Norms, Natures and God

Alexander R. Pruss

Contents

Acknowledgments

Central ideas for this paper were developed as part of the Wilde Lectures in Natural and Comparative Religion at Oxford University, Trinity Term, 2019.

Acknowledgments

I would like that thank ... Nicholas Breiner, Sherif Girgis, Philip Rand, ...??

??Add precise formulation of hypothesis, with various ingredients, and in conclusions add discussion of pieces of hypothesis.

CHAPTER I

Introduction

1. Introductory remarks

I have a human nature or human form that governs my activity, both voluntary and not. Much as the government governs the activity of the people *both* by legislating norms and encouraging people to follow the norms, my nature's governance also has the dual role of setting norms for me and influencing my activity to follow these norms. This nature is something real and intrinsic to me, something that makes me be what I am, a human being.

When extended to other fundamental beings besides humans, the above is the center of Aristotle's metaphysics. I will show that this center is extremely fruitful, providing compelling solutions to problems in ethics, epistemology, the philosophy of mind, semantics, metaphysics and philosophy of science. Many of these are prominent problems that have been the subject of much discussion, such as the problem of priors in Bayesian epistemology or of vagueness in semantics, while others are problems that have not attracted much attention, such as the problem of seemingly arbitrary detail in moral rules. I shall discuss these solutions in Chapters II–IX.

The ability to give unified solutions to an array of problems spread through many areas of philosophy gives one a very good reason to accept the central Aristotelian theses. However, in Chapter X, I will also argue that this center cannot hold on its own, and the way to be an intellectually satisfied Aristotelian, especially after Darwin, is to be a theist as well.

There are several lines of thought readers attracted to the unified Aristotelian solutions may follow. Some may deny that the problems facing the central Aristotelian theses are as serious as I contend. Some may agree that the problems are serious, and regretfully reject

the Aristotelian apparatus, either because they take the cost of the theistic solution to be too great or are unconvinced that the theistic solution works on its own terms. Others may agree that the problems are serious but find some other solution than the theistic one. But some, I hope, will conclude that the Aristotelian solutions are so attractive, and the theistic solution to the problems is sufficiently plausible, that this book provides not only a good reason to accept the Aristotelian center but also to accept theism.

We will be elaborating the metaphysical apparatus of what I have been calling the "Aristotelian center" gradually?? as we move through the problems and details of their solutions. At the same time, not every detail of the solutions needs to be adopted by the reader to find the general Aristotelian strategy compelling. Finally, in Chapter XII we will collect together the needed aspects of the Aristotelian metaphysics and discuss in greater detail the metaphysics needed.

??paths through the book?

In the rest of this chapter, we will do two things. First, I will sketch the central Aristotelian metaphysics in slightly greater detail. Second, I will discuss a neglected science-based argument from the 17th century polymath Marin Mersenne for the existence of God. This argument does not work, I will argue. However, an important thread running through this book will be how "Mersenne problems" analogous to the problems in science raised by Mersenne arise in many areas of philosophy and provide a compelling case for the existence of Aristotelian natures or forms.

2. Aristotelian natures

2.1. A quick introduction. According to Aristotle, reality is fundamentally built out of substances, which are real mind-independent entities. These substances are not limited

to microphysical entities like quarks and photons—indeed, it is not even clear that the microphysical entities are substances at all[1]—and indeed Aristotle takes biological organisms like oak trees and human beings to be paradigm cases of substances.??ref

Each material substance has a form or nature—I will use the terms interchangeably in this book. This form or nature performs a number of roles including unifying the matter of the substance into a single thing, setting norms for the structure and activity of the substance, and guiding the actual development and activity of the object. The nature of the oak tree is not merely an arrangement of its particles, since an arrangement lacks normative force. In living things, the form of the substance is its life or soul: it makes the substance be alive.

Natures are innate to their substances. Nonetheless, this statement underdetermines an important question, namely whether substances of the same sort—say, red oaks—all numerically share one nature or each individual substance has its own nature, albeit in relevant respects??forwardref they are all exactly alike in substances of the same kind. For two things could in principle share something innate to them. It could be that all people have the same soul, much as two conjoined twins could have the same stomach. Aristotle scholarship is divided on the question whether Aristotle believed in "individual forms", one per substance. However, at least one of the advantages of an Aristotelian theory of form will be accentuated if we accept individual forms, as we shall see.??forward Further, there is good philosophical reason to take natures to be individual, as we shall see in ??forward. Thus, I shall take natures to be individual. Nonetheless, if you like shared forms, *many* of the benefits I will draw out for a theory of forms will be ones you, too, can have.

2.2. Aristotelian optimism. Natures not only define how a thing should function, but also actively lead the thing to function in that way. This means there is an inherent bias in each substance towards acting well. This bias leads to Aristotle's optimistic thought that

[1]The fact that in quantum mechanics, one can have a superposition of states with different numbers of particles is evidence that particles are not substances.??

natural states occur "for the most part"??ref, which is quite useful for figuring out what is in fact natural, since the frequency of the occurrence of a state is evidence of its naturalness.

There is, however, a tension in Aristotle's own thought between the above optimism and the pessimistic observation that most human beings are morally bad.??ref Aristotle may be empirically wrong about most people being bad??refs?, but nonetheless exploring the tension will help us understand Aristotelian optimism more clearly as it faces the problem of moral evil.

There are many substances with different natures in the world. The flourishing of some requires involves the languishing of others: the lion's feeding is the gazelle's death. Moreover, a substance's nature directs it to behavior that works well for the substance in its natural niche. But things do not always stay in their niche. Because of this, Aristotle has many resources for explaining why there is a significant set of cases where substances find themselves in unnatural states.[2] But Aristotle nonetheless thinks that misfortune will only be a minority of the cases.

Let us return to the Aristotelian optimism that things function well "for the most part". What is and is not "for the most part" depends on the reference class. Most humans have legs, but most living substances do not. If the reference class of the "for the most part" is all activities of all substances, then human moral behavior forms such a small portion of that class—it is so outnumbered by bacterial reproduction, say—that even if all human moral behavior were wicked it would be unlikely to make a difference with respect to the Aristotelian optimism. However, at the same time, with such a broad reference class, the optimism would be of little use to us in understanding normativity for humans, for humans could simply be an outlier in all respects.

A more optimistic reference class would be all the activities of a particular kind of substance. On this reading, Aristotle would lead us to expect that each kind of substance does well in most of its activities. But moral activity is only a small proportion of the activity of a human. We also breathe, we circulate blood, we repair cells, etc. Leibniz

[2]For further discussion of the harmony between substances, see ??forwardref.

estimated that three quarters??check,ref of our activity is at an animal level. Stalin was a complete moral failure, but still he maintained homeostasis until the age of 74. Human moral activity could, thus, be mostly bad even though most human activity is good. Again, the tension between Aristotle's general optimism and his pessimism about human morals would be resolved.

A yet more optimistic reference class would be a particular major type of activity—say, moral activity or reproduction—of a particular kind of substance. Now we would have the prediction that most human moral activity will be good, and this seems to contradict Aristotle's thesis about typical human moral badness. But even this is not clear. In MacDonald's *The Princess and Curdie*, Curdie has just expressed to the princess's great-great-grandmother a pessimistic thesis that unavoidably most things humans do are bad.

> 'There you are much mistaken,' said the old quavering voice. 'How little you must have thought! Why, you don't seem even to know the good of the things you are constantly doing. Now don't mistake me. I don't mean you are good for doing them. It is a good thing to eat your breakfast, but you don't fancy it's very good of you to do it. The thing is good, not you.'??ref

The old woman makes two important points. First, we should not forget that we perform *many* morally significant actions each day. Curdie ate breakfast. He could have thrown it at his mother, or just ungratefully poured it out on the grown. His eating breakfast was morally good. And we perform many such morally good actions each day. Second, the fact that we perform these morally good actions does not do us much credit, the grandmother insists. I suspect that the reason for her pessimism here is Curdie's lack of the kinds of motivations that would render breakfast-eating positively creditable. But the mere motivation to nourish himself was already good, even if not particularly creditable.

There is a further point we may add. While on a mathematics exam, it might be enough to get 60% to pass, morally speaking it is not enough that 60% of one's actions be good.

If in the morning I kick a neighbor's puppy, at lunch I charge my private meal to a research budget, in the afternoon I plagiarize something from a foreign language journal for inclusion in my book, and in the evening I cheat in order to beat my kid at chess, I am a bad person even if each of these actions is paired with two morally good actions of the eating-breakfast level of goodness. Having a majority of one's actions be good is not nearly enough to avoid being bad.

Thus with the reference class of "for the most part" restricted to moral activities, Aristotle's optimism and pessimism can be both maintained. And the above considerations also show that Aristotle's optimism is quite compatible with realism or pessmism about human morality.

A further optimistic ingredient that we will at times draw on is the idea that the different ways of being well in an organism have a tendency to mutually support each other in a unified kind of way. There will be trade-offs, sometimes tragic ones, but by and large a healthy heart supports healthy lungs, a healthy mind supports a healthy body, courage supports justice, justice supports courtesy, and courtesy supports kindness, all of which tend to make one live a happier life even by hedonistic standards.

Aristotelian ethics is sometimes accused of inferring "ought" from "is". The Aristotelian optimism should plead guilty, but with a mitigating factor. Since things only typically act how they ought, the move from "is" to "ought" is defeasible. Nonetheless, that things act a certain way is some evidence that they ought to do so. This is, in fact, quite plausible apart from Aristotelian assumptions. For there are three possible views on the statistical correlation between "is" and "ought":

(1) There is no correlation whatsoever between how things behave and how they ought to behave.

(2) There is a negative correlation between how things behave and how they ought to behave.

(3) There is a positive correlation between how things behave and how they ought to behave.

Option (1) is quite implausible. That a behavior obtains is surely *relevant* evidence for the question of how a thing should behave.

Further, take at random any two distinct binary features A and B of human beings, say, having green eyes and having won an egg-and-spoon race. Suppose we actually measure the correlation of these two features across all humanity, by calculating the covariance $\mathrm{Cov}[A, B] = P(A \,\&\, B) - P(A)P(B)$, where $P(C)$ is the probability that a random human has C. We might find that the covariance is very small. But given that there are eight billion people, how likely is it that $P(A \,\&\, B)$ should *exactly* equal $P(A)P(B)$ across the population? Surely very unlikely indeed. Thus, very likely, any two distinct binary features are correlated, positively or negatively.

So now the choice is between supposing a positive or a negative correlation between behavior and norm. The hypothesis of a negative correlation does not seem very plausible. Imagine that you are thrown into a situation about which you know nothing, but see a line of people one by one doing something—say, pressing a red button. And now it's your turn. It seems to be an unjustified pessimism to think that your observation provides you with evidence for the hypothesis that you should *not* press the red button, as it would if you thought the correlation beween behavior and norm was negative. If you know nothing better, it's a better bet to be a sheep. And the same, surely, goes for other kinds of things than humans. That a behavior is exhibited is not evidence against the behavior being right. It is very unlikely to be neutral. So it is at least some evidence for the behavior being right.

This argument does not rule out the possibility that the correlation between behavior and norm is very weak, weaker than the Aristotelian optimist needs to get significant evidence for norms from behavior. But it is a start.

Nor need it be the case that "is" is always evidence for "ought". In some cases we may have independent evidence that a particular entity is so defective that certain sorts of its behavior are of no positive evidential value. We shouldn't learn ethics from the depraved. That said, I happen to be sufficiently optimistic that I doubt that there are many people

who are so depraved that they exhibit *no* positive correlation between behavior and norm, even if that correlation is not mediated by right decision-making but only by self-interest.

2.3. Species where most organisms fail to reproduce.

2.4. Who are the humans?

2.4.1. *Rational animals.* It is traditional in the Western philosophical tradition to describe the human being as a rational animal. One can take this further and *define* the human being as a rational animal. A number of modern-day Aristotelians??refs do this and hold that if rational dolphins or octopuses evolved, they would also be human in the philosophical sense, having the same nature as we do. While Aristotle certainly held that we were rational, he did not *define* us by our rational animality. And such a move would fit poorly with Aristotle's hierarchical way of defining species that inspired the Porphyrian tree. For if some but not all possible primates were human and some but not all possible cephalopods were human, then *human* couldn't be a subdivision of vertebrate or of invetebrate. Rather, it would be a species that cuts across multiple genera. This might not bother those of us who do not accept the hierarchical mode of definition[3], but it would be highly problematic for Aristotle.

But there is another serious related problem. The human nature specifies what is normal for human beings. For a rational primate, having four limbs adapted to movement on land is normal; for a rational dolphin, having flippers and a tail would be normal; for a rational octopus, having eight tentacles would be normal. This suggests that there isn't a single human nature that would be shared by rational primates, cetaceans and cephalopods.

However, this argument is much too quick. After all, the variety of normalcy problem seems may well arise even within the biological species *Homo sapiens*, though particular examples are apt to be controversial. The American National Institutes of Health describes lactose intolerance as "an impaired ability to digest lactose"??ref:https://ghr.nlm.nih.gov/condition/lactose-intolerance, which suggests that

[3]However, see Koons??ref for a fascinating defense of the Porphyrian tree.

lactose intolerance is abnormal (we would not talk of an impaired ability to digest cellulose in humans). Yet this alleged impairment is found in the majority of the human adult population worldwide, especially outside of Europe. We could say that speaking of lactose intolerance as an impairment is just a mistake. But there is another possibility: it may be that lactose intolerance is normal for some members of our species and abnormal for others, and the description of it as an impairment is correct when restricted to persons of certain European ancestries. In that case, assuming that all *Homo sapiens* do share the same nature, what makes it abnormal for an adult of Irish ancestry to be lactose intolerant but normal for an adult of Armenian ancestry[4] cannot just be our shared human nature.

