lead us to laws because the procedure employed by the BSA is just an idealization or extension of the procedure actual scientists employ in discovering laws – given they have discovered many of what we regard as laws by such procedures of systemization. Lewis is clear on this aspect of the BSA, in saying:

> *...I take a suitable system to be one that has the virtues we aspire to in our own theory-building, and that has them to the greatest extent possible given the way the world is. (Lewis 1983; 367)*
>
> *The standards of simplicity, of strength, and of balance between them are to be those that guide us in assessing the credibility of rival hypotheses as to what the laws are. (Lewis 1986; 123)*

The advocates of the BSA tend to think it is in fact an advantage of the analysis that it is rooted in methodology of actual science of ours. So defenders of the BSA tend to spend little effort on trying to precisely characterize the notions of simplicity, strength, and best balance on which they rely – because, if there are vagueness in such notions, the blame is not on the BSA but on the actual scientific practice. These notions are to be exemplified in scientific practice and the BSA does not have to fill in the details of the notions; what we have to do is to check how these notions are effectively used in practice of science. The vagueness in the official definitions of the system choice standards and the balance are thought of as tolerable because the concrete details of what simplicity, strength, fit, and best balance involve are to be supplied contextually in particular cases by science itself (Woodward 2014).

4.3.2 Why the BSA Advocates Do Little Work On Precisification: The Hope Thesis

Another reason that the BSA advocates do little work on precisification lies in what I call the Hope thesis (Lewis 1973, 1994). Let me quote Lewis' canonical expression of his hope about nature:

> *Maybe some of the exchange rates between aspects of simplicity, etc., are a psychological matter, but not just anything goes. If nature is kind, the best system will be robustly best-so far ahead of its rivals that it will come out first under* ***any*** *standards of simplicity and strength and balance. We have no guarantee that nature is kind in this way, but no evidence that it isn't. It's a reasonable* ***hope****. (Lewis 1994; 479, bold mine)*

> *…our standards of simplicity and strength, and our standard of the proper balance between them, are only roughly fixed standards. That may or may not matter. We may hope, or take as an item of faith, that our world is one where certain true deductive systems as best, and certain generalizations come out as laws, by* ***any*** *remotely reasonable standards…(Lewis 1973; 74, emphasis original)*

Many commentators on the BSA more or less agree that this hope thesis is reasonable and acceptable, unless there are clear and concrete counterexamples.

> *We can imagine, for example, that our world is such that there are two or more deductive systems which have little in common and which tie for first place on* ***any*** *reasonable account of simplicity… In this case the notion of lawhood would be more subjective than we like to think. I take David Lewis to be saying that in our current state of knowledge we have reason to* ***hope*** *that such cases do not in fact arise in the actual world. And I take actual scientific practice to be a practical expression of this hope… Failure to produce them [concrete counterexamples] would support Lewis' hope. (Earman 1993; 418, bold mine)*

In short, the Hope thesis postulates that nature will kindly arrange itself in such a way that all legitimate system choice procedures will pick the same best system or systems that agree on the laws of our world. I think the Hope thesis is another reason that the BSA advocators, while noting the system choice standards are vague, do not work much on presicifying them. For example, I take it is for the above reason why Schwarz, like many other commentators on the BSA, says “it is not crucial to the best system approach how exactly these details are filled in” (Schwarz 2014; 4). Loewer seems to be in a similar spirit when he says: “No doubt the practice of physics leaves leeway concerning how to evaluate these criteria and how they apply. But it is not implausible that our world is so rich and complicated that *all* reasonable ways of precisifying these notions will result in Best Theories of our world that agree on the laws.” (Loewer 2007; *italic* mine)

In short, the BSA defenders strategy is to announce that it is harmless to leave vagueness in the crucial notions in their analysis like simplicity, fit, strength, and their balance, because 1) it is the business of the actual practice of science to fill in the details; and 2) the best systems will come out robustly best anyway under any reasonable and legitimate precisification of the system choice standards. It is our task to assess the adequacy of these responses. In order to carry out the task, we first need to refine the system choice standards first to carry out the task, at least to the extent they seem reasonably operational.

4.3.3 Chauvinism Objection, Kuhnian Worry for the BSA, and the Refined Hope Thesis

It is often pointed out that the BSA needs the mind-independent system choice standards and the balance metric for inter-standard trade-offs. Otherwise the account would entail what Lewis calls "ratbag idealism" that laws of nature are dependent on how we think about the standards and the balance metric (Lewis 1994; 479).[75] Given the way the BSA relies upon actual standards of science, it may seem to be the case that:

> If our standards were different, then laws would be different.

If this counterfactual is true, then it may follow that laws are subjective. This would be a very undesirable consequence for the BSA. Lewis's first solution (Lewis, *ibid.*) is to *rigidify* the standards and the balance according to the actual and present practice of science; they are *fixed* standards as such. Lewis's second solution, which he thinks is better than the first solution, is the Hope thesis that under any reasonable refinements of the standards the best system will come out to be robustly best. These

[75] See Carroll 1994; 49-55, Roberts 2008; 8-10, for example.

two solutions have become the BSAers' two standard responses to the questions where the system choice standards come from and how to justify them: First, like we saw in §4.3.1, the BSAers delegate the task of precisifying the standards and the balance metric to actual and present practice of science. Second, as we just saw in §4.3.2, the BSAers hope that the best system will be robustly best under any precisification of them – even in case where science does not seem to provide determinate precisification of them.

On the first rigidification solution, the worry about subjectivity goes away. In a possible world where our counterparts have standards different than ours, the best system is still same as our best system because the standards for the best system are rigidified by our standards. While this rigidification resolves the subjectivity problem, it creates yet another problem.

Some criticize that the BSA's rigid reliance on our actual and present practice of science is arbitrary (Armstrong 1983; 67, Carroll 1994; 54). On the rigidification solution, it may start to seem that the BSA relies too much on our actual and present standards. What makes *our present* practice so special in the analysis of lawhood? Why not, for example, invoke the standards of Martian scientists, or of our ancestors or predecessors? It is *chauvinistic* (Carroll 1990; 201) that only our standards should count in the analysis of lawhood. Call this objection the Chauvinism objection to the BSA.

Some BSAers proposed the 'indexical' BSA to avoid the Chauvinism objection (Roberts 1999). Consider the following typical BSA-style law statement:

> The laws of nature are the theorems or axioms of the best system according to *our* standards.

Lewis's rigidification solution was to fix the referent of "our" as our present standards of science. The indexical BSA, in contrast, treats "our" as a kind of indexical terms like 'here', 'now', 'I'. So, at a world like ours that its inhabitants use standards different from ours, the rigidified BSA says their laws are just same as ours, and the indexical BSA says that their laws are the members of the best system according to their own standards.

But the indexical BSA seems to entail undesirable consequences. If our world and the Martian world agree on the Humean base, then, according to the Humean Supervenience thesis we saw in §4.2.3, laws at the Martian world should be same as ours. Suppose further that the Martians are bizarre anti-inductionists. They aim at finding the least inductively successful regularities; they aim at finding the weak, inaccurate, and complex regularities in their theory building. On the indexical BSA, such bizarre anti-inductive regularities are laws at that world. Should the BSAers consider them as laws? I do not think they should. As we saw in §4.2.1, the important motivation for the BSA is that our science has been quite successful in its inductive practice of discovering laws of nature; the best system is the idealized version of science. For the BSA, not just any regularities count as laws; the regularities which play important roles as axioms or theorems in the best systemization count. So in the BSA conception of laws are already incorporated such inductive fruitfulness. I think an appropriate response on the BSA's side to the examples like our bizarre Martians is simply that, while they might call such bizarre regularities "laws" based on their anti-induction systemization methods, they are not laws in any interesting and meaningful sense.

In any case, while the Chauvinism objection helps clarify important aspects of the BSA, it is not a devastating blow to the BSA. This is because Lewis abandons the rigidified BSA and instead adopts the Hope thesis. Consider again the following counterfactual which allegedly was worrisome for the BSA:

> If our standards were different, laws would be different.

This counterfactual is false under the Hope thesis. In a world which is just like ours except our counterparts use different standards for theory choice, the best systemization at that world is same as the best systemization at our world. This is entailed by the Hope thesis that the best system will be robustly best under any reasonable standards (§4.3.2). What is in action here is the term 'reasonable'. The term clearly means something like 'proven to contribute to induction' as hinted in the bizarre Martians example. So, the Hope thesis with its hidden elements specified would be something like:

> **the Hope thesis refined**: The best system will be robustly best under any standards as long as they are contributory to inductive inference.

Refined this way, the Hope thesis addresses the analogue of Kuhnian skepticism (§2.1) for system choice for the BSA. The Kuhnian skepticism is:

> Different scientists may employ perfectly reasonable but very different theory choice standards and balance metric, resulting in very different conclusions.

Given that the BSA relies on actual standards of our science, Kuhnian skepticism applied to the BSA would mean that there can be very different but equally legitimate best systems. Now the refined Hope thesis can block this worry. Kuhn viewed theory choice standards as values, which should admit of individual variance as to how to apply them and how much weight to give them (Kuhn 1977a; 321-2 the "big five"). But these values are not something intrinsically valuable in themselves. They are instrumental values which are considered to have inductive advantages in one way or another (*ibid.*; 334-6).

The objections and responses we saw in this chapter have helped clarify the important role of the Hope thesis for the BSA to block the Kuhnian worry for the BSA and its range over inductively contributing standards.

4.3.4 "What Justifies the Use of Epistemic Standards in Metaphysics?"

There has been raised another problem with the BSA's reliance on actual theory choice standards of science. They are essentially epistemic standards, while the BSA purports to be a metaphysical theory of laws. Typically, the use of epistemic standards of has end-means justifications. For example, use AIC to achieve predictive accuracy, use Bayesian model selection methods to achieve maximum posterior probability, and so on. The standard of simplicity in AIC, for example, plays a role as a means to achieve certain end. The standard of fit in AIC plays a role again defined by the means-ends terms. But since the BSA is a metaphysical theory, such means-ends justifications do not hold. The epistemic roles played by the standards of science have no counterparts in the BSA. So, one might inquire, what justifies the use of epistemic standards in metaphysics?[76]

[76] For an argument in the same spirit, see Woodward 2013a.

Here is what I think would be a standard response on the BSA's side to inquisitions like above. It comes from the idea that the best system in the BSA is a kind of 'ideal science' as we saw in §4.2.1. An 'ideal observer', call her the *BSA Oracle*,[77] who has access to all the facts in the universe, takes standards *from our science* and use them to figure out the best systemization of facts, determine the axioms or theorems in the best system, and declare they are laws. So, the BSA is taking what are considered to be epistemic standards and *elevating* them to the status of standards that *are* 'constitutive of laws of nature'.[78] Despite its apparent reliance on epistemic standards, the BSA is constructing a metaphysics about laws of nature. This understanding of the BSA is in line with what Lewis says:

> *Despite appearances a*nd the odd metaphor, this is not epistemology! You're welcome to spot an analogy, but I insist that I am not talking about how evidence determines what's reasonable to believe about laws and chances. Rather, I'm talking about **how nature -the Humean arrangement of qualities- determines what's true about the laws and chances**. (Lewis 1994; 482-3, e*mphasis mine)*

In other words, lawhood of *P* consists in the Humean mosaic's being arranged in such a way that the BSA Oracle will choose a best system that entails *P*. It is just the 'actions behind the curtain' that her system choice rules are the ones lifted from our epistemic practice.

Depending on what conception of laws one holds, the above line of thought may or may not seem appropriate. It may seem unacceptable to primitivists (§4.1.2). For them, laws are ontologically primitive entities existing independently of us. So, metaphysical theories about laws should try to identify the metaphysical nature of

[77] I borrow the name from Hall (forthcoming)'s example.

[78] Hall forthcoming; p.16

laws and the way they govern the universe; in contrast, epistemic practice in science is only aiming at discovering them. For primitivists, what justifies epistemic standards may be that they help us discovering laws , but since metaphysics about laws have no business in 'discovering' laws, the use of epistemic standards in metaphysics of laws is unjustified.

But the response may seem to make perfect sense to the BSAers and reductionists in general who hold that laws are some regularities. For them, what justifies epistemic standards may be that they help us discovering regularities that we call laws. On this both primitivists and reductionists agree. Now, for the BSAers and reductionists, since laws are just regularities (ones that play certain roles in the best systems), the use of epistemic standards in their metaphysics about laws doesn't need any extra justifications. As Hall (Forthcoming) describes, for example, the BSA is officially metaphysics but unofficially is a kind of extended science. The BSA Oracle tells us what is the best we could ever get if we were to continue to use our standards. And, given that the ultimate aim of our practice of using the standards is to discover laws of nature, what she tells us are law *are* the best results we could ever get.

It is not my aim of this dissertation to make a verdict on whose conception of laws is adequate. I have made it clear that simply accusing the BSA of 'illegally using epistemic standards' might be begging the question against the BSA. So, this does not seem to be a serious challenge to the BSA so far.

However, a more serious challenge seems to arise from a different kind of justification problem. It is not unusual that the use of epistemic standards in science presupposes certain laws. Statistical mechanics is such an example. Then, what if the epistemic standards invoked by the BSA presuppose some laws? According to the BSA, laws are members of the best system, which in turn is to be determined by the standards, which are taken from epistemic practice in science. In short, laws are supposed to be the result of analysis on the BSA. The challenge is that there seem to be cases where the analysis itself has to presuppose laws. Unlike the earlier accusation of the BSA's allegedly unjustified use of epistemic standards, this challenge is on the seemingly circular justification or explanation. On the one hand, in some cases, scientist's use of certain epistemic standards is explained by laws about how the Humean mosaic behaves. On the other hand, why the Humean mosaic behaves in the way it does is to be explained by laws, which on the BSA are to be determined by the best systemization based on those epistemic standards. So we seem to have circular explanation. Recently, a number of commentators on the BSA have discussed this circularity problem, for example, Maudlin (2007), Loewer (2012), Lange (2013), Hicks and Elswyk (2015), and Marshall (forthcoming). In §5.6.2, I will discuss this circularity problem. I will suggest that the recent solutions to the circularity problem are on the right track but those solutions eventually will have to hang onto the Hope thesis.

It serves our purpose of this chapter to make it clear that mere accusation that the BSA cannot justify its use of epistemic standards does not seem to work, and that however a more serious challenge awaits. Now let us continue to precisify the BSA.

4.3.5 Precisifying the System Choice Standards of the BSA: Strength

In the following three sections, I will make an attempt to refine the three system choice standards invoked by the BSA, to the extent that each of the refine standards allow weak orderings of systems in competition, that is, systems can be ranked with respect to each standard. In the current subsection, let us examine strength.

Lewis is not clear how to measure a system's strength. One possible precisification of strength would be to understand it as logical strength. Given Lewis describes systems as true deductive systems (Lewis 1994), it might seem as a good precisification to define strength as logical strength. Logical strength of a system is measured by all of its deductive consequences. That is, the more consequences can be deduced from it, the stronger the system is. On this conception of strength, one might think we can compare strength of two systems by counting the number of deductive consequences from each system. But this will not work; for example, infinitely many consequences can be deduced from a set of propositions. If a proposition P is a consequence of a system, so is P v Q, P v R, P v S, or any disjunction with any other propositions. The list goes on. The strength of systems cannot be compared.

There may be some limited cases in which the relative ranking of strength of two systems still can be determined. System X is stronger than system Y if the set of deductive consequences of Y is a proper subset of the set of deductive consequences of X. For a simple example, if X consists of P, Q, and R, and Y consists of P and Q, then every deductive consequence of Y is also a deductive consequence of X but not vice versa. But this comparison is only possible when there is such a subset relation between the set of consequences of systems in competition.

Another common conception of strength is to measure strength of a system by the degree of informativeness. Lewis often indicates that the strength of a system is to be measured in terms of its informativeness (Lewis 1983, 1994). Lewis does not precisify what informativeness is, but on his description a system is more informative about a world when it says more about the facts in the world –e.g., about what will happen or what the chance of a certain kind of event occurring will be (Lewis 1994; 480). Informativeness is generally understood by many commentators on the BSA as a matter of excluding possibilities or possible worlds. In general, a system is said to be more informative if it rule out more possible ways (possible worlds) than others do. That is, the more possibilities a system excludes, the greater its strength (Earman 1984, Loewer 2004, 2007, Callender and Cohen 2009, Woodward 2014, Hall (Forthcoming)). Some BSAers add more specification on this general notion of informativeness. For example, Earman (1984) suggests that strength should be measured not by sheer information about the facts *per se* but by information about the facts and regularities which can be explained by dynamic laws in conjunction with appropriate boundary conditions. Some suggest that a system should be considered stronger if it allows a wider range of initial conditions and a narrower range of candidate dynamic laws (Hall (forthcoming) and Woodward (2014), for example). But there is a problem with this notion of strength as informativeness. Informativeness is a matter of excluding possibilities but the excluded possibilities are typically infinite. For example, consider the proposition:

P_{10}: The number of planets in solar system is less than 10.

P_{10} rules out the possible worlds in which there are 10 planets, 11 planets, and so on; infinitely many possible worlds are excluded. Then different propositions (or systems, for that matter) typically will have the same degree of strength if strength is a matter of how many possible worlds are excluded.

There might be some limited cases in which relative ranking of strength of two propositions (or systems, for that matter) can still be determined. For example, consider another proposition:

P_{100}: The number of planets in solar system is less than 100.

Then the set of possible worlds excluded by P_{100} is a proper subset of the set of possible worlds excluded by P_{10}. In this case P_{10} comes out to be stronger than P_{100}. But comparisons like this is only possible when there is the subset relation between the sets of the excluded possible worlds by systems in competition.

Lewis and other BSAers seem to take for granted that strength comparisons across systems can be made in terms of objective features of those systems. But we just have seen that the comparisons don't come easy.

4.3.6 Precisifying the System Choice Standards of the BSA: Fit

Let us precisify the system choice standard of fit in this section. Lewis defines a system's fit as the chance that the system in question assigns to the world's total history (1986; 128). It is clear that Lewis has in mind the *likelihood* of a theory, which can be defined as the joint probability of the data (or history of events) *given* that theory. Suppose, throughout the history of a world *w*, a particular coin C_1 has been tossed hundred times and landed forty nine times on heads and fifty one times on tails. Call that history H_w. Suppose further that, of two competing systems S_1 and S_2,

S_1: the chance of C_1 landing on heads is 0.49

S_2: the chance of C_1 landing on heads is 0.50.

Computing the likelihood for each system, we see that the likelihood of S_1 comes out higher than that of S_2 because,

Fit of S_1=Pr(H_w | S1)=$(0.49)^{49}(0.51)^{51}$=8.0479 × 10^{-31}

Fit of S_2=Pr(H_w | S2)=$(0.50)^{100}$=7.8886 × 10^{-31}

So S_1 comes out to fit H_w better than S_2 does, and this result conforms to our intuition. This notion of fit has a serious problem in the case of infinite history. The problem is called the zero-fit problem (Lewis 1980, Elga 2004): in short, when history is infinite, all the systems come out to have equally zero fit, rendering fit as a system choice standard useless. I will discuss in detail the problem and possible solutions to it in §5.5.1.

4.3.7 Precisifying the System Choice Standards of the BSA: Simplicity

As with the two system choice standards we have seen, Lewis does not precisify what he means by simplicity. But he is generally understood as referring to the *syntactic* notion of simplicity; "simple systems are those that come out formally simple" (Lewis 1986; 124). His examples are "a linear function is simpler than a quartic or step function" and "shorter alteration of prenex quantifiers is simpler than a longer one" (Lewis 1994; 479). Let me make a brief note on the syntactic notion of simplicity. It is well known the syntactic notion of simplicity suffers from *language dependence*. Formal simplicity of a theory can vary depending on which language is used for encoding what the theory says. For example, consider Goodman (1983)'s famous example of Grue-Bleen predicate. (Also see Priest (1976) for the language dependence problem in the context of curve-fitting.) For example, a theory saying "*all emeralds are green*" seems formally simpler than a theory saying "*all emeralds* *are green until the year 2050; after than they are blue*." But if we encode what is said by them using Goodman's 'grue' predicates, then the simplicity ranking of the two theory is reversed. Of green and grue predicates, one might say the former is more natural in that they capture natural kind of the world.[79] In this way, we may fix the privileged language and then compare syntactic simplicity of sentences, theories, or systems. This is what Lewis does in defining simplicity of systems; we should measure simplicity of systems formulated in a certain privileged language. Noting the possibility that a wrong choice of language can completely distort the simplicity ranking of systems, Lewis says:

> *We face an obvious problem. Different ways to express the same content, using different vocabulary, will differ in simplicity... In fact, the content of any system whatever may be formulated very simply indeed. Given system S, let F be a predicate that applies to all and only things at worlds where S holds. Take F as primitive, and axiomatise S (or an equivalent thereof) by the single axiom* $\forall x Fx$. *If utter simplicity is so easily attained, the ideal theory may as well be as strong as possible. Simplicity and strength needn't be traded off. Then the ideal theory will include (its simple axiom will strictly imply) all truths, and fortiori all regularities. Then, after all, every regularity will be a law. That must be wrong. (1983, p. 367)*

His remedy to this problem is to fix a language in which systems are to be axiomatized: a primitive vocabulary that refers only to *perfectly natural kinds* (1983; 367-8), which "carve nature at the joints."[80] Lewis doesn't offer much more

[79] As well known, Goodman (1983)'s choice of green predicate is based on its inductive success and projectibility.

