Lund University

MA Thesis in Theoretical Philosophy

Solomonoff Induction:

A Solution to the Problem of the Priors?

Author:
Aron Vallinder

Supervisor:
Staffan Angere

September 24, 2012
30 credits

Abstract

In this essay, I investigate whether Solomonoff's prior can be used to solve the problem of the priors for Bayesianism. In outline, the idea is to give higher prior probability to hypotheses that are "simpler", where simplicity is given a precise formal definition. I begin with a review of Bayesianism, including a survey of past proposed solutions of the problem of the priors. I then introduce the formal framework of Solomonoff induction, and go through some of its properties, before finally turning to some applications. After this, I discuss several potential problems for the framework. Among these are the fact that Solomonoff's prior is incomputable, that the prior is highly dependent on the choice of a universal Turing machine to use in the definition, and the fact that it assumes that the hypotheses under consideration are computable. I also discuss whether a bias toward simplicity can be justified. I argue that there are two main considerations favoring Solomonoff's prior: (i) it allows us to assign strictly positive probability to every hypothesis in a countably infinite set in a *non-arbitrary* way, and (ii) it minimizes the number of "retractions" and "errors" in the worst case.

Acknowledgements

I'm grateful to the participants of the working seminar in philosophy of science at Lund University, where a draft of this essay was presented. George Masterton and Abraham Wolk offered several useful comments, and in e-mail correspondence, Shane Legg clarified many things for me. Finally, I wish to thank Staffan Angere for excellent supervision.

Contents

List of Symbols **3**

Outline **4**

1 Introduction **5**
1.1 Basics of Bayesian epistemology 5
1.2 Subjectivism Without Convergence 8
1.3 Subjectivism With Convergence 9
1.4 Objective Bayesianism: Equivocation 10
1.5 What Should a Solution Achieve? 14

2 Solomonoff induction **19**
2.1 Universal Turing machines 19
2.2 Kolmogorov complexity and Solomonoff's prior 20
2.3 Some properties of Solomonoff's prior 23
2.4 Reparametrization and regrouping invariance 25

3 Applications **27**
3.1 Determining the model space 27
3.2 The Raven paradox 28
3.3 From strings to propositions 30

4 Criticism **32**
4.1 Language dependence 32
4.2 Incomputability 34
4.3 Are the hypotheses computable? 36
4.4 Why simplicity? 36

5 Summary and Discussion **40**

6 Bibliography **43**

List of Symbols

$\mathrm{P}(\cdot)$	Probability function	**??**
A, B, C	Events or propositions	6
X_1, X_2, X_3	Random variables	9
$b(\cdot), c(\cdot)$	Personal probability functions	9
$H(\mathrm{P})$	Entropy of probability function P	11
$\mathcal{E}$	Background knowledge	11
ϵ	The empty string	22
x	A binary string	21
$K(\cdot)$	Kolmogorov complexity	21
p	Program for a Turing machine	21
$M(\cdot)$	Solomonoff's universal prior	21
$x*$	A binary string that begins with x	21
$l(p)$	Length of Turing machine program p	21
μ, ρ	Semi-measures	23
$x_{<t}$	Binary string of length t	24
$x_{1:\infty}$	Infinite binary string	24
$\mathbb{B}^t$	The set of binary strings of length t	25

Outline

The topic of this essay is the problem of the priors for Bayesianism, and in particular whether Ray Solomonoff's prior can be used to solve it.

In this first introductory section, I describe the basics of Bayesian epistemology. Bayesianism comes in many versions, and my taxonomy is based on how these respond to the problem of the priors. According to subjectivism without convergence, any choice of prior is fine. Furthermore, Bayesians are *not* guaranteed to converge on the same posterior probabilities, even as shared evidence accumulates. Subjectivists who believe in convergence agree that any choice of prior is fine, but they also think that there will be convergence among Bayesian agents. According to objective Bayesians, there is a unique prior distribution (or perhaps a unique set of such distributions) which is best. I discuss two branches of objective Bayesianism. The first holds that when assigning prior probabilities, we should equivocate between all hypotheses under consideration. The second holds (roughly) that we should assign prior probabilities in proportion to the simplicity of a hypothesis. This is known as Solomonoff induction. I end the introduction with a discussion of what a solution to the problem of the priors should achieve. By what criteria do we judge proposed solutions? I argue that solutions to the problem of the priors should be evaluated on purely epistemic grounds: how do they aid an agent in the search for truth? I provide a list of different ways in which priors could provide such help. This list provides the basis of my evaluation of Solomonoff's prior.

Section 2 introduces the formal apparatus necessary to understand Solomonoff induction. In 2.1 and 2.2 I review universal Turing machines and Kolmogorov complexity, which then lets us define Solomonoff's prior. In 2.3 and 2.4 I discuss some important theoretical features of Solomonoff's prior. Provided that the correct answer is computable, Solomonoff induction is guaranteed to converge on it under a finite bound, both in deterministic and stochastic settings. In section 2.4 I discuss the fact that Solomonoff's prior meets two desirable conditions: it is (almost) invariant under both reparametrization and regrouping.

The third section looks at some applications of Solomonoff induction. As we will see, the framework offers a good way of determining which hypotheses to consider. In this section I also summarize an application of Solomonoff induction to the raven paradox, as done by Rathmanner and Hutter (2011). The section closes with a discussion of how to bridge the

gap between Solomonoff induction (where probabilities are assigned to binary strings) and standard Bayesian epistemology, where we assign probability to propositions.

In section 4, I tackle numerous potential problems for Solomonoff induction. First there is the fact that in defining Kolmogorov complexity, we must choose a universal Turing machine, which introduces a kind of language dependence. This might threaten Solomonoff induction's status as objective Bayesianism. Second, Solomonoff's prior is neither computable, nor a proper probability measure. This is obviously problematic when it comes to applications. Another potential difficulty is that Solomonoff induction only considers computable hypotheses. But the biggest obstacle is perhaps this: why prefer simplicity? Solomonoff induction is in a sense a formalization of Ockham's razor. But no justification for this principle has been given. Is there one? I argue that we can adopt Kevin Kelly's work on Ockham's razor in the framework of formal learning theory to justify the principle in terms of retraction efficiency.

In the closing section, I review Solomonoff induction in terms of how well it meets the epistemic criteria listed in the introductory section, and also point toward some questions that can be addressed by further research.

1 Introduction

1.1 Basics of Bayesian epistemology

I will be concerned here with Bayesianism as a *normative* epistemology – as a theory of how we *should* respond to evidence. Others propose Bayesianism as a descriptive epistemology, but I will not touch on those approaches here. Thus it will not be a problem if the recommendations offered here fail to capture how people actually respond to evidence.[1]

Normative Bayesianism comes in many varieties. What unites all of the approaches are the following three principles (Easwaran, 2011, p. 321):

- There is such a thing as credence or degree of belief.
- A rational agent must obey the axioms of probability theory.

[1]For a detailed exposition of Bayesian epistemology, see Joyce (2011). Joyce does an excellent job at explaining the basis of the theory, how philosophers have attempted to justify it, and the various problems it runs into. See also Earman (1992) and Kaplan (1998).

- Beliefs should be updated by way of *conditionalization*.[2]

To say that there is such a thing as degree of belief is simply to say that beliefs aren't necessarily all-or-nothing: we may believe some things stronger than others. Probability theory is one way of making such talk more precise. Here is one common formulation of probability theory. The objects of belief are sometimes taken to be propositions, and sometimes taken to be events. If we formulate probability theory in terms of a Boolean algebra, both of these interpretations are valid. We have a Boolean algebra Ω which is closed under countable disjunction and negation. The propositions or events are elements of Ω. A probability function on Ω is a mapping $P : \Omega \mapsto \mathbb{R}$ that satisfies the following:

> *Normality*. For any $A \in \Omega, P(A \vee \neg A) = 1$ and $P(A \wedge \neg A) = 0$.
>
> *Finite Additivity*. $P(A \vee B) + P(A \wedge B) = P(A) + P(B)$.
>
> *Continuity*. If $A_1 \subseteq A_2 \subseteq A_3, \ldots$ is a countable sequence of elements such that $A = \vee_n A_n$, then $P(A_n)$ converges to $P(A)$.

With these axioms in place, it follows that $P(A) = 1$ if A is a logical truth, $P(A) = 0$ if A is a contradiction, and $P(A) \leq P(B)$ if A entails B.

Of course, these axioms only define an abstract mathematical structure. The distinctively Bayesian claim is that our degrees of belief should meet these requirements, and that they should respond to evidence by way of *conditionalization*. Most Bayesians believe that conditionalization should be done by way of Bayes' theorem, which states that

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

As applied to probabilities and conditional probabilities at one time, the theorem follows directly from the definition of conditional probability. The Bayesian claim is that this should also govern how agents *update* their beliefs over time, as new evidence accumulates. Put another way, Bayes' theorem applies diachronically as well as synchronically.[3]

Why be a Bayesian in the first place? Typically, Bayesianism is justified via pragmatic "Dutch book" arguments, which show that unless you follow

[2]Actually, the "objective Bayesianism" defended by Williamson (2010) eschews conditionalization in favor of maximum entropy. More on this below.

[3]Many Bayesians think that *Jeffrey conditionalization* is more appropriate, but the difference will not concern us here.

the recommendations of Bayesianism, you will be vulnerable to a combination of bets that entail a guaranteed loss – a Dutch book (Hajek, 2008). But some commentators are not convinced by such a pragmatic justification. We are concerned with Bayesianism as a normative epistemological theory, and hence we should like an epistemic justification. A justification of Bayesianism as a whole is outside the scope of this essay, but see e.g. Joyce (1998), Leitgeb and Pettigrew (2010a) and Leitgeb and Pettigrew (2010b) for attempts to justify it by considerations of accuracy.[4]

Bayesian approaches have been deployed widely in philosophy and elsewhere. What concerns us here is Bayesian epistemology. We need to make sure we distinguish this from other Bayesian approaches. There is the Bayesian interpretation of probability, which holds that probabilities are rational degrees of belief. There is Bayesian confirmation theory, which uses Bayes' theorem to measure the degree of confirmation a piece of evidence gives to a hypothesis. And there is also Bayesian statistics, an approach to statistics that relies heavily on Bayes' theorem.

Of course, Bayesian epistemology has not avoided criticism. One major objection is the so called "problem of the priors". For the Bayesian machinery to get going, a prior probability distribution must first be in place. That is, before we can begin updating on any evidence, we need to have degrees of belief in propositions we know absolutely nothing about. How should these prior degrees of belief be selected? Are there any restrictions, or are all priors equally admissible? In this essay, I will focus on a particular solution that has been proposed. The idea is roughly that prior probability should be assigned in proportion to the simplicity of a proposition, where simplicity can be given a precise formal measure.