Instead, it might be that our shared human nature encodes some conditional like "If you have the complex feature F, then you should be able to digest lactose through all your life." This conditional applies to humans worldwide, but only a minority of humans actually exhibit F. This feature might, for instance, be genetic coding for life-long lactose digestion, so that those who have this coding ought to be lactose tolerance and those who do not need not be. In that case, even among the Irish lactose intolerance need not always be an impairment: it would only be an impairment amng those who have the genetic coding for lactose tolerance but are nonetheless for some other reason intolerant. Or this feature could specify some other aspect of the genome that correlates with, but does not cause, lactose digestion.

An even more controversial and politically charged example could be sex-linked traits: for some members of our species, having a uterus may be normal, and for others it may be abnormal. Again, our human nature could encode a conditional like "If you have feature G, then you should have a uterus." Note that while it is controversial whether there is such a conditional, holding that there is does not by itself decide important questions about gender identity. For instance, someone who thinks that gender identity is determined by genetics could say that the conditional holds with G being a genetic feature

[4]There is 4% prevalence of lactose intolerance in Ireland and 98% in Armenia. ??ref:https://www.thelancet.com/journals/langas/article/PIIS2468-1253(17)30154-1/fulltext

(say, having two X chromosomes), while someone who thinks that gender identity is determined by personal self-identification could say that the conditional holds with G being self-identification as a woman. They could then both use the conditional to argue for opposite answers to the question of whether hysterectomy for sex-reassignment purposes should be covered by health insurance.

Much less controversial are more temporary conditions that affect what is normal. A high temperature is normal for a sick human but abnormal for a healthy one. Yet we surely don't want to say that getting sick is a substantial change.

The problem of the variety what is normal among logically possible biological species of rational animals could be handled analogously. Human nature could specify that if you are a primate you have four limbs, if you are a cetacean you have flippers and if you are an octopus you have eight tentacles. But while this is theoretically possible, it would mean that human nature would have to encode an infinite number of such conditionals. Indeed, taking this to its logical conclusion, we would have to include conditionals to fit not just rational animals in worlds with laws of nature like ours, but in worlds with a physics radically different different from ours. It seems very implausible to suppose that there is such infinite normative complexity in our nature.

Moreover, the expansive view of humanity results in a very implausible restriction of the space of possible natures. For while Aristotelian harmony may not allow for every combination of normative features—such as an animal that ought to live all its life deep underwater but also ought to breathe atmospheric oxygen—we would expect there to be a broad selection of logically possible natures harmoniously combining different normative features. Thus, just as it is possible to have an animal that is supposed to both echolocate and fly (the bat), and animal that is supposed to both echolocate and be snake-like (perhaps this is unexemplified, but possible), it should be possible to have an animal that is supposed to be both snake-like and rational. But such an animal would not be human, because humans are not supposed to be snake-like (on the expansive view, they can be

non-defectively snake-like, while on narrower view, being snake-like would be a defect). So it should be possible to have non-human rational animals.

It should be noted that even if dolphin and octopus persons are not human, it is very plausible that our human nature would require us to treat them with the respect that persons deserve. After all, even on the broad definition of humanity, disembodied beings like deities or angels would not count as human, and yet many people who have believed in such beings have held that we should show them at least the kind of respect we do for fellow humans, for instance by keeping promises made to them.

2.4.2. *The narrower view.* Even within the narrower view of humanity that would exclude dolphin and octopus persons from being human, we still have the difficult question of where exactly the boundaries of humanity are to be drawn. We can make use of our best ethical intuitions to give us good reason to draw them in a way that includes at least all members of Homo sapiens. There was a small study done by philosophers of the motivations of rescuers of Jews from Holocaust, and the one common factor found in the rescuers was a tendency to identify others as fellow humans. The intuition of these rescuers, as well as of typical anti-racists and anti-sexists in our time, supports taking all members of Homo sapiens to be human.

But there are two further questions. One is about other historical species closely related to us, such as Neanderthals and Denisovans. While we could identify humanity with Homo sapiens, we need not. Probably, we do not know enough about Neanderthal and Denisovan life to see whether it is plausible that one form encompasses the norms for them as well as for us. And fortunately for us this question is of merely theoretical importance.

The second question is whether while we should accept all *adult* members of Homo sapiens as human, we should do so likewise for those biological humans who do not satisfy the philosophers' criteria for developed personhood, such as language or generalized problem-solving skills.??ref:Warren In the case of adults, there is a simple Aristotelian argument for doing so. Adult biological humans who do not fulfill such criteria are abnormal, as is made clear by the fact that if we could treat them medically in a way that leads to

the fulfillment of the criteria, we would have good reason to do so. Indeed, if these biological humans were not of the same kind as us, then such medical treatment by transforming them into beings like us would constitute what Aristotelians call a substantial change, a change from one kind of thing to another. But objects do not survive substantial change. Thus, such treatment would be a killing. That is implausible.

The remaining subquestion is for immature members of Homo sapiens, such as zygotes, embryos and fetuses. Here, the questions become more controversial. On a view that excludes such immature members, we would have to say that their ordinary course of life is that they mature and die *in utero*, giving rise to a human in the metaphysical sense. Yet death seems to be a catastrophic event for a substance, and in the growth of these immature organisms into more mature ones there does not appear to be such a catastrophe. This suggests that these members of the our biological species are also human in the metaphysical sense, but much more needs to be said.[5]

3. Mersenne questions

3.1. Mersenne's argument. Marin Mersenne was a monk, philosopher, theologian and the 17th century equivalent of the arXiv preprint archive—he was a crucial line of communication between a broad variety of thinkers and scientists. He drew on his broad knowledge of the science of the time to offer an argument that begins with many pages of questions, of which the following are representative:

> Who gave more strength to the lion than to the ant? Who made it be that earth is not in the moon's place, and that the planets aren't larger or smaller, closer or further? Who has ordered all the parts of the world as we see them? ... Why is the moon 56 earth-radii away from the earth? Why is the sun 1182 [earth-radii] away from us at its apogee? ... and

[5] ??refs

> why is its distance at perigee not other than 1101 [earth-radii]? ... I could equally ask you about Saturn, and Jupiter, and Mars ...??refs[6]

These "Mersenne questions" go on and on, with a mind-numbing number of examples. And Mersenne has one answer to all these questions, posed in a rhetorical question: "Was it not God?"??ref

The argument sounds similar to fine-tuning arguments for theism which became popular in the late 20th century. These arguments, too, list a variety of physical parameters and offer God as an explanation of them all.??ref But there is a crucial ingredient that the fine-tuning arguments, namely that the parameters listed are needed for intelligent life as we know it, or for some other valuable trait of the universe, like its amenability to scientific investigation.??ref The basic idea behind the fine-tuning argument is, very roughly, that nature is indifferent to value but God cares about value, so the fact that the parameters are valuable provides evidence for theism over naturalism.

It is, thus, natural to look in Mersenne for arguments that it is particularly valuable for the moon to be 56 earth-radii from the earth, but at least in this work, Mersenne does not supply them or even hint at them. Nor is there any argument that it is better that lions are stronger than ants, or that it is better for the moon to orbit the earth rather than the other way around. If Mersenne is giving a fine-tuning argument, the argument is oddly incomplete. And Mersenne's penchant for adumbrating detail at great length makes it unlikely that he has simply omitted such a crucial part of the argument.

Rather, it appears that Mersenne is simply looking for an explanation of the scientific details he cites, sees no prospect of a scientific explanation, and offers theism as the alternative. And indeed it is only in the 20th century with computer models of solar system formation that we have much in the way of plausible answers to Mersenne's questions about the distances between solar system bodies. For instance, the leading theory of lunar formation involves the earth being hit by another body and a large chunk being pushed

[6]The moon-earth distance is approximately correct. The earth-sun distance is an order of magnitude off.

into orbit. Given assumptions about the impact, one can then explain the resulting distance between the earth and the moon. But notice that such an explanation only gives an answers to the Mersenne question about the earth-moon distance at the cost of raising similar Mersenne questions about the parameters of the impact such as the mass distribution of the pre-impact earth, the angle and location of impact, the mass distribution of the impacting body, etc.

But Mersenne has a fatal argumentative flaw. Even if we grant that it is very unlikely that a future science will predict these exact numbers, there is always the possibility of a stochastic explanation, one that does not predict exact values, but supposes a random natural process that generates a set of values at random. Now, if Mersenne had an argument showing that the values of the parameters are suspiciously valuable—say, necessary for intelligent life—then a stochastic explanation might not be as good as a theistic one. From a Bayesian point of view, we might be able to argue that it is extremely unlikely that a random selection of parameters would have such value, but not nearly so unlikely that God would choose such parameters and hence the data supports theism over randomness. But given that Mersenne makes no case that the parameters have anything to recommend them to God for creation, we have no reason to think that the probability of God choosing is these parameters is any higher than the probability of them arising randomly, and hence we have no support for theism.

Suppose, however, that we had a Mersenne-type case where randomness was not a satisfactory explanation. Then there would still be one more problem with the argument. If one is willing to deny the Principle of Sufficient Reason, one could simply say that the parameters are what they are and there is no reason why they are like this—that they are a *brute* fact. This, however, is less satisfying than the stochastic answer, for adverting to brute fact should be a last resort, to be chosen when no explanation is available. But here there is an option, namely theism.

3.2. Appearance of contingency. Mersenne gives a dizzying number of examples, and he seems to relish the sheer appearance of arbitrariness of the numbers like “56” and

"1182". While this has some rhetorical force, it also has argumentative force. The more arbitrary-looking parameters the parameters are, the less epistemically likely it is that they are what they hold of necessity or that good scientific theories will predict their exact values. And the greater the number of parameters, the less likely it is that science can provide an explanation of them all.

The appearance of arbitrariness is evidence of contingency, and contingency calls out for explanation.[7] But at the same time, we have to be careful here. For instance, it might seem arbitrary that protons have (approximately) 1836 times the mass of electrons, but the massses of protons and electrons could well be essential properties of them, so that a pair of particles whose mass ratio were different from 1836 could not be a proton-electron pair. So in some cases, the arbitrary-seeming parameter does in fact hold of necessity. But that does not mean that the Mersenne question disappears. For while the parameter itself is not contingent in these cases, there is contingency "nearby". Even if the masses of protons and electrons are essential properties, it is possible to have particles with similar behavior but other masses, and it will be contingent that the world contains a pair of opposite-charge particles with mass ratio (approximately) 1836 that form atom-like entities similar to the atoms of our world.

The point generalizes: Sometimes the apparently arbitrary parameters can be explained by the necessary features in the essences of things, but in those cases it will often be the case that it is contigent that these essences, rather than other similar ones, are exemplified. In those cases, the appearance of arbitrariness yields an appearance of contingency, and the true contingency is nearby.

There is, however, a further worry here. Consider the apparent arbitrariness of the fact that the ratio of the circumference of a circle to its diameter in decimal notation has 1 and 4 as its second and third digits, respectively. Yet this fact can be wholly mathematically

[7] In ??ref, I have argued for a Principle of Sufficient Reason (PSR) that holds that all contingent facts have an explanation. But even if one rejects the PSR, one should hold that explaining relevant contingencies is a good feature of a theory, one that provides evidence for the theory.

explained by necessary mathematical truths such as that $\pi = 4 - \frac{4}{3} + \frac{4}{5} - \frac{4}{7} + \cdots$.[8] Thus the appearance of arbitrariness of a parameter is merely *defeasible* evidence of contigency in the parameter or even nearby.

We thus have to be cautious: moving from apparent arbitrariness to contingency, whether of the parameter itself or of something "nearby", is always going to be a defeasible and non-deductive move. This is why there is a value in Mersenne's giving as many examples as he does, since non-deductive arguments tend to stack up. But in any case, a number of Mersenne's particular examples, such as the astronomical distance examples, are ones where it would be difficult to believe in a necesity-based explanation without any contingency involved.

In the rest of the book we will find that if we turn our attention away from science and towards philosophy, we will find a myriad of cases like Mersenne's where there are seemingly arbitrary parameters. But these will be cases where a randomness explanation is implausible, bruteness is not satisfactory and the appearance of contingency is undefeated. However, unlike in Mersenne's cases, I won't be arguing—at least not in the first instance—that theism provides the solution. Rather, the solution will be Aristotelian metaphysics of form.

[8]This point is very similar to an argument Hume makes in Part IX of his Dialogues??ref.

CHAPTER II

Mersenne questions in ethics

1. Motivating examples

1.1. The rule of preferential treatment. Let us begin with a more detailed discussion of an example from Thomas Aquinas's discussion of the order of charity. Aquinas thinks, along with common sense, that those who are closer to us have a greater moral call on us. Thus, if it is a question of bestowing the same good on one of two people, where one is more closely related to us, we should benefit the closer one. But Aquinas writes: "The case may occur, however, that one ought rather to invite strangers [to eat with us], on account of their greater want."??ref And then he raises the question of what one should do "if of two, one be more closely connected, and the other in greater want."??ref

We might hope that here Aquinas would give us some clever rule for weighing connection against need. But instead he writes very sensibly: "it is not possible to decide, by any general rule, which of them we ought to help rather than the other, since there are various degrees of want as well as of connection".??ref It is tempting at this point to throw up one's hands and simply say that in these in-between cases there is no fact of the matter as to what should be done, or both options are permissible, or else relativism applies to the case. But that would not do justice to the way we agonize when we find ourselves in such a difficult situation, trying to discover the truth of the matter. (It is interesting to note that the most common real-life moral dilemmas tend to be

like these kinds of cases, rather than highly controversial questions about trolleys, strategic bombing or bioethics much discussed by philosophers.) And indeed Aquinas maintains a realist attitude to the question while simply offering this advice for how to figure out the answer in a particular case: "the matter requires the judgment of a prudent man.??https://www.newadvent.org/summa/3031.htm#article2

We can think of this as the problem of specifying a function $f(r, a, s, b)$ of four variables, two of them, r and s, being degrees of relation and the other two, a and b, being degrees of benefit, where the function takes one of three values corresponding to whether it is obligatory, permissible but not obligatory or impermissible to bestow a benefit of degree a on a person with relation of degree r to the agent in place of bestowing a benefit of degree b on someone related to degree s.