[80] Lewis doesn't specify what determines 'natural' kinds; and it deserves an independent discussion on what is an adequate way to specify it. See Hall 2010 for an extensive survey of a number of possible precisifications of 'natural'.

explication than this, but it is obvious that he thinks simplicity is an objective feature of the set of expressions in such a privileged language (Loewer 1996).[81]

In addition to the syntactic notion of simplicity, Lewis and other BSA advocates also invoke 'counting' in their construal of simplicity. Systems with fewer assumptions, axioms, theorems, or fewer postulated entities, are considered simpler in the BSA (Lewis 1983, 1994). Whichever of these different senses of simplicity is taken, Lewis believes that simplicity is not resting on subjective matter; simplicity is an objective property of systems, assuming that the *natural kinds* in question are determined by facts of the matter. Let me discuss two problems with the simplicity as above invoked by the BSA.

Subjectivity of Simplicity

First, as opposed to Lewis's belief, there may be significantly subjective factors to simplicity. Carroll (1994) provides an example in which simplicity seems to be a matter of psychology. His idea is to view "... simpler than ..." relation as a triadic relation, which subjective factors. A sentence "A is simpler than B" is elliptical for "A is simpler than B for C", with C being an epistemic agent, a task, and the like. Consider the two hypotheses:

[81] Some BSA theorist adopt a pluralisitic pragmatic approach to the BSA. For instance, Cohen and Callender (2009) argue that, since it is impossible to make inter-system comparisons independently of the basic predicates employed by the systems in question, the BSA has to allow for plural best systems, each of which would be the best *relative to* the choice of the basic predicates. That is, systems can be compared only with respect to a certain system of predicates. This relativism is not harmful, they argue, because it is in accordance with scientific practice. Two scientists holding different ontological stances about the basic kinds in the world could agree with each other about what kind of predicates are to be used to describe the observable data. They could then relativize (i.e., reformulate) what they think is the best system of the said predicate kind for the observables, and then determine which system is better, i.e., better relative to that predicate kind. This approach may be able to handle the problem at hand.

$y = \sin x$

$y = x - x^3/6 + x^5/120.$

Both hypotheses are fitting equally well the following data represented as X in the following figure.

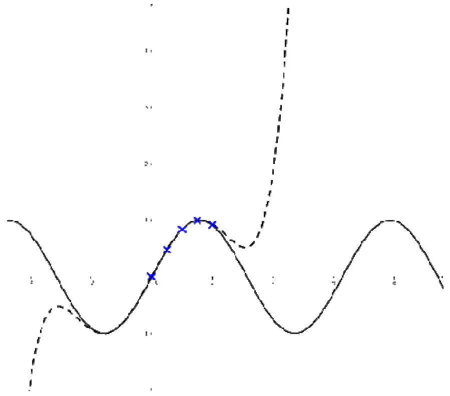

Fig 4. Carroll’s Lefty and Righty; Two Equally Fitting Curve*s*

Imagine that scientists in two cultures, namely Righty and Lefty, are comparing these two hypotheses against the given set of data. In Righty, scientists discovered truths about the relation between the angles and lengths of sides of triangles, and trigonometry is taught at an early age. For them, trigonometric equations are easier than polynomial equations. The opposite is true of scientist in Lefty. They are excellent at algebra but weak in geometry. The Lefty seem to have legitimate reason to rank simplicity of the trigonometric hypothesis above simplicity of the polynomial one (and choose the former as the overall winner given that both fit the data equally well) and the Righty seem to have equally legitimate reason to rank their simplicity in the reverse way. This example shows, Carroll claims, that there are psychological factors to simplicity.

Counterexamples like this don’t seem to be devastating to the BSA, though. First, if “… is simpler than …” is an ellipsis of “… is simpler than … for …” as Carroll claims, then it may just mean that there are more than one system choice standards invoked by the BSA, which happen to be under one umbrella term 'simplicity'. In other words, 'Simplicity-Lefty' generates a weak ordering of systems in its own and 'Simplicity-Righty' generates a weak ordering of systems in its own as well, rather than one weak ordering of systems under 'Simplicity'. Second, the BSAers still have a resort to appeal: the Hope thesis (§4.3.2). The hope would be: Nature will kindly arrange itself such that, under any refinement of simplicity like above, there will be a clearly winning best system – even when different senses of simplicity generate vastly diverging orderings of systems. It is open to question how plausible such hope is. But I think the BSAers have some answer: scientists often invoke different kinds of simplicity in theory choice but overall they have been successful in finding regularities that we consider as laws. So, they may argue, the Hope is not too far-fetched. A more difficult problem awaits, however.

Conflicting Standards of Simplicity[82]

As we saw above, in standard characterizations of the BSA, simplicity of a system concerns not only the simplicity of an axiom or theorem but also the number of axioms or theorems in the system. The problem would arise when the two standards of simplicity are in conflict. Consider the following case. System S_1 says there are four elementary forces in our world. Each of these forces may be represented as an axiom in the system. Let us say each of such axioms may be defined as a parametric model, and the number of parameters in each axiom is just one or two. In contrast, S_2

[82] I owe this part of discussion to Aidan Lyon.

says there is only one elementary force in our world, represented as an axiom, which is defined as a very complex parametric model, say, with 10 parameters.

S$_1$: Force$_1$=ax_1+bx_2; Force$_2$=cx_3; Force$_3$=dx_4; Force$_4$=ex_5+gx_6.

S$_2$: Force$_1$=$ax^{10}+by+cz+dx^9+ex^8+gx^7+my^2+ny^3+ky^4+hy^5$.

Additionally suppose S$_1$ and S$_2$ fits history of at a world equally well. The problem is that it seems impossible for the BSA to determine simplicity ordering of S$_1$ and S$_2$, when S$_1$ has four axioms and each axiom is very simple, and S$_2$ has one axiom which is very complex. Determining which system is simpler requires a trade-off ratio between simplicity in terms of the number of axioms and simplicity in terms of the number of parameters; there seems no reasonable, consistent way of trading them off. Furthermore, the Hope thesis cannot block this problem either because both systems fit history equally well. Unlike the subjectivity problem of simplicity, the BSA needs resources to solve this problem other than the Hope thesis.

4.3.8 Precisifying the System Choice Standards of the BSA: Balance

As with the other system choice standards, Lewis does not specify how to balance different system choice standards. First let us see what kind of balance metric is required for the BSA. The BSA needs a principled method for comparing the value of a certain gain in one virtue with a certain loss in another. So a good refinement of the required balance metric should allow for trade-offs between standards like "*This much* loss of simplicity can be compensated by *that much* gain of fit", "adding certain initial conditions to one system results in the new system that is only a *little less* simple but *vastly more* informative" (Loewer 2007; 305, *italic* mine).

As we saw in §4.2, Lewis (1986) thinks such a balance metric can come from the actual practice of science. In discussion of the Kuhnian worry and the Arrovian threat for theory choice, we saw in §2.4 that statistical model selection methods might provide some principled balance metrics. They seem to provide very specific exchange ratio between different standards as we saw in §3.2 and §3.4, for example:

AIC-rule: Choose the model *M* which has largest AIC score.

AIC score (*M*): Maximum Log likelihood of (*M*) – number of parameters of M

BIC-rule: Choose the model *M* which has largest BIC score.

BIC score (*M*): Maximum Log likelihood of (M) – (log *n*)(number of parameters of M)

If the likelihood and the number of parameters of a model can be understood as the model's fit and complexity, respectively, then the above methods seem to provide sufficiently specific inter-standard trade-off ratios. Presuming what works for theory choice will also work for system choice, the BSAers might expect that the principled balance metrics with the sufficient level of specifications required for the BSA may come from statistical model selection methods as well. Whether this expectation comes true or not is a part of the questions we are going to investing in the next chapter.

Conclusion: Towards the Arrovian Impossibility for System Choice

In this chapter we first surveyed different philosophical accounts of laws of nature. We then assumed the task of precisifying the Best System Account of laws. Some important items we have examined include: the range of the Humean Supervenience thesis, the role of the Hope thesis to block the Kuhnian worry for the BSA, a standard approach to the problem of justifying the use of epistemic standards in metaphysics, the best system as an extended, ideal science, and the heavy reliance on the concept of 'balance' between the system choice standards.

All these will be connected to the subject of the next chapter: the analogue of the Arrovian impossibility in the domain of system choice for the BSA. The BSA invokes an aggregation procedure of different system choice standards and especially it requires an exchange ratio between the standards being aggregated. So, the BSA might be susceptible to the Arrovian impossibility, if the conditions for the Arrovian theorem are met. In the next chapter, we will investigate whether the conditions apply to system choice for the BSA.

Chapter 5: The BSA and the Arrovian Threat

Introduction

As we saw in §2.1, Kuhn's worry[83] about theory choice (Kuhn 1977) was that different scientists may legitimately employ quite different theory choice procedures and reach conflicting conclusions. As we saw in §4.3, the system choice algorithm for the BSA is to be determined by the theory-choice practice of scientists.[84] Therefore, if Kuhn is right about there being different but legitimate theory choice procedures, there may be different but legitimate system choice procedures for the BSA. This seems to pose "the threat that two very different systems are tied for best" and "in this unfortunate case there would be no very good deservers of the name of laws." (Lewis 1994; 479) As we saw in §4.3.2, the Hope thesis (Earman 1993, Lewis 1994, Loewer 1996) is an attempt to block this worry by postulating that nature will kindly arrange itself in such a way that all legitimate system choice procedures will pick the same best system or systems that agree on the laws of our world.

In this chapter, I shall discuss a new worry about system choice. In social choice theory, as we saw in §2.2, Arrow's impossibility theorem (Arrow 1951/1963) says that there cannot exist any preference aggregation procedure satisfying Arrow's rationality conditions of **U**, **P**, **I**, and **D**. The result of this theorem is that any

[83] What Kuhn had in mind may not be necessarily a 'worry', given he says there are number of advantages for theory choice procedures to leave some diversity and indeterminacy (Kuhn 1977; 331-2). If we require there to be theory choice 'algorithms', then such diversity and indeterminacy might be a worry. On the related point, see §2.1.2. Also see Morreau (2015).

[84] It is an important question whether the BSA is a descriptive or prescriptive theory, i.e., whether the BSA should use (prescriptive) or is using (descriptive) the same choice procedure as the procedure scientists actually use to make theory choices. This question deserves a separate discussion. See Woodward (2014) for a critical discussion of descriptive and prescriptive adequacy of the BSA.

aggregation procedure satisfying the first three conditions will fail to satisfy **D**, the non-Dictatorship condition. We discussed the possibility of applying Arrow's theorem to theory choice in §2.3. If theory choice procedures are procedures of aggregating different theoretical virtues, and if the conditions **U**, **P**, and **I** are satisfied, the result of Arrow's theorem is that theory-choice algorithm will be dictatorial. In Chapter 2 and Chapter 3, we explored some possible escapes from the Arrovian result in theory choice. The focus of this chapter is on the question whether the Arrovian result carries over to system choice for the BSA. The Arrovian result for the BSA, if it obtains, will be that there is one criterion whose ranking of systems dictates overall system choice regardless how well they do with respect to the other criteria.[85] This seems like a serious threat to the BSA. In the following section, I will provide a detailed plan to assess the threat.

5.1 Threat Assessment: The Plan

The Arrovian threat for the BSA is that, if the certain conditions for Arrow's theorem (**U**, **P**, and **I**) apply to system choice, the system choice algorithm will be dictatorial (**D** fails).[86] For example, if fit is a dictatorial criterion implied by the Arrovian result, it means the most fitting system will always win, regardless how complex or how uninformative it is. The system(s) that is picked by such dictatorial criterion can hardly be seen as a result of certain 'balancing' procedure of multiple criteria as prescribed by the BSA. Furthermore, the Arrovian result cannot be blocked by the Hope thesis (§4.3). For example, suppose the Humean pixels in our world happen to

[85] Notice that the Arrovian result does not imply the existence of a 'dictatorial system'; what it implies is the existence of a 'dictatorial criterion'. These are completely different claims.

[86] See §2.2.3 for the formal statements of the conditions and the theorem.

be ‘kindly’ arranged in such a way that one system is dominantly better than its rival systems with respect to all of the system choice criteria. In this case, the Arrovian impossibility implies that each criterion is a dictator, by definition. This might seem as innocuous dictatorship. But, on the strong Humean Supervenience (§4.2.3), there can be other worlds where we should worry about dictatorial system-choice algorithms, where, for example, one criterion comes out to be a dictator.[87] So, dictatorship should be a genuine concern for the BSAers. Thus the Arrovian result seems to undermine the BSA to the extent it has a balancing process as integral to the analysis of lawhood.

This seems to raise a serious challenge to the BSA, so in this chapter I will examine whether the conditions of Arrow’s theorem apply to system choice. Let me lay out the plan for this chapter.

One might attempt to escape from the Arrovian threat by relaxing one or more conditions of Arrow’s theorem in system choice.[88] In this chapter, I will explore the plausibility of such attempts. First, in §5.2, I will discuss the possibility of relaxing the unrestricted domain condition (**U**) for system choice. As we saw in §2.4.2, **U** does not apply to theory choice but the Rich domain condition (**R**), a weaker version of **U**, may apply to theory choice. It will be argued that the same applies to system choice. I will claim that **U** does not apply to system choice either but **R** does. We will also note

[87] The strong HS says that any two worlds which agree on history should also agree on laws. On the BSA, if the system-choice algorithm is dictatorial, there can be cases where the strong HS fails. We will have further discussion on the relation between the weak HS and the Arrovian impossibility in §5.3.

[88] This does not necessarily open up a sure-fire escape route. For example, when **U** is relaxed, some other variants of the Arrovian impossibility still obtain, if R, the weaker counter part of **U**, obtains. We will have discussion of it shortly.

that relaxing **U** to **R** does not block the Arrovian result. As we saw in §2.4.2 and §2.4.3, literature in social choice and theory choice (Parks 1976, Hammond 1976, Kemp and Ng 1976, Roberts 1980, Rubinstein 1984, Feldman and Serrano 2008; Morreau 2015, Okasha 2015) has suggested that, even if **U** is weakened to **R,** a variant of Arrow's impossibility theorem obtains provided the strong neutrality condition (**SN**), a stronger version of **I**, is met. This will lead us to the question whether **SN** applies to system choice.

In §5.3, I will discuss **SN** in connection with the Humean Supervenience thesis. As we saw in §4.2.3, there are strong and weak versions of the Humean Supervenience (HS) thesis, the underlying principle for the BSA. While many BSAers believe the strong HS is the appropriate one, Lewis's original, weak HS is a contingent thesis which only applies to the range of worlds like ours. Arrow's impossibility theorem (1951/1963) was originally derived in the multi-profile framework (§2.2). So, the BSAers might respond to the Arrovian result for system choice by falling back to the weak HS thesis as a way to block such a multi-profile framework for system choice. They might hope this would open an escape from the Arrovian result for the BSA. I will note that this move is analogous to some early reactions to Arrow's impossibility in social choice literature. However, as we saw in §2.4.3, the conjunction of **SN** and **R** (along with **P** and **D)** yields an analogue of the Arrovian impossibility even in the single-profile framework. Assuming **R** is met in system choice (§5.2), we will be led to the question whether **SN** can be dropped in system choice. I will argue that **SN** is a desirable property of consistency for system choice.

In §5.4 and §5.5, I will discuss the possibility of relaxing the condition **I**. As we saw in §2.2.3, condition **I** is logically equivalent to the conjunction of the two conditions: Independence of Irrelevant Utilities (**IIU**) and Ordinal Non-Comparability (**ONC**). I will discuss each of them in the context of system choice since an escape from the Arrovian result will open up if we can abandon either one of **IIU** and **ONC** (Hammond 1991, 2004). As for inter-criterial comparability in system choice, we will first need to examine the cardinal measurability of the system choice criteria invoked by the BSA. This is because if they can be measured on cardinal scales, then our concern will have to be about Cardinal Non-Comparability (**CNC**).[89]

In §5.4 I will discuss the possibility of abandoning **IIU**, in comparison to **SN**. I will suggest that **IIU** is a desirable property of system choice procedure.

In §5.5, I will investigate the possibility of cardinalizing fit (§5.5.1), strength (§5.5.2), and simplicity (§5.5.3), the three criteria invoked by the BSA. I will conclude that none of the criteria seems cardinally measurable. Even if they were, as we saw in §2.4, cardinality without comparability cannot open up an escape from the Arrovian impossibility (Kalai and Schmeidler 1977, Sen 1970) [90]. This will lead us to a search for a form of inter-criterial comparability.

In §5.6, I will propose a variant of the BSA as an attempt to make a case for inter-criterial comparability between fit and simplicity: the A-BSA, an implementation of the BSA with Akaike Information Criterion discussed in §3.2. I will examine how

[89] This is also because we need to block some common but misguided responses to the Arrovian impossibility like "Why not measure them on cardinal scales?"

[90] If ONC is replaced with the logically weaker condition Cardinal Non-Comparability (**CNC**), the other conditions being met, the Arrovian impossibility still obtains for SWFL (Social Welfare Functional). See Sen 1970, Hammond 1986, for example.

this implementation turns out to be. I will also attempt to implement the BSA with Bayesian Information Criterion discussed in §3.4. However, it will be shown that the proposed variants face counterexamples, due to i) the context gap between statistical model selection and system choice, and ii) the assumption made by those statistical methods about the existence of 'true curve', which is inconsistent with the BSA of laws. I will conclude this chapter by suggesting that the attempts examined in this chapter do not have a good outlook.

I have laid out the plan for assessing the Arrovian threat. Following the plan, let us begin with the possibility of relaxing **U**, unrestricted domain.

5.2 Possibility of Relaxing the Unrestricted Domain Condition

In this section, I will discuss the possibility of relaxing the unrestricted domain condition, **U**, in system choice. In social choice, **U** says that any social choice algorithm should be able to handle all logically possible profiles of individual rankings. In theory choice, the analogue of **U** is that any theory-choice algorithm should be able to handle all possible profiles of rankings of theories with respect to theoretical merits such as simplicity, fit, informativeness, and so on. The analogue of **U** for system choice would be:

Unrestricted Domain (**U**): The domain of the system-choice rule is the set of all logically possible profiles of orderings of systems with respect to fit, strength, and simplicity.

U might appear to apply to theory choice and system choice. For example, as we saw in §2.3, Okasha (2011) suggests that **U** applies to theory choice as there should be no *a priori* restrictions on what profiles are admissible and what not to theory choice rule.[91] Likewise, it might seem for system choice that the analogue of **U** is appropriate because there should be no metaphysically privileged way to impose any restriction on what system ranking profiles are to be allowed. But, as we saw in §2.4, **U** does not seem to apply to theory choice. Once we fix the sense of simplicity to use to rank theories and the language to describe the theories in question, simplicity ranking of two theories is invariant regardless of data (Morreau 2015). If **U** is inapplicable to theory choice like this, then it means the admissible profiles are restricted in theory choice.