It should be noted that the term "prior probability" is sometimes used in different ways. Often, the prior probability of A refers to the agent's probability $P(A)$ *prior to updating on a particular piece of evidence* B. Say that a physicist performs an experiment to estimate the value of some constant. Based on what he's read, and on previous experiments he has performed, the physicist already has a prior probability $P(A)$. Since he is already aware of some evidence $\{B_i\}_{i=1}^{k}$, we really have that $P(A) = P(A \mid B_1 \wedge B_2 \wedge \cdots \wedge B_k)$. This leads us to the other use of the term "prior

[4]Another common proposal is to use *Cox's theorem*. The theorem shows that, under certain conditions, every way of representing a rational belief function is isomorphic to a probability. Of course, those who aren't already Bayesians are free to – and often do – reject the needed conditions. See section 2.2.2 of Joyce (2011) for a more detailed discussion of Cox's theorem.

probability": the probability of *A prior to any updating whatsoever*. This thesis is mostly concerned with the latter problem, but the two are not completely separate. We can use this difference to distinguish between *strict subjectivism* and *empirically informed subjectivism*. The strict subjectivist holds that, so long as the agent's beliefs conform with the axioms of probability theory, there are no restrictions on $P(A \mid B_1 \wedge B_2 \wedge \cdots \wedge B_k)$. The empirically informed subjectivist, on the other hand, believes that the pieces of evidence $\{B_i\}_{i=1}^{k}$ provide at least some further restrictions on $P(A \mid B_1 \wedge B_2 \wedge \cdots \wedge B_k)$, but agrees with the strict subjectivist that any value that respects the axioms of probability theory is allowed for $P(A)$.

There are many forms of Bayesian epistemology. The typology below is based on how they respond to the problem of the priors.

1.2 Subjectivism Without Convergence

For the most extreme subjectivists, prior probabilities must only meet the minimal requirement of obeying the axioms of probability theory. (Chalmers, 1999, p. 133) expresses a worry about this form of subjective Bayesianism:

> Once we take probabilities as subjective degrees of belief [...] a range of unfortunate consequences follow. The Bayesian calculus is portrayed as an objective mode of inference that serves to transform prior probabilities into posterior probabilities in light of given evidence. Once we see things this way, it follows that any disagreement in science must have their source in the prior probabilities held by the scientists. But these prior probabilities themselves are totally subjective and not subject to critical analysis. Consequently, those of us who raise questions about the relative merits of competing theories [...] will not have our questions answered by the subjective Bayesian, unless we are satisfied with an answer that refers to the beliefs that individual scientists just happen to have started out with.

But it does not follow that any disagreement in science must have its source in the prior probabilities held by the scientists. What follows is that any disagreement between *ideal Bayesians* must have their source in the priors. Of course, this might be enough to incriminate subjective Bayesianism. If subjective Bayesianism is the normative model for epistemic reasoning, and this model explains disagreement by referring to priors that may be chosen at random, the problem remains. A natural question then arises: just how much of a difference do the prior probabilities make? In the next

section we consider an argument that in the long run, differences in prior probabilities will tend to wash out as more and more evidence accumulates.

Some subjectivists agree with Chalmers' conclusion, but without considering it a reductio. Instead, they will simply contend that this is the best we can do, and that there's no need to hope for something more.

1.3 Subjectivism With Convergence

Other subjectivists think that we can do better. Sure, any prior goes, but in the end the choice of prior will not matter much. As enough evidence accumulates, Bayesians will converge on the same answer regardless of their priors. Such convergence results exist in various forms. The following exposition builds on Joyce (2011).

Assume that two agents have priors b and c, such that $b(A) > 0$ and $c(A) > 0$ for some hypothesis A.[5] Both agents go through a potentially infinite sequence of learning experiences, involving random variables X_1, X_2, $X_3, \ldots$, which take on a finite number of values. We further assume that both agents agree that the data statements are independent and identically distributed (i.i.d.), conditional on both A and $\neg A$. To say that such data statements are independent conditional on the hypothesis means that if we already know the hypothesis to be true, adding knowledge of previous data statements will not change the probability of any given future data statements. Let d_j denote a data sequence. We are considering the data statement $X_k = x_k$, with $k > j$. Independence conditional on A and $\neg A$ then means that $b(X_k = x_k \mid A) = b(X_k = x_k \mid A \wedge d_j)$ and $b(X_k = x_k \mid \neg A) = b(X_k = x_k \mid \neg A \wedge d_j)$, respectively. By our requirement above, the same also holds if we replace b with c. To say that the data statements are identically distributed conditional on A (or $\neg A$) means that if we know A (or $\neg A$), all of the random variables $X_1, X_2, X_3, \ldots$ will have the same probability distributions. We must also have that b and c are identically distributed, i.e. that $b(X_k = x_k \mid A) = c(X_k = x_k \mid A)$ for each random variable X_k. The same should also hold if we replace A with $\neg A$.This means that both agents must agree about how likely various observations are, conditional on A being true, and conditional on A being false. Given these assumptions, it can be shown that $b_j(h)$ and $c_j(h)$ will converge to the same value with probability one according to both b and c.

[5] In the interest of keeping the exposition simple, I only discuss the case with two agents. But everything in this paragraph also applies when we have more agents.

While applicable in some cases, the i.i.d. assumption is very strong, and there are several cases where it fails to hold. Apart from the i.i.d. requirement, we must also have that all agents agree about the possible hypotheses (i.e. which hypotheses they assign prior probability greater than zero). Furthermore, the agents must agree on the possible data sequences. In any case where at least one of these conditions is not met, convergence is not guaranteed, and we are back to strict subjectivism.

Given all of these strong assumptions that are required, I think it is fair to say that in most real-life cases, the convergence results will not be applicable. Hence if we are to avoid strict subjectivism, something more needs to be said.

1.4 Objective Bayesianism: Equivocation

Williamson (2010) defends a form of objective Bayesianism that is characterized by three norms:

> **Probability**: The strength of an agent's beliefs should be representable by a probability function.
>
> **Calibration**: The agent's degrees of belief should satisfy constraints imposed by her evidence.
>
> **Equivocation**: The agent's degrees of belief should otherwise be sufficiently equivocal.[6,7]

The Probability norm is clearly shared by all forms of Bayesianism. Many Bayesians – and not only objectivists – also accept the Calibration norm. For instance, this is what distinguishes strict subjectivists from empirically informed subjectivists. The Equivocation norm, however, is distinctive of objective Bayesianism. Bot not all objective Bayesians need accept the Equivocation norm. The norm specifies *one* rule for fixing prior probabilities, but there are many other possibilities.

[6]Note that Williamson often uses the term "objective Bayesianism" to refer to the position that accepts all of these three norms. By contrast, the way I use the term it means that there is a uniquely best prior probability distribution (or set thereof) – the correct priors are objectively determined. This is compatible with Equivocation, but also with other methods of assigning prior probabilities.

[7]Equivocation has of course been in use long before Williamson (2010)'s book. Laplace made use of the "principle of insufficient reason", and Keynes called it the principle of indifference. I focus on Williamson's approach here because his detailed exposition makes the position easier to engage with.

Williamson thinks we should make the equivocation rule more precise by using the *maximum entropy principle*. The entropy of a probability function P is given by

$$H(P) = -\sum_{\omega \in \Omega} P(\omega) \log P(\omega). \tag{1}$$

Here, Ω is a finite domain of mutually exclusive elementary outcomes. By elementary, Williamson means that we assign truth value to all atomic propositions, and then form a conjunction. Thus if we have n atomic propositions we have 2^n atomic states. These atomic states are basic in the sense that all other probabilities can be defined in terms of them. Say that we have some background knowledge $\mathcal{E}$, consisting of a set of propositions that the agent takes for granted. This gives us a set $\mathbb{E} \subseteq \mathbb{P}$ of probability functions that are compatible with $\mathcal{E}$. The principle of maximum entropy then tells us to pick a probability function $P_{\mathcal{E}}$ such that $P_{\mathcal{E}} \in \{P \in \mathbb{E} : P \text{ maximizes } H\}$. Ideally, $P_{\mathcal{E}}$ will be uniquely determined, but this is not always so. In such cases, Williamson thinks we are free to choose. [8]

How does the maximum entropy principle work in practice? Say that we have a very simple language, consisting only of the elementary proposition A. We then have two atomic states, A and $\neg A$. Every point on the dotted line in figure 1 represents a possible way of distributing probability between P and $\neg P$.[9] The solid curve represents the entropy H of these possible probability distributions. With only two options, entropy is given by $H(P) = -P(A) \log P(A) - (1 - P(A)) \log(1 - P(A))$. As you can see, the closer to the middle of the dotted line we get, the higher the entropy.

Williamson's objective Bayesianism attempts to solve the problem of the priors via equivocation: when nothing else is known, you should equivocate between the hypotheses under consideration; that is, give them all equal probability. The idea is that this is the best way to represent your uncertainty: if you gave one hypothesis more weight than another one, that would indicate that you had some evidence favoring the former. But you do not, and therefore you should give all of them the same probability.

At any step in time, an agent that follows Williamson's branch of objective Bayesianism will use the propositions she takes for granted to calibrate her probabilities as much as these propositions allow, and then equivocate using the maximum entropy principle. This way, there is no clear distinc-

[8] For the purpose of applying the maximum entropy principle, $0 \log 0$ is defined to be 0.

[9] I have here recreated a figure used by (Williamson, 2010, p. 49).

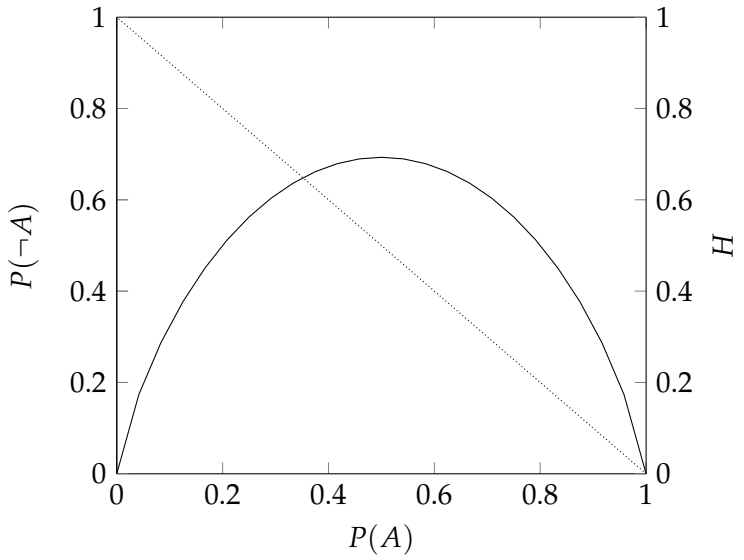

Figure 1: Probability distributions over two propositions (dotted line) and their entropy (solid line).

tion between prior and posterior probabilities. It turns out that this way of updating is actually inconsistent with standard Bayesian conditionalization, and that there are cases where Williamson's updating can be applied and Bayesian conditionalization cannot. There are four cases in which the maximum entropy principle and Bayesian conditionalization yield different results: (i) when you update on a sentence that is not in the original language, (ii) when you update on a sentence that is not "simple" with respect to the sentences the agent takes for granted,[10] (iii) when the set of sentences the agent takes for granted after updating are inconsistent, and (iv) when a conditional probability doesn't satisfy the constraints imposed by sentences the agent takes for granted. In many of these cases it is not clear how conditionalization should be applied. By contrast, maximum entropy updating can be used in all of them.[11]

This essay is not an investigation of Williamson's objective Bayesianism, but I will note three potential problems for his approach. The first is that, since maximum entropy updating will sometimes disagree with Bayesian conditionalization, the updating procedure cannot be justified by a prag-

[10] A sentence A is simple with respect to the sentences the agent takes for granted if the only constraint it imposes with respect to these sentences is that $P(A) = 1$.