In fact, the problem of a rule of preferential treatment is much more complicated than the above indicates. First, the *kinds* of benefit and relation also matter: "we ought in preference to bestow on each one such benefits as pertain to the matter in which, speaking simply, he is most closely connected with us."??ref So the function will depend not merely on quantitative features but qualitative ones. Second, although Aquinas does not mention it here, the evaluation will no doubt depend on various features of the circumstances. And, third, in practice instead of choosing between two certain benefits, we are choosing between two probability distributions over the space of possible benefits.

Now, as Aquinas admits, we do not know what the moral evaluation function for choices between benefits to different people is. But abstractly speaking there is some such function, even if we do not know what it is, just as there is a function that assigns to each person alive now the number of hairs they now have, even though we cannot specify any of the values of the function. And we have good reason to expect the moral evaluation function to be very complicated. Indeed, probably the only serious proposal for a relatively simple function f here is the utilitarian suggestion that $f(r, a, s, b)$ yields obligation when $a > b$, mere permission when $a = b$ and prohibition when $a < b$. But this utilitarian

suggestion betrays the intuition that the degrees of relation r and s, much less the kinds of benefit and relation, are relevant to the moral evaluation.???refs

Indeed, the function is apt to look arbitrary. Fix the degrees of relationship to be one's parent and a total stranger, and fix a specific and certain financial benefit of $1000 to one's parent, and fix the circumstances. Then as we vary the financial benefit to the stranger from zero to infinity, we will presumably initially have a requirement of benefiting the parent (it would be wrong to give $1 to a stranger instead of $1000 to a parent in ordinary circumstances), then a permission either way, and then a requirement to benefit the second party. There will be boundaries between these regions of logical space, and these boundaries will look as arbitrary and contingent as the boundaries between different tax brackets. Like the tax brackets, some proposals for boundaries will be *clearly* unreasonable, but there will be many proposals that appear reasonable. And whatever the actual boundaries will look arbitrary.

Of course, seemingly arbitrary numbers can come out of an elegant and simple rule: it seems arbitrary that the fifth and sixth digits of π are 5 and 9 respectively, but there is an elegant mathematical explanation. But apart from the utilitarian proposal, we do not have any at all plausible simple proposal for f.

These seemingly arbitrary boundaries in the order of charity raise call out for an explanation at least as much as the exact distance between the earth and the moon does. Just as it seems implausible that the distance between the earth and the moon *must* be exactly what it is, it seems implausible to think that the boundaries must be exactly where they are—unless the utilitarian is right about f being very simple.

In fact, the ethics case calls out for an explanation even more than Mersenne's scientific examples did. For we might be able to swallow the earth-moon distance being a contingent and brute unexplained fact. But a brute fact seems unfitting for a moral rule. A claim that it just so happened, with no explanation at all, that you should ϕ undercuts the moral force of the alleged moral obligation. We expect anything seemingly arbitrary in our moral norms to have an explanatory ground.

To further argue for this point, consider a version of Divine Command Theory on which obligations are divine commands, and God rolled indeterministic dice to decide which actions to command, and by chance God's commands coincided with our common-sense morality, though they could just as well as well have commanded cruelty and dishonesty. A Divine Command Theory on which it is mere chance that cruelty is forbidden rather than commanded provides an unacceptable answer to the Euthyphro problem.?? Intuitively, a set of injunctions that is as arbitrary as that cannot constitute morality. But this point generalizes beyond divine command theory. Suppose that that we have some preferential treatment rules that are brute and contingent, and could just as well have enjoined on us the anti-utilitarian rule that we should always prefer the lesser benefit. Then whatever these rules are, they do not constitute morality, but at best happen to agree with morality in content.

Thus, even if there is some bruteness in the rules of preferential treatment, the rules in our world must be generated in a way that makes rules such as the anti-utilitarian rules not be among the possible outcomes. But this makes it very unlikely that the rules would be brute. For what force would limit the brute rules to avoid unacceptable options? Such a view of limited bruteness would be akin to a view on which banana peels can come into existence *ex nihilo*, but not where we might trip over them.

It is important to remember that the Mersenne question here is a metaphysical question: What explanatory grounds are there for why this rule, rather than some competitor, holds? The epistemic question may well have a virtue-theoretic answer like Aquinas's: if we acquire the requisite virtues, we will be able to judge particular cases fairly reliably, and until then our best bet is to ask the advice of virtuous others.

But before I continue the discussion of the possible explanation for the above ethical Mersenne question, let me follow Mersenne's lead and multiply the examples, in order to defend against potential answers that only work in some cases, and to make clear how widespread the problem is.

1.2. Risk and uncertainty. Some people—perhaps you—would accept a 92% chance of winning a thousand dollars at the cost of an 8% chance of losing ten thousand. I wouldn't. I say that both I and they are reasonable. On the other hand, someone who (in ordinary circumstances) rejects a 99.9999% chance of winning a thousand dollars at the cost of a 0.0001% chance of losing ten thousand and someone someone who accepts a 10% chance of winning a thousand dollars at the cost of a 90% chance of losing ten thousand are unreasonable. It is well known that attitudes to risk vary between people, and while there are unreasonable attitudes, it is very plausible that there is a broad range of reasonable attitudes.??refs So, as we vary the probabilities of wins and losses, we move between cases where accepting the risk is unreasonable, to cases where both accepting and rejecting are reasonable, to cases where rejecting is unreasonable.

This, once again, raises the Mersenne problem of why the transitions between the various evaluative categories lie where they do. And of course things are more complicated than described above. The rational evaluation function will depend not just on the probabilities involves but also on the values of the potential gains and losses.

While in the previous case, utilitarianism provided a neat but implausible solution, so too in this case, expected utility maximization provides a neat but implausible solution. On expected utility maximization, you are rationally required to accept a chance p of a good of degree α despite a chance q of a bad of degree β against a status quo of value zero just in case the expected utility $p\alpha + q\beta$ is strictly positive; when it is zero, you are permitted but not required; and when it is negative, you are not permitted. One problem with this solution is it requires all goods to be neatly quantifiable (cf. the next example for difficulties related to that). But the more serious problem is that it requires an implausibly negatively judgmental attitude towards ordinary people's attitudes to risk.

Indeed, here is a plausible trio of theses about risk that are incompatible with expected utility maximization:

(1) There is no upper bound on possible finite utilities.

(2) A decade of the worst tortures the KGB could think of has a finite negative utility.

(3) There is no possible good G of finite utility such that one would be rationally required in accepting a certainty of a decade of the worst tortures the KGB could thing of one for a one in billion chance of G.

For as long as $(1/1000000000)\alpha + \beta > 0$, where α is the value of G and β is the (highly negative) value of the tortures, one would rationally required to accept the deal on expected utility maximization, and by (1) and (2) there exists a possible G that makes $(1/1000000000)\alpha + \beta$ strictly positive. Hence, we should reject expected utility maximization, and and absent expected utility maximization, it is likely that the rationality evaluation function for risk will be messy and arbitrary-looking.

The most plausible thing for the apologist for expected utility maximization to reject is the no-upper-bound thesis (1). Here is one way an argument for such a rejection might go. First, there is a maximum intensity of goods that our brain can handle. Second, goods become significantly less valuable as they are repeated, decreasing in such a way that the sum of the values of any goods you could have over an arbitrarily long life has an upper bound.??refs

But the repetition thesis is only plausible when boredom and other memory-based phenomena are in play. Suppose you have lived for a very long time. Then you suffer from partial amnesia: you have lost all episodic memory of your past meals and of your past pinpricks. You are offered what you are reliably informed is the most delicious and wholesome dessert every prepared by the best chef on earth, a dessert which you are told you've eaten some large number n times in the past, and you may eat the dessert at the cost of a one in ten chance of a small pinprick. It's clearly worth it, regardless of what n is. So now suppose this happens to you every day of a very long life. The marginal value of each such dessert (i.e., the amount it contributes to total lifelong utility), absent memories of past desserts, must be at least one tenth of the marginal disvalue of the pinprick, at least given expected utility maximization. But the disvalue of the pinpricks clearly does not tend to zero with forgotten repetition. Hence, the value of the desserts does not tend to zero. And hence for any finite utility bound, enough such desserts will exceed the bound.

For a different example, not involving radically large utilities, imagine this Star Trek plot. Captain Kirk visits a planet inhabited by two intelligent alien species, all on par with respect to moral value: there are $1,000,000$ oligons and $2,000,000,000$ pollakons. Unfortunately, an asteroid is heading for the planet. If nothing is done, it will hit the planet in such a way as to wipe out all the pollakons. The only thing Kirk can do is to fire phasers at the asteroid. Spock has calculated that if this is done, the asteroid's track will be redirected in such a way that it will wipe out all the oligons. Kirk asks whether that will help the pollakons? Spock's answer is that probably not: there is a $999/1000$ chance that all the pollakons will still die, but there is a $1/1000$ chance that they will all survive.

It seems very plausible that Kirk should not fire phasers. And it is even more plausible that Kirk is not required to fire phasers. He should not sacrifice the oligons for a small chance of saving the pollakons. But the expected utility of firing phasers is $-1,000,000 + (1/1000)(2,000,000,000) = +2,000,000$ lives.

We can also argue against expected utility maximization by considering the following case. Suppose that on every day n of eternity, with $n \geq 1$, you are offered the opportunity to pay half a unit of utility in exchange for playing a game with a $1/2^n$ chance of winning 2^n units of utility. By expected utility maximization, you would value the value of the game at $(1/2^n) \cdot (2^n) = 1$ units of utility, and at a price of $1/2$ units, it would be worth playing.

But consider what will almost surely happen if you adopt the policy of following expected utility maximization and playing the game, where "almost sure" is the technical term that probabilists use to describe an event that happens with probability one (such as getting heads at least once if you toss a fair coin infinitely many times). The sum of the probabilities of winning on the different days is finite: $1/2 + 1/4 + 1/8 + \cdots = 1 < \infty$. The Borel-Cantelli Lemma??ref then says that almost surely you will win only a finite number of times.[1] In other words, almost surely, there will come a day after which you will win no

[1] We can give an elementary proof of this fact in the case at hand (the proof generalizes to the general case). Let W_n be the event that you will win at least once after day n. Then $P(W_n) \leq 2^{-(n+1)} + 2^{-(n+2)} + \cdots = 2^{-n}$.

more. At that point, you may well be ahead, having won more than you paid. But the sum of what you won is finite, and from then on you will just lose half a unit of utility every day. Eventually, there will come a day when your losses will overtake your winnings, and from then on, you will just fall further and further behind every day.[2]

The very unhappy situation of playing infinitely many times and eventually starting to lose every time is the almost sure result of following expected utility maximization on each day. We can compare this to the neutral situation of refusing ever to play, and getting zero each day, or the situation of accepting the expected utility maximizing gamble for a number of days, until the probability of winning becomes really small, and refusing from then on.

It is worth noting that this is not just a paradox involving the aggregation of infinitely many utilities, except in the trivial sense that infinitely many zeroes make a zero (i.e., there is overall no benefit from playing the game once you stop winning). Almost surely, after a finite number of days, the expected utility maximizer falls behind the consistent refuser, and every day after that, the expected utility maximizer is further and further behind, like someone who got a subscription to a streaming service and forgot to either use or cancel it. And all these amounts are finite, and a finite, albeit unknown, distance into the future.

We can also consider an interpersonal version of the story. Suppose we have (countably) infinitely many people, numbered $1, 2, ...$, and person n is offered the chance to pay half a unit of utility in exchange for a chance $1/2^n$ of winning 2^n units. As before, by expected utility considerations it's worth it. So, if everyone is an expected utility maximizer,

---

Let W be the event that you win infinitely many times. Then $P(W) \leq P(W_n)$ for every n, since if you win infinitely many times, you must win on infinitely many days after day n, and so you must win on at least one day after day n. Since $P(W_n) \leq 2^{-n}$, we have $P(W) \leq 2^{-n}$ for every n. But probabilities cannot be negative, and the only non-negative real number x such that $x \leq 2^{-n}$ for every n is zero. So $P(W) = 0$, and hence almost surely W does not happen, so that almost surely you win only finitely many times.

[2]The example above uses exponential growth. More moderate growth will work, as long as the sum of the probabilities is finite. Thus, we could say that on day n the prize is $n(\log(n+2))^2$ and the probability of winning the prize is the reciprocal of that, since $\sum_{n=1}^{\infty} 1/(n(\log(n+2))^2) < \infty$.

everyone will pay. But by the Borel-Cantelli Lemma, almost surely, only finitely many people will win. Thus, almost surely, we will have infinitely many people pay a cost of half a unit each, and finitely many people win some finite amount. This is a disastrous situation, with a negative infinite overall utility. Almost surely, it would be much better if everyone refused to play, or only those who had a "non-negligible" chance at winning played.[3]

??refs:PrussBlog2011,Zhao, Wilkinson??ref:https://philpapers.org/rec/WILRAA-16

It appears that expected utility maximization cannot be rationally required. But it is the only clearly non-arbitrary solution to the problem of deciding under uncertainty.

In addition to Mersenne questions about risk and prudential rationality, there will be Mersenne questions about risk and morality. For instance, what risks we may morally impose on others in exchange for a good to ourselves depends in a complex way on one's relationship to these others, the probability of the risk, the degree to which these others accept the risk, the benefit to self, and so on. When I drive, I risk killing other drivers, their passengers, pedestrians by the side road, and so on. But the probability of these awful outcomes is very small, and typically other people on or by the road have accepted reasonable risks (or have had them accepted by proxies, in the case of children), so these dire but unlikely outcomes typically do not render it impermissible for me to go to the grocery store to pick up ice cream.[4] But when the risk is higher, say because I am tired and sleepy after a long day and hence less likely to be a safe driver, the matter becomes less clear. At some point, as the risk increases, it becomes impermissible to go to the grocery store for ice cream. A particularly thorny set of issues arises in the special case of balancing the risk that the innocent are punished with the risk that the guilty go free. And we have the Mersenne question of why the switchovers happen where they do.