What of **U** for system choice? Is **U** applicable or inapplicable to system choice? It is worth quoting Lewis at this point. Lewis (1994; 479) says:

> *The worst problem about the best-system analysis is that when we ask where the standards of simplicity and strength and balance come from, the answer may seem to be that they come from us… I used to think rigidification came to the rescue... But now I think that is a cosmetic remedy only. It doesn't make the problem go away, it only makes it harder to state. The real answer lies elsewhere: if nature is kind to us, the problem needn't arise. I suppose our standards of simplicity and strength and balance are only partly a matter of psychology. It's not because of how we happen to think that a linear function is simpler than a quartic or a step function; it's not because of how we happen to think that a shorter alternation of prenex quantifiers is simpler than a longer one; and so on. Maybe some of the exchange rates between aspects of simplicity, etc., are a psychological matter, but not just anything goes. If nature is kind, the best system will be robustly best-so far ahead of its rivals that it will come out first under any standards of simplicity and strength and balance.*

Here Lewis may seem to say that system rankings can come out differently under ‘different standards of simplicity’ (and of the other criteria as well) but nature will be kind in such a way there will be a clearly winning system under ‘any of standards of simplicity’ (the Hope thesis; see §4.3). Passages like the above might give an

[91] In response to Morreau (2015), Okasha (2015) concedes that **U** does not apply to theory choice.

impression that Lewis himself conceded the possibility of reversible ranking of simplicity of systems (and of the other criteria as well). If the simplicity ranking were reversible, for example if simplicity could rank X above Y at one time and Y above X other times, then it may seem desirable that the domain of system choice admits all of such possible profiles of simplicity of systems.

But this is a false impression. If there are ‘different standards of simplicity’, it simply means that different choice criteria are invoked. (We may use labels like Simpliciy-1, Simplicity-2, … for each of the ‘different standards’ of simplicity to avoid confusion.) It is often claimed that different scientists can legitimately rank simplicity of theories in different ways. It is also often claimed that simplicity is a matter psychology, like Lewis himself conceded to some extent. Claims like these may have contributed to the false impression. As we saw in §4.3, typical examples for the so-called reversibility of simplicity ranking are misguided (Carroll 1994, for example). If different scientists rank simplicity of theories differently, it means that they invoke different theory choice criteria, which happen to have the same name ‘simplicity’. This is analogous to the case where voters with the same name, say John Doe, rank alternatives differently; we would not describe their different orderings of alternatives as “John Doe’s ordering of the alternatives is reversed.” If this observation is correct, then simplicity ranking of systems is rigid. Given this rigidity of simplicity, the domain of system choice is somewhat restricted.[92]

The Hope Thesis as Restricting Domain for System Choice

[92] Of course it doesn’t follow from this that the Arrovian impossibility is blocked for system choice. It still remains to see if this kind of restriction is sufficient to block it. Until then, we still have to examine the other conditions for the Arrovian impossibility in system choice.

In addition to the rigidity of simplicity ranking of systems, the BSAers might suggest that the Hope thesis also serves as a restriction on the range of the admissible profiles in system choice. Lewis does not specify the exact range of his Hope thesis, but in the above quote he seems to exclude some 'troubling' ranking profiles by saying "If nature is kind, the best system will be robustly best-so far ahead of its rivals that it will come out first under any standards of simplicity and strength and balance." One may consider this Hope thesis as a sort of restriction on what profiles are admissible. For example, nature might kindly arrange itself in such a way that profiles like Condorcet paradox (§2.2) would not arise. Or, for another example, nature might be arranged in such a way that it can block the profiles in which systems are tied in all ways; the Humean mosaic will be kindly arranged so that there will emerge tie breakers. Or, the Humean mosaic might be very subtly arranged in such a way that some decisive information about overall goodness of systems would be revealed when system criteria are measured on cardinal scales. So the hope goes.

Whichever refinement it takes, the Hope thesis as domain restriction like above does not guarantee an escape route from the Arrovian impossibility. At best, what can be drawn from the Hope thesis is that, among many possible ways that nature can be kind to us, one way is that nature might restrict the domain of system-choice rules. After all, the Hope thesis is a very general and unarticulated 'hope' that the system-choice algorithm will generate a clear winner; it remains silent about how the system-choice algorithm should operate. Recall Lewis's primary aim in proposing the Hope thesis is to guarantee there comes a sure winner in a race for the best system (§4.3.2 and §4.3.3). In contrast, the Arrovian result for system choice, if it obtains, questions the way the algorithm operates; the result would imply that the winner is picked by a dictatorial criterion. So the Arrovian threat is on the 'balancing' process in its analysis of lawhood; the Hope thesis is just aimed at producing a winner in one way or another.

Furthermore, even if mother nature is indeed kind enough to restrict the domain, it does not mean that the Arrovian threat can be avoided. As we know from the literature on social choice (§2.4.2 and §2.4.3, also see the subsection below), even if **U** is not satisfied, a variant of the Arrovian impossibility obtains if the domain of the social-choice algorithm *rich* enough. In the context of system choice, then, the BSAers would have to add to their Hope thesis that mother nature will be extra kind so that she would sufficiently restrict the domain of system-choice rules such that the domain is very impoverished. Invoking hope after hope in this way seems like an *ad-hoc* maneuver. Also, we may have good reasons to believe that the domain of system-choice rules should be rich. In either case, the Hope thesis as domain restriction does not seem to save the BSA.

The Rich Domain Condition for System Choice

As for theory choice, Morreau (2015) argues that **U** does not apply to theory choice, while noting that the weaker condition, rich domain (**R**) may apply to theory choice. Roughly put, according to **R**, what should be unrestricted in theory choice is the *patterns* of unfixed placeholders for theories. **U** says all logically possible profiles of the same specific theories should be admitted. But simplicity, as we have just seen, does impose significant constraints on their ranking profiles because once it is set which candidate theories or systems are to be considered their simplicity rankings *could not* be different than the rankings in the actual profile. But the *pattern* itself should not be restricted; that is, it should still remain true that simplicity may rank some theory above another theory. This condition is called Rich domain (**R**) condition.

As we saw in §2.4, in the literature on social choice theory, it is now agreed that **U** is not required for the Arrovian impossibility; something strong enough is required, that is, the domain should be diverse enough (Kelly 1978; ch7, Pollak 1979; 76-7, Campbell and Kelly 2002; 64-5, for example). For example, the so-called Pollak diversity condition requires that, for any logically possible profile over three 'hypothetical' alternatives (*x*, *y*, *z*), then there exist three alternatives (*a*, *b*, *c*) such that the profile restricted to that triple coincide with the profile over the hypothetical triple. Morreau (2014a, 2015) elegantly subsumes Pollak's pioneer work and similar studies on diverse domain conditions under *Rich domain* (**R**) condition. Morreau defines a *pattern* as a list of weak orderings of some set of logical variables (not actual alternatives). A profile is said to *realize* a pattern if there is a matching between a set of variables in the pattern and a set of alternatives in the profile. Morreau's Rich domain condition is:

Rich domain: A domain is *rich* if for every suitable pattern *P* of three variables, there is some profile in this domain that realizes *P*.

In words, a domain is rich if orderings of three alternatives are showing patterns which coincide in one way or another with all possible orderings of three hypothetical, not actual, variables. According to **R**, what should be unrestricted in theory choice is the patterns of variables. This rich domain, along with the other suitably modified analogue[93] of Arrow's condition, is enough to give a rise to a variant of the Arrovian impossibility. The literature on social choice and theory choice (Parks 1976, Pollak 1979, Hammond 1976, Kemp and Ng 1976, Roberts 1980, Rubinstein 1984, Feldman and Serrano 2008; Morreau 2014a, 2015) agrees that even if **U** is replaced by **R** (or some similar conditions in early literature)**,** a variant of Arrow's impossibility theorem obtains provided the strong neutrality condition (**SN**), a stronger version of **I**, is met. That is, simply weakening **U** to **R** in theory choice does not open up an escape route from the Arrovian result.

Turning to system choice, we saw that **U** does not apply to system choice. Does the analogue of **R** apply to system choice, based on the same consideration as above? According to official statement of the BSA (Lewis 1983, 1994; see §4.2), some systems come out to be simpler than others; some more informative than others; and fit of systems may vary as well. This rich pattern, in which systems are ranked in one way or another by the three system choice criteria, seems to have to be allowed in system choice.

Morreau (2014a) gives an illustrating example of rich domain in theory choice. Given that there are generally less restrictions on choice rules in system choice than in theory choice, we do not have reason to doubt that a similar construction could be done in the case of system choice as well. Let us consider a simple example. When it comes to three *hypothetical alternatives*, or, say, *variables* (*x, y, z*), there are only four weak orderings of them: (1) $x \approx y \approx z$, (2) $x > y > z$, (3) $x > y \approx z$, and (4) $x \approx y > z$. A suitable pattern *P* would be one of them. Suppose we are dealing with deterministic

[93] It is Strong neutrality condition (**SN**). Shortly we will have discussion on it.

laws of nature, i.e., invoking the two system choice criteria of simplicity and informativeness. Now, for example, let $P = [x > y > z, z > x \approx y]$. What we would need is a case where the orderings of the candidate systems (*a*, *b*, c) satisfy the two component orderings of *P* with respect to the two system choice criteria. Say *a* is a system like 'theory of everything' which talks about all fundamental properties and their instantiations on the Humean mosaic, *b* and *c* are systemizations of facts only about biological entities while *b* presupposes fewer number of ontological basic kinds than *c* does. If we plug *a* in *z*, *b* in *x*, and *c* in *y*, then we have: $a > b \approx c$ with respect to informativeness; and $b > c > a$ with respect to simplicity. We can see examples for other choices of *P* can be generated in a similar manner. So it seems that **R** is applicable to system choice as well.

In this section, we have discussed the possibility of relaxing **U**. It turns out that the analogue of **U** for system choice is inapplicable but the weaker condition **R** still applies to system choice. This weakening alone does not open up an escape from the Arrovian impossibility. Following the plan laid out in §5.1, let us turn to the Strong Neutrality condition for system choice.

5.3 Neutrality for System Choice: Multi-Profile versus Single-Profile Approaches and the Humean Supervenience

In this section, we will ask if the strong neutrality condition applies to system choice. As we saw in §2.4, literature in social choice and theory choice (Parks 1976, Hammond 1976, Kemp and Ng 1976, Roberts 1980, Rubinstein 1984, Feldman and Serrano 2008; Morreau 2015) has shown that, even if **U** is replaced by its weaker counterpart **R,** a variant of Arrow's impossibility theorem obtains provided the strong neutrality condition (**SN**), a stronger version of **I**, is met. This leads us to the question whether **SN** applies to system choice.

Let me discuss this question in connection with the Humean Supervenience (HS) thesis we examined in §4.2. It says that truth supervenes on what there is, and what there is is just a vast mosaic of local particular matters of categorical facts and the spatiotemporal relations among them. That is, all the facts, nomic and non-nomic, supervene on the Humean mosaic. It follows from the HS thesis that there cannot be difference in laws without difference in local matters of facts.

As we saw in §4.2, there are weak and strong version of the HS thesis. Lewis (1994; 475)'s weaker HS thesis about laws says:

> For any two *worlds like ours* which agree with the spatio-temporal distribution of fundamental qualities, laws are the same.

In contrast, the strong version of the HS states that

> For *any* two worlds which agree with the spatio-temporal distribution of fundamental qualities, laws are the same.

Both defenders and critics of the Humean conception of laws generally take the strong version as the official HS thesis about laws. (Hall 2012, Beebee 2000, Roberts 2001, Earman and Roberts 2005; Armstrong 2004).

At this point, let me recapitulate the multi- and single-profile approaches in social choice discussed in §2.4.3. Arrow's impossibility theorem was originally derived in the multi-profile framework, which involves possible profiles of preference orderings of individual voters that are different from their actual orderings. However, in actual situations, there is only one profile: the actual profile of how individuals actually prefer the alternatives. So, it may seem that, in a given actual situation, social choice rule only needs to generate a social ordering for one fixed, actual profile. Based on this consideration, some claimed that Arrow's nihilistic conclusion should be rejected (Little (1952), Samuelson (1967)) because the conditions imposed by Arrow are defined in the multi-profile framework. According to these objectors of Arrow, individual preferences are given and social choice procedure only need to determine the best alternative given those individual preferences; and if individual preferences change then we just have “a new world and a new order” (Little 1952; 423-424). Requiring social choice rule to be sensitive to all logically possible profiles like Arrow did is just “an infant discipline of mathematical politics” rather than that of appropriate welfare economics, hence we should “export Arrow from economics to politics” (Samuelson 1967; 42).

Turning back to system choice, the analogue of the multi-profile framework in system choice would mean that non-actual profiles of rankings of systems would be used. It seems that, on the weak HS thesis, we only need to care about actual profiles of systems because the weak HS only concerns our actual world (and similar worlds). (And profiles in similar worlds should be same as actual profile anyway, according to the Hope thesis.) That is, on the weak HS thesis, the BSA doesn't need to care what happens in counterfactuals; what matters is how things are in actual world. So, the defenders of the BSA facing the Arrovian threat might be tempted to fall back to the weaker HS thesis. On the weak HS thesis, system choices need to be made given how systems ‘actually’ fare with respect to the actual system choice criteria. Then, adopting the weak HS, they might claim that the analogue of Arrow’s theorem does not obtain because the theorem relies on multi-profile framework and there is no need to worry about the multi-profile framework on the weak HS.

Unfortunately for these BSAers, however, even under the weak HS there is a variant of the Arrovian impossibility: the single-profile one. In social choice, in response to the complaints about the multi-profile framework like above, literature in social choice theory in the late 1970's and early 1980's showed the single-profile variants of Arrow's theorem obtains for a fixed preference profile if the profile is diverse enough and the intra-profile counterparts of Arrow's inter-profile conditions are met (Fishburn 1973, Parks 1976, Hammond 1976, Kemp and Ng 1976, Pollak 1979, Roberts 1980, and Rubinstein 1984; See Suzumura 2002 and Feldman and Serrano 2008 for historical overview). It is now agreed that there are single profile analogues of all the results given in the multi-profile framework, provided suitably constructed single-profile conditions are met (Pollak 1979; 86, Sen 1977; 1564, Rubinstein 1984; 726).

So, our imagined BSAers, who argue that the system choice rule for the BSA only need to yield the best system based on how competing systems fare with respect to actual history, against fit, simplicity, and strength, now face the single-profile variant of the Arrovian impossibility. As we saw in §2.4, the conjunction of **SN** and **R** (with the other conditions being met) leads to the Arrovian impossibility even in the single-profile framework. We just saw that **R** is motivated in system choice (§5.2), now we are led to the question whether **SN** can be dropped in system choice. The analogue of **SN** for system choice would be:[94]

Strong Neutrality (**SN**): For all w, x, y, z in the set of alternative systems X, and for all profiles R, R' in the domain of system choice rule,

If, for every system choice criterion i, [xR_iy iff zR'_iw] and [yR_ix iff wR'_iz], then [xRy iff $zR'w$] and [yRx iff $wR'z$].

There is no definitive answer to the question whether the analogue of **SN** holds in system choice, but we know what kind of questions we have to ask in order to find it out. Suppose the system choice rule, whatever it is, is established for the BSA. That is, it is settled which choice criteria to use and how to balance different criteria. Suppose the system choice rule ranks S_1 over S_2. For this particular pair-wise competition, S_1 comes to be the better system than S_2 under the choice rule (whatever it is). Now imagine S_3 and S_4 exhibit the same ranking pattern as S_1 and S_2. For example, S_1 is simpler than S_2, so is S_3 than S_4; S_2 is more informative than S_1, so is S_4 than S_3, and so on. That is, the pair S_1 and S_2 and the pair S_3 and S_4 show the same criteria-ranking pattern. In this case, should the system choice rule yield the same ordering for S_3 *vs* S_4 pairwise competition as it did for S_1 *vs* S_2? We assumed it ranks S_1 over S_2, so assuming they share the same ranking pattern, should it be the case that

[94] Morreau (2015)'s formal definition of **SN** is identical with the formal definition of Neutrality found in some social choice literature. For example, Bossert and Suzumura (2010) give the formal definition of Neutrality which is identical with Morreau's definition of **SN**. But, see, for example, d'Aspremont and Gevers (2002, pp. 493–494) for the formal definitions of Intraprofile Neutrality (IAN) and Strong Neutrality (**SN**), which clearly indicate that the latter is the stronger condition than the former. This is, roughly put, because SN requires social choice rule to be consistent over different pairs across different profiles and IAN requires consistency over different pairs within a profile. Given IAN is a special case of SN (when two profiles are equated), SN is the stronger imposition on a social choice rule than IAN.

the rule should also rank S_3 over S_4, consistently? Is this kind of consistency needed for the BSA? Our answers to questions like this will tell us whether SN is applicable to system choice for the BSA. The answer seems to be that this kind of consistency is a desirable property of the system choice procedures for the BSA.

The BSAers might think **SN** can be dropped in system choice. They might suggest that we utilize cardinal information about systems. For example, while the pair S_1 and S_2 and the pair S_3 and S_4 show the same ordinal criteria-ranking pattern, S_4 might be vastly more informative and only minimally less simple than S_3, while S_1 is vastly simpler and minimally less informative than $S_{2.}$ In that case, even if S_1 is ranked above S_2 under system choice rule, S_4 may have to be ranked above S_3; **SN** would not apply. But this line of response assumes a certain kind of inter-criterial comparability. We will discuss the inter-criterial comparability in §5.6. Unless the inter-criterial comparability is established, this answer does not work out.

Another possible answer may be given in a contextualist perspective. The answer would go as follows: In the above example, S_1 and S_4 might systematize *more interesting parts* of the world than S_2 and S_3 do, in which case S_4 should be ranked over S_3. The system for Earth biology and the system for Martian biology might be equally strong and simple, but Earth biology delivers information about what is more interesting to *us*. So Earth biology is the preferable system.

This line of response wouldn't work out because systems are supposed to work on all the pixels in the Humean base. Contextualist answers like above may make sense in the case of theory choice, where the 'context of interest' can be well-implemented. But given that the BSA is a metaphysical theory about laws which tries to remove subjective elements in the analysis of lawhood, this answer does not seem appropriate. If my diagnosis is right, then **SN** condition is motivated in system choice. We have discussed **U** and **R** (§5.2), and **SN** (§5.3). The upshot so far is that **R** and **SN** seem to apply to system choice. Let us turn to the condition of Independence of Irrelevant Alternative (**I**).

5.4 Neutrality and Independence

As we saw in §2.2, the condition **I** has the 'irrelevant' and 'ordering' aspects. Formally, **I** is logically equivalent to the conjunction of the two conditions: Independence of Irrelevant Utilities (**IIU**) and Ordinal Non-Comparability (**ONC**) (Sen 1970). Let us focus on **IIU** in this section. In social choice, **IIU** says that

Independence of Irrelevant Utilities: the social ordering over the set of a pair of alternatives depend only on individuals' utility functions restricted to that set.

IIU has been widely accepted in social choice theory (Kemp and Ng 1987, Hammond and Fleurbaey 2004). My discussion of **IIU** below will be mostly done in comparison to **SN**.

Does **IIU** apply to system choice? Before answering this question, let us review the discussion on **SN** we saw in §2.4.2, the stronger condition than **I**. **SN** is more stringent than **I**. **I** requires consistency for each pair of alternatives separately. Figuratively speaking, in social choice, **I** means that when the social welfare function aggregates individual orderings, it should take each pair of alternatives separately, paying no attention to preferences for alternatives other than the pair in question. **I** requires consistency between two profiles over a pair each time; it leaves possibility that different pairs might be treated differently. For example, when two profiles $\langle R_i \rangle$

and $\langle R'_i \rangle$ coincide on *x* and *y*, and the individuals exhibit the exactly same pattern of preference orderings on *z* and *w* as they do on *x* and *y*, what **I** requires is that the social preference ordering of two profiles should agree on the pair of *x* and *y*, and agree on the pair of *z* and *w*, separately. But **I** does not require that the social ordering of two profiles are same across the pair of *x* and *y* and the pair of *z* and *w*. As it should be clear now, it is **SN** that precisely requires such consistency across different pairs. **SN** is a fairly strong condition, and in social choice, it has the effect of forbidding social choice procedure from using non-utility information. In theory choice, the analogue of **SN** would require that theory choice procedure should only use information about how well theories fare with respect to the theory choice criteria; for example, the *identity* of theories should not enter the procedure. It seems unclear whether **SN** applies to theory choice. On the one hand, we can think of some examples where other kinds of information seem to be allowed in theory choice procedure. If two theories, for example a descent of Darwinianism and a descent of Creationism, are in competition, scientists may take into consideration information about the theoretical lineage of the two theories. Or, scientists working in different branches of science may judge theories in different contexts of interest. On the other hand, it seems to desirable that theory choice procedure is as 'neutral' and consistent as possible, for theory choice to be rational in the most common and intuitive sense of the term. So, we don't seem to have theoretical or empirical ground for outright rejection or acceptance of **SN** in theory choice.