[11] See Williamson (2010), p. 78 for a more detailed discussion of the difference between maximum entropy and Bayesian conditionalization.

matic Dutch book argument. An agent using maximum entropy updating will sometimes be vulnerable to a Dutch book.

The second problem has to do with the justification of the equivocation norm. The principle is sometimes justified by symmetry considerations. If we do not have any evidence relevant for whether or not A is true, the norm tells us to set $P(A) = P(\neg A) = 0.5$ which will in fact add further information by describing a state of incomplete evidence with a single function. Starting from a state of complete ignorance, equivocation gives us the precise form of the distribution, rather than considering a set of possible distributions.

The third problematic feature of Williamson's objective Bayesianism is that it is language-relative. Here is an example from Halpern and Koller (2004). Assume first that an agent's language only has one propositional variable, C, which asserts that a particular book is colorful. The agent has no further information, so by equivocation she sets $P(C) = P(\neg C) = 0.5$. Now assume instead that we have a richer language with three propositional variables. R denotes that the book is red, G that it is green, and B that it is blue. With no further evidence, equivocation yields $P(G) = P(\neg G) = 0.5$, and similarly for R and B. $\neg C$ is equivalent to $\neg R \wedge \neg G \wedge \neg B$. But from the way we have just assigned prior probabilities, it follows that $P(\neg R \wedge \neg G \wedge \neg B) = \frac{1}{2^3} = \frac{1}{8}$. Thus the prior probability of a hypothesis will depend on the language we use to represent it. So much for objectivity, one might think. Yet Williamson argues that the terms that exist in natural language give us some guide to how the space of outcomes could usefully be partitioned.

A peculiar feature of Williamson's use of the equivocation norm is how it interacts with the calibration norm. First, Williamson thinks we should calibrate our credence with the available evidence. After this, any remaining freedom in the choice of credence is taken care of by equivocating as much as the evidence allows. Say that we are considering hypothesis H and its negation $\neg H$. We then learn evidence which imposes the constraint $\mathrm{P}(H) \in [0.6, 0.9]$. At this stage, we use equivocation to set $\mathrm{P}(H) = 0.6$, because this is the probability function compatible with our evidence that is closest to the equivocator $\mathrm{P}_{Equiv}(H) = \mathrm{P}_{Equiv}(\neg H) = 0.5$. Another natural reading of equivocation would be to, in light of the evidence considered above, set $P(H) = (0.6 + 0.9)/2 = 0.75$, but Williamson prefers the previous option.

The topic of this essay, Solomonoff induction, is another form of objective Bayesianism. Whereas Williamson uses equivocation, in Solomonoff induction we should assign prior probability proportional to the simplicity of a hypothesis. To the best of my knowledge, philosophers have barely dealt with Solomonoff induction at all, or at least not in print. There are a few brief mentions here and there, but no systematic treatment has been given of the philosophical aspects.[12] In section 2 I introduce the formal basics of Solomonoff induction.

1.5 What Should a Solution Achieve?

Given all of this disagreement, one could take a step back and ask what a solution to the problem of the priors should achieve. Typically it will be claimed that some set of priors are more justified than others – in the case of objective Bayesianism – or that all priors are equally justified – in the case of subjective Bayesianism. Thus, a solution to the problem will tell us which set of priors, if any, is the most justified. Of course, not everyone's standards of justification are the same, and so people might want different things from a solution. In what follows, I will speak of truth instead of justification. I am interested in finding out whether Solomonoff induction offers a connection between prior probabilities and truth that is lacking in other frameworks. I won't argue that prior probabilities should be evaluated in this light, but simply start my investigations with this assumption. Thus if some set of priors offer a better connection to truth than others, objective Bayesianism is correct. If there is no such set of priors, instead subjective Bayesianism is correct.

How could a prior be connected to truth? There are several possibilities. It is important to get clear on these, so that we are all evaluating priors by the same lights. We can distinguish between three main categories of potential connections between prior probability and truth. First, it may be that the prior is more truth-like to begin with: by using this prior, the agent will already from the start be closer to the truth than if she'd used a different prior. Second, it may be that the prior allows an agent converge on the truth quicker than other priors. Third, the agent should be able to use the prior in practice. Here are some candidates for evaluating priors, grouped under these three categories:

1. **Truth-likeness**

   (a) Every true hypothesis (or a majority of them) are assigned higher prior probability than their respective negations.

[12] See for instance Ortner and Leitgeb (2011) or Kelly (2004).

(b) The average distance to truth is smaller than 0.5.

2. **Convergence on Truth**

   (a) It lets us consider more hypotheses. The more hypotheses we consider (i.e. assign probability strictly greater than zero), the smaller risk we run of ruling out the true hypothesis.

   (b) The prior guarantees convergence on the true hypotheses in the limit of observation.

   (c) The prior guarantees that agents get "sufficiently close" to the true hypothesis after a certain number of updates.

3. **Applicability**

   (a) The prior is computable.

   (b) It can be used by computable but idealized agents.

   (c) It can be used by actual humans.

These criteria can be in conflict. For instance, increasing the number of hypotheses under consideration may make a problem less tractable for an actual human reasoner. Similarly, by increasing the number of hypotheses under consideration, we will assign lower prior probability to some other hypotheses, and these may turn out to be true.

I have only included criteria for evaluating priors from the viewpoint of an individual epistemic agent. But viewed collectively, it may very well be the case that some other choice is better. Even if it turns out that some prior distribution is objectively best for an individual, it may be that having a certain amount of disagreement better advances science, and hence leads to better posteriors for the individual as well, provided she adheres to the opinion of mainstream science.

If there is no prior distribution (or set of such distributions) which is clearly superior in all of these respects, but, say, some prior that does better along one dimension, and some that does better along another, we will have to decide which one to give more weight. But which one to give more weight may be obvious if we have particular applications in mind. For instance, if we are looking for something to use in practice, a prior that is uncomputable clearly won't do, even if it scores high on other items.

If it is impossible to find a prior that does better than at least one other in at least one of these respects, and worse in none of them, I will take

that to show that subjective Bayesianism is correct. But note that we can have varying degrees of objectivity. In the most extreme form of objective Bayesianism, the correct prior is uniquely determined: there is one single prior probability distribution which is strictly superior to all the others. At the other end of the spectrum, extreme subjectivism holds that all priors are equally good, provided that they conform with the axioms of probability theory. Between these two extremes, there are several possibilities. For instance, one could take extreme subjectivism and add the requirement that all hypotheses are assigned strictly positive probability. This restriction is typically seen as too weak to qualify the resulting view as a form of objective Bayesianism. But we needn't worry too much about terminology at this point. The interesting question is how strong constraints there are on prior probabilities, regardless of whether these constraints are strong enough to justify the label "objective."

Let's look at the desiderata in more detail. The two items listed under truth-likeness are clearly unrealistic. If we had a method that let us assign probability > 0.5 to all and only the true hypotheses, we might as well go all the way and give them probability 1. But these are *prior* probabilities, and such knowledge cannot be acquired *a priori*. If it were, there wouldn't be any need for inductive reasoning in the first place. The recommendation to assign probability 1 to all and only true hypotheses is not particularly helpful. Thus any satisfactory solution to the problem of the priors must provide us with a method that can be used to assign probabilities *a priori*. Call this the *helpfulness condition*. Even if we relax the first item in truth-likeness and hold only that a majority of the true hypotheses must be assigned higher prior probability than their respective negations, the helpfulness condition cannot be met. The same is true of the second item.

If we rule out all forms of truth-likeness, any potential connection between prior probability and truth must have to do with how the prior probability helps the agent in getting closer to truth. Put another way, the question of how to assign prior probability is methodological, not metaphysical. Such items are listed under the heading convergence on truth.

Item 2a has to do the scope of belief. Since we cannot get from probability 0 to any positive probability by Bayesian conditionalization, assigning probability 0 to a hypothesis means that we are ruling it out *a priori*.

Item 2b has to do with ideal learnability. We want to, at least in the limit,

be able to believe fully, i.e. assign probability 1, to those hypotheses that are true. Note that 2b is logically stronger than 2a. If we are guaranteed to converge on the true hypothesis in the limit, whatever the true hypothesis happens to be, this means that we must assign non-zero probability to every logically consistent hypothesis. But the reverse doesn't necessarily hold: even if we consider all logically consistent hypotheses, we needn't converge on the truth in the limit. Item 2c deals with the same issue as 2b, but from a less idealized perspective. What counts as "sufficiently close" is of course vague, but can be made more precise if we hook it up with a particular decision procedure.[13]

The last three items, listed under applicability, have to do with increasing degrees of practical usefulness. If a proposed solution is uncomputable and so fails to meet 3a, it is a solution only in a very abstract sense. However, even uncomputable priors – such as the one considered in this essay, Solomonoff's prior – may be illuminating if they for instance place an upper bound on how well a Bayesian agent can perform. In some cases, uncomputable priors can be approximated, and such approximations may provide practical guidance. Item 3b gets us one step closer to applicability. If this criterion is met, the prior could be used to construct an artificial reasoner, for instance.

Some priors, while computable, require such computational resources that they cannot be used by humans. Thus item 3c requires that they are easy to use. If two priors score alike in all other aspects, but differ in how computationally tractable they are, then the one that is easier to use is to be preferred, if we are interested in practical applications by humans.

It should be noted that all three items listed under applicability can also be used by a subjective Bayesian to argue for the choice of a certain prior. After all, if prior probabilities can be picked as we like, why not go with ones that are easy to compute with? Thus the fact that there clearly are priors that are more easy to use for humans to use than other priors, will not in itself count in favor of objective Bayesianism. Instead, it should be used to choose a prior from the set of priors that perform well on other criteria, provided that there is such a set.

There is an important interaction between the alleged convergence results

[13]For instance, say that an agent only takes some risky action when her degrees of belief in the relevant propositions are sufficiently high. We would then want her to be able to reach this degree of belief in reasonable time.

and objective Bayesianism. The stronger the convergence results, the less important it becomes to determine the objectively best prior, if such there be. If the convergence results are very strong – and do in fact apply in everyday cases – the choice of prior won't matter much. But as we saw in the discussion of these convergence results in section 1.3, there is reason to think that they do not apply very often.[14] There is also the possibility of a form of objective Bayesianism which says that our priors should be such that the convergence results apply. But unless there are independent reasons for favoring such priors, this reasoning would be circular. Moreover, due to the nature of the convergence results, such priors would also require *a priori* coordination among agents – we saw in the discussion in section 1.3 above that agents must have the same likelihoods. But it appears impossible for agents to actually achieve such coordination without communication, and so the proposal doesn't meet the helpfulness condition.