[3]It is worth noting that exponential growth is not necessary for the examples to work. All we need is that there is a chance p_n of winning a prize of $1/p_n$, and that $\sum_{n=1}^{\infty} p_n < \infty$. While for ease of calculation above I let $p_n = 1/2^n$, one can have much more moderate shrinkage, such as $p_n = 1/n^2$ or even $p_n = 1/(n(\log(n+2))^2$.

[4]I leave open the question whether concerns about global warming render it impermissible.

Expected utility utilitarians[5] will have a nice answer to this problem. But utilitarianism, as already noted??ref, has many highly counterintuitive implications.

????add: moral risk?

1.3. Orderings between goods. Under ordinary circumstances, it would not be reasonable to choose to be a mediocre mathematician rather than a superb musician. But suppose one's choice is whether to be a superb musician or a superb mathematician? Here we are dealing with incommensurable goods and either choice is reasonable.

But now let's ask this general question: Is it is reasonable to choose to be a mathematician of quality α rather than a musician of quality β? Again, we have a function that takes a number of variables, including α and β and the circumstances, and tells us whether (a) it reasonable to opt to become a mathematician but not reasonable to opt for music, or (b) both are reasonable, or (c) opting for music is reasonable but opting for mathematics is not. And, just as before, it is very plausible that the function is extremely complex.

The problem obviously generalizes to all the many kinds of pairings of incommensurable goods there are. In each case, there will be some function of many variables encoding the correct rational evaluation of the situation/, and we will have the Mersenne question of what grounds the fact that this function, rather than one of the infinitely many others, encodes the correct rational evaluation.

We also have Mersenne questions here that involve qualitative rather than quantitative comparisons. Other things being equal, social pleasures are better than solitary ones. This seems rather arbitrary. What makes it be so?

In the preferential treatment and moral risk examples, utilitarianism offered a nice solution. But the problem of incommensurable goods is also going to be a problem for any plausible utilitarianism. Utilitarianism comes in two varieties, depending on whether the good is pleasure or the good is satisfaction of desire. As Mill famously noted, it is essential to the plausibility of utilitarianism that one be able to make a distinction between

[5]As opposed to actual-outcome utilitarians who evaluate actions morally based on the actual utilities that would result from an action.??refs

lower and higher pleasures, so as to get the common-sense conclusion that it is better to be Socrates unsatisfied than to be a satisfied pig.

But once one makes the distinction between lower and higher pleasures, or lower and higher desires, incommensurability quickly shows up, since different kinds of pleasures and desires do not simply come in a linear ranking. Let's suppose that you get more enjoyment and satisfaction of the desire for truth out of mathematics and more enjoyment and satisfaction of the desire for music out of music, and let us suppose (contrary to typical situations) that your choice of life will not affect anyone else. Then it seems right to say that the mathematical and musical lives are incommensurable even on utilitarianism. But even if they are not incommensurable, but equal or one is better than the other, we still have a Mersenne problem as to what level of quality of mathematical life exceeds, equals or falls below what level of quality of musical life. And in fact it will be more complex than that, in that the quality of a mathematical or musical life is clearly multidimensional.

One might try to get out of this by hoping for some precise definition of the degree of pleasure or the strength of a desire. Perhaps there is a neural correlate of the degrees of pleasure or the strengths of desire that can be quantified in a single number. But such an approach is likely to lead to the swinish utilitarianism that Mill wisely rejects. For presumably the neural correlate can be manipulated directly, and the pig could be given pleasures which, in terms of neural intensity, exceed the highest of Socrates' refined joys, and could be made to have a degree of intensity of desire for its swill far exceeding Socrates' desire for virtue.

Moreover, any neural approach is likely to fall prey to questions of cross-species comparison. While pig and human brains are similar, they are not the same, and states of pleasure and desire are likely to be merely analogical. It is clear that some comparisons between human and porcine goods are possible: a tiny human pleasure is worth less than a great porcine one. As one increases the human pleasure and/or decreases the porcine one, there will come cases where neither of the two is to be preferred, and then eventually cases where the human pleasure is to be preferred over the porcine one. But where exactly

the cross-over points are is not something we can just read off the neural correlates. And things get even messier when we compare humans to possible beings that have no brains, such as intelligent robots (if these are possible) or aliens with very different biochemistry.

And even if one could give some such precise formulation, we would still have the Mersenne problem of why *this* formulation corresponds with true value rather than some other.

1.4. Intersubject aggregation of value and population ethics. Whether or not consequentialism is true, there are some questions which need to be settled in a consequentialist way, for instance questions where the stakeholders are strangers one has no special obligations towards and where there are no deontic considerations. If we aim for the consequentialistically best outcome for more than one person, we will need a way of aggregating value between these subjects. A particularly difficult set of cases comes up when the number of subjects in existence varies between the options.

Perhaps the most straightforward option is to *add up* utilities across the affected population. This faces several problems. The most famous is Parfit's repugnant conclusion??ref. Any finite scenario full of highly fulfilled people can be beat by a scenario with a much larger number of people whose level of fulfillment is minimal. Yet it seems implausible to think that we should aim at vastly multiplying human population at a minimal level of fulfillment.

A technical problem is the following. Utilities are normally considered to be defined "up to positive affine invariance"??ref. They can be rescaled and shifted without changing anything. This means that for any positive number α and any number β, if we consistently replace every utility x with $\alpha x + \beta$, then we have not changed anything. But if we add utilities across a variable number of individuals, although multiplying all utilities by a positive factor α makes no difference to our aggregate decisions, the addition of a constant β to every utility can make a difference. For instance, suppose we have a choice between ten individuals each enjoying a utility of 15 each and twelve individuals enjoying a utility of 10 each. As it stands, on the additive model of aggregation, the ten individual option

has a total utility of $10 \cdot 15 = 150$ and the twelve individual option has a total utility of $12 \cdot 10 = 120$, thereby yielding a preference for the ten individual option. But if we take into account the affine invarience with $\beta = 25$ and $\alpha = 1$, then the utilities enjoyed by the individuals in the two scenarios become $15 + 25 = 40$ and $10 + 25 = 35$, yielding totals of $10 \cdot 40 = 400$ and $12 \cdot 35 = 420$, flipping the preference in favor of the larger population option.

Essentially, the technical problem is that we need a "zero point" for utilities. If we have such a point, then increasing the number of people with utility above zero will always improve the outcome, while increasing the number of people below it will make things worse. Now, while there may be clear cases—Einstein's life was above the zero point, but a life of constant torture is below—defining a zero point precisely in terms of the vast multitude of various good- and bad-making features of a human life is apt to involve a large number of Mersenne questions.

The most natural alternative to adding utilities is averaging them. As Parfit has noted, this leads to the another unpleasant conclusion: if the average of some nation's utility is even slightly below the average utility for all human beings, then we would be better off if the people in that nation didn't exist.[6]

Furthermore, averaging only escapes the repugnant conclusion for humans given assumptions about what the rest of reality is like. Suppose that it turns out that humans are outnumbered by non-human persons by a factor of ten, and the overall average utility in reality is terrible, as most persons live lives of horrific misery, so that the average utility of a non-human person is -100 while that of a human is 10. This makes

[6]The problem becomes perhaps even more vivid if we include non-human animals as subjects. For it may well be that no non-human animal enjoys a utility greater than that of an average human. One way to this conclusion is the intuition that no non-human animal is such that we should save its life instead of an average human's. Another way is to reflect on the fact that an average human enjoys goods—say, moral or cultural ones—that are qualitatively higher than those of any non-human animal. On any averaging view, then, it seems it would be good to allow all non-human animals on earth die out. (Antinatalists may of course disagree with the intuitions in this footnote.??refs)

the average utility overall be $(-100 \cdot 10 + 10)/11 = -90$. Now suppose we have a choice between two options. On the first option, we can make every human enjoy a utility of 100, which is a life of deep fulfillment, without changing the number of humans. On the second option, we can multiply human population by a factor of 10, but make each person's utility be a very miserable -50. On the first option, overall average utility becomes $(-100 \cdot 10 + 100)/11 = -81.8$. On the second option, it becomes $(-100 + (-50))/2 = -75.0$. In other words, it is worthwhile to multiply human population as long as humans are even a little less miserable than the average among non-humans, and depending on what options are available, multiplication of human misery, with sufficient increase of population, may be better than making all humans happy. This is even more repugnant than the original repugnant conclusion, since we are "improving" things by making people not merely slightly well-off but by making them actually miserable, just not as miserable as the average.

The last point vividly illustrates a well-known problem with averaging: what decision is right depends on epistemically inaccessible facts about intelligent life outside earth.??ref One can, of course, solve this by restricting the averaging to our own species and hoping that we won't meet aliens. But we can still get some version of this repugnant conclusion simply within the human species. We can imagine a situation where three quarters of humans are found in a repressive nuclear state, on average have an utterly miserable flourishing level of -100, and there is no way for those of us outside that state to remedy the situation. Suppose instead we have a choice between two policies for the rest of the world: keep the population of that part of the world constant and bring everyone to enjoying approximately a utility of 20 or triple the population and bring everyone down to a pretty miserable -35. The resulting averages will be $(-100 \cdot 3/4) + 20 \cdot 1/4 = -70.0$ and $(-100 + (-35))/2 = -67.5$. As in the alien case, we have a nasty version of the repugnant conclusion: lots of people at a low level of happiness—indeed, a level of significant misery—outweigh a moderate amount of people at a high level of happiness.

It is worth noting that in the case of averaging, there is a decision point about aggregating across time. Averaging the utility across all subjects at all times is mathematically straightforward if the the number of subjects across time is finite. If it is infinite, there are many additional complications.[7] However, averaging across all subjects at all times could seriously exacerbate some of the repugnant conclusion worries. For instance, if non-rational conscious animals count, then the vast majority of conscious animals on earth across all time may be relatively primitive, say on the level of lizards and squirrels, and hence not capable of much flourishing. As a result, the average level of happiness may easily end up being even lower than if we just average at the present time, thereby making it more practical to increase average flourishing by vastly multiplying subjects, human or not, with low levels flourishing.

The other intuitively natural option is to average simultaneous utilities at each time, and then integrate them across time. This runs into at least three special problems. First, we have a difficulty of how to account for a multiverse whose universes do not share a common time-line. Second, we need a privileged sequence of reference frames—a privileged foliation—to define the simultaneity of utilities. Third, and perhaps most problematically, many aspects of the flourishing or languishing of humans are difficult to localize in time. Examples include having a *diversity* of cultural experiences over a lifetime, having one's goals be posthumously fulfilled or frustrated, or failing to ever find a good friend.

It is possible that some simple function without free parameters can be used to aggregate utility across people in a way that avoids paradoxes. But it seems very unlikely.

[7]*If the infinity is countable, then any averaging will have to involving a limiting procedure. For instance, if u_i is the utility of the ith subject, then we can aggregate by computing $\lim_{n\to\infty} \sum_{i=1}^{n} u_i$. However, the value of that limit is likely to depend on the order of the subjects. If all the subjects are in the same spacetime, we might try to order the subjects temporally. That itself carries some decision points. Do we order subjects temporally by the beginnings of their existence, the middles, or the ends? And in relativistic spacetimes, temporal ordering will be relative to a foliation, and so a privileged foliation will be needed. Though we might get lucky and the limit might be the same for any reasonable ordering. If, however, the subjects are in different spacetimes—as in a multiverse scenario—then finding an appropriate for the limit is even harder.??ref

Intuitively, we have some sort of diminishing returns as we increase the number of people with a minimal level of happiness, but the diminishment is not the simple "one over the population count" multiplier of the averaging solution. It seems very plausible that an aggregation function that does not generate some repugnant conclusion will be rather complex and will have multiple free parameters. And then we can ask what grounds these parameters' having the values they do. ??see if literature proves there is no function

The above assumed, for ease of modeling, that we could reduce the utility enjoyed by each individual to a number. For reasons discussed in Section??ref, this is itself unlikely. Our flourishing seems to consist of a number of incommensurable goods such as friendship, understanding, play, etc. Moreover, even if we reduce flourishing to pleasure or desire satisfaction, it is likely that we have incommensurable pleasures and incommensurable desires. All this greatly complicates any aggregation procedure, multiplying the number of its free parameters.

1.5. A miscellany of other Mersenne questions. There are many other cases which involve thresholds or transitions that appear to be arbitrary.

On strict deontological views, one shouldn't torture one innocent person to save any number of lives. But of course it would be permissible to gently prick someone with a pin to save even one life. Somewhere between the pinprick and the torture is a transition. What makes the transition be where it is?

On threshold deontological views, it is wrong to torture one innocent to save a small number (say, one or two) of lives, but it is permissible to do so to save a very large number (say, a billion). Again, we have a transition to be explained.[8] And note that even if one is a strict deontologist about torturing the innocent, likely one is a threshold deontologist about some other things. Thus, one may think it's permissible to save an innocent life but not permissible to lie to get a deserved (but on other grounds) salary raise, and hence there needs to be an explanation of the grounds of the transition from permissibility to

[8]I am grateful to Philip Swenson for this example.

impermissibility. Or one may think it is permissible to trespass on a neighbor's property to save a cat's life but not to save a grasshopper's. Probably everyone who isn't a full-blown consequentialist is a threshold deontologist about some things.

The Principle of Double Effect allows one to foreseeably cause bad effects that it would be impermissible to cause intentionally, as long as these bad effects are not intended either as ends or means. For instance, it seems permissible to bomb Hitler's headquarters even if one finds out that an innocent prisoner is held captive there. But of course there needs to be a proportionality condition imposed on this: the good achieved, say the end of a war, must be proportionate to the bad, say the death of the prisoner. It would be wrong to demolish an old building while knowing that there is a child playing inside: the good of having a lot to build on is not proportionate to the death of the child. So there will be some function of variables including harms and benefits that specifies when the benefit is proportional to the harm in Double Effect contexts. In fact, there will be other variables, such as one's relationships to those harmed and those benefited.