What of the system-choice analogue of **SN**? For the BSA, systems are true, deductive systemizations of categorical facts. **SN** seems motivated in system choice. We saw some cases like above where 'irrelevant' information might be used in theory choice, but there seems no such cases for system choice – systems have no identity; there is no context of interest for systems. Now, turning to our question for this section, what of the system-choice analogue of **I**, in particular, the 'irrelevant' aspect of **I**? The system-choice analogue of **IIU** would be that

IIU for system choice: the ordering over the set of a pair of alternative systems depend only on those two systems' scores with respect simplicity, fit, and informativeness, restricted to that pair.

In words, the ordering of two systems *x* and *y* should be only determined by the scores of *x* and *y*; the score of the irrelevant system *z*, for example, should not enter. I do not have a conclusive argument for **IIU**.[95] But given the above considerations which seem to motivate the system-choice analogue of **SN**, and given that **SN** is much stringent than **I**, I suggest that the system-choice analogue of **IIU** is a minimum requirement for the system-choice rules for the BSA. Furthermore, there are conditions like **ONC** that have been extensively discussed as a possible candidate to relax in order to avoid the Arrovian impossibility. Until we find a clear case against **IIU**, I suggest, we should investigate the possibility of relaxing the other conditions first. Following the plan laid out in §5.1, let us turn to the possibility of cardinal measure of the system choice criteria invoked by the BSA.

[95] Of course, by equating x to z and y to w in the following statement of **SN**, we obtain **I**. So, if **SN** is met, its special case **I** is met. *Strong Neutrality* (**SN**): For all w, x, y, z in the set of alternative systems X, and for all profiles R, R' in the domain of system choice rule, If, for every system choice criterion i, [$x\mathrm{R}_i y$ iff $z\mathrm{R}'_i w$] and [$y\mathrm{R}_i x$ iff $w\mathrm{R}'_i z$], then [$x\mathrm{R}y$ iff $z\mathrm{R}'w$] and [$y\mathrm{R}x$ iff $w\mathrm{R}'z$].

5.5 Exploring Escape Routes from the Arrovian Threat: Cardinal Measurability

As we saw in §2.4, Arrow’s ordinalism (Arrow 1951) is that social choice procedure should only deal with information about ordinal preference over alternatives. This ordinalism approach is ‘informationally impoverished’ as it allows very little information about intensity of preference. Sen (1970, 1986) proposed the ‘information enrichment’ approach to the possibility of social choice, in which social choice procedures are allowed to use more enriched information than ordinal information about individual preference, provided intensity of preference can be cardinally measured. The cardinal measurability of theory choice criteria was discussed in §2.4.4. In this section, I will examine cardinal measurability for each of the system choice criteria of the BSA: fit, strength, and simplicity. Let us turn to the criterion of fit first.

5.5.1 Possibility of a Cardinal Measure of Fit

Let us consider cardinal measurability of fit. As we saw in §4.3.6, on Lewis’s definition (Lewis 1994), a system's fit is measured by the probability that it confers to the complete history of chance events in a given world. On this definition, fit is cardinally measured on the closed interval of real numbers from zero to one, with non-arbitrary zero point. However, there are a number of problems for using Lewis’s definition as a cardinal measure of fit of systems. In this section, I will discuss the zero-fit problem and the typicality solution to the zero-fit problem (§5.5.1.1), and the incompleteness problem and some possible solutions to it (§5.5.1.2). I will examine if these solutions provide cardinal measures of fit.

5.5.1.1 The Zero-Fit Problem and Suggested Solutions

We should be careful about Lewis's definition of fit. Systems assign zero probability to the history of events when there are infinitely many events or the outcome space of the event is infinite. On Lewis's definition of fit, if history is infinite, all the systems come out to have equal fit of zero. This is called the *zero-fit* problem (Lewis 1980, Elga 2004). For example, imagine actual history consists of a sequence of infinitely many flips of a coin:

History: HTHTHTHTHTHTHTHTHTHT…

Suppose we are comparing different systems which make different claims about what the chance of the coin's landing heads *Ch*(H) is:

S_1: *Ch*(H) = 1/2

S_2: *Ch*(H) = 1/6.

The problem with Lewis's definition of fit is that both S_1 and S_2 will assign zero probability to history. This is because there are infinitely many occurrences of the outcome H in history and multiplying infinitely a value less than 1 by itself is zero. On Lewis's definition of fit, then, all systems come out to have equal fit of zero in the case of infinite history. This is the zero-fit problem. Lewis (1980, 1994) disregarded the problem.[96] Elga (2004) proposes that fit of a system for a world should be measured by how 'typical' the system views the world. Call this proposal the typicality solution. Let us examine the proposal in detail.

[96] Lewis was aware of this problem: "…the fit between the system and a branch would be the product of these chances along that branch; and likewise, somehow, for the infinite case. (Never mind the details if, as I think, the plan won't work anyway.)" (Lewis 1980 postscript, p. 128) He simply dismissed the problem of infinite case because he was engaged with *reductio* here. But he uses the same definition for his official characterization of the BSA (1994), where he does not mention the problem of infinite case. See Bostrom (1999) for critical discussion of it.

First let us examine what 'typicality' means in the typicality solution. For example, a system which claims *Ch*(H)=1/2 will view sequences like … HTHTHTHT … as more typical than sequences like … HHHHTHHHHT…, while another system which claims *Ch*(H)=5/6 will regard the latter sequence more typical. The more typical a system views a world, the greater fit the system has to the world. If the fit of systems are measured with respect to one proposition true only at a world of infinite history, then there arises the zero-fit problem. Instead of one proposition, Elga's suggestion goes, we[97] should choose a set of 'test' propositions which pick out some important features of history and compare the probabilities assigned to those test propositions by the systems in comparison. The test propositions should be in simple language. In the present example, H_i: the *i*th toss landed heads is a good candidate.

Elga's formal definition of fit is as follows.

> **Fit$_{\textbf{Typicality}}$**: System X fits better than system Y iff the chances X assigns to the test propositions are predominantly greater than the corresponding chances that Y assigns. (Perhaps X assigns a higher chance than Y to every test proposition. Or perhaps X assigns higher chances than Y overall.) (Elga 2004; 72)

As it stands, "X fits better than Y" expresses an ordinal fit-ranking of X and Y. We are exploring the possibility of a cardinal measure of fit, in particular, on an interval scale.[98] As we saw in §2.4.1, measurement on an interval scale carries meaningful

[97] To avoid the risk of sounding too anthropomorphic, "we" may be replaced with the "BSA Oracle" introduced in §4.3.4.

[98] This doesn't have to be an interval scale. If it were shown that fit can be measured on a ratio scale, it would be even better for the BSAers. A ratio scale with a meaningful zero point, carries information that an interval scale carries and also more fine-grained information about ratios of

information about the degree of differences between the measured items. If the fit of systems were to be measured on an interval scale, it should be possible to make meaningful comparisons of intervals between fit of systems, for example:

$$\mathrm{Fit}(X) - \mathrm{Fit}(Y) > \mathrm{Fit}(Z) - \mathrm{Fit}(W)$$

$$\mathrm{Fit}(X) - \mathrm{Fit}(Y) < \mathrm{Fit}(Z) - \mathrm{Fit}(W)$$

$$\mathrm{Fit}(X) - \mathrm{Fit}(Y) = \mathrm{Fit}(Z) - \mathrm{Fit}(W)$$

$\mathrm{Fit}(X) - \mathrm{Fit}(Y) = 2 \times [\mathrm{Fit}(Z) - \mathrm{Fit}(W)]$, and so on.

The operative terms in Elga's definition, "predominantly" and "overall", are left imprecise. Let us examine a number of possible precisifications of the typicality definition as a cardinal measure of fit.[99]

Let H_i be the base test proposition from which compound test propositions can be constructed, e.g., "H_1 or H_2", "H_1 and H_3", "H_8 and not-H_9" and so on. If there is one test proposition of finite length to consider, Elga's typicality solution works well as a cardinal measure of fit. For example, with respect to the test proposition T: H_1 & H_2

the measured quantities of items. But the 'meaningful zero point' is not tenable for the criterion of fit in system choice. This is because, for the BSA, fit is a measure of accuracy of probabilistic systems and different probabilistic systems are all compatible with the same history. The only way a system can have zero fit is when history is infinite. However, as we will see in this section, the case of infinite history cannot be satisfactorily solved for the BSA.

[99] There is rather a general problem associated with the use of the notion of simplicity in Elga's definition. Elga suggests that test propositions should be true and simple and that simple sentences in the form "$\exists\forall\phi$" should be used to describe typicality (Elga 2004; 71). This is a syntactic conception of simplicity, and he does not explain on what basis the form counts as simple or why it should be used. (Williams 2008; n.27). As we saw in 3.3.4, a syntactic notion of simplicity, if it were to be a useful comparison criterion, has to invoke one or another kind of 'privileged' language. It deserves a separate discussion whether there can be principled ways of selecting such reference language for the criterion of fit but it is worth mentioning that even in the very first step for the task of measuring fit certain choices have to be made and we need some criteria for making such choices.

& H_3 & H_4 & not-H_5 (meaning the sequence "HHHHT"), suppose we consider the following systems:

S$_1$: *Ch*(H) = 5/6

S$_2$: *Ch*(H) = 1/2

S$_3$: *Ch*(H) = 1/6

S$_4$: *Ch*(H) = 4/5.

The fits of these systems come out to be:

$\text{Fit}(S_1) = Pr(T|S_1) = (5/6)^4(1/6) = 0.08037$

$\text{Fit}(S_2) = Pr(T|S_2) = (1/2)^5 = 0.03125$

$\text{Fit}(S_3) = Pr(T|S_3) = (1/6)^4(5/6) = 0.00064$

$\text{Fit}(S_4) = Pr(T|S_4) = (4/5)^4(1/5) = 0.08192.$

This result conforms to our intuition about ordinal ranking of the fit of systems with respect to the given test proposition. Furthermore, this notion of fit works well as a cardinal measure in the finite cases like this example. Intuitively speaking, S$_4$ fits slightly better than S$_1$ fits T, S$_2$ fits quite better than S$_3$, and S$_4$ fits greatly better than S$_3$, all of which make comparisons of fit intervals between systems, and the above result conforms to our intuition about these comparisons. So far it seems to work well as a cardinal measure.[100]

Elga's typicality solution involves the use of more than one test propositions. We need a way of measuring fit of systems with respect to a set of propositions. An

[100] Lewis's original definition of fit works in the exactly same way, if actual history is finite and only one history (i.e., actual) is to be considered. It deserves an independent discussion which assumption is more plausible as to whether actual history is finite or infinite, but here I just point out that if history is finite Lewis's original definition works as a cardinal measure of fit. If it is infinite, then we need a solution to the zero-fit problem which can also provide a cardinal measure of fit, which is the target of the current investigation.

analogy would help with the present investigation. In computer science, there are three common ways of measuring performance of a given algorithm with respect to more than one cases: measure by its best-case, worst-case, or average-case performance. Analogously, we may measure fit of a given system with respect to a set of test propositions either by the highest probability, or by the lowest probability, or by the average probability it assigns to the test propositions.

Let $\boldsymbol{T}$ be the set of test propositions and T_i be an element of $\boldsymbol{T}$. Then the fit measure of a system based on the best-case, worst-case, and average-case would be:

$$\mathbf{Fit}_{\mathbf{Best}}(S) = \Pr(T_k|S) : \{k \mid \Pr(T_k|S) \geq \Pr(T_i|S) \text{ for all } T_i \in \boldsymbol{T}\}$$

$$\mathbf{Fit}_{\mathbf{Worst}}(S) = \Pr(T_k|S) : \{k \mid \Pr(T_i|S) \geq \Pr(T_k|S) \text{ for all } T_i \in \boldsymbol{T}\}$$

$$\mathbf{Fit}_{\mathbf{Average}}(S) = \Sigma\Pr(T_i|S)/n \ (n = |\boldsymbol{T}|)$$

Each of these notions of fit might seem to work as a cardinal measure of fit. But there are number of problems for them to be used as cardinal measures of fit. Since there are infinitely many test propositions,[101] all systems will have same $\text{Fit}_{\text{Average}}$ because the average of probabilities assigned to infinitely many test propositions is zero. There are problems for Fit_{Best} and for $\text{Fit}_{\text{Worst}}$. In the infinite case, all systems will come out to have same Fit_{Best} of 1. For example, in the infinite case, the system S_1: $Ch(H)=1/2$ assigns probability 1 to the test proposition "there are as many H as T in

[101] Elga puts no constraint on the number of test propositions in $\boldsymbol{T}$. The basic idea of the typicality solution is to measure fit not by the probability of a particular outcome but by the probabilities of some suitable features ('typical' patterns, for example) represented by certain outcomes. Elga doesn't specify what counts as suitable features (he gives an example: if the coin is fair, the suitable properties we would expect to typically see are 'there are as many heads as tails, in the long run', 'the relative frequency of H is ½ in the limit', 'the pattern HTH will appear as often as THT in the limit', and so on). Assuming history is infinite, since the boundary between what's typical and not typical is vague, there is no principled way to pinpoint the cutoff point for what and how many propositions are to be included in $\boldsymbol{T}$.

the limit" and S_2: *Ch*(H)= 2/3 assigns probability 1 to the test proposition "there are twice many H as T in the limit". Furthermore, in the limit, both systems can assign the same probability 1 to the first test proposition (and to the second). There is no metaphysically privileged way to allow the first test proposition in ***T*** but not the second, or vice versa. This is because there is no metaphysically privileged way to calculate the limit of sequence.[102] So both S_1 and S_2 come out to have the same Fit_{Best}. A similar problem arises for $\text{Fit}_{\text{Worst}}$. In the infinite case, S_1 assigns probability 0 to the test proposition "all heads", so does S_2. As before there is no principled way to exclude that particular test proposition from ***T***. This again is because there is no metaphysically privileged way to calculate the limit of sequence. Both S_1 and S_2 come out to have the same $\text{Fit}_{\text{Worst}}$. In short, the Fit_{Best} is 'too good' and the $\text{Fit}_{\text{Worst}}$ is 'too bad'. These problems render the proposed notions of fit useless as cardinal measures of fit.

There may be other possible precisifications that might work. For example, one might define fit of S as the sum of the probabilities S assigns to the test propositions in ***T***. Or, one might define it as the average probability S assigns to the 'most' test propositions in ***T***. But their prospects don't look good. For the first precisification, the sum of their probabilities will diverge to infinity since there are infinitely many test propositions. For the second precisification, it is indeterminate how to set the cut off number or percentage for what counts "most", since there are infinitely many test propositions. The upshot is that, since there is no metaphysically privileged way to

[102] See Lyon (ms.) for a number of examples on this point.

address the problems associated with the infinite number of test propositions, the prospect for cardinal measure of fit doesn't look good.[103]

So far we have examined the zero-fit problem and the typicality solution along with its variants. Unfortunately, none of them seems to provide a satisfactory cardinal measure of fit.

5.5.1.2 The Incompleteness Problem and Suggested Solutions

Suppose we may set aside the zero-fit problem as Lewis (1980) did. There is another concern about the possibility of cardinal measurement of fit. In the context of theory choice, scientific theories are often incomplete, for example, accounts of crucial parts of the theories are missing, basic concepts are vague, and so on. When theories are indeterminate and imprecise like this, measuring fit of such theories may be impossible because there can be many hypothetical ways of completing and precisifying the theories in question. For example, Morreau (2014a) expresses skepticism of cardinal measurability of fit as "without resolving such indeterminacy … there can be no saying exactly how well a theory fits available data."

[103] Probably the most promising line of thought would be to confine ourselves to finite number of test propositions of finite length of sequences. For example, randomly choose a certain number of test propositions from T, truncate them to finite length of sequences, and measure the average (the best, the worst, or what have you) of the probabilities a system S assigns to the chosen test propositions. This suggestion can be viewed as a sort of statistical hypothesis testing; in the proposal under consideration, we test different systems in terms of how typical they view a given finite set of truths (hence the name 'test' proposition), just as we test different hypotheses by how likely they view the observed data set. I suspect that this is the best the BSAers can do to measure fit cardinally in the case of infinite history. But this suggestion won't come without problems. First, just as there is a risk of overfitting when we choose a hypothesis based only on how well it fits a given data (§3.1.3), the current proposal will face an overfitting problem. Secondly, it is indeterminate where the truncation in question should take place. Thirdly, related to the second problem, two intuitively very differently fitting systems can come out to have the same fit if the two systems agree on the remaining portion of history after truncation and vastly disagree on the discarded portion of history by truncation.

Similarly for system choice, one might worry that the only meaningful information about fit of incomplete systems may be just information about ordinal fit-ranking of systems. Even if cardinal measure of fit were possible for each hypothetical completion of a given system, the worry goes, it is indeterminate exactly how well the system in question fits history of a world given there are many hypothetical completions of such systems. From this, one might conclude, all admissible cardinal information is simply reduced to ordinal information because a system can only be meaningfully said to fit better than its rivals when the former possesses higher degrees of fit than the latter in all possible hypothetical completions. Call this problem the incomplete problem.

The BSAers may propose some ways to cardinally measure fit of incomplete systems. For example, we may make comparison of fit *profiles*, which contains information about the degree of fit with respect to each possible way of completing the incompleteness. This may be done in a way similar to our earlier attempt in the previous section to use the typicality fit as a cardinal measure. The attempt involved measuring fit of systems with respect to multiple cases. Likewise in the current context, we may attempt to measure it with respect to multiple ways of completing the incomplete. For example, we may measure fit of a given incomplete system with respect to its *best* possible completion; or with respect to its *worst* possible completion; or measure *the average* fit over all possible ways of completing the system in the question. Accordingly, there may be three possible definitions of fit of incomplete systems: $\text{Fit}_{\text{Best-Completion}}$, $\text{Fit}_{\text{Worst-Completion}}$, and $\text{Fit}_{\text{Average-Completion}}$. Let C_i be a possible way of completing an incomplete system S; C^S be the set of all possible ways of the incomplete system S. Let H be actual history and Pr(H|S; C_i) be the probability assigned to H given the system S, competed in the way of C_i. The three suggested definitions of fit would be:

Fit(S)$_{\textbf{Best-Completion}}$= Pr(H|S; C_k) : {*k* | Pr(H|S; C_k) ≥ Pr (H|S; C_i) for all $C_i \in C^S$}

Fit(S)$_{\textbf{Worst-Completion}}$= Pr(H|S; C_k) : {*k* | Pr(H|S; C_i) ≥ Pr (H|S; C_k) for all $C_i \in C^S$}

Fit(S)$_{\textbf{Average-Completion}}$= Σ_iPr(H|S; C_i)Pr(C_i|S)

There are some problems for these notions of fit. First, Fit$_{\text{Best-Completion}}$ may be 'too good'. Suppose history consists of some chancy events of a particular kind, say, atomic decay of Un^{346} atoms. Suppose systems S_1 and S_2 are indeterminate about this kind of event. Among many possible ways to complete each system in question, the way which yields the best fit is to make the system simply list every instance of this kind of events at every time it occurs. Completed that way, both S_1 and S_2 will come out to have the same Fit$_{\text{Best-Completion}}$ of 1. But notice that this move would have a huge cost of simplicity, and most likely such a hugely complex system won't be a good candidate for the best system. So, Fit$_{\text{Best-Completion}}$ seems to be next to useless as a system-choice criterion.

Secondly, Fit$_{\text{Worst-Completion}}$ may be 'too bad'. In the above example, the worst completion would be to make the system assign zero chance to the event at time *t*, like S:$Ch_t(Un^{346})=0$, for every time point *t*. The probability it assigns to history would be zero if there is just one instance of the decay event of U^{346} in history. For example, let T mean it decays, F be it doesn't decay, let the time-series sequence in question be FFFTF, then Pr(FFFTF|S) = $(1.0)^4(0.0)^1 = 0$. Systems in competition will come out to have the same Fit$_{\text{Worst-Completion}}$ of 0.