With the above criteria in mind, let's look at how Williamson (2010) tries to justify his use of the maximum entropy principle. First, he goes through several previous proposals, and concedes that these are unlikely to convince anyone who is not already on board with maximum entropy. One such argument assumes that our degrees of belief should be constrained by evidence, but otherwise maximally non-committal. Such degrees of belief are given by the maximum entropy principle. But even if one agrees that this principle does give the maximally non-committal degrees of belief, why should one think that such degrees of belief are desirable in the first place?

In the end, his main argument is an argument from caution: having more extreme degrees of belief than those advocated by the maximum entropy principle will trigger action more often. As Williamson is aware, this is a pragmatic justification. But he argues for a strong link between belief and action.

However, the maximum entropy probability function is not the most cautious in every situation, nor is it the uniquely most cautious on average, as Williamson shows. Instead, Williamson argues that the maximum entropy probability function is the most cautious where it matters most: when it comes to risky decisions. In such decisions, we typically have a high trig-

[14]While it doesn't appear to be a very common position, at least in theory one could hold both that the convergence results apply widely, *and* that there is a uniquely correct set of priors.

ger level: since so much is at stake, we want to be very certain that we are correct before taking action. When the trigger level is high, the maximum entropy probability function may still be the most cautious one.

One might agree that caution is good in many cases. But why should caution enter into your credence? The fact that an outcome would be bad is not a reason to give it high probability. Instead, this looks like the very definition of pessimism. The problem with making the link between rational degrees of belief and action too tight, is that considerations that are clearly non-epistemic will enter the picture. Thus we may reject the Calibration norm, on the ground that in many cases, being overconfident will help you better achieve your goals. That is: even though beliefs are connected to actions, the aims of epistemic and instrumental rationality may nevertheless differ.

Because he doesn't use conditionalization, the number of hypotheses that are assigned non-zero prior doesn't matter in Williamson's framework. Even if a true hypothesis is initially given probability zero, Williamson's objective Bayesian can still converge on it. Thus Williamson's framework performs well on item 2a.

2 Solomonoff induction

Hutter (2007) describes Solomonoff induction as the combination of Epicurus, Ockham, Bayes, Turing and Kolmogorov. That is, Solomonoff induction respects Epicurus' idea of keeping all potential explanations that are consistent with the data, while also following Ockham's razor in the sense of giving greater weight to simpler hypotheses. The probabilities assigned to hypotheses are updated by Bayes' theorem, and the whole thing is computed by Turing machines, which also allows us to quantify simplicity by way of Kolmogorov complexity.[15]

2.1 Universal Turing machines

To define Kolmogorov complexity we first need to make sure we know what a universal Turing machine is, so here is a brief recap.[16] A *Turing machine* (TM) consists of the following: a finite set Q of states, with an

[15]See Hutter (2007) for a rather technical introduction to the topic, and Rathmanner and Hutter (2011) for a more accessible one. Solomonoff (1964a) and Solomonoff (1964b) are the first published sources of the theory.

[16]For a broader discussion in a philosophical context, see Barker-Plummer (2011)

identified starting state, and a finite set Σ of symbols – the alphabet. At any particular step s of the execution

- $Q_s \in Q$ is the state the Turing machine is in,
- $\sigma_s : \mathbb{Z} \mapsto \Sigma$ is a function that describes the contents of each of the cells of the tape. The function maps the index (an integer) of a cell to a particular symbol of the alphabet.
- h_s is the index of the cell being scanned.

We also have a table of transition functions $\delta : Q \mapsto Q$ such that if $\delta(S) = T$ then

- σ_T is the same as σ_S everywhere apart from h_S (and possibly there too).
- If $\sigma_S(h_S) \neq \sigma_T(h_S)$, then $h_T = h_S$, otherwhise $|h_T - h_S| \leq 1$.

The first of these constrains the transition function so that it may only change the state of the current cell. The second constrains the transition function so that, if it does change the state of the current cell, the index must remain the same. If it does not change that state, it may move at most one step in either direction.

A transition function is defined by the quadruple $\langle Q_s, \sigma_s(h_s), Q_t, A \rangle$. If the machine is in state Q_s and the current cell contains the symbol $\sigma_s(h_s)$, move into state Q_t, taking action A. If there arises a situation with no unique transition rule, the Turing machine halts. Otherwise, it finds the transition rule that fits the current situation and carries on.

A *universal Turing machine* (UTM) is a Turing machine that can simulate the behavior of any Turing machine. If a UTM receives input starting with a specification of a particular Turing Machine T, followed by instructions for T, it will produce the same result as T would when given the same instructions. This is the notion we need to define Kolmogorov complexity.

2.2 Kolmogorov complexity and Solomonoff's prior

Consider the following binary strings.

1111111111

1100100111

Both are ten digits long, and are equally likely to represent the outcome of ten flips of a fair coin. But there is a visible difference in the complexity of describing the two. In English we can describe the first string as "1 ten times." No similarly short description is available for the second one. This idea of description complexity is the basis of Kolmogorov complexity. But natural language is imprecise, so we turn instead to universal Turing machines to quantify complexity.

Suppose that we have a fixed UTM. For any given string, there will be several programs we could run on the UTM that would generate that string. The key idea is that the complexity of a string is given by the length of the *shortest* program that will generate the string. The length of a program is measured by the number of bits it contains. Thus if $x = x_1 x_2 \dots x_n$, where each x_i is either 0 or 1, then we have that $l(x) = n$.

With this in place, the Kolmogorov complexity of an infinite string x is given by

$$K(x) := \min_p \{\text{length}(p) : U(p) = x\}, \tag{2}$$

where U is a given UTM, and p a variable that ranges over the set of programs such that when U is applied to them, produce x as the beginning of their output.[17] If no such program p exists, we set $K(x) := \infty$. For finite strings x, we define $K(x)$ to be the length of the shortest program that outputs x and then halts. We may also define the *conditional* Kolmogorov complexity:

$$K(x \mid y) := \min_p \{\text{length}(p) : U(y, p) = x\}$$

Here we get the length of the shortest program that outputs x given that it receives y as an extra input.

Suppose that we want to generate a prior probability distribution over binary strings. One way of doing so would be to respect Ockham's razor and give higher prior probability to simpler strings. There's a way of doing this which is closely related to Kolmogorov complexity, *Solomonoff's prior*:

$$M(x) := \sum_{p : U(p) = x*} 2^{-l(p)}. \tag{3}$$

The sum is over all programs whose outputs start with x – denoted $x*$ – when they are applied to U, and $l(p)$ denotes the length of program p. As you can see from the definition, we do not just take the Kolmogorov

[17] To be precise, we need to assume that U is a *prefix UTM*, which means that no valid program for U is a prefix of any other.

complexity of x as our prior. Instead, we sum over *all* programs whose output start with x. This way, how much a program p contributes to the value of $M(x)$ depends on its length $l(p)$. If $l(p)$ is large, $2^{-l(p)}$ will be small. Thus shorter programs will contribute more to the value of $M(x)$ than long ones do.[18]

Why consider programs whose outputs start with x, and not only programs that output x and then halt? We are interested in making predictions about how the string that starts with x continues. One possibility is of course that it doesn't, which is taken care of by the programs that halt at this step. But if the string does continue, we look at programs that generate further output.

All of this may look fine and interesting, but it turns out that M is not a proper probability measure. To see why, we first need some notation. We let ϵ denote the empty string. By definition, every string begins with ϵ. Thus any probability measure must meet the requirement that $P(\epsilon) = 1$. This is not the case with M, where we instead have that $M(\epsilon) \leq 1$. Furthermore, we have that $M(x) \geq M(x0) + M(x1)$, where x is some initial segment of an infinite binary string, and $x0$ and $x1$ denote the string x with a 0 or 1 appended at the end, respectively. Since 0 and 1 are the only letters of the alphabet, this means that the value given to x by M may be larger than the sum of all possible continuations of x. Functions that meet these requirements are called *semi-measures*. All probability measures are semi-measures, but not all semi-measures are probability measures. This is because probability measures have the additional requirement that the inequalities above are equalities. One can think of a semi-measure that is not a probability measure as a kind of deficient probability measure where the probabilities do not add up as they should. If a semi-measure is computable, one can easily turn it into a probability measure by normalization (i.e. multiplying all terms by a constant so that they will sum to 1). Alas, M is not computable.[19] This can clearly be problematic. In fact, it violates one of the desiderata I listed in section 1.5 on what a solution to the problem of the priors should achieve. I will discuss this in the criticism section.

[18]There are other priors, such as the Solomonoff-Levin prior (see e.g. (Legg, 2008, p. 32)) that are mathematically very closely related to M. But for purposes of clarity, I'll leave those out of this discussion.

[19]Despite the fact that M is not a probability measure, authors typically speak of M as assigning *probabilities* to strings. For convenience I will use this language as well, hoping that it doesn't serve to cover up what may be a serious problem with Solomonoff's prior. In section 4.2 we return to the fact that M isn't a probability measure, to see just how serious a problem it is.

Solomonoff's prior is not computable, but it is *lower semi-computable*, meaning that it can be approximated from below by a computable function. The reason for why Solomonoff's prior is not computable is that when trying to decide whether a given program p outputs the string x, we will run into the halting problem.

The length of the shortest program will critically depend on which UTM is chosen. It can be shown that for any arbitrarily complex string x, as measured against the UTM U, there is another UTM U', for which the string has Kolmogorov complexity 1. We therefore need a method for choosing the particular UTM that is to be used.

So far we have only spoken of the Kolmogorov complexity of strings. But we shall soon make use of the Kolmogorov complexity of a semi-measure. This can be defined in the following way. We take all lower semi-computable semi-measures, and give them an index. The index i of a lower semi-computable semi-measure μ_i is in effect a description: given x and i there exists a Turing machine T such that $\mu_i(x) = \lim_{k\to\infty} T(x, i, k)$. That is, T can approximate the value of $\mu_i(x)$ for any x. To define Kolmogorov complexity of a semi-measure μ_i we take the Kolmogorov complexity of its index:

$$K(\mu_i) := \min_p \{\text{length}(p) : U(p) = i\}.^{20}$$

When Solomonoff's prior is applied to an inductive problem, the framework is known as Solomonoff induction.[21]

2.3 Some properties of Solomonoff's prior

Perhaps the main reason why several computer scientists are excited by Solomonoff's prior is that it has many nice theoretical properties. In particular, the three theorems discussed in this section are often given as reasons for why Solomonoff induction is a good framework.