We need to show *respect* for intelligent beings. This respect includes such things as not killing them when they are innocent and non-aggressive, not eating them (except perhaps in extreme circumstances), not acting as if they were fungible, treating them as ends rather than as mere means, and so on. But what is an intelligent being? First, we have a distinction between an individual and a kind based concept of intelligence: on the former, a being is intelligent to the extent that it currently has certain intellectual powers; on the latter, a being is intelligent to the extent that it is of a kind that should have certain intellectual powers. But whichever we choose, and plausibly there are principled reasons to choose one rather than the other[9], we still have a Mersenne question as to the degree of intellectual power—whether actual or proper to the kind—that is needed for us to have duties of respect. Intellectual powers, after all, clearly come in degrees, and if at some point respect is called for, we need an explanation of why that point shows up where it does.

[9]Though they will be highly controversial, since a significant part of the debate about the moral status of the unborn turns on this.

The question of what in fact the degree of intellectual powers is needed for respect is one that we actually face with regard to our treatment of higher mammals on earth, and that we currently only face hypothetically with regard to extraterrestrial life. It is an important question. But, as usual, the Mersenne puzzle isn't that of determining what the fact is, but of what makes an answer be an answer, especially in light of the appearance that any threshold will be arbitrary.

A state presumably comes about when the people sufficiently agree to form a state, whose laws then typically need to be obeyed. But what constitutes such an agreement? Suppose that it is the agreement of a majority of those adults who make a reasonable effort to make their opinion heard (say, by voting). But where does the transition between a child and an adult lie? What effort is reasonable to require and what is unreasonable? And if we are to speak of the majority of the people, of *which* people? Presumably, the people inhabiting a given land. But there are many overlapping areas that can be considered the "given land", and in different overlapping areas majority opinions may be different. Moreover, what counts as inhabiting? (Suppose, for instance, someone lives different parts of the year in different places.) There is a vast multitude of questions to be answered by a majoritarian or any other account of the institution of a government, each question facing a Mersenne puzzle. And there are similar questions about the dissolution: Plausible, when a government becomes sufficiently unconcerned about the wellbeing of the people, it becomes illegitimate. But why does this transition happen where it does?

Punishment should not be disproportionate to a crime. *Lex talionis* provides a neat and elegant account of this: the criminal get done to them the same thing as they did. But at the same time, *lex talionis* is not in general plausible. It is morally abhorrent to torture torturers or rape rapists. And even if we accept such abhorrent extremes, there are cases where punishment simply has to switch types of harm. If a thief does not have enough honest property of their own to make possible a deprivation equal to what they stole, imprisonment is a plausible alternate currency, but there is no simple and elegant formula providing an exchange rate between the value of stolen property (perhaps corrected for

the economic state of the victim) and length of imprisonment, unless it turns out—as is quite unlikely??back/forwardref—that that there is a precise way of quantifying personal utilities. Maybe there is no simple and elegant formula, but a complex one. If so, it is puzzling what grounds it. And even if there is no specific formula, there are bounds of moral acceptabiloty: a day's imprisonment for stealing and destroying a car is insufficient; a lifetime's imprisonment for stealing a book is excessive. And what grounds these bounds?

Next, observe that standards of consent necessary to permit one's being treated a certain way vary widely depending on the treatment. There are multiple dimensions in which we can measure the "strength" of a consent requirement: how well informed the consenting party needs to be, what age or level of intellectual development does the party need to have, what proxies if any can offer consent on the party's behalf, how unpressured the consent needs to be, how clearly formulate the consent needs to be, whether the consent must be specific to the case or whether prior blanket consent suffices, etc. Under ordinary circumstances, no consent—at most, lack of refusal—is needed for a pat on the shoulder. The permissibility of major surgery, however, has a consent requirement of significant "strength" along many of the above axes. On the other hand, the permissibility of sex has a consent requiremnt of even greater "strength" along some of the above axes—thus, while proxy consent and prior blanket consent can suffice for major surgery, they do not suffice for sex.[10] The mapping between the form of treatment and the multidimensional strength of consent is of great complexity, and has an appearance of significant arbitrariness. What grounds it?

We have a number of normative powers that we exercise through communicative acts with specific illocutionary force. With promises, we create obligations for ourselves, and

[10]It is tempting to explain this in terms of the fact that surgery—or at least the sort of surgery for which proxy consent suffices—benefits the patient regardless of the patient's consent, while sex is only beneficial when consented to. But this is arguably false. Parents can validly consent to an organ transplant between their children, even if the donor is not expected to benefit on balance (though generally there is a benefit from having one's sibling alive!).

with commands, we create obligations for others. With requests, we create reasons for others, and with permissions, we remove reasons for others. The scope of our normative powers has limitations, though where exactly the limitations lie can be a matter of controversy. Perhaps the clearest case is requests, where as a society we have developed a broad spectrum of levels of insistence, signaled by linguistic and extra-linguistic cues. Normally, a more insistent request creates stronger reasons and a less insistent one creates weaker ones. We do not appear to have any lower limit on the strength of reasons we can create solely by requests. We can always add yet another "But, please, don't let me impose on you" to weaken the request. However, there is a contextually-variable upper limit on the strength of reasons created by a request. One can roughly measure the stregth of these reasons, say, by the cost to the requestee at which fulfillment becomes pretty unreasonable. The strength of the reasons to fulfill a request is then a function of the intrinsic reason provided by the requester's needs (if someone is starving, one has reason to offer them food even if they don't ask for it), the degree of insistence, and the relationship between the two parties.[11] Normally, then, the more insistent the request, the more it skews the requestee's reasons in favor of the requested action, but there is a limit to how far one can skew the strength of reasons away from the no-request *status quo*. If I am not actually in poverty, requesting money from strangers for my personal pleasure cannot create a very strong request-based reason, no matter how much I ask for it.[12]

We have similar limits on the strength of reasons coming from the exercise of other normative powers. The example of promises is particularly interesting here in that we might think that our normative powers are strongest here since we use them to bind our own wills. But notice that while I can probably make a promise to defend my friend's life where the promise creates a reason whose strength is such that I am obligated to seriously

[11]Note that the relationship may itself have been modified by the fact of the request. A friendship might be damaged by an unreasonable or rude request, and strengthened by the vulnerability revealed in a disclosure of need.

[12]Though of course I might create a prudential reason to give me the money to shut me up, or out of fear that I am a mugger, but that's not a request-based reason in the sense I am talking about.

risk my own life, by promising to come to my friend's party I cannot create a reason strong enough to stand against serious risk to my life, unless there is something very special about that party.[13] Even in promises, there are serious limits to how far we can affect the reasonableness of our decisions.

Social convention sets certain aspects of the mapping from communicative acts exercising normative powers to the normative effects. It determines which normative powers are exercised in which words or gestures, and how various communicative and extra-communicative features affect the strength of reasons. But social convention works within the limits discussed above. Social convention can *say* that if I spit on my hand and shake hands after promising to come to your party, then I am obligated to come to the party even if it costs me my life, but in fact this action would not create any such reason, since skewing the intrinsic reasons thus far is just as beyond our normative power as running a one minute mile is beyond our locomotive power. But while we can find a physical explanation for the locomotive limits on an individual human, the normative limits raise Mersenne questions about all of their free parameters—and they have many, since there are multiple normative powers and in each one the limit will depend on multiple contextual factors.

Political philosophy also provides a number of examples of seemingly arbitrary parameters. Consider the constitution problem. A state has a written or unwritten constitution specifying what must happen for legislation to be valid and hence authoritatively binding on the citizens. But how is a constitution instituted? One theory is that it happens by the consent of the people.??Aquinas But obviously for any state of sizeable size it will be false that all the people have consented: some have not made their opinions heard and some have been overruled. Requiring "consensus" or a supermajority raises the question of exactly how many dissenters can be tolerated, and once that question is answered we have a Mersenne question as to what grounds that cut-off being where it is. Requiring a simple majority or plurality involves one less free parameter: the cut-offs in having more than half

[13]Perhaps the party is the best hope for reconciling with someone who is dying of cancer, and my friend cares deeply about the relationship.

of the voters or having more voters than any alternatives seem non-arbitrary. However, even a majority or plurality based system leaves questions about other parameters. Does one need a quorum of the governed? It would not seem right, for instance, to have a vote on a constitution on a day where the bulk of the population is unable to get the polls due to a hurricane or a war, and as a result only a small number of unrepresentative citizens can express their opinion. But if a quorum is required, then of course we have a Mersenne question as to exactly what constitutes quorum.

Voting cut-offs and quorum are fairly easy to quantify. But what about the question of who the people giving their consent are? Presumably, small children should not be eligible. But where do we draw the line between small children and paradigmatic adult deciders? Any age-based line raises several Mersenne question—one about the numerical age cutoff and multiple questions about how age is measured (from fertilization, implantation, brain development, beginning of the birth process, completion of the birth process, etc.)? A cut-off based on mental capacity, on the other hand, involves many parameters that need to be set, because there is no single measure of mental capacity, and so one needs to have multiple measures with their respective weights. Moreover, we have a decision point on whether those who do not get a vote have proxies voting for them (e.g., parents) and, if so, who counts as whose proxy.

Or consider such details as how well-publicized the constitutional consultation needs to be, how clearly spelled-out the constitution needs to be for people's vote on it to be valid, and who gets to decide which options are presented to people?

Many of the above questions only make sense in the case of a formal consultation process of a sort that has occurred rather rarely over the course of human history. If we are not to think the vast majority of polities to be illegitimate, the account needs to allow for implicit consent, maybe of the sort involved in social customs. But there things become much less clear. There will almost always be some citizens who regard the state as illegitimate—indeed, some will regard any state as illegitimate. For many people, acceptance of a political system's legitimacy is not an simple binary question, but something

that comes in degrees and has contexts. Imagine that two thirds of the population has a credence of two thirds that the political system is legitimate, and the remaining third of the population has a credence of ten percent in the system's legitimacy, and they express these credences in their actions. Should we then look at the average credence of the relevant citizens (e.g., those of age)—0.48 in my example above—and see if it meets some cut-off? If so, the cut-off will raise Mersenne questions. Moreover, people often do not just have a single credence it the all-or-nothing legitimacy of the political system, but rather have different credences regarding different aspects of legitimacy: is this a state that has the right to use violence against its citizens to enforce laws, is it a state that has a right to levy taxes, to draft citizens to defend it or do other work for it, etc.

A set of Mersenne questions similar to those raised by the constitution problem is raised by the dissolution problem: the question of when it is that a political regime becomes illegitimate, and the respects in which it may be illegitimate (thus, perhaps, the traffic regulations of the Nazi state were legitimate). One might think of dissolution as resulting from the state's failure to keep its side of the social bargain. But there has probably never been a state that kept its side in every respect. It is only gross failure that implies illegitimacy, but that raises a Mersenne question about the degree of grossness.

Further, we have a set of Mersenne questions regarding who lies within the scope of the state's authority. Typical human states have authority largely but not entirely defined geographically—for an exception, consider ships under a country's flag that move on the high seas.??check Geography itself raises Mersenne questions. Suppose we say that a state in the future has authority over the same territory as it now occupies. But what counts as "the same territory"? We live on a planet that is constantly changing its shape, whether at a large scale due to movements of tectonic plates or a small scale due to our own digging. When a tectonic plate shifts, how does territory shift? There are, of course, precise ways to answer these questions. We define latitude and longitude in terms of coordinates on a mathematically idealized oblate spheroid approximating the earth. We might then define

"same territory" in terms of these coordinates. But there are infinitely many ways of mathematically modeling the earth at any given time and infinitely many ways of matching up that model to the physical soil of the earth over time. And it seems unlikely that the geography of a planet is in the end what defines a state. It may well be the case that in the future a significant proportion of the earth's population lives on space stations or in the asteroid belt, and we will have ship of Theseus kinds of questions about identifying the sameness of a space station over time, and difficult questions about defining segments of the asteroid belt.

One might say that a state, or its people in their constitution, gets to define its own understanding of who counts as among the governed, and so a state can opt to define the governed as those occupying a certain portion of a certain specific diachronic oblate spheroid model or to define the governed as those within a certain specific distance of a particular landmark. But there must be limits to a state's normative power to define these boundaries, since otherwise a state could simply swallow up territory by mere fiat. Imagine, for instance, if France defined its territory in terms of all land within five hundred kilometers of the Eiffel tower, and then French agents conquered the world by secretly taking small bits of the tower and distributing them worldwide so that no place was more than 500 km away from one or more of them. It is very plausible that a state has some freedom in how it defines its boundaries, but that freedom is limited. And the range of ways of defining the scope of authority seems like the sort of thing that would have many different parameters without privileged values, and hence raises many Mersenne questions.

One may wonder why questions about the legitimacy and scope of a political system are being raised as part of a discussion of ethical questions. There are two reasons. First, we have a moral duty to obey the commands of a legitimate state. Second, only those acting on behalf of a legitimate state are morally permitted to make certain onerous demands on the population, especially ones backed up by threats of violence.

Some readers will disagree with a number of the examples I gave. Double Effect, for instance, is quite controversial, and philosophical anarchists will deny that any state is

such that one morally ought obey its commands as such. But it seems likely that a number of the remaining examples will still compellingly raise Mersenne problems. And the list above is not exhaustive: the reader should be able to generate more items.

2. Arbitrariness

Whatever the values of the parameters in the ethical Mersenne questions are, these values appear likely to be such that if we knew their exact values, we would find them arbitrary. In physics, some hold out a hope that the fundamental constants in the fundamental laws of nature may be "nice numbers" like 2, π, $\sqrt{2}$ or e. It seems intuitively even less plausible that things would so turn out in ethics.