What of $\text{Fit}_{\text{Average-Completion}}$? First, note the term $\Pr(C_i|S)$ in the definition. It represents the probability of C_i being the 'true' completion given system S. The problem is how to determine its distribution over all C_is. If S were a theory, and if the current matters were in the context of theory choice, $\Pr(C_i|S)$ could be determined by some prior assumptions about S. It could be determined by some historical background about S. Or, observing some portion of data will let us know about the best estimate value of C_i and how the estimate values would be distributed.[104] While these considerations for determining $\Pr(C_i|S)$ might be available if this were the business of theory choice, no such prior assumptions are available to system choice for the BSA. Secondly, given there are infinitely many (or astronomically many) possible ways of completing the incomplete systems, $\text{Fit}_{\text{Average-Completion}}$ of most systems will come out to be indiscernibly very low, rendering it next to useless as a discriminating criterion for system choice.

For these reasons, the prospects for the proposed notions of fit don't look good. There may be other possible notions that might work. For example, one might attempt to measure fit of S by the sum of the probabilities in each possible completion S assigns to history. Or, one might define it as the average probability S assigns to 'most' of the hypothetical completions of the incomplete. But their prospects don't look good either; we saw in §5.5.1.1 the same kind of attempts have failed. For the first attempt, the sum of their probabilities will diverge to infinity since there are infinitely many ways of completing the incomplete. For the second attempt, it is indeterminate how to set the cut off number or percentage for what counts 'most' since there are infinitely

[104] Analogous to Maximum Likelihood Estimate in model selection. See §3.1.3.

many completions. The upshot is that, since there is no metaphysically privileged way to address the problems associated with the infinite number of completing the incomplete, the prospect for cardinal measure of fit doesn't look good. We saw the attempts in §5.5.1.1 didn't work for the same reason.

So far we have discussed the possibility of cardinal measurement of fit. Lewis's original definition of fit suffers from the zero fit problem and the incompleteness problem. In §5.5.1.1 we have investigated whether solutions to the zero fit problem can provide a cardinal measure of fit. In §5.5.1.2 we have investigated whether solutions to the incomplete problem can provide a cardinal measure of fit. The upshot is that their prospects do not look good.

Now let us move on the other system choice criteria. Following the plan I have laid out in §5.1, let us investigate the possibility of cardinally measuring strength of systems for the BSA.

5.5.2 Possibility of a Cardinal Measure of Strength

The second system choice criterion invoked by the BSA is strength. In this section, let us examine the possibility of cardinally measuring strength. As we saw in §4.3.5, Lewis (1983, 1994) characterizes strength as informativeness. Lewis does not precisify what informativeness is, but on his description a system is more informative about a world when it says more about the facts in the world –e.g., about what will happen or what the chance of a certain kind of event occurring will be (Lewis 1994; 480). The question in this section is how to measure 'how much' information a system carries about a world. Informativeness is described by the commentators on the BSA as a matter of excluding possibilities or possible worlds. In general, a system is said to be more informative if it rule out more possible ways (possible worlds) than others do. That is, the more possibilities a system excludes, the greater its strength (Earman 1984, Loewer 2004, 2007, Callender and Cohen 2009, Woodward 2014, Hall (Forthcoming)). Some BSAers add more specification on this general notion of informativeness. For example, Earman (1984) suggests that strength should be measured not by sheer information about the facts *per se* but by information about the facts and regularities which can be explained by dynamic laws in conjunction with appropriate boundary conditions. Some suggest that a system should be considered stronger if it allows a wider range of initial conditions and a narrower range of candidate dynamic laws (Hall (forthcoming) and Woodward 2014, for example). What these different conceptions of strength have in common is that it involves the business of 'counting' possible worlds in measuring strength.[105]

One might argue that we can come up with a way of measuring a system's informativeness on a cardinal scale. For example, we may define the informativeness of a system by how many possible worlds are ruled out by that system (Hall (forthcoming); 12)[106]. The more worlds excluded by a system, the more informative that system is. The maximally informative system, for example, would be one that perfectly describes each and every pixel of the Humean mosaic of actual world in maximum detail, to the extent that all of the propositions describing the pixels are

[105] In §4.3.5, we have examined problems for the general notion of strength. Here we are only concerned with cardinal measurability of strength.

[106] Hall discusses this notion of strength in *reductio*; later he proposes an alternative notion of strength. But it also involves 'counting' in it.

true only in the actual world. The (reverse) measure of informativeness, then, might be measured on a cardinal scale with a meaningful zero point.

But there are problems with this conception of strength as a cardinal measure. Counting the number of excluded possibilities or possible worlds is impossible because there are infinitely many of them. As we saw in §4.3.5, ordinal comparisons of fit of systems may be possible when there are set inclusion relations between the systems; but no such case is available for cardinal measurability.

When systems talk about chance of events, measuring strength becomes more complicated. First, history of chancy event is compatible with different systems claiming different chances about the event in question. (Hall (forthcoming);12, Schwarz 2014;.6) How to measure strength in probabilistic case? When S_1 says *Ch*(H)=0.5 and S_2 says *Ch*(H)=0.9, which system can be said to be stronger? Further, how to measure their 'degree' of strength? Which system says more about the facts concerning the chance of the event H? (Note that the current investigation is not about fit. If the actual frequency of the event H turns out to be roughly 50-50, then of course S_1 fits better. But this doesn't necessarily mean that it says 'more' about the event of the event H.)

Probably the most promising approach to the case of chancy events and probabilistic systems would be to connect strength to *entropy*. For example, consider the following systems each of which, at time *t*, makes a claim about the chance of the next coin flip outcome being H:

S1: Ch(H) = 0.5

S2: Ch(H) = 0.501

S3: Ch(H) = 0.9

S4: Ch(H) = 0.9999999 ......

S5: Ch(H) = 1

We may say these systems are listed in order of decreasing entropy. S1 is very little informative in the sense that it says virtually nothing about the next outcome – S1 says probabilities of the outcomes are uniformly distributed, so it gives no meaningful information about the outcome of the next coin flip. S2 is slightly more informative than S1 in that at least S2 gives us more information than just assuming the next outcome will be completely random. S3 seems to carry more information about what the next coin flip will be. It gives us more certainty than S1 and S2 in the outcome of the next coin flip being H. S5 is maximally informative because it removes any uncertainty about the next coin flip outcome. S4 is equally informative given it converges to 1.

This approach seems on the right track. But note that this conception of strength as it stands only gives us an ordinal scale of strength; it is not clear yet if the strength interval between S1 and S3 are twice as the strength interval between S3 and S5. Also, although S1 doesn't carry much information about the next coin toss, it still says something about the coin (e.g., the coin is fair). Then, on Lewis' definition, S1 possesses some degree of strength. It doesn't seem that strength as reverse of entropy has a meaningful zero point.

We have examined cardinal measurability of strength. The entropy notion of strength seems to be on the right track when it comes to probabilistic systems. When it comes to deterministic systems, as long as the suggested measure of strength invokes counting infinity, the prospect doesn't look good.

5.5.3 Possibility of Cardinal Measure of Simplicity

The third system choice criterion invoked by the BSA is simplicity. Whether simplicity is cardinally measurable is a difficult case to account for. But we have some clues of what the answer might be. Lewis (1994) states that a linear function is simpler than a quartic or a step function and a shorter alternation of quantifiers is simpler than a longer one.

One might think this conception of simplicity is cardinally measurable. For example, suppose we define simplicity of a system as $1/n$, where n is the number of parameters in it (for now let us set aside the case of zero number of parameters), provided the system in question is a parametric model.[107] Then the model LIN is simpler than Poly-2, and Poly-3 is simpler than Poly-5. Surely this conception of simplicity can be measured an ordinal scale; but it is unclear if we can meaningfully say the simplicity interval between Poly-3 and Poly-5 is twice the simplicity interval between LIN and Poly-2.

Whichever is the case, discussing simplicity only in the domain of parametric models is next to useless for the BSA. In standard characterizations of the BSA (as we saw §4.3.7), simplicity of a system concerns not only the simplicity of a single proposition

[107] Defining simplicity in this way faces many problems to begin with, the language dependence problem and the subjectivity problem, to name a few. The language dependence problems arise for simplicity measure of mathematical curves as well, as simplicity of the curves in question can come out differently depending on the choice of coordinate system. In §4.3.7, we discussed the problem of language dependence for syntactic conception simplicity (Goodman 1983, Priest 1976); and the subjectivity-relativity problem for simplicity (Carroll 1984, Craig and Callender 2009, Halpin 2003) Our current focus in this chapter is just on whether simplicity is cardinally measurable.

or model, but also the number of axioms or theorems in the system, the length of axioms or theorems, and so on. For example, consider system S_1 says there are four elementary forces in our world. Each of these forces may be represented as an axiom in the system. Let us say each of such axioms may be defined as a parametric model, and the number of parameters in each axiom is just one or two. In contrast, S_2 says there is only one elementary force in our world, represented as an axiom, which is defined as a very complex parametric model, say, with 10 parameters. Suppose S_1 and S_2 explain the facts in our world equally well. S_1 has four axioms and each axiom is very simple. S_2 has one axiom which is rather complex. In the standard statement of the BSA, both the number of axiom and the number of parameters are mentioned in its characterization of simplicity. So, even if each sense of simplicity were cardinally measurable, it seems indeterminate how to cardinally measure simplicity of S_1 and S_2 overall, let alone ordinal comparison of simplicity of them. The prospect for cardinal measurability of simplicity doesn't look good.

In this section we have examined a number of attempts to measure system choice criteria on cardinal scales. Their prospects do not look good. To begin with, the key concepts in the BSA are vague. The BSAers tend to regard this as 'practical' problem (§4.3.1). They seem to believe that those key concepts will become clearer as science improves and that there will be sufficient level of comparability so that a metric for balance can be prepared (see, for example, Hall (forthcoming); *cf.* Woodward 2014). Furthermore, even if the system choice criteria were somehow found to be cardinally measurable, as we saw in §2.4.4, cardinal measurability alone cannot open up an escape route. Especially regarding the question of inter-criterial comparability, we saw that a well-specified trade-off rule between standards is needed. Let us turn to that possibility. For theory choice, some statistical model selection criteria (AIC, for example) seem like good candidates, at least in a limited domain, for the inter-criterial comparability between simplicity and fit. Following the plan, let us now turn to the question if the same is applicable to system choice for the BSA.

5.6 Searching for Inter-Criterial Comparability: the A-BSA

In this section, I will propose a variant of the BSA for which there seems to exist some form of inter-criterial comparability. In §5.6.1, I will propose a variant of BSA, in particular, implemented with Akaike Information Criterion (AIC), as a possible form of inter-criteria comparability in system choice for the BSA. In §5.6.2, I will discuss some problems with the implementation, focusing on the plausibility of normality assumption in system choice. I will raise a problem with normality assumption in system choice: the circularity problem. In §5.6.3, I will diagnose the source of the circularity problem as being due to the failure of the BSAers to properly recognize the context gap between metaphysics and epistemology. In §5.6.4, I will provide a counterexample in which the gap causes the A-BSA to fail to yield the best system appropriately. Lastly, in §5.6.5, I will attempt to generalize my counterexample to the BSA overall. The conclusion of this section would be that, to the extent the BSA import the inter-criterial balance method from statistical methods such as AIC or BIC, the inter-criterial comparability in system choice will not obtain.

5.6.1 The BSA Implemented with Akaike Information Criteria

Scientists often avail themselves of certain metrics for trading off standards. For example, consider the model selection methods we have examined in Chapter 3 – AIC, in particular. We may be able to supplement the BSA with AIC, that is, let the BSA Oracle (§4.3.4) trade off simplicity and fit using the exchange ratio as it is expressed in AIC (or, if we were to use BIC, have her use that ratio in BIC). Let us call this variant of the BSA implemented with AIC the *A-BSA*.[108] The BSAers might hope that such a comparability expressed in AIC applies to system choice. If the A-BSA works, the hope goes, we might make a case for Cardinal Unit Comparability (**CUC**). This, if it works, will surely open up an escape route from the Arrovian threat to the BSA.

There may be some problems with the implementation. AIC as a statistical tool has limited applicability, and some of these limitations may be relevant to the project of constructing the A-BSA. The assumptions and limitations of AIC were discussed in Chapter 3. This section will focus on the problems of implementing AIC to the BSA. Before we move on to the next subsection, let me briefly discuss some obvious, rather mundane problems with the A-BSA. AIC is a model selection criterion and model selection is a component of model-based reasoning. It is an epistemic project aimed at acquiring knowledge about the world through learning about models that are designed to approximate the target.[109] Due to the epistemic limitations inherent in it, model-based reasoning needs to be constructed upon a series of decisions of, for example, the frame of discernment (Giere 1979), model space, design of experiment, and so on.

[108] And call the BSA implemented with BIC the *B-BSA*. Mostly I will discuss the A-BSA, but it will be shown that the same conclusion is drawn from both cases.

[109] Friggman and Hartman (2012).

But the BSA is a metaphysical analysis of lawhood, whose domain is the entire Humean base. So there seems to be a gap.

Furthermore, systems in the BSA are supposed to be summaries regarding all pixels of the Humean mosaic, and are hence not limited to a particular portion of the mosaic or a particular kind of pixel. AIC is considerably limited in this respect. First, it is well-known that all optimization algorithms will fare equally well if their performances are averaged over all problem space. That is, if one method fares better in some types of problem sets, then it will likely perform less well in other types. The same goes for AIC. Forster (2001) argues that AIC outperforms in one problem set and underperforms in another. In some cases, AIC is not applicable at all. For example, exponential models or sin-wave models are commonly used in scientific practice, but AIC is not suitable for estimating the predictive accuracy of these models. Furthermore, AIC cannot deal with the problem of extrapolation. Forster (2000) also describes AIC as addressing the problem of interpolation, while it is not appropriate as a model selection criterion for extrapolation. All these seem to suggest that, while the A-BSA might be appropriate as an analysis of laws in special science, it doesn't seem an adequate analysis of fundamental laws,[110] because AIC as an epistemic method bears certain limitations.

But, as we saw in §4.3.4, the accusation of the BSA's allegedly unjustified use of epistemic standards is probably committing the question-begging against the BSA. Given the BSA's main advertisement being that the BSA Oracle is doing idealized version of science, to the extent that scientists do somehow choose theories by

[110] Some BSAers bite the bullet by taking this move. See Callender and Cohen 2009 for example.

balancing conflicting virtues, it might be said a system choice can be made in similar way by implementing some practice in science. This itself should be harmless, the BSAers would claim. For now let us grant that this is true. A more serious challenge awaits the A-BSA, however.

5.6.2 Normality Assumption and the Circularity Problem

As we saw in §3.2.4, the derivation of AIC assumes that the distance between the point representing the set of true values of parameters and the point representing hypothesized parameters on the parameter space will be normally distributed due to the central limit theorem (CLT) (Akaike 1973; 273, Akaike 1974; 718, Kieseppä 1997). However, the use of CLT cannot be made in isolation; it should accompany assumptions about the target phenomena it is being used to explain.[111] This seems to mean that the A-BSA would require similar assumptions. But we have no reasons to believe that the Humean pixels (which is the domain of the A-BSA) will be arranged in a way that is suitable to the use of CLT. The BSAers cannot make an *a priori* assumption about how the Humean pixels are arranged because an adequate balance metric for the BSA is one which is applicable to the categorical facts, without being loaded with any *a priori* metaphysical constraints.

It might seem empirically true that we often observe phenomena of normal distribution in nature.[112] Maybe the BSAers want to draw from this empirical observation that therefore the entire Humean pixels are normally distributed.[113]

[111]See, for example, Lyon (2014).

[112] As a matter of fact, it is not true that we observe normal distribution often. See Lyon (2014) for the argument that what often appears as normal distribution is in fact log-normal distribution.

[113] One obvious, rather general problem with this is that, for all we know, what we have observed, which probably is a tiny portion of the entire spatio-temporal arrangement of the pixels, may not

Maybe they want to point to some underlying cosmic principle, for example the 2nd law of thermodynamics,[114] as what is responsible for the observed normal distributions.

But this doesn't seem to be a viable option for the BSAers because such principles or laws should come *as a result of their analysis*, not the other way around. Suppose we ran the best system analysis using the A-BSA, and as a result we obtained something like the 2nd law. But at the same time the BSA should rely on this law to explain the adequacy of the implementation of the AIC. There seems to be a problem of circularity here. In fact this kind of the circularity problem has been voiced for a while. Let us examine the circularity problem and some proposed solutions in recent literature. For the sake of simplicity, we are going to simply suppose the A-BSA is the correct analysis of lawhood in the Humean picture; AIC is what scientists actually use in general; and the use of AIC has been very successful in our practice of inductive inference. Suppose furthermore the Humean mosaic is in fact arranged as normality assumption says it is.[115]

5.6.2.1 The Circularity Objection

As we saw in §4.2, according to the Humean conception of laws, laws supervene on the totality of the Humean mosaic and there is nothing metaphysically above and beyond it. Laws are regularities that appear as axioms or theorems of the best systems. So, that *P* is a law is determined by, hence explained by (almost) all the

be a good representative of the unobserved. This is a problem for everyone, including not only Humeans but also primitivists and eliminativists about laws. In this dissertation I will focus on the problems specific to plausibility of the BSA and its variants like one under discussion.

[114] Johnson 2004 connects CLT and the 2nd law, for example.

[115] If there is a better candidate for the title of the actual and inductively successful method than AIC, then we may substitute if for AIC in the following discussion; it won't affect my argument.

pixels of the Humean mosaic. But at the same time laws explain or help explain instances of regularities. For example, suppose "*F=ma*" is a law on the BSA (imagine we are in a Newtonian world); that is, the BSA Oracle collected all the instances and summarized them under the regularity "*F=ma"* in her best system. Now, we would appeal to that law to explain why a free-falling bowling ball of a mass of *m* hits the ground with a certain amount of force, along with some other facts e.g., absence of interferences, measurement methods, sometimes the initial conditions of the universe, and so on. Then there appears to be an obvious circularity. The Humean mosaic (partly) explains laws and laws (partly) explain the Humean mosaic; hence the mosaic (partly) explains itself. This is absurd.

This objection, the *circularity objection*, to the BSA has been voiced for a while. Maudlin says:

> *If the laws are nothing but generic features of the Humean Mosaic, then there is a sense in which one cannot appeal to those very laws to explain the particular features of the Mosaic itself: the laws are what they are in virtue of the Mosaic rather than vice versa. (Maudlin 2007; 172)*

Lange also makes the same objection:

> *If the Humean mosaic is responsible for making certain facts qualify as laws, then the facts about what the laws are cannot be responsible for features of the mosaic. (Lange 2013; 256)*

The circularity objection may be formulized as follows[116]:

(P1) Laws are generalizations. (HUMEANISM)

(P2) The truth of generalizations is (partially) explained by their instances. (GENERALIZATION)

[116] I draw on Hicks and Elswyk (2014)'s formulization with slight modifications.

(P3) Laws explain their instances. (LAWS)

(P4) If A (partially) explains B and B (partially) explains C, then A (partially) explains C. (TRANSITIVITY)

(C1) The natural laws are (partially) explained by their instances. (P1 & P2)

(C2) The instances of laws explain themselves. (P3, P4, & C1)

If the argument is sound, then clearly the Humean laws entails the absurd consequences like (C2). Now let us examine Loewer's solution and Lange's rejoinder on this problem.

5.6.2.2 Loewer versus Lange: Metaphysical and Scientific Explanation

Loewer (2012) attempts to defend the BSA from the circularity objection. Loewer contends that the objection fails to make a distinction between two different kinds of explanation:

> *I claim that this objection rests on failing to distinguish metaphysical explanation from scientific explanation. On Lewis' account the Humean mosaic metaphysically determines the ... laws. It metaphysically explains (or is part of the explanation together with the characterization of a Best Theory) why specific propositions are laws. This metaphysical explanation doesn't preclude ... laws playing the usual role of laws in scientific explanations. (Loewer 2012; 131)*

Loewer's solution here is relying on the distinction between two different kinds of explanation: In the picture of the BSA, the explanation in (P2) is *metaphysical* explanation, while the explanation in (P3) is *scientific* explanation. Therefore, on Loewer's solution, the circularity objection argument commits the fallacy of equivocation. So the objection fails, Loewer concludes.