We can measure the error of a semi-measure ρ by the negative logarithm of the probability it assigns to the sequence x that actually occurs: $-\log \rho(x)$. Since the outcome that actually occurs is unknown, it makes sense to compare the worst-case outcomes of two semi-measures. If the error of a semi-measure μ is at most a constant times larger than the error of another

[20] See (Legg, 2008, p. 33) for a more detailed discussion of how to assign Kolmogorov complexity to semi-measures.

[21] The standard textbook on Kolmogorov complexity is Li and Vitanyi (1997).

semi-measure ρ, we say that μ dominates ρ. How does Solomonoff's prior $M(x)$ fare in this regard? It dominates *all* lower semi-computable semi-measures.

Theorem 1. *Solomonoff's prior $M(x)$ dominates all lower semi-computable semimeasures in the sense that*

$$c \cdot 2^{K(\mu)} \cdot M(x) \geq \mu(x)$$

if μ is a lower semi-computable semimeasure. (Hutter, 2004, p. 46)

The term $c \cdot 2^{K(\mu)}$ measures the largest amount by which the dominated semi-measure $\mu(x)$ may outperform $M(x)$. As is clear, the term $c \cdot 2^{K(\mu)}$ depends only on the Kolmogorov complexity of the semi-measure μ, and not on the string x. The higher the Kolmogorov complexity of μ, the more it may outperform $M(x)$. Given the definition of $M(x)$, this is natural: $M(x)$ is biased toward low Kolmogorov complexity.

The next two theorems establish bounds on the prediction error of M. Assume first that the true distribution μ is deterministic. This means that at each step in the sequence, all of the distribution is concentrated on a single character of the alphabet (i.e. 0 or 1 in the binary case). Given such a distribution, the following theorem holds (Hutter, 2004, p. 47):

Theorem 2.

$$\sum_{t=1}^{\infty} |1 - M(x_t \mid x_{<t})| \leq \frac{1}{2} \ln 2 \cdot K(x_{1:\infty})$$

Here $M(x_t \mid x_{<t})$ is the probability that the complete string is x_t given that it starts with $x_{<t} = x_1 \ldots x_{t-1}$, a particular string of length $t-1$. $K(x_{1:\infty})$ is the Kolmogorov complexity of the infinite string $x_{1:\infty}$. If $x_{1:\infty}$ is computable, $K(x_{1:\infty})$ will be finite. If this is the case, the infinite sum on the left hand side is finite, and hence the terms $|1 - M(x_t \mid x_{<t})|$ must converge to zero as $t \to \infty$. This means that as more and more digits of the string x_t are revealed, the prediction of how it will continue converges on the correct answer. Moreover, the speed of this convergence depends only on the Kolmogorov complexity of the infinite string $x_{1:\infty}$.

This result can be generalized to the case of arbitrary computable semi-measures, in addition to the deterministic case we just considered. Assume that the true (computable) objective probability distribution is μ. We then have the following theorem (Hutter, 2004, p. 48):

Theorem 3.

$$\sum_{t=1}^{\infty} \sum_{x_{<t} \in \mathbb{B}^{t-1}} \mu(x_{<t}) \Big(M(0 \mid x_{<t}) - \mu(0 \mid x_{<t}) \Big)^2 \leq \frac{1}{2} \ln 2 \cdot K(\mu) + c < \infty.$$

The notation is quite dense, so let's try and unpack it. The term $\mu(x_{<t})$ denotes the objective probability (the μ-probability) that the binary string starts with $x_{<t}$. $M(0 \mid x_{<t})$ is the M-probability that the next digit is 0, given that the string begins with $x_{<t}$. Similarly, $\mu(0 \mid x_{<t})$ is the μ-probability that the next digit is 0, given that the initial string is $x_{<t}$. The term $K(\mu)$ denotes the Kolmogorov complexity of the objective probability distribution μ, and c is a constant. Since we have assumed that μ is computable, it follows that $K(\mu)$ is finite, and consequently that $\frac{1}{2}\ln 2K(\mu) + c$ is also finite.

For each binary string $x_{<t}$, there is an objective probability $\mu(x_{<t})$ that true infinite string begins with this string. The fact that the squared difference is multiplied with this term $\mu(x_{<t})$ means that proportionally greater weight is given to those binary strings $x_{<t}$ that have higher objective probability. The inner sum adds the term $\mu(x_{<t})(M(0 \mid x_{<t} - \mu(0 \mid x_{<t}))^2$ for all 2^{t-1} binary strings of length $t-1$. The outer (infinite) sum repeats this process for all natural numbers.

Again, the only way for an infinite sum to be finite is if its terms tend to zero. In our case, this means that the difference $M(0 \mid x_{<t}) - \mu(0 \mid x_{<t})$ must tend to zero as $t \to \infty$ with μ-probability 1. So Solomonoff's prior M will converge to the true objective probability distribution.

That is: the total number of prediction errors over an infinite sequence is bounded by a finite constant that depends only on the nature of the true objective probability distribution. Thus, for probability distributions with higher Kolmogorov complexity, the error bound will be larger, reflecting the bias toward simpler hypotheses.

Now, what is meant by an "objective" probability distribution here? According to Hutter, a distribution μ is objective if this is the distribution from which the true sequence is drawn. We can think of this in terms of limiting frequency, or as the probability distribution used by some other agent to generate a binary string.

2.4 Reparametrization and regrouping invariance

Typically, there will be many ways to describe the possible events that are to be assigned probabilities. In some cases, using different descriptions and then equivocating will lead us to assign different probabilities to

what are in fact the same events. Consider the following example, adapted from van Fraassen (1990).[22] A factory produces cubes with side-lengths between 1 and 3 cm. What is the probability that a randomly selected cube has a side-length between 1 and 2 cm? One might be tempted to think that equivocation gives the answer 0.5: if all side-lengths are equally probable, we simply take $\frac{2-1}{3-1} = 0.5$. But the factory can also be described as producing cubes with volumes between 1 and $3^3 = 27$ cm^3. A side-length between 1 and 2 cm corresponds to a volume between 1 and 8 cm^3. So, according to this description, equivocating should lead us to assign probability $\frac{8-1}{27-1} = \frac{7}{26}$ to the same event. What gives?

One option is to become a subjective Bayesian. If the result of equivocation depends so critically on how we describe the outcomes, then perhaps no prior is better than any other. As we have seen, however, Williamson (2010) accepts this language dependence and maintains that the language an agent uses gives her some useful information about the world, which is then used to justify the use of this particular language. His branch of objective Bayesianism is objective in the sense that once we have decided on the particular language to use, the probability distribution will be objectively determined. A third option is to look for a prior distribution that is not vulnerable to the problem.

It turns out that Solomonoff's prior isn't vulnerable, as shown by Hutter (2007). The technical term for this is *reparametrization invariance*. Strictly speaking, Solomonoff's prior does not satisfy reparametrization invariance: rather, it is invariant up to a multiplicative constant. This means that if we replace x with $f(x)$, there is a constant c (which depends on f but not on x) such that $M(x) = c \cdot M(f(x))$ Furthermore, the result established by Hutter only holds for "simple" transformations, i.e. transformations f such that their complexity $K(f)$ is constant (i.e. doesn't depend on x).[23]

Whereas reparametrization involves a bijective function so that all instances of the transformed parameter correspond to one and only one instance of the original parameter, this need not be the case in regrouping. An example might help here. Suppose that we are to determine whether a coin is fair or biased. We might represent this by two hypotheses: {fair,

[22]Similar phenomena are known as the *Bertrand paradox*. In the original formulation, we are asked to consider an equilateral triangle inscribed in a circle. A chord of the circle is chosen at random. What is the probability that the chord is longer than a side of the triangle?

[23]We can define the Kolmogorov complexity of such transformations f in much the same way as we did for lower semi-computable semi-measures above.

biased}. But equally, we could represent it with three hypotheses: {fair, heads biased, tails biased}. In the first case, equivocating between the hypotheses gives a probability of $\frac{1}{2}$ to the coin being fair. In the second case, the same method assigns a probability of $\frac{1}{3}$ to the same thing. In this case of regrouping, the same result holds: for simple group transformations (i.e. when we transform one way of grouping the hypothesis to another grouping using a transformation with constant complexity $K(f)$), M is invariant up to a multiplicative constant. Both reparametrization invariance and regrouping invariance are about avoiding a form language dependence: we do not want the probability that we assign to a given hypothesis to depend on the language that is used to represent it.

It should be noted that a proponent of equivocation does in fact have a response to the cube scenario considered at the beginning of this section. We should equivocate neither over side-length nor over volume, because these are arbitrary units of measurement. Instead, we should equivocate in a way that makes our probability assignment invariant. In this case, the solution is $\frac{\ln 2 - \ln 1}{\ln 3 - \ln 1} = \frac{\ln 2}{\ln 3}$. If we consider the volume instead, we get $\frac{\ln 8}{\ln 27} = \frac{\ln 2^3}{\ln 3^3} = \frac{\ln 2}{\ln 3}$. So we apply equivocation not to the side-length or volume, but to their logarithm. The same method has been used to deal with similar problems. The main idea behind this solution is that before assigning prior probabilities, we should identify the relevant symmetries. This is supposed to be justified *a priori*, but Joyce (2011) discusses cases where the units of measurement could in fact matter.

3 Applications

In this section we'll look at three things. First, we shall see that Solomonoff's prior solves some problems to do with which hypotheses to consider from the beginning, which is difficult for many other priors. Second, I'll give a more concrete illustration of how Solomonoff induction might work in practice. Rathmanner and Hutter (2011) apply the framework to the raven paradox, and I will borrow their example. Third and finally, I'll look at how one might get from the binary strings that Solomonoff induction considers, to the propositions that are typically taken to be the objects of credence in Bayesian epistemology.

3.1 Determining the model space

By telling us how to assign prior probabilities, Solomonoff induction solves another problem as well: it determines which hypotheses to consider. The

prior M gives non-zero probability to every possible computable binary string. Moreover, as we saw in theorem 3, M can get arbitrarily close to any computable semi-measure on binary strings.

The fact that M gives non-zero probability to every possible binary string solves another, related problem: how to assign probability to the sentences of a countably infinite language. For a finite set of discrete outcomes, we use a probability mass function to assign a non-zero probability to every possible outcome. For a variable that takes on continuous values, we use a probability density function. But for an infinite set of discrete outcomes, many standard ways of assigning probability have problems. For instance, equivocating between each outcome in an infinite set leads us to assign zero probability to every outcome. This even violates the probability axioms: the condition of continuity isn't met. Assume we have elementary outcomes $A_1, A_2, \ldots, A_n$. We define the sequence $\{B_n\}_{n=1}^{\infty}$ by setting $B_1 = A_1$ and $B_n = B_{n-1} \vee A_n$. Then we have that $\{B_n\}_{n=1}^{\infty}$ is a countable sequence such that $\vee_n A_n = \Omega$. By definition, $P(\Omega) = 1$. Convergence requires that for every real number $\epsilon > 0$, there exists a natural number n_0 such that for all $n > n_0$, $|1 - B_n| < \epsilon$. But this is not the case, so continuity is not met.