And even if the parameters turned out to be such "nice numbers", that would itself be a very surprising fact, because while such numbers seem very natural in physics, they seem rather less natural in ethics. Imagine that you should benefit your parent over a sibling just in case the ratio of benefits is no lower than $1 : \sqrt{2}$. That would itself seem arbitrary. It seems that whatever the numbers turn out to be, they will have an appearance of arbitrariness and of contingency.

3. Continuity

Many of the examples involve thresholds, such as the amount of intelligence needed for respect or the degree to which a government needs to care for the common good to have auhtority. It is plausible to reject the idea that there are discrete thresholds, and instead hold that there are continuous functions, say a function $r(x)$ specifying the degree of respect required to be shown to a being with intelligence of degree x.

But then instead of explaining one threshold, one needs to explain the whole complex shape of the "respect function". On the most naive version of this, intellectual power will be graphed along one axis and respect on another, which will raise Mersenne questions about the slopes of the graph, the positions of the inflection points, and so on. But of

course in reality, both intelligence and respect have many dimensions, so what we have is a complex function of many arguments and whose values are multidimensional.

In general, moving from thresholds to continuous functions only multiplies the degrees of freedom that call out for explanation.

4. The human nature solution

On our Aristotelian picture, the nature of an organism grounds norms about what the organism's structure and behavior should be. In particular, the nature of the organism will ground many arbitrary-seeming norms, such as those governing the range of appropriate sizes of Indian elephants, the migratory behaviors of monarch butterflies, and the lengths of human femurs. Having the nature makes the organism be the kind of organism it is, and imposes on it the associated norms.

In the case of humans, the behaviors include voluntary ones, and so it is unsurprising that there are norms governing these as well. And just as there are many parameters governing bodily structure and sub-voluntary behavior, there are many parameters governing moral behavior, all grounded in the form.

At the same time, Aristotelian optimism provides us with evidence as to what the parameters approximately are. The actual bodily structures of humans give defeasible evidence as to what normative human bodily structure is and the actual behaviors of humans give defeasible evidence of moral norms. And in both cases, we have ways of identifying healthier or more virtuous paradigms, using the optimistic idea??backref that the various ways of doing well tend to hang together with some degree of unity, and the structure and behavior of such paradigms gives us further evidence as to the norms.

Admittedly, there appears to be a disanalogy between health and virtue. We might use a Mahatma Ghandi or a Mother Teresa to figure out moral norms, but we wouldn't use an Usain Bolt or a Serena Williams to figure out physical norms. One explanation of the difference is that Bolt and Williams have highly-developed traits that are specialized to a forms of life quite different from that of the typical human—namely, the life of a

professional athlete—while Ghandi and Teresa's excellences in justice, fortitude and mercy are as important to our life as to theirs.

All this raises the question of why the form includes these norms and not others. Here there is an easy answer available. The form is at least partly defined by the norms it includes. Thus, Mersenne's question about the lion and the ant when reformulated into normative terms, as the question of why the lion's strength *ought to* be greater than the ant's, is easily answered: this follows fron defining features of what make lions be lions and ants be ants.

The appearance of arbitrariness and of contingency in the ethical Mersenne problems is somewhat misleading: it is like the appearance of arbitrariness and contingency in the fact that water is H_2O or that carbon atoms have six protons. Water couldn't have a different chemical structure and carbon atoms couldn't have a different number of protons. But it is also an important truth here that there could be other substances that could have a different chemical structure or a different number of protons. Similarly, *we* couldn't have other norms of preferential treatment than the ones written on our nature, but there could be—and perhaps in this vast universe are—other intelligent animals with other such norms.

5. Other solutions

We thus have many Mersenne questions pointing to arbitrary-seeming parameters in ethical rules. I will now argue that a broad spectrum of ethical theories and solultions are unlikely to yield good answers to the Mersenne questions or else raise new Mersenne questions of their own.

5.1. Kantianism. Kantianism is an attempt to derive moral rules from the very concept of objective rationality. Famously, this leads to difficulties in accounting for the substantive content of rules. For instance, from the point of view of objective rationality, it is difficult to generate a presumption in favor of causing pleasure and against causing pain. The more tightly connected a moral rule is to the specifics of the human condition and of the circumstances, the more difficult it will be for the Kantian to account for it. But the Mersenne

questions above thrive precisely on such detail. Consider, for instance, the improbability of a good Kantian account of how much we should, other things being equal, favor siblings over cousins, or of why proxy consent is sufficient for surgery but insufficient for sex. The "logical distance" between the high level principles, like the categorical imperative to treat others as ends and never as mere means or to act according to universalizable rules, and such specific moral content appears unlikely to be bridgeable. Thus, precisely those cases that we have seen to raise compelling Mersenne problems make Kantianism an implausible ethical theory.

Of course, such appearances can be deceiving. One might well have antecedently thought that the relatively simple axioms of set theory are unlikely to generate the richness of mathematical theorems that we have seen to come from them. So it would be good to go beyond an intuition of "distance".

There are at least four ways to do that. First, proceed by intuitions regarding a specific example. Consider two different moral rules regarding to the relative treatment of siblings and cousins. One rule says that benefits to siblings are to be slightly preferred to benefits to first cousins and the second says that first cousins and siblings are to be treated on par. Neither rule requires us to treat anyone as a mere means or takes away from treating people as ends. Both rules are universalizable. So we are not going to be able to derive one rule rather than the other from Kantianism as originally formulated by Kant.

Second, we can makes use of a heuristic as to the validity of arguments. One heuristic I employ in checking whether a numbered argument given by undergraduate students is valid, i.e., whether its conclusion logically follows from its premises, is to see if the conclusion of the argument contains any substantive terms that do not appear in any of the premises. If it does, it is in practice unlikely that the argument is valid, though of course there are possible exceptions. If the premises are contradictory, then the logical rule of explosion makes every conclusion a valid consequence. And it could also be that the conclusion is disjunctive and the substantive term that did not occur in the premises occurs in one disjunct while another disjunct follows from the premises (though I have yet

to see this happen in a student paper). An argument from premises about the nature of rationality as such with a conclusion about specific familial relationships or about specific human activities such as sex or surgery fails the heuristic, and hence is unlikely to be valid. And the cases do not seem to be like the most common exceptions—the premises are not contradictory and the conclusion is not disjunctive.

Third, all or most of the examples that raised Mersenne questions have an appearance of contingency to them, in a way that does not fit with the hypothesis that they derive from necessary principles about the nature of rationality. One way to formulate this contingency is to note that many of the rules are ones that we would not expect to apply to other intelligent species. If we came across an alien species that regarded familial ties as somewhat more or somewhat less important than we think permissible for humans, we should not judge them immoral. It would not surprise us if other intelligent animals—perhaps ones occupying other niches—were rationally or morally required to take greater or smaller risks than we.[14]

Finally, we have an epistemological argument. While clearly we do not know the exact values of the parameters in the Mersenne questions, we have some approximate knowledge, as already indicated above in a number of the cases. We clearly did not come to this approximate knowledge by logically deriving it from Kantian first principles. Nor did we even do so by means of an intuition that they follow from these principles. For I take it that we do not in fact have an intuition that, say, the preference for siblings over cousins follows from Kantian principles. If anything, we have an intuition that it does not. So, it seems that if these rules in fact follow from Kantian principles, it's just a coincidence that our beliefs about the parameters are correct, a coincidence that makes the beliefs be mere justified true belief rather than knowledge. But the beliefs are knowledge. So, the Kantian explanation does not work.

[14]One thinks, for instance, of the Klingons and Kelpians from the Star Trek universe, respectively.

The epistemological argument has some force, but not that much. First, the argument is related to the highly controverted literature on evolutionary debunking arguments.??refs,add?? Second, a theistic reader has an easy way out of the argument: God knows what values of parameters in fact logically follow from Kantian principles and could either directly instil in us correct beliefs about them or ensure that we evolve in a way that yields such true beliefs.

5.2. Act utilitarianism. The main problem with act utilitarianism is that it generates incorrect moral claims. It says that a healthy patient whose organs can save three others can be killed when doing so doesn't have any other countervailing consequences such as making others more callous. It says that if you and I are loners who make no contribution to society, but I own a dog and you don't have any pets, then you have a duty to sacrifice your life for mine, to save my dog from being ownerless; and if neither of us has a pet, but you enjoy chocolate a little more than I do while everything else is equal, then I have a duty to sacrifice my life for you, since your life would include slightly more utility.

Moreover, as we saw in ??back, for utilitarianism to be plausible and not swinish requires a hierarchy of goods, and there will be Mersenne questions regarding that.

Finally, even hard-nosed desire-fulfillment or hedonistic utilitarianism will be unlikely to be exempt from Mersenne questions. There are multiple mental state concepts that could be argued to correspond to the words "desire" and "pleasure".

When the psychotherapist tells Jones that she always unconsciously wanted to kill her mother, is that a "desire" in the sense of desire-fulfillment utilitarianism or not? A case can be made either way, and this decision point generates a degree of freedom for the theory, and hence a Mersenne question as to why it is one sort of "desire" or the other that counts as defining the good. In fact, reflection the complexity of human life as seen in literature??ref:ColinAllen? shows that there are likely to be many "desire"-type concepts, differing along multiple dimensions, and hence generating a multiplicity of Mersenne question. And there will be multiple ways of quantifying the strength of a desire.

And as for pleasure and pain, we will again have a broad variety of concepts and a multiplicity of ways of quantifying them. This can perhaps best be seen if we think about the mental life of possible and actual non-human sentients. Does a particular state of an earthworm count as a pleasure? It is unlikely to be exactly like a state of ours. There will likely be many ways of classifying mental states across species, and on some the worm's state will be a pleasure and on others it won't. So we have a degree of freedom in our act utilitarianism as to what we count as pleasure or pain in non-humans. And even within humans there are complex questions. Consider for instance masochism or the subtle morose "satisfaction" of the pessimist who sees everything going downhill. There are likely to be different ways of classifying states as pleasures or pains, and the hedonistic utilitarian will have a Mersenne question as to why one rather than another classification is the one that defines ethics.

5.3. Rule utilitarianism. On rule utilitarianism, instead of requiring that each action optimize total utility, it is required that each action follow rules that are themselves optimized for total utility. Rule utilitarianism's main advantage is held to be that its escape from the counterintuitive consequences of act utilitarianism. The rule not to kill the innocent may well be the optimal rule for us, even if in a lifeboat situation it would maximize utility for the two stronger people to kill and eat the weaker third.

Rule utilitarianism could not only neatly explain the apparently arbitrary specifics of the moral rules, but could also explain the appearance of arbitrariness and contingency in a way that, say, Kantianism is unlikely to. For the optimization procedure that would define the moral rules would be a vast and complex one, taking into account the impact of the actions falling under the rules both in the short and the long run, both on humans and on non-humans. It is unsurprising if a complex optimization procedure produces results that seem arbitrary but are in fact carefully chosen to their end. A computer-optimized airplane wing will have precise angles and bends that cannot really be explained without running through the whole computation.

Moreover, rule utilitarianism is less prone than Kantianism to make our limited but true beliefs about the moral rules be merely coincidental. For we have evolved biologically and mimetically in the service of survival and reproduction, and because of the contingent connections between these goods and other aspects of utility, evolution put pressures on us that directed our moral beliefs in a truthful direction. There are deep and difficult questions whether this is enough to make the connection between our beliefs and the truth be sufficient for knowledge??refs, but there is more hope here than on the Kantian side.

However, famously, rule utilitarianism divides into two varieties, depending on exactly what the rules are optimized for. On ideal rule utilitarianism, the rules are such that everyone's successfully following them would be optimal, even if in fact they are too difficult for us to follow. Ideal rule utilitarianism, however, is widely held to reduce to act utilitarianism, since if everyone were to actually follow the rule of maximizing utility, that would be optimal with respect to maximizing utility. But act utilitarianism has already been put aside.??backref

Non-ideal rule utilitarianisms, on the other hand, inject a note of realism into the optimization procedures. For instance, what might render a set of rules correct is that if everyone were to *try* to follow them, optimal results would result. This already raises a Mersenne question. For trying is something that comes in degrees, and it is very likely that different rules will be generated when we optimize for the utility resulting from everyone's trying hard to follow them than if if we optimize for the utility resulting from everyone's trying with minimal effort. And there will be a vast number of intermediate cases, so there will be a Mersenne question of what grounds the fact that α, say, is the right degree of effort for defining the optimization procedure that generates the moral rules.

Furthermore, specifying the degree to which the hypothetical agents try to follow the moral rules is not enough to specify the optimization procedure. For instance, one has to specify the level of intelligence of the hypothetical agents, their non-moral interests and the environment, which yields multiple Mersenne questions as to what the requisite levels of these for the hypothetical optimization procedure are.

The only way to avoid such questions is to simply require the counterfactual world to match our world in the respects, but this runs into two problems. First, we would normally expect a world where all agents try to follow the moral rules to have agents that have different non-moral interests, higher levels of intelligence since such a world would have a much more just educational system than ours and hence would nurture children into greater intelligence, and a rather different natural environment. If we try to keep the three factors fixed while having the hypothetical agents try to follow the moral rules, we are likely to get some very unlikely counterfactual results, just as keeping too much of our world fixed in a counterfactual situation results in the odd claim that if Oswald did not kill Kennedy, Kennedy would have been buried alive. Second, we have to say that if our history had gone slightly differently, so that (say) the distribution of intelligence in the general population were slightly different, the optimization procedure would have generated different rules, and hence different moral rules would have been true. Indeed, on this view we would get the very strange idea that what we morally do can affect morality itself.

Besides this, there are other non-ideal aspects that we should probably introduce. Some of our important moral rules discuss how we should deal with culpable malefactors. But in a world where everyone tries to do the right thing, depending on the strength of trying, there might well be *no* culpable malefactors, or at least very few. And it is unlikely that moral rules optimized for such a very different situation would be likely to be the right ones for us. So we probably need to optimize the rules with respect to a hypothetical situation where not everyone tries to follow them. And that raises Mersenne questions as to how many people in the hypothetical case follow these rules, and what the others do with their lives.