In response, Lange (2013) argues that Loewer's solution fails to save the BSA from the problem. He proposes a refined version of (TRANSITIVITY). Lange argues that his refined principle enables the circularity objection to run even if Loewer's suggested distinction is taken. Let *explain*$_M$ be metaphysical explanation and *explain*$_S$ be scientific explanation. Lange's refined transitivity principle is:

> (TRANSITIVITY*) If A (partially) explains$_M$ B and B (partially) explains$_S$ C, then A (partially) explains$_S$ C.

If (P4) in the above argument is replaced with Lange's (TRANSITIVITY*), then even with Loewer's distinction between *explain*$_M$ and *explain*$_S$, the circularity occurs. In short, Lange's claim is that the Humean conception of laws still suffers from the circularity problem because scientific explanations are transmitted across metaphysical explanations.

5.6.2.3 The Hope Thesis Again: Too Much onto The Hope

Recently, Hicks and van Elswyk (2015) argued against Lange that (TRANSITIVITY*) is false. Drawing on Bennett (2011)'s notion of 'building relations', Hicks and van Elswyk argue that there is significant difference between scientific and metaphysical explanation: the two have different 'backing relations'. Metaphysical explanations are backed by non-causal relations. Starting with small parts and properties, more parts and properties are built upon. Different building materials may be used; different relationship might hold between them; in this way, the backing relations in metaphysical explanations come in wide variety. In contrast, scientific explanations are backed by very limited back relations. Typically, causal or nomic relations back up scientific explanations. So, if *explain*$_M$ and *explain*$_S$ invokes substantially different backing relations like this, then (TRANSIVITY*) is deemed false (Hicks and van Elswyke; 439). In the case of the Humean conception of laws, the backing relation for "the Mosaic explains$_M$ laws" is simply a truth-making relation. Suppose *P* comes out to be a law on the BSA. Then, the fact that the mosaic is arranged in the way it is is simply what makes true the statement that *P* is a law; it's 'truth-making'. In contrast, the backing relation for "Laws explains$_S$ the Humean mosaic" is not a truth-making relation. Laws do not explain the truth that the Mosaic is arranged in the way it is; they explain why the Mosaic behaves with regularity and uniformity. Therefore, they conclude, the distinction between metaphysical and scientific explanation is well-motivated and there is no problem of circularity for the Humean laws.

The BSAers may draw on the distinction between the two kinds of explanation like above to answer the questions concerning the required normality assumption for the A-BSA. They may argue that the Humean mosaic explains$_M$ the 2nd law's being a law and the 2nd law explains$_S$ why the mosaic behaves as if it is normally distributed; so the validity of the A-BSA is saved.

I think this line of response is on the right track and probably the best option for the BSAers, but they seem to need an extra safety guard to ensure the backing relation for metaphysical explanation doesn't go awry. Hicks and van Elswyke said that the backing relation for "the Mosaic explains$_M$ laws" is a truth-making relation. But it is in fact the Mosaic plus the best systemization of the Mosaic that jointly make the statement "*P* is a law" true. For example, what if we used the BIC-implemented-BSA instead of the A-BSA? Could *P* not be a law if BIC scores and AIC scores of the systems in competition diverge too much?

At this point one might recall the Hope thesis as discussed in §4.3.2 and §4.3.3. The BSAers would respond: Even in a possible world where the counterparts of our scientists use BIC in their practice of science, the same system(s) as ours will be picked as the best systems. So, *P* would still be a law. What allows the BSAers to make such a response is, once again, the Hope thesis. This is the extra safe guard I mentioned above.

There I think goes a warning sign. It starts to seem that too much is hanging onto the hope. The Hope thesis is the BSAers' last resort – when they hit the dead-end in attempting to solve a problem. For example, Elga, in attempting to solve the zero-fit problem, says

> *On this proposal, it can certainly happen that two systems are incomparable with respect to fit. That is no special worry—the best-systems analysis already depends on the hope that some system will be robustly best, as regards the tradeoff between simplicity, strength, and fit. It is no great cost to add an additional hope: that this robustly best system possess a fit profile that holds its own against the profiles of its competitors on any reasonable way of judging when one profile assigns higher chances overall than another. (Elga 2004; 72)*

Now the BSAers seems to have to make a similar remark in attempting to resolve the circularity problem; the nature will kindly arrange itself such a way that, on any reasonable implementation on the BSA, the backing relation of truth-making for metaphysical explanation for the BSA will be robust. The upshot of this subsection is that the BSAers have some available solution to the circularity problem but there seems to be a concern that too much is depending on the Hope thesis. Let us turn to another concern about the BSA: the gap between the contexts of the BSA Oracle doing metaphysics and scientists doing epistemology.

5.6.3 “Mind the Gap” Between the Contexts of Metaphysical and Epistemological Analysis of Laws

As we saw in §3.1, AIC’s exchange ratio between simplicity and fit[117] serves as an *epistemic constraint* on creatures like us. The BSA Oracle, however, doesn’t seem to need to be constrained in the same way because she has as evidence all of the facts in the Humean base in the complete history at the world. Recall that the goal of using AIC is to maximize predictive accuracy by penalizing models for complexity, in order to manage the risk of overfitting. So, our qualm is that the context in which AIC gives such a specific trade-off metric is fundamentally different from the context in which the BSA Oracle mindlessly uses it.

As we saw in §4.3.4, simply accusing the BSA of smuggling epistemology into metaphysics misfires; the accusation is probably begging the question against the BSA. Carroll, himself a primitivist, seems to beg the question when he says “[I]t may very well be true that the right way for us to discover what propositions are laws of nature is via balancing the standards of simplicity, informativeness, and fit. These standards are epistemologically relevant to lawhood. However, they are not metaphysically relevant to lawhood.” (Carroll, 1994; 54). At first blush, criticisms like Carroll’s which rely on the distinction between the context of discovery of laws and the context metaphysics of laws of might seem to work. But we saw that, at least in the picture of the Humean conception of laws, the Humean conception of laws itself is just that laws are axioms or theorems of the extended, ideal version of our actual science; the Humean conception consists in the ascent of the discovery

[117] Simplicity in a specific sense (number of parameters) and fit in a specific sense (log-likelihood).

methods to constituents of laws themselves, in 'the final theory' or 'theory of everything' (§4.2.1).

What I am going to discuss in this section is a different challenge from the above one. Simply put, the challenge here is what the BSA Oracle does is not 'ideal' if she were to use the epistemic standards from the practice of science; it would be less than ideal, in fact. Let me explain.

5.6.3.1 The Context Gap

Recall our discussion of statistical model selection methods in Chapter 3. Among many candidate models ranging from a very simple model (e.g., LIN) to a very complex model that perfectly fit the data (e.g., POLY-99), in spite of the temptation to select the perfectly fitting model, we ought to restrict our choice to a moderately fitting model if we want to minimize the risk of overfitting. This restriction is imposed in the context of our epistemic limitations that we need to make an inference about the entire population from the observed data. *Precisely how much* we should restrict our choice is given by how much simplicity and fit contribute to predictive power. So there is a means-end contextual explanation for the very specific mathematical recipe for balancing simplicity and fit expressed in AIC. In contrast, the BSA Oracle has availed herself of the *entire* history of the actual world. In her context, there is no risk of overfitting. She doesn't have any specific end to achieve when she determines the best system. In short, the specific trade-off metric of AIC is responsive to the specific context in which it was mathematically derived. The BSA is, however, in a different context. So, we may expect that the BSA's use of the specific trade-off metric in such a different context might create some problems. Consider what a scientist would do when he has access to all the data. A scientist whose goal is the minimization of informational loss would cease to use AIC when all of the data is in. This is because, when all of the data is in, there is no more information to be lost, so the continued use of AIC would result in unnecessarily penalizing the complexity of models. In the next section I will give a concrete example as such.

Woodward (2014; 111-12) makes similar critical remarks about the use of simplicity in the BSA. The role assigned to the simplicity in the BSA and the role assigned in the curve-fitting problem are different. In curve-fitting, simplicity guides us to choose a model that has better predictive accuracy (in the Akaikean framework) or the one with greater posterior probability given the observed data (in the Bayesian framework)[118]. In either case, it helps scientists' inductive tasks. Why bother these for the BSA, when all the data is in?

5.6.3.2 A Response: Closer to the True Model

A possible defense from the A-BSA's side may be something like the following.

Even though the BSA Oracle faces no risk of overfitting, she still has good reason to use the AIC's recipe; if she is interested in finding a model closest to the 'true' model, then she would still be justified in using AIC as the balance metric.

As we saw in Chapter 3, choosing the model with the best AIC score is equivalent to choosing the model which has minimum the K-L distance. So scientists interested in

[118] It might be argued that, at least in the Bayesian framework, the BSA Oracle has reason to bother simplicity; for example, she may need to assign higher priors to simpler systems because simpler theories have done better in science. But this seems unacceptable in the BSA's own light. The systems in the BSA, unlike models or theories, are all equally true (at least according to the standard version of the BSA); the difference between systems is just how the fundamental facts are summarized in each system. Given this, their *prior* probabilities are to be same across systems. Nothing in the BSA's metaphysical outset determines the probability distribution for the systems' prior probabilities.

making accurate future prediction are well-motivated to use AIC because minimizing the K-L distance is a way of maximizing predictive accuracy.[119] Then, scientists working in the Akaikean framework are instrumentalists about scientific theories. Now, in response to the concern about the gap between the contexts of scientists and the BSA Oracle, a defender of the A-BSA might argue as follows: It is true that the Oracle would be not interested in predicting future events, however, she could still be interested in finding a model that is *closest to the true model*. For example, Sober (2008) argues that scientists would be justified to use AIC even when their goal is not making accurate prediction but finding the model closest to the true model, for

> *...specific curves have different Kullback–Leibler distances to that true curve. Models are instruments for finding curves that are close to the truth and models are compared with each other to determine how well they advance that goal. (Sober 2008, p.98)*

In the footnote for the above quote, he says

> *The relation of AIC to Kullback–Leibler distances provides an easy answer to the question of why one should care about AIC estimates if one has no interest in using fitted models to predict new data. One still might care about finding fitted models that are close to the truth when K-L distance is used to measure closeness.*

[119] To recapitulate: Any *information* criterion (AIC, BIC, DIC, and the like) for statistical model selection is based on the idea that one should choose a model that minimizes *K-L divergence*. And this divergence can be treated as a distance as it satisfies the conditions for it being a distance metric (See §3.2.1). Suppose the true (or quasi-true) probability distribution function for a random variable X is p(X). Suppose the model we fit to the data (i.e., our hypothesis) is some hypothesized probability distribution, namely, q(X). Then the *K-L distance* is the measure of the *average* (i.e., expected) distance between p(X) and q(X). Let x_i be the observed values. The K-L distance:

$$D(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

The above is for the case of discrete random variables, but it can be easily generalized to the continuous case. The intuitive idea of the K-L distance is that it is a measure of the 'average distance' of a model from the true model. Using information theoretical terms, from which the notion of the K-L distance originated, it is the measure of the inverse of information loss (*entropy*) when estimating true p.d.f p(X) using q(X), with respect to the data x_i.

That is, one may be motivated to use AIC if she is interested in approximating the true model – even if one is not interested in making predictions through models.[120] In a similar vein, then, the defender of the A-BSA might argue that the BSA Oracle is well-motivated in supplementing the BSA with AIC because she is interested in finding a model closest to the true model.

5.6.3.3 "No Such Thing as the True Model for the BSA"

The above line of response requires a more careful assessment. The opponent of the BSA might claim that the above response cannot save the BSA. Recall the K-L distance is the distance between the true model and a hypothetical model. So, the objector might argue, the Akaikean framework assumes that the true model is out there, behind the data, generating or governing the observed data. But, the objection would continue, the presence of such true model is not consistent with the metaphysical outset of the BSA because in the picture of the BSA there is no such thing as the law or the true model which *guides* or *governs* how the facts are to be arranged. So the response relying on the true model fails, it concludes.

It is true that the above notion of the true model is incompatible with the BSA, hence with the A-BSA. But the A-BSAers do not need to be committed to the existence of the true curve or true model. After all, the BSAers can take an approach that the Humean pixels are arranged *as if* there were the true model *g* that would have generated for the observed data; but in fact all there is nothing but a summary of the certain patterns of the pixels. Let me provide an example. Imagine a simple world in which there is only one measurable kind of property. Let Y be a variable representing its quantity of the pixel of the mosaic on which it is instantiated. Suppose the BSA

[120] Assuming the distance here is defined in terms of the K-L distance.

Oracle carries out an extremely tedious task of recording all the occurrences of Y values throughout the entire Humean mosaic. The resulting work would be very detailed but too complex. A bit smarter move would be to compute and report an arithmetic average of all the values – in this case she has one number summary of all the data – very simple, but not maximally informative. Even much smarter work would be to not only report the mean but also the variance of all the data, in the form of *N~(mean, variance)*, if the Y values appear normally distributed throughout the mosaic. Among the three above, the last one seems to be an efficient summary of all the data. Name this summary of the pattern *g*. So this *g* is not the true model or anything like that which governs the data; it is just an efficient summary. Now the A-BSAers seem to have some solution to the problem above. The BSA Oracle using AIC is interested in finding a model which is closest to *g*, a 'quasi' true curve, which in fact is just an efficient summary of instances. There is no need to assume the 'governing' true curve. In this way, the BSAers might defend the use of AIC and resolve the problem of the context gap under discussion.

This solution comes with a cost. It has an effect of getting rid of chance and chancy law from the BSA and moving back to Lewis's original BSA which only invokes strength and simplicity. It is not the aim of this dissertation to discuss the adequacy of such a move. For the purpose of this section, it suffices to show that the BSAers seem to have a solution to the context gap problem.[121]

[121] There seem other philosophers of science who are seriously considering taking this approach. For example, Roberts (ms.) hints at the proposal of what he calls 'nomic frequentism', which views probabilistic laws as just laws about frequency, which would result in turning back to the original version of the BSA.

In the next section, however, I will provide a concrete counterexample to the above solution. The result would be that the BSA will be in a kind of dilemma: either biting the bullet or giving up the refinement under discussion but still maintaining that the main idea of the BSA is tenable, however vague it is. I will argue that neither of them is satisfactory.

5.6.4 A Counterexample: the A-BSA Fails To Pick the Best System

Let me give an example in which the A-BSA fails to pick the best system. This failure will be due to the BSA's failure to mind the context gap discussed in the previous section.

A Counterexample

Suppose there are only two measurable qualities in a world: *x* and *y*. They are instantiated at some pixel points of the Humean mosaic.[122] For the sake of simplicity, let us consider the relation between *x* and *y* in the usual curve-fitting context. Let's suppose there is some trend in the relation between *x* and *y* which may be represented in a form of functions like $y=f(x)$. That is, for each value of *x*, the corresponding *y* value is $f(x)$, *f* being a function which can be represented as a mathematical curve on the *x*-*y* plane. Let us additionally suppose *y* values usually fall right on the $f(x)$ but sometimes they fall above or below it. Some systems may claim there is a certain deterministic relationship between *x* and *y*: $y=f(x)$ (of course these systems will come out to be). Some other systems may claim there is a probabilistic relation between *x* and *y*: for values of *x*, the corresponding *y* values are concentrated around $f(x)$ forming certain probabilistic distribution. For example, system S may claim that *y* has

[122] This is Lewis's characterization of the Humean mosaic (Lewis 1986). For alternative characterization, see Loewer 1996 (*cf.* Maudlin 2007).

a normal distribution with variance of σ^2 and mean of $f(x)$. That is, for each value of x,

System S: $y \sim N(f(x),\sigma^2)$.

Fit of system S may be defined as follows, n being the number of the pixels on the Humean mosaic in each of which x and y are instantiated:

Fit(S): $p(y_1, y_2, . . . , y_n \mid S, x_1, x_2, . . . x_n)$.

Let us also suppose that some systems make different polynomial degrees of $f(x)$. For example (for the ease of presentation, we are going to assume that the variance is same across different systems), consider the following systems:

$S_{\text{POLY-3}}$: $y \sim N(\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3, \sigma^2)$, $\alpha^3 \neq 0$.

$S_{\text{POLY-5}}$: $y \sim N(\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 x^4 + \alpha_5 x^5, \sigma^2)$, $\alpha^5 \neq 0$.

$S_{\text{POLY-99}}$: $y \sim N(\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_{98} x^{98} + \alpha_{99} x^{99}, \sigma^2)$, $\alpha^{99} \neq 0$.

Suppose the Humean mosaic is arranged in such a way that all the y values are distributed *as if* they are distributed exactly by

$g(y)$: $y \sim N(\hat{\alpha}_0 + \widehat{\alpha_1} x + \widehat{\alpha_2} x^2 + \cdots + \widehat{\alpha_{98}} x^{98} + \widehat{\alpha_{99}} x^{99}, \sigma^2)$.

($\hat{\alpha}_i$ refers to the value for each parameter which yields the maximum likelihood with respect to the all the relevant pixels in the Humean mosaic.)

In short, the mosaic is arranged such that the best member[123] of $S_{\text{POLY-99}}$ has *the least* K-L distance from $g(y)$. Note that we are not committed to the existence of true curve of any sort. It is just that the mosaic is arranged *as if* it was generated by some probability distribution which is maximally close to $g(y)$.[124] This means the best

[123] That is, the set of parameter values which yields maximum likelihood in the sense above.

[124] Construed this way, the Humean mosaic under consideration is compatible with distributions other than $g(x)$. But, among many of such possible and compatible hypothetical distributions, the

member of $S_{\text{POLY-99}}$ has the *minimum information loss*[125] from $g(y)$, with respect to the Humean mosaic. Now suppose the logarithm of fits of the above systems came out as:

logFit($S_{\text{POLY-3}}$) = -500

logFit($S_{\text{POLY-5}}$) = -400

logFit($S_{\text{POLY-99}}$) = -350

(That is, $S_{\text{POLY-3}}$ fits least and $S_{\text{POLY-99}}$ fits best. Also note that on our definition of fit, the value will come out very small, which will render logFit as a negative value.) Then AIC scores for these systems come out as:

AIC Score($S_{\text{POLY-3}}$) = logFit($S_{\text{POLY-3}}$) – 3 = -503

AIC Score($S_{\text{POLY-5}}$) = logFit($S_{\text{POLY-5}}$) – 5 = -405

AIC Score($S_{\text{POLY-99}}$) = logFit($S_{\text{POLY-99}}$) – 99 = -449

This means that the A-BSA has to choose $S_{\text{POLY-5}}$ as the best system, as it marks the highest AIC Score. But we assumed that the Humean mosaic is arranged such that $S_{\text{POLY-99}}$ would have the least K-L distance to $g(y)$. In the context of statistical model selection, avoiding the perfectly fitting models is a way to achieve better predictive accuracy.[126] But in the context of system choice, all the data is in, therefore there is nothing to predict. So systems seem to be unnecessarily penalized for complexity if we were to adopt the A-BSA.

one which has the same parameter values as $g(y)$ is the one that has the minimum information loss, given the mosaic.

[125] See §3.2.1 for the K-L distance as a measure of information loss.

[126] For example, see Chapter 7 in Burnham and Anderson 2000

5.6.5 Generalizing the Problem

To the problems like the above, I can think of two possible lines of response from the BSAer's side. The first line of response would be that it is simply the fact that it is chosen by the A-BSA that (the best member of) $S_{POLY\text{-}5}$ is a law of nature on *x* and *y*; the very fact that it is chosen such is what constitutes its lawhood. In short, it is a law because the best system says so, whichever way it is determined to be the best. The second line of response would be that, while the A-BSA fails in the counterexamples like mine, the main idea of the BSA is still on the right track that laws are the elite members of the best systemization. Admittedly, the system choice standards are vague. But the task of refining them should appeal to the empirical considerations of theory choice criteria and there still might be some empirical criteria we haven't considered which will enable us to dodge the counterexamples like mine. I contend that neither of these two responses is successful.

On the first line of response, I think it conflicts with the ultimate motivation of Lewis and the other BSAers. As we saw in §4.2, the Best System may be understood as an ideal theory. By extending actual scientific practice to the ideal case in which all the data is in, we might reach what scientists have been ultimately aiming for, given the successes of physics to date (Lewis 1983, 1994, Loewer 2012, Woodward 2013, Schwarz 2014). Now, in my counterexample above, what scientists would be ultimately looking for is *g*(*x*). But the A-BSA fails to pick the corresponding system as the best system.[127]

[127] It might be insisted like "$S_{POLY\text{-}5}$ is what you (scientists) would get if you extended your actual scientific practice to the entire Humean mosaic. So here is, *in that sense*, what you are looking for." I don't think this is satisfactory given that $g(x)$ is intuitively the ideal goal that scientists would want to reach.