This is not the case with M: each of the countably infinite number of binary strings will have a non-zero probability. Thus, someone who favors equivocation in the finite case will have to switch a method for assigning prior probability in the countably infinite case. Not so for Solomonoff's prior, which works in both cases.

3.2 The Raven paradox

Hempel's "raven paradox" is one of the most famous problems in confirmation theory, so it might be instructive to consider how Solomonoff induction deals with it. The statement "All ravens are black" is logically equivalent to "Every non-black thing is a non-raven". Observing a black raven is evidence for the first statement. Observing a red apple is evidence for the second statement. But since they are logically equivalent, anything that is evidence for one of them must also be evidence for the other. Hence any observation of a non-black non-raven will support the hypothesis that all ravens are black. This is the paradox: how could such an observation support that conclusion? Rathmanner and Hutter (2011) apply Solomonoff induction to the problem. They proceed as follows.

We are considering two predicates: B for black, and R for raven. This

gives us the following observation language: $\{BR, \overline{B}R, B\overline{R}, \overline{BR}\}$. For each of these, there is a parameter which represents their proportion of the total population: $\vec{\theta} = (\theta_{BR}, \theta_{\overline{B}R}, \theta_{B\overline{R}}, \theta_{\overline{BR}})$. These parameters must all sum to 1. A complete hypothesis is an assignment of values to the parameters that meets this requirement. The hypothesis space can be represented by a 3-simplex, as in figure 2. In this figure, BR denotes an observation sequence consisting only of black ravens, $\overline{BR}$ denotes an observation sequence consisting only of non-black non-ravens, and so forth. We are interested in the partial hypothesis given by "All ravens are black". This is the shaded area in figure 2. Can we confirm this hypothesis with Solomonoff induction?

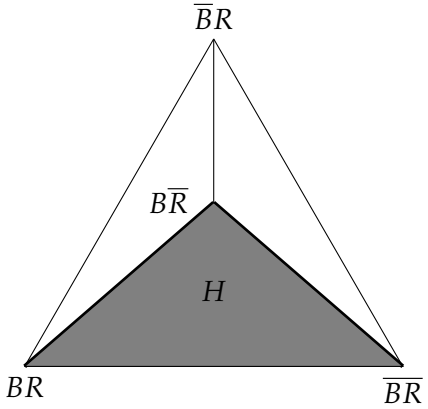

Figure 2: The hypothesis space for the black ravens problem.

The prior assigned depends on the Kolmogorov complexity of the parameters. Hutter and Rathmanner argue that 0 and 1 are the simplest numbers in the interval $[0, 1]$. Therefore, Solomonoff's prior will favor hypotheses that have parameter values of 0 and 1. These are the hypotheses corresponding to each vertex in figure 2 above. Since they are the simplest hypotheses, they will have the highest priors. Hypotheses that lie on one face of the simplex will also be favored, because in these cases, one of the parameters must be zero. This is the case in H. Now H is consistent with any observation stream that doesn't contain any non-black ravens. That is, $H = \{x_{1:\infty} : \forall t\, x_t \neq \overline{B}R\}$. Hutter and Rathmanner claim that Solomonoff's prior converges on the correct answer, that is

$$\lim_{n\to\infty} M(H \mid x_{1:n}) = 1 \text{ if } x_{1:\infty} \text{ is sampled from any } \vec{\theta} \in H.$$

However, this result is a direct consequence of theorem 2, which states that M is guaranteed to converge on the right answer in the limit. Rathmanner

and Hutter *do not* show whether M also gets the absolute and relative degrees of confirmation correct. How much does the observation of a black raven confirm the hypothesis, and how much does the observation of a non-black non-raven confirm it? One common Bayesian solution (Good, 1960) is to admit that while observations of non-black non-ravens do in fact provide some degree of confirmation, that degree is much smaller than the degree provided by observations of black ravens. Whether such a solution also follows from Solomonoff induction is not clear from their example. Thus the treatment given by Rathmanner and Hutter (2011) doesn't deal so much with the *paradox* of the ravens. Instead, they use the example to provide a more concrete illustration of how Solomonoff induction might deal with a particular inductive problem.

One could also imagine a different application of Solomonoff induction to the paradox. If we choose a UTM so that "All non-black things are non-ravens" gets a higher Kolmogorov complexity than "All ravens are black," the latter hypothesis would get higher prior probability than the former. Since the two hypotheses are logically equivalent this could clearly be problematic, but it may nevertheless explain some of our reasoning about the paradox. More generally, the fact that we in this fashion can use Solomonoff induction directly to propositions, rather than to sets of propositions closed under logical equivalence, might be an advantage of the framework. It means that the scope of the framework is wider, even though it would likely be problematic from a normative viewpoint to recommend that we assign different probability to logically equivalent propositions.

3.3 From strings to propositions

In cases like the raven paradox above, it is fairly straightforward to move between the formal framework of strings on the one hand, and hypothesis on the other. At every step in time, there were four possible observations: $\{BR, \overline{B}R, B\overline{R}, \overline{B}\overline{R}\}$. The hypothesis "All ravens are black" was identified with all observation sequences that lack any instance of $\overline{B}R$.

Solomonoff himself clearly intended, or at least foresaw, this kind of epistemological application. In one of the original papers, he writes (Solomonoff, 1964a, p. 14):

> Suppose that all of the sensory observations of a human being since his birth were coded in some sort of uniform digital notation and written down as a long sequence of symbols. Then

> a model that accounts in an *optimum* manner for the creation of this string, including the interaction of the man with his environment, can be formed by supposing that the string was created as the output of a universal machine of random input.[24]

But what about more involved cases? What, for instance, is the Kolmogorov complexity of the hypothesis that the Christian God exists? Of course, the Kolmogorov complexity of anything will depend on which UTM we use to measure it. But assuming we have a fixed UTM, how does one assign Kolmogorov complexity to a hypothesis like the one just mentioned? Stated just like that, the hypothesis is of course hopelessly vague. But once we get more specific, there are several things that are relevant. All else equal, sensory observations indicating the existence of miracles, or that you have reached the afterlife should increase the probability that God exists. If we assume that in the limit of observation we can tell whether or not God exists, then we could identify the hypothesis with the set of sequences that entail His existence. Once this is done, defining Solomonoff's prior for the hypothesis is straightforward.

Assume that propositions, or sensory observations, can be identified with binary strings in this fashion. If these propositions or sensory observations are finite or countably infinite, this is unproblematic. If the number of sensory observations possible at a given time is some finite number n, we could have a UTM with an alphabet consisting of n symbols.[25] Thus any possible sequence of sensory observations could be represented by a string of such symbols. Extending the alphabet in this way is only done for convenience: it does not increase the computational power of the Turing machine. Therefore another way to go would be to identify each possible sensory observation with a (finite) binary string, and represent a sequence of sensory observations with a concatenation of such strings. But how should it be done, exactly? How do we decide which binary string to associate with which sequence of sensory observations?

We have seen that the probability assigned to a binary string by Solomonoff induction depends on the choice of UTM. And in turn, the probability assigned to a sequence of sensory observations depends on how we choose to represent these observations with symbols. Thus we have two separate cases of language dependence.

[24] The idea brings Carnap's *Aufbau* to mind, and Solomonoff was in fact a student of his.

[25] We need only require that the number of sensory observations the subject can discriminate between is finite.

4 Criticism

4.1 Language dependence

In section 2.4 we saw that Solomonoff's prior is invariant under both reparametrization and regrouping, up to a multiplicative constant. But there is another form of language dependence, namely the choice of a universal Turing machine.

There are three principal responses to the threat of language dependence. First, one could accept it flat out, and admit that no language is better than any other. Second, one could admit that there is language dependence but argue that some languages are better than others. Third, one could deny language dependence, and try to show that there isn't any.

For a defender of Solomonoff's prior, I believe the second option is the most promising. If you accept language dependence flat out, why introduce universal Turing machines, incomputable functions, and other needlessly complicated things? And the third option is not available: there isn't any way of getting around the fact that Solomonoff's prior depends on the choice of universal Turing machine. Thus, we shall somehow try to limit the blow of the language dependence that is inherent to the framework. Williamson (2010) defends the use of a particular language by saying that an agent's language gives her some information about the world she lives in. In the present framework, a similar response could go as follows. First, we identify binary strings with propositions or sensory observations in the way outlined in the previous section. Second, we pick a UTM so that the terms that exist in a particular agent's language gets low Kolmogorov complexity.

If the above proposal is unconvincing, the damage may be limited somewhat by the following result. Let $K_U(x)$ be the Kolmogorov complexity of x relative to universal Turing machine U, and let $K_T(x)$ be the Kolmogorov complexity of x relative to Turing machine T (which needn't be universal). We have that

$$K_U(x) \leq K_T(x) + c_{TU}$$

That is: the difference in Kolmogorov complexity relative to U and relative to T is bounded by a constant c_{TU} that depends only on these Turing machines, and not on x.[26] This is somewhat reassuring. It means that no

[26] See Li and Vitanyi (1997, p. 104) for a proof.

other Turing machine can outperform U infinitely often by more than a fixed constant. But we want to achieve more than that. If one picks a UTM that is biased enough to start with, strings that intuitively seem complex will get a very low Kolmogorov complexity. As we have seen, for any string x it is always possible to find a UTM T such that $K_T(x) = 1$. If $K_T(x) = 1$, the corresponding Solomonoff prior $M_T(x)$ will be at least 0.5. So for any binary string, it is always possible to find a UTM such that we assign that string prior probability greater than or equal to 0.5. Thus some way of discriminating between universal Turing machines is called for.

Hutter (2004) argues for the "short compiler" assumption. We should pick our Turing machine from a class which is such that, for any two Turing machines T_1 and T_2 in it, there will always be a *short* program on T_1 that can interpret all T_2 programs. Of course, "short" is ambiguous and needs to be quantified. But regardless of how we quantify it, for any Turing machine T there will be a set of Turing machines that satisfy the short compiler assumption with respect to T. So which members this set has depends not only on how we quantify "short", but also on the choice of the "reference machine" T. How do we decide which Turing machine to use? The recommendation Hutter (2004) offers is that we simply agree upon a fixed reference universal Turing machine and stick with it. Is it possible to do better than that, or are we stuck with this element of arbitrary choice? Rathmanner and Hutter (2011) write that

> the practical and theoretical benefit of having some final fixed reference point outweighs the importance of making this fixed reference point “optimal” in some sense, since it has little practical impact and appears to be philosophically unsolvable.

What’s more, the dependence on the choice of UTM is critical only for short strings. As the length of the string grows, predictions of how it will continue become more and more independent of the particular UTM that is used. By taking prior information in to account, we needn’t consider short strings.

As I discussed in the previous section, when applying Solomonoff induction to real world problems, we need to devise some encoding for translating sentences of natural language into binary strings. This encoding also gives rise to language dependence.