In short, ideal rule utilitarianism is implausible, while developing the non-ideal rule utilitarian project raises multiple Mersenne questions as to the details of what is to be fixed in the hypothetical situation.

5.4. Social contract. Social contract theories ground ethical rules in agreement between agents. We can divide this based on whether the agreement is actual or hypothetical.

Actual agreement theories face obvious problems. First, it is highly implausible to think of the typical agent in society as having *actually* agreed to live by moral rules, apart from special cases such as a pious person vowing to God to sin no more. Second, actual agents can agree to live by unjust rules, even rules unjust to themselves, and such rules would not constitute morality.

Contemporary social contract theories tend instead to be based on duties grounded in hypothetical agreement between agents in situations of ignorance.??refs Anyone who has been in a long committee meeting knows that actual agreement between agents can result in complex rules with much apparent arbitrariness, and it would be unsurprising if hypothetical agreement were similar. Thus far, social contract fits our data well.

But the hypothetical agreement condition involves multiple parameters such as how smart the hypothetical agreers are (and there are multiple dimensions of intelligence), what exactly are they ignorant of, how many of them are there, what are their attitudes towards risk and uncertainty, etc. We have here an explanation of the Mersenne parameters in terms of other Mersenne parameters, and the problem remains fully entrenched.

The risk and uncertainty point is worth emphasizing. Some hypothetical agreement theorists think that rational agents would only agree to rules that do not treat anyone inhumanly.??refs But a rational agent who is more accepting of risk will be willing to tolerate rules that create a minority group that is treated inhumanly if the risk of being a member of that group is sufficiently small—i.e., if the group is a small enough fraction of the general population—and the the benefits to the majority are sufficiently large. There will be types of inhuman treatment and levels of risk that it would not be rational to accept for the sake of a high probability of a large benefit, but the lines between these and the ones that it would be rational to accept do not seem derivable from any plausible set of basic principles of rationality.

Granted, typical Kantian constructivists will insist that certain kinds of inhuman treatment would never be rationally acceptable. But now consider the Mersenne questions about these kinds of treatment. For instance, destroying the autonomy of another person

might be taken never to be rationally acceptable. But a minor limitation on another's autonomy clearly is acceptable for a sufficiently great good: if the only way to save a country from nuclear destruction by evil enemy would be to acquiesce in the enemy's demand that everyone wear jeans on Friday, then this limitation on sartorial autonomy should be enforced. Somewhere there is a line between minor limitations of autonomy and such deep destruction of autonomy that could not be tolerated no matter the price. The only "natural" place to draw this line would be at *complete* destruction of autonomy. But if it is only complete destruction of autonomy that is prohibited by the Kantian, then this does not place a sufficient constraint on the rules that could be accepted. For instance, the enslavement of persons would not be prohibited, as long as the enslaved persons were still capable of some autonomous agency, no matter how minor.

Furthermore, even prohibiting treatment that completely annuls someone's autonomy will not avoid Mersenne questions in the vicinity. For we will have probabilistic questions. Is it permissible to perform an action that has a 99.9lines to draw seem to be at 100autonomy prohibits nothing: any action we perform can fail. A prohibition on an action that has a 0prohibits everything: I scratch my head, and there is a tiny chance that due to some weird sequence of events this causes an earthquake that leads to you getting hit on the head by a beam that results in your life being reduced to a vegetative level. And while 50much more reasonable, in some difficult cases it is excessively restrictive. If a child is certain to die within a day, and is suffering from horrific pain that can only be relieved by a drug that has a 50of the day, administering the drug can be permissible.

5.5. Virtue ethics. Aquinas himself invoked the virtuous agent as providing at least the epistemic path to an answer to the preferential treatment question. We could also take virtue ethics to provide an answer to the Mersenne question: What makes these parameters, rather than others, hold is that the virtuous agent's patterns of behavior are thus and so parameterized.

But this of course simply shifts the problem to that of why the virtuous agent's patterns of behavior are parameterized as they are. The best answer to that question appears to be the one given in the Aristotelian tradition which grounds this in the agent's nature.

5.6. Divine command. On divine command ethics, the right is what is commanded by God. Divine command ethics, like social contract and rule utilitarianism, carries with it significant hope for explaining the apparent arbitrariness in ethical parameters. We would not be surprised if the laws coming from an infinitely intelligent and good legislator had significant complexity that to us would look like arbitrariness.

It may initially seem the divine command ethics runs into the same problem of pushing the Mersenne questions back to the question of why God legislated these parameters and not others. But notice that the Mersenne problems I have been discussing are *grounding* questions. Even if God's legislation were completely arbitrary in a way that ultimately violated the Principle of Sufficient Reason, on divine command ethics we would have a *ground* for the parameters in preferential treatment and other ethical rules being what they are. To say that we should prefer siblings over first cousins in a ratio of $1.7 : 1$ because God commanded so is to give a ground for the obligation, even if that ground itself needs an explanation. Compare the moral prohibition on adding cyanide to friends' drinks. There would be something absurd if that prohibition were ungrounded. But it has a ground, or at least a partial ground: cyanide is fatal to humans. Imagine now that there was in fact no possible explanation of why cyanide is fatal to humans. Nonetheless, the grounding problem for the moral prohibition would have been solved by citing the danger of cyanide.

In this way, our ethical grounding Mersenne problem is quite different from Mersenne's merely explanatory problem. In Mersenne's case to explain why the distance between the earth and the moon is what it is in terms of other parameters of earlier states of the solar system does not make significant progress. But when we have given a plausible

ground to the moral obligation, we have indeed made progress. Mersenne's original argument depends for its plausibility on a fairly general Principle of Sufficient Reason.??ref-on-PSRr Here we just use a heuristic principle that moral truths with an appearance of arbitrariness need a deeper ground.

Moreover, the divine command theorist has nice answers available to the question of why God chose these rules. For instance, God could be an act consequentialist and could have optimized the rules to produce the best consequences, including perhaps such consequences as the value of following and disvalue of breaking moral rules in addition to first order values and disvalues like pleasure and pain. We would expect a complex optimization to produce results with an appearance of arbitrariness. A sailboat hull computer-optimized to minimize drag is likely to have many parameters that look arbitrary to those who do not know how it was generated.

At the same time, we still have some serious Mersenne grounding problems. The plausibility of divine command ethics rests in the idea that God is a legitimate authority and legitimate authorities need to be obeyed. This suggests that logically prior to divine command ethics there is some sort of a proto-ethical general rule about obedience to legitimate authority. That rule itself will have to have parameters specifying which authorities are legitimate and what the scope of their authority is. And we will have the Mersenne problem of grounding these parameters.

Moreover, even if we do not have such a general rule about all authority, but a specific rule about divine authority, this will still raise some Mersenne problems. For, as Aquinas noted??ref, legislation only has a claim on our obedience when it is appropriately promulgated. And promulgation is a complex concept involving thresholds and parameters. It is not necessary for promulgation that all those subject to the legislation have heard of it. But it is not enough for the legislators to meet secretly, and write the legislation on a stone buried on public land. Intuitively, we need the legislation to be reasonably accessible to those governed by it, but there are many parameters hidden behind the word "reasonably", and we need grounds for them all.

Nor is it even the case that the promulgation condition on God's commands is met in a really clear way, so that all that would suffice is some proto-rule that has a really strict and non-arbitrary promulgation condition like that everyone governed knows of the rules. For any such strict condition is likely to have in fact been violated by God's commands, since there is no agreement on what God's commands are—or even on there being a God.

What is worse, when we focus on the Mersenne cases in ethics, it unclear that divine commands instituting the parameters would even satisfy a fairly modest promulgation that requires those who try really hard to be able to find what the legislation is when it is relevant to life. There surely are cases where we have tried really hard to figure out what is the right thing to do and we didn't succeed. Perhaps it could be argued that we didn't try "hard enough", but now we are the true Scotsman territory.??more?

6. Other attempts at escape

6.1. Particularism. One might try to escape the Mersenne questions by opting for particularism. On particularism, while there may be general rules like "Other things being equal, don't torture people", the application of these general rules to specific situations is not rule-governed. Hence, there won't be a rule specifying when one, say, favors a sibling over a cousin. Instead, there are particular facts about what to do in particular situations.

However, particularism only multiplies the Mersenne questions. For whereas on rule-based systems we had Mersenne questions about why the parameters in the rules had the values they do, now we will have Mersenne questions about why in particular actual circumstances C_1 we should act one way while in slightly different particular actual circumstances C_2 we should act a different way.

Furthermore, plausibly, there will still abstractly speaking be a function that assigns to each circumstances a hypothetical determination of how one would be obligated to act in that circumstance. There may, of course, be no formula specifying the function, but that does not affect the Mersenne question of why this function rather than another, perhaps similar one, is correct.

6.2. Brute necessity. Perhaps we could say that it is a brute, unexplained but necessary truth that the answers to the ethical Mersenne questions are as they are. The boundaries lie where they do, but there is no special ontology behind them: it's just a necessary truth that we should prefer parents to cousins, that an armed up-rising up against a regime responsible for Nazi-style atrocities is permissible while only non-violent protest against the faults of modern-day Canada is permitted, and so on.

Of course, brute necessities should never be a first resort in theorizing, but sometimes they might be acceptable as a final resort. Consider Mersenne-type questions one could ask about set theory. If the Zermelo-Fraenkel with Choice (ZFC) Axioms for set theory are consistent, then for every natural number n they are compatible with the hypothesis CH_n that there are exactly n cardinalities strictly between the cardinality of the natural numbers and the cardinality of the real numbers (the hypothesis CH_0 is the famous Continuum Hypothesis).??ref:check Suppose it turns out that in fact CH_{15} is true. We would have an excellent Mersenne question as to why it is CH_{15} that is true, but the mind boggles as to what could be a satisfactory answer to that question, much as it does in the ethical questions. Perhaps the truth of CH_{15} could be a brute fact, albeit a necessary one since it seems implausible that mathematical truths be contingent (though see ??Pruss for an Aristotelian metaphysical story on which they might be).

Some brute necessities can perhaps be admitted in ethics. For instance, if CH_{15} is necessarily true, then it is necessarily impermissible for us to punish someone for falsely informing us that CH_{15} is true. This impermissibility would derive from the impossibility of CH_{15} being false (and hence the impossibility of falsely informing someone of that it's true) and the impermissibility of punishing people for actions that they did not do. (It is possible, of course, to insincerely inform someone of a necessary truth. But that's a different wrong action, even if equally bad.)

But truly ethical brute necessities are deeply implausible. Here is one way to see this. Suppose there is a sequence s of one or more English sentences expressing your favorite set of fundamental and necessarily ture ethical norms. For instance s might be the single

injunctions "Love your neighbor as yourself" or "Maximize total pleasure minus pain of all sentients", or it might be a longer list. Encode s into a sequence of decimal numbers in some natural way, for instance by encoding each symbol in s into a three decimal digit ASCII number. It is widely believed—though it has not been proved—that π is a normal number, so every possible sequence of digits occurs in it. If so, then the decimal encoding of s occurs somewhere inside π—and even if not, it may well still do so. Suppose that the decimal encoding of s occurs in π as the nth through $(n+m)$th digits. Now consider this metaethical theory: (**??**)To do the right thing is to follow the English injunctions in three decimal-digit ASCII encoding between the nth and $(n+m)$th digits of π. Call this π-metaethics. On the hypothesis that the fundamental ethical injunctions are necessary and can be expressed in English, some version of π-ethics has the correct normative content. But, nonetheless, no version of π-metaethics has any plausibility. For there is no plausible normative connection between an injunction being found inside π and its being binding on us.

Admittedly, if we in fact found a sequence of English injunctions near the beginning of π (say, starting with the tenth digit), we would have some reason to follow them. But the reason would be something like this: The best explanation for why these injunctions are found in π is found in a being or beings that in some way incomprehensible to us can control mathematical truths or, more plausibly, the evolution of our linguistic systems, and there is good pragmatic reason to follow the commands of such beings. Perhaps they have our good in mind, perhaps they will get mad if we don't follow their commands, or perhaps they are trying to inform us of the true ethics. But nonetheless π-metaethics would be false. The reason these injunctions would apply to us wouldn't be that they are found in π, but something else, such as that a being with practical authority commanded them to us or a being with epistemic authority informed us of them.

In other words, a metaethics where the ethical claims are grounded in something intuitively of no relevant to our moral activity, such as the content of the digits of π, is not plausible. To be a candidate for a grounds of ethical claims, a thing needs to be ethically

compelling. For a more controversial illustration of this point, consider that no collection of the traditional attributes of God (omnibenevolence, creation, omniscience, omnipotence, etc.) is such as to make it plausible that the commands of a being with those attributes are what ethics is (??ref:MacIntyre??), and this is a strong reason to doubt divine command metaethics.

But now take some attempt at founding an arbitrary-seeming ethical principle on a non-compelling ground, say the digits of π, and remove the ground altogether. Removal of the ground surely does not make the story any better. Someone who said that what explained why we should favor siblings over cousins by a margin of twenty percent by saying that it is thus written starting with the nth digit of π would be ethically ridiculous (though if n is small, finding the injunction might be some evidence for its correctness). But suppose we drop the spurious π-based ground: surely the ungrounded ethical claim is no better off than the spuriously grounded one.

There may be ethical truths that are not themselves grounded. But these truths should be compelling ethically—perhaps the Golden Rule is like that—and not have an appearance of arbitrariness. And there may be arbitrary-seeming truths in ethics, but they are not fundamentally ethical.

6.3. A two-step vagueness strategy. It is very tempting to dismiss the Mersenne questions above with a two-step strategy. In each case, we first give non-arbitrary grounds for an approximate and vague determination of the parameters involved. Thus, while it is implausible to think that, say, social contract theory will generate a precise answer to the preferential treatment question, it is reasonable to think it will generate claims like: "Benefits to siblings are to be *somewhat* preferred to benefits to cousins." And, then, we simply note that the Mersenne question as to the grounds of the exact dividing line has the false presupposition that there is an exact dividing line—instead, we have insuperable vagueness.