The second line of response would be that the BSA should continue to refine its vague system choice standards from empirical methods of inter-criterial balancing – until it finds the successful one. The prospect for this line of response doesn't look good either, because any refinement of the standards will likely to face the same kind of problems like my counterexample. Counterexamples to any newly proposed refinement can be easily replicated following my earlier recipe. All is needed is to find a case where the refinement in question is based on inductive practice in science. The system choice criteria, and their trade-off recipe in AIC or other statistical model selection methods are set in the way they are for certain inductive purposes: the maximum predictive accuracy for AIC, and the maximum posterior probability for the Bayesian methods, for example. As long as these model selection methods include a correction term for over-fitting or for any potential issues inherit in induction, the implementation of them into the BSA will suffer from the counterexamples like mine. The lesson from the context gap problem and attempted solutions seems to be the following. The BSA, and its main theme about balancing different theoretical virtues, might appear to be on the right track because the BSA does not completely sharpen up the system choice rule and the standards. Once it sharpens them up in one way or another, it will be susceptible to the counterexamples like above as long as the sharpening in question comes from inductive method which recognizes and includes a correction term for inherent limitations of induction.[128]

When vagueness remains as to the system comparison criteria in the BSA, then (as we saw in earlier sections) the only meaningful way to say one system is better than

[128] This suggests there might be another solution for the BSA, if it appeals to empirical methods which do not include specific error correction term.

another with respect to such imprecise criterion is to show it is so in all possible ways of precifisying it, which is effectively equivalent to saying it is so in its ordinal ranking. Then there is the Arrovian threat for the BSA.

Conclusion

In this chapter, I have examined various possible escape routes from the Arrovian threat for the BSA. In §5.2, we saw that **U** does not apply to system choice either but **R** does. But I also noted that, even if **U** is weakened to **R,** a variant of Arrow's impossibility theorem obtains provided the strong neutrality condition (**SN**), a stronger version of **I**, is met. In §5.3, I discussed **SN** in connection with the Humean Supervenience (HS) thesis. I argued that **SN** applies to system choice. While the HS thesis seems to reject the multi-profile framework for system choice, assuming **R** is met in system choice (§5.2), since **SN** applies to system choice, the single-profile variant of the Arrovian impossibility seems to obtain in system choice. In §5.4 I suggested that **IIU** is an indispensable property of system choice procedure. In §5.5, we have discussed a number of possible attempts to make a case for the cardinal measurability of fit, strength, and simplicity, the three criteria invoked by the BSA. We concluded that most of the attempts failed. In §5.6, I proposed a variant of the BSA as an attempt to make a case for inter-criterial comparability between fit and simplicity. I concluded that its prospect doesn't look good mostly due to the gap between epistemological and metaphysical justification of implementing inductive method to the analysis of lawhood. The Arrovian threat for the BSA seems still real and imminent.

Chapter 6: Other Possible Escapes

Introduction

This dissertation investigated the analogue of the Arrovian impossibility in system choice for Lewis's Best System Analysis. In this final chapter, we will review the contributions of this dissertation and discuss future research directions. First, in §6.2 I will discuss the concept of benevolent dictatorship in system choice. In social choice, it seems that dictatorship is an undesirable property of social choice procedure. But in system choice, there might be a possibility of non-harmful or even beneficial dictatorship. In §6.3, I will discuss the statistical model selection method called Minimum Length Description Principle (MDL). The method might be in line with the BSA, so we will briefly review its outlook for system choice. Finally, in §6.4, I will summarize the contributions of my dissertation, draw overall conclusions about the Arrovian threat for the BSA. This will close this dissertation.

6.1 Other Possible Escapes: Non-Harmful Dictatorship, Lexicographical Dictatorship, and Threshold Priority in System Choice

One of the results of §5 is that the system-choice analogue of the Arrovian theorem seems to obtain; there seems a formal analogy between social and system choice and a number of possible escape routes we have explored were unsatisfactory.

In this section, I will discuss other possible escapes that might have good prospects: relaxing Non-Dictatorship condition. The system-choice analogue of Non-Dictatorship would be:

Non-Dictatorship: There is no system-choice criterion *i* such that for all profiles in the domain of system-choice function *f* and for all pairs of alternative systems *x* and *y*, if xP_iy, then xPy. In words, there should be no criterion such that the system which is strictly better another system with respect to that criterion will always win regardless how the systems in comparison fare with the other system choice criteria.

So far we have taken it for granted that dictatorship in theory choice and system choice is an undesirable property of choice procedure. Now let me deal with the following question: Is there a possibility of non-harmful dictatorship?

Let me discuss the possibility of the 'benevolent' dictatorial criterion in system choice. In social choice, Arrow (1963) compares oligarchic political systems where a small elite group of individuals make the social choices with full democratic systems where every individual has natively an equal portion in the social choice procedures. According to those who support the former kind of political system, for example Plato, a society in which an ethically ideal observer (or a small group of such observers) makes social choices would achieve the higher level of social welfare than fully democratic society does. But Arrow views this as an untenable ideal because "power always corrupts; and absolute power corrupts absolutely." (Arrow 1963; 86) Indeed, in reality, dictators always label themselves as "benevolent dictators" but none of them really deserves the label.

But when it comes to theory choice (and system choice as well) we might have good candidates for the role of a benevolent dictatorial criterion in theory choice, or at least a decisive criterion which receives lexicographical priority. For example, van Fraassen (1980)'s empirical adequacy might be such a benevolently dictatorial theory-choice criterion. Arguing none of the existing philosophical theories about laws provides an adequate account, van Fraassen says that the aim of science lies in empirical adequacy, not truth. A scientific theory is empirically adequate if it truthfully says about the *observable* features of the world, that is, if it "saves the phenomena" (1980; 12). For him, therefore, the only proper theory-choice standard is empirical adequacy, and things that we usually count as theoretical merits are in fact good-making features which ultimately contribute to the empirical adequacy. Strength might be a good candidate for the lexicographically prior criterion. According to Woodward (2014), the BSA is not descriptively adequate because it fails to properly capture actual scientific practice of theory choice. The way 'Ockham's razor' principle operates in science is that increasing complexity is permitted only when doing so sufficiently increase strength, but not otherwise; sacrificing strength is not permitted even if doing so sufficiently increase simplicity – that is not how theory choice is made in actuality, Woodward observes. While the BSA emphasizes on the optimal trade-off between simplicity and strength, Woodward argues, actual theory choice procedure puts lexicographical priority or threshold priority on strength. That is, when a theory comes out strictly better with respect to strength, the theory is better in the overall ranking (lexical priority). Or, as something close to lexical priority, for a theory to be considered as good it first must have a sufficient level – meeting a certain 'threshold' level– of strength. If the theory in question does not possess sufficient level of strength, it cannot be compensated by gain in simplicity, however great it is (threshold priority). If two theories A and B are on par with respect to strength, then simplicity may factor in as a tie-breaker. Theory-choice procedures like this which puts priority on certain theory-choice criteria does not satisfy the theory-choice analogue of condition **D** (in the case of strength with lexical priority, whenever a theory is strictly better than other theories with respect to strength, it comes out to be better overall), or condition **I** (in the case of the 'strength threshold', two profiles which agree on the rankings of two theories with respect to the theory choice criteria might disagree on the overall ranking of the two depending on whether the theories are above or below the strength threshold and how strength trades off with simplicity). If such a lexical or threshold priority on strength is a correct picture of how theory choice procedure works in fundamental physics, and if the BSA is to adopt the same procedure as its system choice procedure, then it may open up an escape from the Arrovian impossibility. But this is still a seminal idea as we still need to investigate how this picture of putting priority on certain choice-criterion would work out with probabilistic systems on the BSA, and also how strength and simplicity trades off in case of threshold-priority of strength. As we saw in §5 and §6, measuring strength cardinally or even ordinally is not a straightforward matter, let alone their comparability. This idea requires further research.

6.2 Other Possible Escape: Minimum Length Description Principle in System Choice

Different model selectin criteria use different definitions of simplicity and fit. For example, the model selection principle called Minimum Description Length Principle (MDL) (Rissanen, 1978; Grünwald *et al.*, 2005) take a fundamentally different approach to model selection problems. In MDL, the goal of statistical inference is to find regularity in the data, and regularity is identified with "ability to compress." The underlying idea is that there are different ways to 'summarize' the observed regularities in the data sets, and the shorter the summary is, the better.

The following example (from Grünwald 2005) would help understand the underlying ideas of MDL. MDL is interested in developing a method for *learning* the laws and regularities in data. Consider the following three sequences. Assume that each sequence is 10000 bits long. Just the beginning and the end of each sequence is listed:

(1) 00010001000100010001 . . . 0001000100010001000100010001

(2) 01110100110100100110 . . . 1010111010111011000101100010

(3) 00011000001010100000 . . . 0010001000010000001000110000

The first sequence (1) looks regular. Apparently ‘0001’ is being repeated. If one were to predict what the future data will be like, it would be reasonable to base her prediction on such a regularity, pretending there is a law behind this sequence and it will govern the future sequence as well. The second sequence (2) seems to have no regularity behind it. So, we cannot seem to find any law-like regularity here, nor can we make any reasonable prediction other than that the future data will be just ‘random’. The third sequence (3) shows some regularity in the relative frequencies of 0s and 1s. There seem to be approximately four times as many 0s as 1s. It looks more regular than (2) but less regular than (1). If one were to predict future data in the case of (3), one would probably makes a prediction in the form of probabilistic claims such as “Chance of ‘1’ is 0.2”. In any case, regularity in the data can be used to make predictions; the more regular the data is, the more deterministic predictions one can make about the future data.

In the framework of MDL, regularities in the data means the data can be compressed. The descriptions of compressed data are given in terms of some description method; the most commonly used example for a description method is a general-purpose computer program language like C or Pascal. A description of the set of data *D* is then any computer program that prints *D* and then halts. For the three sequences above, we may write a program

```
i = 1 to 2500; print “0001”; next; halt
```

which will print sequence (1). This is far shorter than the original sequence. That is, (1) is highly compressible. In contrast, the shortest program that will print (2) would be something like this:

```
print
“011101001101000010101010...101011101011101100010110001
0”; halt
```

Basically this is just a repetition of the actual sequence. There is nothing to compress in (2), because there is no detectable regularity in virtue of which we can compress the data. The third sequence (3) lies in between the first two. It cannot be compressed as compactly as (1) but surely the shortest program that can print (3) will be shorter than the length of the actual sequence (3).[129]

This idea of connecting regularity and compressibility[130] is the underlying idea of MDL. MDL says that we should pick the hypothesis which itself can be described in

[129] For mathematical proof for generalization of this kind of result, see Grünwald 2005; 27.

[130] The idea of connecting simplicity and compressibility might seems to be highly language dependent. But one does not need to worry too much about the language dependence problem in MDL framework. According to the so-called *invariance theorem* (Kolmogorov 1965, and independently Solomonoff 1964), for any two general-purpose programming languages A and B, and any data sequence D, the difference in the length of the shortest program for printing D in language A and B is always smaller than a constant *c*. (Grünwald 2005).

short length and also can describe the data compactly (i.e., by describing the data compactly, one effectively compress the data). More formally,

> **MDL principle** (Rissanen 1978): the best hypothesis H to explain the set of data D is the one that minimizes the sum of:
>
> (1) *L(H)*: The length of the description of the hypothesis itself, plus
>
> (2) *L(D|H)*: The length of the description of the data *D* when the data is described with the help of the theory.

The first term can be understood as complexity of *H*, and the second term as goodness-of-fit of *H*. Intuitively speaking, the second term tells us goodness-of-fit of the theory because, the better it fits the data, the fewer bits we would need to describe the data *given* the theory. For example, in order to 'describe the data' as in (2), we would need to describe the discrepancies between the values predicted by the theory and the actually observed value; but we would not describe what the theory predicts about the data, which is the job of (1) (Grünwald 2005).

Let us examine a simple example. Imagine a polynomial context where the data can be plotted on the *x*-*y* plane and we have rival polynomial models to explain the data. Our familiar curve-fitting problem would be such a context. Suppose we consider the following hypotheses:

> Hypothesis *A*: $y = 2x^2 + 3x + 4 +$ [error]
>
> Hypothesis *B*: $y = 3x + 1 +$ [error]

Assume the error term above is a normally distributed noise term. Construed this way, each hypothesis above defines probability distribution for *y* values given *x*.[131]

[131] See §3.2.1 for how each curve can be regarded as a probability distribution.

Now we need to define the two terms $L(D|H)$ and $L(H)$ in the statement of MDL principle; that is, we need to define codes for encoding the two terms. First, $L(D|H)$ in this example can be best understood if we define it as $-\log P(D|H)$, where $P(D|H)$ probability mass or density of D given H (Grünwald 2005; 28). Hence, the better H 'fits' the data, the shorter description length $L(D|H)$. For hypothesis A, the data points fall on the curve $y = 2x^2 + 3x + 4$ do not need extra description, because they are precisely the values expected by A. The data points that do not fall right on that curve do need extra description, for example, in terms of Sum of Squares. So, if A fits the given data better than B does, then $L(D|A)$ will be shorter than $L(D|B)$. What of $L(H)$? Earlier forms of MDL allowed any form of coding to encode H but soon it was realized that allowing any code has the same problem as language-dependence for curve-fitting; depending on which code we use, we may get vastly different lengths for $L(H)$. A wide variety of codes are suggested for that term to minimize the code-dependence problem. See Grünwald (2005) for an extensive survey of them.[132] They deserve further research. In our simple example, a suitable set of codes will be the ones what would yield $L(A)$ be longer than $L(B)$. Encoded by such a code, MDL tells us to compare the following MDL lengths of the hypotheses and choose the one with the smaller MDL score.

MDL for A: $L(A) + L(D|A)$

MDL for B: $L(B) + L(D|B)$

Note that this MDL expresses a specific trade-off ratio between simplicity and fit, defined in particular ways. For the purpose of this section, let us now turn to some

[132] A modern, more refined version is based on the concept of *universal coding*.

important features of MDL that have close relevance to the BSA. Grünwald 2005 is probably the clearest exposition of the underlying philosophy of MDL, so let me draw on him in the following.

No Need for 'Underlying Truth'

Rissanen, the main pioneer of MDL, says:

> *We never want to make the false assumption that the observed data actually were generated by a distribution of some kind, say Gaussian, and then go on to analyze the consequences and make further deductions. Our deductions may be entertaining but quite irrelevant to the task at hand, namely, to learn useful properties from the data. (Rissanen 1989; 15)*

What is important for us is that MDL does not need for 'underlying truth.' As we saw in §3.2 and §3.4., the common statistical model methods like AIC and BIC crucially relies on the 'true curve'. For example, K-L divergence, the key concept in AIC and BIC, is defined as the distance between the true hypothesis and the fitted hypothesis, Derivation of AIC and BIC rely on the assumption that the parameter values of fitted hypotheses is expected to centered around the 'true' parameter values. We saw in §5.6.3 that this feature causes troubles when we implement the BSA with those statistical methods. In contrast, as we can see in the above quote from Rissanen, the main idea of MDL is that we should focus on what we can learn from the data as all we have is the data. In the picture of MDL, our inductive inference should only be based on the data, not on the assumption of some underlying true state of nature. According to Rissanen, the goal of inductive inference should be to 'squeeze out as much regularity as possible' from the given data.[133] So, the main task for statistical

[133] Vitány and Li make the same claim (1997; 351)

inference is to separate ‘structure’ (i.e., the regularity, the ‘meaningful information’) from ‘noise’ (i.e., the ‘accidental information’) in the given data.

Clearly, the above feature of MDL seems to sit very well with the BSA. As we saw in §4.2 and §4.3, the BSA views laws as regularities in the best systematization of the facts at a world. It does not assume some metaphysically heavy laws as primitivism does. Also, it views laws are just results of extending our best practice inductive inference to the ideal case (that is, the case where ‘all the data is in’). So, there is no such thing as ‘true curve’ or ‘true status of nature’ on the BSA. If we implement the BSA with the trade-off recipe expressed in MDL, we would no longer have the problems for the A-BSA or the B-BSA.

So, MDL seems to have a good outlook as an escape from the Arrovian result because it makes different system-choice criteria commensurable, while not falling for the problems that AIC or BIC has. But we would still need a further examination of its philosophical assumptions. In principle, MDL views the data as messages and we need to make a priori assumptions about the nature of the source of code. In particular, MDL assumes that regularities described in shorter length would help us better conduct scientific investigation. But, as Adriaans (2008; 164-5) says:

> *The extreme regularity of the universe could be a ‘local’ condition accidentally observed by us. In terms of modern information theory: every infinite random string has an infinite number of regions of extreme regularity. If we transpose this idea to the analysis of our world we might just accidentally live in such a regular region in a purely random universe…*

If we live in such a purely random universe, then MDL would not be motivated because we seem to have no reason to favor simplicity. However, if we assume that even such a universe will be cooperative to us, in the sense that nature will first show us the data of a kind from which we can squeeze out regularities in short descriptions. In that case, the hypothesis favored by MDL would be the right one in that it confers high probably on the given data. This is called the *cooperative universe* hypothesis (Adriaans 2008). In the light of the project of implementing the BSA with MDL, the cooperative universe hypothesis might be just another guise of the Hope thesis. If this is correct, then once again we have to hang onto the hope.

Conclusion

In this section, I will first give a somewhat pessimistic outlook on the BSA based on the results from the previous chapters. But my conclusion would not be just pessimistic; I will also suggests that there might be still a way to save the BSA from the problems I raise in this dissertation.

If my arguments in Chapter 5 are sound, then the BSA is threatened by the Arrovian impossibility. The BSAers might attempt various escapes from the impossibility result. The might attempt to fall back to the weak version of Humean Supervenience (HS) thesis and claim that the multi-profile framework of Arrow's impossibility theorem is blocked on the weak HS thesis. But we saw in §5.3 that there are single-profile variants of the Arrovian impossibility, provided the domain of the system-choice rules is rich and the rules satisfy Strong neutrality (**SN**). The BSAers might argue that Rich domain (**R**) is not motivated in system choice; but the conclusion of §5.2 was that we have reason to think that it is motivated. Or, they might attempt to drop **SN**, but as we saw in §5.4, in the context of system choice, **SN** is well motivated. Or, they might attempt to abandon Independence of Irrelevant Alternative

(**I**) condition. But as we saw in §5.4, the ‘irrelevance’ aspect of I is motivated in system choice.

Maybe the BSAers hang their hope onto cardinality and comparability. Careful examinations conducted in §5.5, however, suggest that it is doubtful that the system-choice criteria are cardinally measurable. Being charitable to the BSAers, it might even granted that we would be able to find a better refinement of the system-choice criteria such that they come out to be cardinally measurable. Unfortunately, this alone cannot save the BSA from the Arrovian impossibility result unless the system-choice criteria are commensurable. So, the BSAers now might turn to the cases in statistical model selection literature where they can find cases for inter-criterial comparability. The two common methods, AIC and BIC, however, cannot help them as we saw in §5.6. This was because (i) there are the context gap between statistical model selection and system choice, and (ii) AIC and BIC in principle rely on the existence of the ‘true status of nature’, which is in conflict with the BSA’s conception of laws of nature discussed in Chapter 4. MDL has the better prospect as a way to make a case of cardinal comparability as we saw in §6.3 because it does not assume the existence of the true status of nature. But, it does not come free; the BSA implemented with MDL might have to rely on the Hope thesis.

This leads me to make final comments on the BSA’s reliance on the Hope thesis (§4.3). To all the problems above, the BSAers last resort would be the Hope thesis. Here is a long list of things the Hope thesis would entail. The nature will kindly arrange itself such that: the domain of the system-choice rules will be severely restricted; different cardinal measures of the system-choice criteria will agree on the best system;, different ways of trading off the criteria, for example the BSAs implemented with AIC, BIC, MDL, or any other statistical model selection methods will eventually agree; the zero-fit problem (discussed in §5.5) would be resolved, the circularity problem would be avoided (discussed in §5.6), and the list is open-ended. As I showed throughout the dissertation, virtually for every single problem the BSA faces, eventually their responses have to rely on the Hope thesis.