4.2 Incomputability

When I introduced Solomonoff induction in section 2, I explained that the prior M is incomputable. This is obviously problematic when it comes to applications.

In normative ethics, it is often assumed that ought implies can. On this view, if a theory of normative ethics implies that an agent ought to take a certain action, it must also be possible for her to do so. The same is sometimes claimed for epistemology as well: if an epistemological theory implies that an agent should be in a particular credal state, then it must be possible for her to be in that state. If Solomonoff induction is the correct normative epistemology, then we should be able to do what it tells us to do. The problem here is twofold: actual humans cannot accurately follow Solomonoff induction, but neither can even idealized humans, since M is incomputable. The first of these is not unique to Solomonoff induction: all standard forms of Bayesianism require logical omniscience, which is obviously a requirement humans cannot meet. So insofar as one takes this worry seriously, several normative theories other than Solomonoff induction are also ruled out.

Solomonoff's prior *can* be approximated with smaller and smaller errors, but at no point in the approximation can we estimate the size of the error. However, if is often possible to know how much closer to Solomonoff induction one computable measure is than another.

But I think that even if one takes the epistemic "ought implies can" very seriously, Solomonoff induction may still be valuable. If the framework lives up to its other promises, one can view it as an upper bound on how well inductive agents can perform in the absence of luck or other factors that could give prior probabilities substantial truth-likeness. In the words of Solomonoff (1997, p. 83) himself, it is a kind of "gold standard" for inductive systems.

It is important here that we keep the desiderata listed in section 1.5 in mind. Since Solomonoff induction was known to be incomputable from the start, it was clearly never intended to give this kind of direct action-guidance. Thus the framework was developed for other purposes, and shouldn't be judged based on how well it does something it was never intended to do.

One might think that the problematic incomputability would go away if

hypercomputation is physically possible. Alas, hypercomputation cuts both ways. Hypercomputers are devices – so far only theoretical – that can compute functions which are not Turing computable.[27] For instance, Turing himself established that the halting problem is undecidable. A hypercomputer, however, would be able to decide the halting problem for a Turing machine. But every hypercomputer has its own halting problem which it cannot decide. So if it turns out that some form of hypercomputation is physically possible, this shows that we do not have good reason to restrict ourselves only to Turing computable functions. Instead, we should consider all functions that are computable by the most powerful form of hypercomputation that is physically possible. But as we have seen, this form of hypercomputation would also have its own halting problem, and so the resulting Solomonoff prior would be uncomputable even by this hypercomputer.

We saw earlier that Solomonoff's prior is not a proper probability measure, since it sums to less than 1. What effect does this have? For starters, it means that it becomes harder to obtain a pragmatic justification in the form of a Dutch book argument. Such arguments attempt to show that unless an agent's credences conform to a set of rules – in this case, the probability axioms – she will be vulnerable to a Dutch book, i.e. a set of bets that result in a guaranteed loss regardless of the outcome.

Another problem with the fact that Solomonoff's prior is not a proper probability measure is that it becomes harder to use decision-theoretically. By assigning *probabilities* to hypotheses, we can calculate utilities. This is the case in the von Neumann-Morgenstern utility theorem, where it is assumed that individuals face options called lotteries. Each lottery consists of a set of mutually exclusive outcomes which are all assigned probability that sums to 1. By noting an agent's preferences among such lotteries, we can describe the utilities she ascribes to the individual outcomes. But since semi-measures do not necessarily sum to 1, the framework cannot be applied in this case.

However, the fact that Solomonoff's prior is not a proper probability measure is more of a side effect of its not being computable. If it were computable, we could normalize it to a computable probability measure without any trouble. And as I briefly noted earlier, we can in fact normalize it, but if we do so it will no longer be possible to approximate it from below. What's more, any computable approximation of Solomonoff induction can

[27] See e.g. Ord (2006) or Copeland (2002) for an overview of hypercomputation.

also be normalized, so neither of these two problems will arise in practice.

4.3 Are the hypotheses computable?

Solomonoff's prior is itself not computable. But by explicitly anchoring the framework of Solomonoff induction in universal Turing machines, we are in effect assuming that the hypotheses under consideration are Turing computable. That is, every hypothesis we may take into account must be representable by a binary sequence that is computable by a Turing machine. Does this mean that we assume that every process in the universe is computable? No. Even though there are several "computational universe" hypotheses floating around, one need not make such an extravagant assumption. With regards to uncomputable processes, we can just ask: are there methods other than Solomonoff induction that give better results? If the answer is no, this would be sufficient to justify the assumption. This amounts to an assumption that, for all practical purposes, the things we are concerned with are in fact computable.

4.4 Why simplicity?

Solomonoff's prior gives higher probability to simpler strings, according to the formula given in equation 3. In effect, this amounts to a kind of Ockham's razor – indeed, this principle is often mentioned in connection with Solomonoff induction. In its most well known formulation, Ockham's razor states that "entities are not to be multiplied beyond necessity." There are a few different ways of spelling this out, however (Baker, 2011). First, philosophers typically distinguish between two kinds of simplicity: elegance and parsimony. Elegance has to do with the formulation of the hypothesis, whereas parsimony has to do with the entities postulated by the hypothesis. There is often a tension between these two: by postulating more things, we may be able to formulate a theory in a simpler way. And conversely, by restricting the entities we postulate, the formulation of a theory might become more complex. In Solomonoff induction, simplicity is identified with a weighed sum of program lengths. But the lengths of these programs depend on the choice of UTM. Thus in a sense we can identify the choice of UTM with parsimony, and the length of programs with elegance. By choosing a UTM with a strong enough bias, we can make any hypothesis appear elegant, as measured by program length.

Furthermore, Ockham's razor can be stated either as an epistemic or as a methodological principle. In its epistemic version, Ockham's razor states that other things being equal, it is rational to place higher credence in the

simpler of two hypotheses. In the methodological version, the principle states that for scientific purposes, it is preferable to adopt the simpler one as a working hypothesis. If nothing else, the methodological principle can be justified by the fact that we are cognitively limited agents, and simpler theories are easier to deal with. In the context of Solomonoff induction, we are concerned with Ockham's razor as an epistemic principle. Given how idealized the framework is, it would be difficult to try and justify it by appealing to the cognitive limitations of humans.

So the question then becomes: how can we give an epistemic justification of Ockham's razor, in terms of the desiderata listed in section 1.5 on what a solution to the problem of the priors should achieve. According to theorem 2 in section 2.3, the total number of prediction errors over the length of an infinite sequence is bounded by a constant, provided that the infinite sequence is computable. Similarly, theorem 3 in section 2.3 establishes a bound on the total prediction errors in the stochastic case. Perhaps this could be used as a starting point for a justification of Ockham's razor: if the constant is smaller than any such other constant achievable by other methods, this would count in favor of Ockham's razor. However, the size of the constant depends on the Kolmogorov complexity of the true sequence in the deterministic case, and on the Kolmogorov complexity of the true semi-measure in the stochastic case. Thus if the true environment (whether deterministic or stochastic) has a high Kolmogorov complexity, the constant will be large. Thus it is easy to find methods that will outperform Solomonoff induction when the true environment has a very high Kolmogorov complexity. So using these two theorems to justify Ockham's razor is a circular endeavor. It amounts to claiming that we should assume that the world is simple because if it turns out that the world is in fact simple, this assumption leads to good results. This point is noted by Kelly (2004), who claims that if we adopt Solomonoff's prior, we are in effect just assuming that Ockham's razor is correct. The authors of the leading textbook on Kolmogorov complexity also admit this:

> a priori we consider objects with short descriptions more likely than objects with only long descriptions. That is, objects with low complexity have high probability, while objects with high complexity have low probability. (Li and Vitanyi, 1997, p. 260)

Is it possible to find a non-circular justification of the epistemic Ockham's razor? On one view, defended by Kevin Kelly, it is not that simpler hypotheses are *a priori* more likely to be true. However, starting out by assuming the simpler theory means that your worst-case performance in

terms of how many times you change your mind is better than if you don't. In Kelly's words, Ockham's razor is justified by its *truth-finding efficiency*. Kelly actually proves this, but he's working not in the framework of Bayesianism, but in that of formal learning theory.

In formal learning theory, we are considering finite or infinite evidence sequences, much like the binary strings in Solomonoff induction. An empirical hypothesis is a proposition whose truth-value relative to an evidence stream is completely determined by that evidence stream. Thus even the truth-value of universal statements like "All ravens are black" could at least in theory be determined by an infinite evidence stream. An inductive method is a function that maps finite data streams to hypotheses, or to suspension of judgment. This is perhaps the main difference between formal learning theory and Bayesianism: in formal learning theory there is only full belief and suspension of judgment, whereas Bayesianism considers degrees of belief. An inductive method converges on a hypothesis H on a data stream e by time n just in case for all times $m \geq n$, the function outputs the same hypothesis H. A *discovery problem* is a pair $\langle \mathbf{H}, \mathbf{K} \rangle$, where K is a set of data streams representing background knowledge, and $\mathbf{H}$ is a mutually exclusive set of hypotheses that covers K. An inductive method *succeeds* on a particular data stream in K if it converges to the correct hypothesis when applied to this data stream. It *solves* the discovery problem if it succeeds for all data streams in K.

When comparing methods, we are interested in their respective costs. Kelly assumes that we restrict ourselves to methods that converge on the correct answer in the limit. This is essentially the same as item 2b in my list of desiderata. Given this restriction, there are essentially two kinds of costs: errors and retractions. The number of errors is the number of times the method outputs some hypothesis H' other than the true hypothesis H. The number of retractions is the number of times the method takes back an earlier answer H' and outputs another one. How well a method performs in terms of the number of errors and retractions will of course depend on what the correct hypothesis is. So instead of looking at the number of errors and retractions in the general case, Kelly compares methods by their worst-case performance. In Kelly (2007), he proves that following Ockham's razor gives you the best possible worst case, both in terms of errors and retractions. When it comes to retractions, the idea is in rough outline as follows. Say that we are considering whether all ravens are black. If we begin by assuming the simpler hypothesis – that all ravens are in fact black – we will change our mind only when we have seen a counterexample: a

nonblack raven. However, if we instead assume that there is exactly one non-black raven and are presented with a long sequence of black ravens, we will eventually have to change our minds or otherwise risk converging on the wrong answer. But after we have changed our minds, we might very well observe a nonblack raven, thus being forced to change our minds again.

Can this result be used to justify the bias toward simplicity inherent in Solomonoff induction? Kelly (2004) writes:

> Let α be a fixed quantity strictly between zero and one half such that values lower than α are 'small' and values higher than $1 - \alpha$ are 'large'. A Bayesian agent can be said to retract when her posterior probability drops from a high to a low level on some answer to the question at hand. With this slight modification, the U-turn argument also applies to Bayesian agents whose posterior probabilities really converge to the truth [...] Moreover, Bayesians with a prior bias toward simple theories will tend to be retraction-efficient, since the high prior probability will remain if the simplest theory is true, will 'wash out' in favor of the next-to-simplest theory if that is true, and so forth, for a total of k retractions in the kth simplest answer.