I suspect that the best solution for the BSA facing the Arrovian impossibility result could be found in the notion of non-harmful dictatorship, which would not require too much reliance on the Hope thesis. That is, the BSA might have to give up the notion of 'balancing' in its analysis of laws of nature. The Humean notion of laws may not require such a balancing procedure of different criteria, to being with. Maybe, for example, van Fraassen's 'saving the phenomena' might be the one and only criterion that the BSA should concern – this would be in line with the Humean perspective on laws of nature.

Bibliography

Adriaans, Pieter & van Benthem, Johan (2008). *Handbook of Philosophy of Information*. Elsevier.

Adriaans, Pieter. (2008). “Learning and the Cooperative Computational Universe.” In P. Adriaans and J. van Benthem (eds.) *Handbook of Philosophy of Information*. Elsevier: 133-167.

Akaike, Hirotugu. (1974). "A new Look at the statistical model identification." *IEEE Transactions on Automatic Control,* 19: 716-723.

Akaike, Hirotugu. (1973). “Information Theory and an Extension of the Maximum Likelihood Principle', in B. N. Petrov and F. Csaki (eds.), *2nd International Symposium on Information Theory*, Budapest, Akademiai Kiado: 267-81.

Armstrong, David M. (1983). *What is a Law of Nature?* Cambridge, UK: Cambridge University Press.

Armstrong, David M. (2004). *Truth and Truthmakers*. Cambridge University Press.

Arrow, Kenneth. (1951/1963). *Social Choice and Individual Values.* New York: Wiley.

Arrow, Kenneth J. (1984). “Social Choice and Justice.” vol. 1 of *Collected Papers of Kenneth J. Aarow*. Belknap Press, Cambridge, MA.

Arrow, Kenneth Joseph, Amartya Sen, and Kotaro Suzumura, eds. (2002). *Handbook of social choice and welfare*. Vol. 1. Gulf Professional Publishing.

Barberà, Salvador, Peter Hammond, and Christian Seidl, eds. (2004). *Handbook of utility theory*. Vol. 2. Springer Science & Business Media.

Baumann, Peter (2005). Theory Choice and the Intransitivity of 'Is a Better Theory Than'. *Philosophy of Science* 72: 231-240.

Beebee, Helen. (2000). "The non-governing conception of laws of nature." *Philosophical and Phenomenological Research* 61: 571-594.

Bennett, Karen (2011). “Construction area (no hard hat required).” *Philosophical Studies* 154: 79-104.

Bird, Alexander (1998). *Philosophy of Science*. University College London Press.

Blackorby, Charles, Walter Bossert, and David Donaldson. (1984). “Social Choice with Interpersonal Utility Comparisons: A Diagrammatic Introduction.” *International Economic Review* 25: 327–356.

Bossert, Walter, and Kotaro Suzumura. (2010). *Consistency, Choice, and Rationality*. Cambridge: Harvard University Press.

Bostrom, Nick. “A critique of David Lewis’ theory of chance?” Unpublished, Url: http://www.analytic.org, Accessed May 2015.

Burnham, Kenneth P., and David R. Anderson. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Callender, Craig, and Jonathan Cohen. (2009). A better best system account of lawhood. *Philosophical Studies* 145:1 - 34.

Campbell, Donald E., and Jerry S. Kelly. (2002) "Impossibility theorems in the Arrovian framework." *Handbook of social choice and welfare,* vol. 1: 35-94.

Carroll, John W. (1987) “Ontology and the Laws of Nature.” *Australasian Journal of Philosophy* 65: 261-276.

Carroll, John W. (1994). *Laws of Nature*. Cambridge: Cambridge University Press.

Carroll, John W. (1990). “The Humean tradition.” *Philosophical Review* 99:185-219.

Cartwright, Nancy. (1980). “Do the Laws of Physics State the Facts?” In M. Curd & J. A. Cover (eds.), *Pacific Philosophical Quarterly*: 865-877.

Cartwright, Nancy. (1983). *How the laws of physics lie*. Oxford: Oxford University Press.

d'Aspremont, Claude, and Louis Gevers. (1977). "Equity and the informational basis of collective choice." *The Review of Economic Studies* 44: 199-209.

d'Aspremont, Claude, and Louis Gevers. (2002) "Social welfare functionals and interpersonal comparability." *Handbook of social choice and welfare* vol. 1: 459-541.

Dretske, Fred I. (1977). "Laws of nature." *Philosophy of Science* 44: 248-268.

Earman, John. (1984). "Laws of Nature: The Empiricist Challenge", in Bogdan (ed.), *D.M. Armstrong*, Reidel, Dordrecht, Holland.

Earman, John. (1986). *A Primer on Determinism*. Dordrecht: D. Reidel Publishing Company.

Earman, John. (1993). "In defense of laws: Reflections on Bas van Fraassen's laws and symmetry." *Philosophy and Phenomenological Research* 53: 413-419.

Earman, John. (2004). "Laws, symmetry, and symmetry breaking: Invariance, conservation principles, and objectivity." *Philosophy of Science* 71: 1227-1241.

Earman, John and Roberts, John T. (2005). “Contact with the nomic: A challenge for deniers of Humean supervenience about laws of nature part I: Humean supervenience.” *Philosophy and Phenomenological Research* 71:1–22.

Elga, Adam. (2004). “Infinitesimal chances and the laws of nature.” *Australasian Journal of Philosophy* 82: 67 – 76.

Feldman, Allan M., and Roberto Serrano. (2008). "Arrow's impossibility theorem: Two simple single-profile versions." *Harvard College Mathematics Review*, Vol. 2, No. 2, 2008: 46-57

Fishburn, Peter C. (1973). *The Theory of Social Choice*. Princeton University Press.

Fleurbaey, Marc, and Peter J. Hammond. (2004). "Interpersonally comparable utility." *Handbook of utility theory.* vol. 2. Springer US: 1179-1285.

Forster, Malcolm R. (1995). "The golfer's dilemma: a reply to Kukla on curve-fitting."*The British journal for the philosophy of science* 46: 348-360.

Forster, Malcolm R. "Model selection in science: The problem of language variance." *The British journal for the philosophy of science* 50.1 (1999): 83-102.

Forster, Malcolm R. (2000). "Key concepts in model selection: Performance and generalizability." *Journal of mathematical psychology* 44: 205-231.

Forster, Malcolm. (2001). "The New Science of Simplicity." In A. Zellner, H. A. Keuzenkamp, and M. McAleer (eds.): *Simplicity, Inference and Modelling*, Cambridge University Press: 83-119.

Forster, Malcolm & Sober, Elliott. (1994). "How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions." *British Journal for the Philosophy of Science* 45 (1):1-35.

Frigg, Roman and Stephan Hartmann. (2012). "Models in Science", *The Stanford Encyclopedia of Philosophy* (Fall 2012 Edition), Edward N. Zalta (ed.), URL= <http://plato.stanford.edu/archives/fall2012/entries/models-science/>.

Gaertner, Wulf. (2009). *A Primer in Social Choice Theory, revised edition*. Oxford: University Press.

Geanakoplos, John. (2005). "Three Brief Proofs of Arrow's Impossibility Theorem", *Economic Theory*, 26: 211–215.

Giere, Ronald N. (1999). *Science Without Laws*. University of Chicago Press.

Goodman, Nelson (1983). *Fact, Fiction, and Forecast*. Harvard University Press.

Grünwald, Peter D. (2005). "Introducing Minimum Description Length Principle" in Grünwald, P. *et al.*:3-21.

Grünwald, Peter D., In Jae Myung, and Mark A. Pitt, eds. *Advances in minimum description length: Theory and applications.* (2005).

Hall, Ned. (2012). "David Lewis's Metaphysics", *The Stanford Encyclopedia of Philosophy* (Fall 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2012/entries/lewis-metaphysics/>.

Hall, Ned. (forthcoming). "Humean Reductionism about Laws of Nature". <philpapers.org/rec/HALHRA>, accessed May 2015.

Halpin, John F. (2003). "Scientific law: A perspectival account." *Erkenntnis* 58:137 - 168.

Hammond, Peter J. (1976). "Equity, Arrow's conditions, and Rawls' difference principle."*Econometrica: Journal of the Econometric Society*, 44: 793-804.

Hammond, Peter J. (1991). "Interpersonal Comparisons of Utility: Why and How They Are and Should Be Made." In Elster, J. and Roemer, J. E. (eds.), *Interpersonal Comparisons of Well-Being.* Cambridge University Press, Cambridge: 200–254.

Hempel, Carl. (1983). "Valuation and Objectivity in Science," in *Physics, Philosophy, and Psychoanalysis: Essays in Honor of Adolf Grunbaum.* Robert S. Cohen and Larry Laudan (eds.) Dordrecht, The Netherlands: Kluwer: 73–100.

Hicks, Michael Townsen & van Elswyk, Peter (2015). "Humean laws and circular explanation." *Philosophical Studies* 172: 433-443.

Johnson, Oliver. (2004). *Information theory and the central limit theorem*. London: Imperial College Press.

Kalai, Ehud and David Schmeidler. (1977) "Aggregation Procedure for Cardinal Preferences: A Formulation and Proof of Samuelson's Impossibility Conjecture," *Econometrica*, XLV, 6:1431–38.

Kelly, Jerry S. (1978). *Arrow Impossibility Theorems*. New York: Academic Press.

Kemp, Murray C., and Yew-Kwang Ng. (1976). "On the existence of social welfare functions, social orderings and social decision functions." *Economica,* vol. 43: 59-66.

Kieseppä, Ilkka A. (1997). "Akaike information criterion, curve-fitting, and the philosophical problem of simplicity." *The British journal for the philosophy of science* 48: 21-48.

Kieseppä, Ilkka. A. (2001a). "Statistical model selection criteria and Bayesianism." *Philosophy of Science*: S141-S152.

Kieseppä, Ilkka. A. (2001b)."Statistical model selection criteria and the philosophical problem of underdetermination." *The British journal for the philosophy of science* 52: 761-794.

Kieseppä, Ilkka. A. (2003). “Akaike's Theorem and Bayesian Methodology.” In A. Rojszczak, J. Cachro & G. Kurczewski (eds.), *Philosophical Dimensions of Logic and Science*. Kluwer Academic Publishers: 117--137.

Kolmogorov, Andrei N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission.* 1: 1–7.

Konishi, Sadanori, and Genshiro Kitagawa. (1996). "Generalised information criteria model selection." *Biometrika* 83: 875-890.

Konishi, Sadanori, and Genshiro Kitagawa. (2007). *Information criteria and statistical modeling*. Springer Science & Business Media.

Kruze, Michael B. (1996). *Scientific Rationality*. (Dissertation). UMI no. 9710026. University of Wisconsin-Madison, MI.

Kuhn, Thomas. (1969). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kuhn, Thomas. (1970). “Postscript—1969”, in *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press: 174–210.

Kuhn, Thomas. (1977a). "Objectivity, Value Judgment and Theory Choice". In Kuhn (1977b): 320−39.

Kuhn, Thomas. (1977b). *The Essential Tension*. Chicago: University of Chicago Press.

Kullback, Solomon, and Richard A. Leibler. (1951). "On information and sufficiency." *The annals of mathematical statistics* 22: 79-86.

Lange, Marc. (2000). "Natural laws in scientific practice." Oxford University Press, New York.

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

Lange, Marc. (2013). "Grounding, scientific explanation, and Humean laws." *Philosophical studies* 164: 255-261.

Laudan, Larry. (1977). *Progress and its Problems*. Berkeley: University of California Press.

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

Lewis, David. (1980). “A subjectivist’s guide to objective chance.” Reprinted with postscripts in Lewis 1986.

Lewis, David. (1983). "New work for a theory of universals." *Australasian Journal of Philosophy*, 61:343–377.

Lewis, David. (1986). *Philosophical Papers*, Vol. II. Oxford: Oxford University Press.

Lewis, David. (1994). "Humean supervenience debugged." *Mind*, 103:473–490.

Li, Ming, and Paul MN Vitányi. (1997). *An Introduction to Kolmogorov Complexity and Its Applications.* 2nd edition. New York: Springer-Verlag.

List, Christian. (2013). "Social Choice Theory", *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2013/entries/social-choice/>.

Little, Ian M.D. (1952). "Social Choice and Individual Values." *Journal of Political Economy* 60: 422–432.

Loewer, Barry. (1996). "Humean supervenience." *Philosophical Topics* 24: 101-127.

Loewer, Barry. (2004) "David Lewis's Humean theory of objective chance." *Philosophy of Science* 71: 1115-1125.

Loewer, Barry. (2007). "Laws and natural properties." *Philosophical Topics* 35: 313-328.

Loewer, Barry. (2012). "Two Accounts of Laws and Time." *Philosophical Studies* 160:115-137.

Lyon, Aidan. (2014). "Why are Normal Distributions Normal?" *British Journal for the Philosophy of Science* 65: 621-649.

Marshall, Dan. (2015). "Humean laws and explanation." *Philosophical Studies*: 1-21.

Maudlin, Tim. (2007). *The Metaphysics Within Physics*. Oxford University Press.

McMullin, Earnan. (1987). Explanatory success and the truth of theory. In N. Rescher, Ed., *Scientific Inquiry in Philosophical Perspective.* Lanham: University Press of America: 51-73.

Mill, John Stuart. (1843) "A system of logic." *Collected Works of John Stuart Mill, Bd* 7 (1865): 388-392.

Miller, David. (2006). *Out Of Error: Further Essays on Critical Rationalism*. Aldershot: Ashgate.

Morreau, Michael. (2010). “It simply does not add up: Trouble with overall similarity.” *Journal of Philosophy* 107 (9): 469-490.

Morreau, Michael.(2014a) "Mr. Fit, Mr. Simplicity and Mr. Scope: From Social Choice to Theory Choice." *Erkenntnis* 79, 1253-1268.

Morreau, Michael (2014b), "Arrow's Theorem", *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2014/entries/arrows-theorem/>.

Morreau, Michael. (2015). "Theory choice and Social choice: Kuhn vindicated." *Mind* 124: 239-262.

Mulaik, Stanley A. (2001). “The curve-fitting problem: An objectivist view.” *Philosophy of Science*, 68 (2): 218-241.

Mumford, Stephen. (2004). *Laws in Nature*. London: Routledge.

Mumford, Stephen, (2005), “Laws and Lawlessness.” *Synthese* 144: 397-413.

Okasha, Samir. (2011). "Theory choice and social choice: Kuhn versus Arrow." *Mind,* 120: 83-115.

Okasha, Samir. (2015). "On Arrow’s Theorem and Scientific Rationality: Reply to Morreau and Stegenga." *Mind* 124: 279-294.

Parks, Robert P. (1976). "An Impossibility Theorem for Fixed Preferences: A Dictatorial Bergson- Samuelson Welfare Function." *Review of Economic Studies* 43: 447-450.

Penrose, Roger. (2004). *The Road to Reality: A Complete Guide to the Laws of the Universe*. New York: Vintage Books.

Pollak, Robert A. (1979). "Bergson-Samuelson social welfare functions and the theory of social choice." *The Quarterly Journal of Economics* 93: 73-90.

Popper, Karl. (1959). *The Logic of Scientific Discovery*, translation of *Logik der Forschung.* London: Hutchinson.

Popper, Karl. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge.* London: Routledge.

Priest, Graham. (1976). “Gruesome simplicity.” *Philosophy of Science* 43: 432-437.

Psillos, Stathis. (2002). *Causation and Explanation*. Chesham: Acumen

Ramsey, Frank P. (1928) “Universals of law and of fact.” In D. H. Mellor (ed.), *Philosophical Papers*, Cambridge: Cambridge University Press: 140–4.

Ramsey, Frank P. (1929) “General propositions and causality.” In D. H. Mellor (ed.), *Philosophical Papers*, Cambridge: Cambridge University Press: 145–63.

Riker, William H. (1982). *Liberalism against populism: A confrontation between the theory of democracy and the theory of social choice*. San Francisco: Freeman.

Rissanen, Jorma. (1978). "Modeling by the shortest data description." *Automatica* 14: 465–471.

Roberts, John T. (2008). *The law-governed universe*. New York: Oxford University Press.

Roberts, John T. (1999). “‘Laws of nature’ as an indexical term: A reinterpretation of Lewis's best-system analysis.” *Philosophy of Science* 66: S502-S511.

Roberts, John T. (2001). “Undermining undermined: Why Humean supervenience never needed to be debugged (even if it's a necessary truth).” *Proceedings of the Philosophy of Science Association* 2001: S98-S108.

Roberts, Kevin. (1980). "Interpersonal comparability and social choice theory." *The Review of Economic Studies*, vol. 47, no.2: 421-439.

Roberts, Kevin. (2005). "Social choice theory and the informational basis approach." *Economics Series Working Papers* 247, University of Oxford, Department of Economics.

Rubinstein, Ariel. (1984). "The single profile analogues to multi profile theorems: Mathematical logic's approach." *International Economic Review* 25: 719-730.

Ruby, Jane E. (1986). "The Origins of Scientific 'Law'". *Journal of the History of Ideas* 47:341.

Samuelson, Paul. (1967). "Arrow's Mathematical Politics", in S. Hook (ed.), *Human Values and Economic Policy*, New York: New York University Press: 41–52.

Schaffer, Jonathan. (2008). "Causation and Laws of Nature: Reductionism." In *Contemporary Debates in Metaphysics,* J. Hawthorne, T. Sider, and D. Zimmerman, (eds.), Oxford: Basil Blackwell: 82-107.

Schwarz, Gideon. (1978). "Estimating the dimension of a model." *The annals of statistics* 6: 461-464.

Schwarz, Wolfgang. (2014). "Best System Approaches to Chance." In A. Hájek and C. Hitchcock (eds.), *The Oxford Handbook of Probability and Philosophy*, forthcoming.

Sen, Amartya K. (1979). "Utilitarianism and welfarism." *The Journal of Philosophy*, 76 (9): 463-489.

Sen, Amartya K. (1970). *Collective choice and social welfare*. San Francisco: Holden-Day.

Sen, Amartya K. (1970b). "The impossibility of a Paretian liberal." *The journal of political economy*, 78 (1): 152-157.

Sen, Amartya. (1977). “On Weights and Measures: Informational Constraints in Social Welfare Analysis.” *Econometrica* 45: 1539–1572.

Sen, Amartya K. (1986). ‘Social Choice Theory’. In K. J. Arrow and M. D. Intriligator (eds.), *Handbook of Mathematical Economics*, vol. III. Amsterdam: North-Holland.

Sen, Amartya K. (1999). "The possibility of social choice." *American Economic Review* 89: 349-378.

Sober, Elliott. (2002). “Instrumentalism, Parsimony, and the Akaike Framework.” *Proceedings of the Philosophy of Science Association* 2002 (3): S112-S123.

Sober, Elliott. (2008). *Evidence and evolution: The logic behind the science*. Cambridge University Press.

Solomonoff, Ray. (1964). A formal theory of inductive inference, part 1 and part 2. *Information and Control* 7: 1–22, 224–254.

Stegenga, Jacob. (2013). “An impossibility theorem for amalgamating evidence.” *Synthese* 190: 2391-2411.

Suzumura, Kotaro. (2002). "Introduction” In Kenneth Arrow, Amartya Sen and Kotaro Suzumura (eds.) *Handbook of Social Choice and Welfare.* vol. 1. Amsterdam: Elsevier/North-Holland.

Tooley, Michael. (1977). “The Nature of Laws.” *Canadian Journal of Philosophy* 7: 667–698.

van Fraassen, Bas. (1980). *The Scientific Image*. Oxford: Oxford University Press.

van Fraassen, Bas. (1989) *Laws and Symmetry*. Oxford: Oxford University Press.

Weinberg, Steven. (1992). *Dreams of a final theory: The search for the fundamental laws of nature*. New York: Pantheon Books.

Williams, Robert. (2008). “Chances, Counterfactuals, and Similarity.” *Philosophy and Phenomenological Research* 77: 385-420.

Woodward, James. (2013). “Laws, causes, and invariance.” In Mumford, Stephen, and Matthew Tugby (eds.) *Metaphysics and science*: 48-72

Woodward, James. (2014). “Simplicity in the Best Systems Account of Laws of Nature.” *British Journal for the Philosophy of Science* 65: 91-123.

Zwart, Sjoerd D. and Franssen, Maarten. (2007). “An impossibility theorem for verisimilitude.” *Synthese* 158: 75-92.