We can define Bayesian error in a way analogous to Kelly's definition of Bayesian retraction in the above quote. Again, let α be some real number strictly between 0.5 and 1. If the distance between the correct value of a hypothesis H (i.e. 0 or 1) and the credence placed in H is greater than α, we say that the agent is in error with regards to H. Thus every step of time at which the distance to truth is greater than α will count as one error.

Why should we care about such Bayesian retractions and errors? More specifically, where does minimizing retractions and errors fit into the list of desiderata? The closest candidate is item 2c, which states that the agent should get "sufficiently close" to the true hypothesis after a certain number of updates. Given that all methods under consideration are guaranteed to converge on the correct answer in the limit, a larger number of errors indicates that you have spent more time further away from the truth. Similarly, a large number of retractions indicate that your credence has fluctuated a lot.

One potential problem with applying Kelly's reasoning to the present case is that we are dealing with a semi-measure rather than a probability measure. Since the value of a semi-measure need not range the full spectrum

between 0 and 1, a lack of belief cannot always be represented by 0.5. We could get around this by considering a normalized version of Solomonoff's prior (which would then no longer be lower semi-computable).

5 Summary and Discussion

We have seen that Solomonoff induction yields a form of objective Bayesianism where simpler hypotheses are assigned greater prior probability. I went through some of its important properties: it is guaranteed to converge on the correct answer in the limit under a bound that depends only on the Kolmogorov complexity of the true environment, provided that the environment is computable. Furthermore, Solomonoff induction satisfies both reparametrization and regrouping invariance, up to a multiplicative constant. I discussed how the framework allows us to deal with a countably infinite set of hypotheses, and applied it to a toy confirmation problem. But there were many potential problems: Solomonoff induction depends critically on the choice of UTM, it is not computable but can only deal with computable hypotheses, and it requires that we justify the bias toward simplicity. In light of all this, how does Solomonoff induction fare with regards to the desiderata I listed in section 1.5?

I dismissed the two items listed under truth-likeness, as these require the impossible: that we are somehow able to assign higher prior probability to true hypotheses. That brings us to convergence on truth. On item 2a, about considering more hypotheses, Solomonoff's prior performs impressively well. It allows us to assign non-zero probability to every hypothesis in a countably infinite set. However, in such a case, staying sufficiently close to Solomonoff's prior is a mathematical necessity if we want to assign non-zero probability to all of a countably infinite set of hypotheses. Assume for instance that we map binary strings to natural numbers in the following way: $(\epsilon, 1), (0, 2), (1, 3), (00, 4)$, etc. To the kth string we assign probability

$$P(x_k) = \frac{6}{\pi^2}\frac{1}{k^2}.$$

This sum converges and is normalized, so that

$$\sum_{k=1}^{\infty} P(x_k) = 1.$$

For many individual strings x_k, $P(x_k)$ and $M(x_k)$ will be very different. For instance, the string consisting of one hundred zeroes will have a very high k, and a correspondingly low $P(x_k)$. But for many choices of UTM,

the resulting value of $M(x_k)$ will be quite high. However, since both priors converge to the same value,[28] we have that

$$\lim_{k\to\infty} |P(x_k) - M(x_k)| = 0.$$

This shows that regardless of which complexity measure we use, when a string becomes complex enough according to one measure, it will also be complex according to another. However, the differences that do exist between various priors may of course be very important when it comes to making predictions.[29]

Item 2b concerned convergence on the truth in the limit. Solomonoff induction does satisfy this, but so do many other possible priors. With regards to item 2c we saw that Solomonoff induction does get sufficently close to the true sequence or the true probability distribution in finite time, provided these are computable. However, the bound depends on the Kolmogorov complexity of the environment, and so Solomonoff induction will perform poorly when the environment is complex.

As we have seen, Solomonoff's prior is not computable, so the three items listed under applicability are out. If we are looking for a prior that can be used by agents, whether idealized or actual, Solomonoff's solution is not on the table. But the same is also true even when we consider most computable approximations of Solomonoff induction. The so-called AIXI model is a theoretical AI agent that uses Solomonoff's prior. In one computable approximation of AIXI, we only consider programs of a certain length l. The larger we choose l, the closer the result will be to M. However, the computation time grows exponentially with l, and thus even computable approximations quickly becomes unfeasible to use.[30]

How does the dependence on the choice of UTM interact with these desiderata? To some extent, this dependence makes the accomplishment less impressive. However, even if we consider all priors generated by all different UTMs, we still have a proper subset of all possible priors. Moreover, any prior in this subset assigns non-zero probability to every computable environment. As we saw in section 4.1 on language dependence, the difference in the Kolmogorov complexity assigned to a particular string x by two different universal Turing machines is bounded by a constant that depends only on these two Turing machines, and not on x.

[28] I'm assuming here that we are working with a normalized version of M, M_{norm}.

[29] I'm grateful to Shane Legg for pointing this out.

[30] See Hutter (2004) for a detailed exposition of AIXI and its computable approximations.

In summary: when it comes to a countably infinite number of hypotheses, Solomonoff's prior is justified by the fact that it allows one to assign positive probability to all hypotheses. However, this criterion does not single out Solomonoff's prior, but rather a set of priors that do not deviate too much from it. But other ways of achieving this often have something arbitrary about them. In the case where we base the prior on Ockham's razor, we can at least try to give some independent justification of the choice. This brings us to the other proposed justification, which is based on Kevin Kelly's work on Ockham's razor in formal learning theory. This application of results in formal learning theory, presented in the previous section, is still rather schematic, and much remains to be done to see whether this argument is strong enough to justify Solomonoff's prior. At the very least, I think we should be weakly objectivist, in the sense that we should assign strictly positive probability to every hypothesis. Not doing so means that we are in effect ruling out some hypotheses *a priori*, because conditionalization can never make us move from probability 0 to any other probability. Since Solomonoff's prior does meet this, it is at least a member of the set of allowable priors.

Solomonoff induction has barely been treated by philosophers at all, but chances are it can be applied to many other problems in epistemology and philosophy of science. Take Nelson Goodman's new riddle of induction, for instance (Goodman, 1983). In the standard formulation, we have seen a sequence of green emeralds, and no emeralds of any other color. We want to use this observation to support the hypothesis that all emeralds are green. But Goodman points out that we can define a new predicate *grue* to mean green before some future time t, and blue after that time t. So far, the observed emeralds are consistent with both hypotheses. Why is it that the hypothesis that all emeralds are green is confirmed, while the hypothesis that all emeralds are grue isn't? One common intuitive response is that grue is an artificial predicate, since it is defined in terms of a disjunction. But Goodman notes that this will not work: if we begin with grue and the similarly constructed bleen, we can define green and blue in terms of these. However, this response neglects the fact that green and grue may have different Kolmogorov complexity. If we assume a computational theory of mind and assign each person a UTM, chances are that the Kolmogorov complexity of green will be much lower than that of grue, and the resulting Solomonoff prior for "All emeralds are green" will be higher than that of "All emeralds are grue." Thus observations of green emeralds will better confirm the former hypothesis, because it has a higher prior probability.

6 Bibliography

Alan Baker. Simplicity. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2011 edition, 2011.

David Barker-Plummer. Turing Machines. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2011 edition, 2011.

Alan F. Chalmers. *What Is This Thing Called Science?* Hackett Publishing Company, Indianapolis, 3rd edition, 1999.

B Jack Copeland. Hypercomputation. *Minds and Machines*, 12(4):461–502, 2002.

John Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. The MIT Press, Cambridge, M.A., 1992.

Kenny Easwaran. Bayesianism I: Introduction and Arguments in Favor. *Philosophy Compass*, 6(5):312–320, 2011.

I. J. Good. The Paradox of Confirmation. *The British Journal for the Philosophy of Science*, 11(42):145–149, 1960.

Nelson Goodman. *Fact, Fiction, and Forecast*. Harvard University Press, Cambridge, MA, 1983.

Alan Hajek. Dutch book arguments. In Paul Anand, Prastanta K. Pattanaik, and Clemens Puppe, editors, *The Handbook of Rational and Social Choice*. Oxford University Press, Oxford, 2008.

Joseph Y Halpern and Daphne Koller. Representation dependence in probabilistic inference. *J. Artif. Intell. Res. (JAIR)*, 21:319–356, 2004.

Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability*. Springer, Berlin, 1 edition, 2004.

Marcus Hutter. On Universal Prediction and Bayesian Confirmation. *Theoretical Computer Science*, 384:33—48, 2007.

James M. Joyce. A nonpragmatic vindication of probabilism. *Philosophy of Science*, 65(4):575–603, 1998.

James M. Joyce. The Development of Subjective Bayesianism. In Dov M. Gabbay, Stephan Hartmann, and John Woods, editors, *Inductive Logic*, volume 10 of *Handbook of the History of Logic*, pages 415–475. North Holland, Oxford, 2011.

Mark Kaplan. *Decision Theory as Philosophy*. Cambridge University Press, Cambridge, U.K., January 1998.

Kevin T. Kelly. Justification as Truth-Finding Efficiency: How Ockham's Razor Works. *Minds and Machines*, 14(4):485–505, 2004.

Kevin T. Kelly. Ockham's Razor, Empirical Complexity, and Truth-Finding Efficiency. *Theoretical Computer Science*, 383(2–3):270–289, 2007.

Shane Legg. *Machine super intelligence*. PhD thesis, Department of Informatics, University of Lugano, 2008.

Hannes Leitgeb and Richard Pettigrew. An Objective Justification of Bayesianism I: Measuring Inaccuracy. *Philosophy of Science*, 77(2):201–235, 2010a.

Hannes Leitgeb and Richard Pettigrew. An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy. *Philosophy of Science*, 77(2):236–272, 2010b.

Ming Li and Paul Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, 2nd edition, 1997.

Toby Ord. The many forms of hypercomputation. *Applied Mathematics and Computation*, 178(1):143–153, 2006.

Ronald Ortner and Hannes Leitgeb. Mechanizing Induction. In Dov M. Gabbay, Stephan Hartmann, and John Woods, editors, *Inductive Logic*, number 10 in Handbook of the History of Logic. North Holland, Oxford, 2011.

Samuel Rathmanner and Marcus Hutter. A Philosophical Treatise of Universal Induction. *Entropy*, 13(6):1076–1136, 2011.

Ray J. Solomonoff. A Formal Theory of Inductive Inference. Part I. *Information and Control*, 7:1–22, 1964a.

Ray J. Solomonoff. A Formal Theory of Inductive Inference. Part II. *Information and Control*, 7:224–254, 1964b.

Ray J. Solomonoff. The Discovery of Algorithmic Probability. *Journal of Computer and System Sciences*, 55(1):73–88, 1997.

Bas C. van Fraassen. *Laws and Symmetry*. Oxford University Press, Oxford, 1990.

Jon Williamson. *In Defence of Objective Bayesianism*. Oxford University Press, Oxford, 2010.