An initial concern with the two-step strategy is to worry whether other ethical theories can actually generate sufficient non-arbitrary grounds that have the degree of precision

that we think really is there. This concern has two variants. One involves cases where we know what the facts generating the Mersenne questions are. Kantianism, for instance, is unlikely to generate even a vague morally-relevant distinction favoring siblings over cousins, and yet we know there is such a distinction. The problem of ranking types of goods generates difficult Mersenne questions as to what grounds comparisons that we know are there, such as that fundamental philosophical truths are more valuable than the pleasures of chocolate. The second variant of the concern involves cases where we agonize over what to do. Our agonizing is a sign of our intuition that there is an answer to a moral problem, albeit one we cannot discern. While we may not be seeking for absolute precision, and may be willing to accept some level of vagueness, in a number of cases we seek for more precision than the various alternatives to the form-based theory can ground.

Suppose the initial concern can be allayed in both of its forms, perhaps by clever development of a theory that does generate the vague moral claims and by biting the bullet and admitting that moral agonizing is out of place in these vagueness cases. There is still another question: how do we account for the vagueness here. There are three main contemporary accounts of vagueness: (a) non-classical logic, (b) supervaluationism and (c) epistemicism.

On non-classical logic approaches to vagueness, one typically increases the number of truth values beyond two. Consider an ethical Sorites series, where we fix some circumstances C and then say:

(A_0) Giving \$1000 to a stranger is better than giving \$0 to one's parent.

Now for each positive integer n, the following material conditional sounds plausible:

(A_n) If giving \$1000 to a stranger is better than giving \$$n$ to one's parent, then giving \$1000 to a stranger is better than giving \$$(n+1)$ to one's parent.

From A_0 and A_1, one concludes by *modus ponens* that giving \$1000 to a stranger is better than giving \$1 to one's parent. From this and A_2, by *modus ponens* one concludes that this is true even if what one gives one's parent is \$2. Continuing onward, once we get to A_{2000},

we conclude that it's better to give $1000 to a stranger than $2000 to one's parent, which is false. Thus, we need to reject one of the premises A_n. Presumably it's one with $n > 0$, since A_0 is clearly true. But a material conditional $p \to q$ is false just in case p is true and q is false. Hence, if A_n is false for $n > 0$, we have:

(4) Giving $1000 to a stranger is better than giving $\$n$ to one's parent and giving $1000 to a stranger is not better than giving $\$(n+1)$ to one's parent.

And that is exactly the kind of sharp transition that the vagueness theorist wishes to deny.

The non-classical approach to vagueness typically involves a logic with many truth values, e.g., a truth value for every number between 0 (fully false) and 1 (fully true). Then the statement:

(B_n) Giving $1000 to a stranger is better than giving $\$n$ to one's parent

is true for $n = 0$ (note that B_0 is just A_0), but becomes less and less true as n increases. If we have a large enough number of truth values, we can accept this at face value.

But, surely, B_1 and B_2 are simply true, too. On the other hand, B_{999} and B_{1000} are simply false. So it does not seem to be the case that we always have *strict* decrease of truth value with increasing n. And hence whereas in the classical logic reading we had one transition to be explained, from true to false, now we have at least two: from truth to truth values intermediate between true and false, and from intermediate truth values to falsity. And the transitions appear to be just as arbitrary as before. Thus we have doubled the number of Mersenne questions. And if we say that the second-order questions are also taken into account with multivalent logic—say, it's being the case for some n that B_n is neither true nor false that—then the multiplication of questions increases even more.

Perhaps, though, one can dig in one's heels and insist on strict decrease of truth value. But the precise assignment of intermediate truth values—say, B_{505} getting a truth value of $T_{0.51}$—also calls for an explanation. Thus it seems we have a vast multiplication of Mersenne questions. But there is a response to this argument: ??refs argues that the exact truth values are a mere feature of the logical model and all that has reality is their ordering.

And the ordering of the truth values is, perhaps, quite non-arbitrary in that B_m is truer than B_n precisely when $m < n$. But the insistence that the ordinal properties of truth values is what has reality still does not escape the multiplication of Mersenne questions. For consider a different set of ethical questions involving a threshold. For instance, let C_x say that one has a duty to obey the orders of a government that cares to degree x about the common good, for some method of x of quantifying care about the common good, where, say, $x = -1$ corresponds to the Nazi German state and $x = 1$ corresponds to modern Finland. Then C_{-1} is pretty false C_1 is pretty true. But even if all we insist on is the ordering of truth values, then we will still have a vast, perhaps infinite, number of Mersenne questions like:

(5) At what value n does B_n become less true than $C_{0.24}$?

For clearly B_0 is truer than $C_{0.24}$ while B_{2000} is falser.

?? higher levels of multivalued logic

The most common response to vagueness these days is supervaluation. The terms of a sentence can have multiple precisifications, with a different truth value corresponding to a different choice of precisifications. "Bob is bald" may be true if we precisify "bald" as having less than half a cubic centimeter of scalp hair and false if we precisify it as having fewer than a meter of hair. Then we have vagueness. When, on the other hand, a sentence is true (respectively, false) under all precisifications, we say it is super-true (super-false).

In the ethical examples, such as whether it is better to give $200 to a stranger or $100 to a parent in circumstances C, presumably the supervaluationist escape from Mersenne questions will be that no matter how far we precisify C, the statement will be vague due a vagueness in ethical terms such as "better" or "right" or "wrong" which have multiple precisifications yielding different truth values for the ethical claim. For instance, it may be better$_{17}$ to give the double amount to the stranger but not better$_{40}$. Indeed, on a view like this, we will have cases (precisely specified by means of the monetary amounts and the circumstances C) where for some precisifications of "better" it will be better to give to the parent and for others it will be better to give to the stranger.??explain-better

Just as in the multivalued logic case, this multiplies Mersenne questions. For where previously it looked like we have a transition from its being true that it's better to favor the stranger to its being not true, now we have two transitions: from its being super-true that it's better to favor the stranger (say, when the amount of benefit to the stranger is extremely large) to its being vague whether it's better to favor the stranger to its being super-false that it's better to favor the stranger. And supervaluating at the next level up—say, supervaluating "super-true"—only multiplies the Mersenne questions more.

But there are some additional problems for the supervaluationist response. A standard objection to supervaluationism in general is that it implies that it is super-true that there is a sharp boundary of "bald": for, given any precisification "bald$_i$", there is a sharp boundary for it. In doing this, supervaluationism explicitly forces the denial of its governing intuition that there are no sharp boundaries.

Finally, the application of supervaluationism to ethics is itself deeply problematic. It is truism that we have reason to do what is better. Truisms had better be super-true. This implies two possibilities with regard to the truism. Either for every precisification "better$_i$" we have reason to do what is better$_i$, or else we need to precisify "reason" and "better" in lockstep when we precisify the truism, so that for every i it will be true that we have reason$_i$ to do what is better$_i$. Neither option is satisfactory.

If for every i we have reason to do what is better$_i$, given the existence of infinitely many precisifications here, it seems that the choice whether to favor the stranger and the parent is governed by infinitely many reasons on both sides. This infinite multiplication of reasons is implausible. Moreover, there is no overall winner here—no reason all things considered—for if there were, then we could raise our Mersenne question with regard to the overall winner, and we would be no further ahead. But saying that there is no on-balance reason here denies the intuition that cases near the boundary are hard cases, that it is a difficult question to figure out whether to favor the parent or the stranger, since as soon as one can see that one is in the vague region, one could just conclude that neither action is on balance required by one's reasons.

But if there are infinitely many ways to precisify "reason", none of them privileged, then this undercuts the very idea of our life being governed in a non-arbitrary way by rationality. It seems entirely arbitrary whether we follow reasons_{17} or reasons_{40} in our lives. Many questions of rationality turn into purely verbal questions as to how "reason" is to be precisified. And the same goes for related terms like "morality" and "virtue". This does not seem to do justice to the non-arbitrariness that is central to a realist conception of reason, morality and virtue. The point here is similar to the one raised in ??backref regarding π-metaethics: it would be arbitrary to require obedience to the commands that are found starting with the billionth digit of π, rather than the commands found in some other location.

Finally, consider an epistemicist theory of vagueness according to which there is a true semantic theory that assigns to each term the precise meaning it has in the light of the patterns of our use of that term, but neither that theory nor the empirical data on the patterns of language use are available to us in sufficient detail to settle the meaning of vague terms. Thus, there is a precise fact as to how much hair one can have and yet have "bald" apply to us, a fact grounded in the patterns of our use of the word "bald", but it is a fact that is not accessible to us. Similarly, there is a precise meaning of "right", "better" and similar ethical terms, a fact grounded in the patterns of our linguistic usage. If the transition between bald and non-bald occurs between 98 and 99 hairs, there is nothing mysterious about the fact that someone with 98 hairs is bald and someone with 99 is not, just as there is nothing mysterious about the fact that a backless chair is a stool.

But the problem here is exactly the same as the last problem with supervaluationism. Ethical questions are turned into purely verbal questions. Just as on supervaluationism, there is a multiplicity of concepts closely corresponding to our words "right", "better" and "reason". On supervaluationism, none of these concepts was privileged, which turned ethical questions into purely verbal questions, undercutting the idea that our lives are to be governed by reasons and morals. On epistemicism, there *are* privileged concepts that exactly correspond to the words, but they are privileged purely linguistically—it just so

happens that these privileged concepts better fit with our usage under the correct semantic theory. We get an unacceptable arbitrariness on which if our linguistic practices were somewhat different, we would be using the word "better" differently, and there would be nothing less natural about that usage. If so, then our actions' being governed by the better or the right, rather than by some variant property, would be entirely arbitrary.

In summary, non-classical logic violates classical logic, which should only be a last resort, and further multiplies rather than resolving Mersenne questions. Supervaluationism likewise multiplies Mersenne questions. And, perhaps most seriously, both supervaluationism and epistemicism as applied to ethics turn ethical questions into purely verbal ones, undercutting a robust realism.

6.4. Anti-realism. Retreating from realism in ethics to error theory does, of course, remove all the Mersenne problems in ethics. But the cost is high: it is incorrect to say that genocide is wrong. Moreover, since some of the Mersenne problems involve not just morality but also prudential reasoning, this requires one to deny the correctness of standard prudential reasoning. But perhaps the most serious problem with the error theoretic solution is that we will have parallel Mersenne problems in other normative areas, such as epistemology (??forward) and semantics (??forward), and the cost of error theory there is very high indeed: indeed one will no longer be able to correctly say that one *ought to* accept error theory.

A more moderate solution is to opt for a form of ethical relativism. Relativism, of course, suffers from serious and standard objections.??refs Perhaps the most obvious is that it justifies an ultra-conservative approach: for if what I (in the case of individual relativism) or my society (in the social variant) thinks is guaranteed to be true, then I or society has no reason to take variant views into account, since if you disagree with me or my society, you're guaranteed to be wrong (from my or my society's point of view).

Moreover, relativism is itself prone to Mersenne questions. Consider individual relativism first on which a moral claim is true just in case one believes it. The Mersenne

question here will be most obvious if one opts for a view on which belief reduces to having a credence above some probabilistic threshold, say 0.95. For then the relativist view comes down to the thesis that a moral claim is true just in case one assigns it a credence of at least 0.95. But that seems arbitrary. Why should one be obligated to do what one has credence of 0.95 in, but not obligated what one has credence of 0.93 in? So we have a threshold problem.

Many, however, resist the reduction of belief to a credential threshold. But if we do not so reduce belief, we should then see belief as just one positive doxastic state among many, such as surmising, being inclined to think, believing, being confident that, and being sure that. Moreover, a little reflection shows that such classifications are too coarse grained to do justice to the richness of our mental life. So we have a Mersenne question: Why are moral claims made true by my believing them rather than my surmising them or being sure of them?

And thinking that the problem here just involves degrees of confidence is probably neglecting much complexity in the human mind. There is likely a continuum between fully believing and merely acting as if one believes. Why does moral truth show up in the continuum where it does? Or think of the case when the psychotherapist diagnoses one with a subconscious belief. Either such define moral truths or they do not, and whichever it is, we have a Mersenne question as to why. And consciousness itself may come in degree.

Furthermore, a narrow relativism that just makes those moral claims that we actually believe is very implausible. Suppose I believe that it is wrong to eat animals, and I know that cows are animals, but I do not actually draw the conclusion that it is wrong to eat cows. On such a narrow relativism, it would be wrong for me to eat animals but it would not be wrong for me to eat cows, even though I know them to be animals. This is incredible. So we want to extend moral truth at least to things that clearly follow from my moral beliefs. But probably we do not want to extend it to things that follow in ways that are far beyond our ability to know. For, first, if do extend it thus far, then a Kantian might end up counting as a relativist, since the Kantian may think that moral truths are necessary truths, and that

necessary truths follows follow from everything. And, second, it seems that this loses sight of the internalist motivations of relativism. But if we restrict moral truth to things that follow *sufficiently easily* from our beliefs. And we will have a Mersenne question of grounding where the line of sufficient easiness lie.

If our relativism is of the social sort, we will have analogues to the above Mersenne questions raised by belief and consequence. And we will have more Mersenne questions There is a complex and difficult literature on how to attribute doxastic states to a community. A reasonable reading of that literature is that there is a multiplicity of concepts that can be expressed with a phrase like "The committee believes that p." For instance, belief by the vast majority of the committee members is enough on the more reductive concepts, while on more procedural versions of the concepts the committee's belief requires some sort of a joint procedure, such as a vote. There will be many answers here, corresponding to a broad spectrum of takes on what a community's beliefs is. And a social relativist will then have a Mersenne question as to why moral truth is defined by the particular take in question.

The second set of Mersenne question arises from the question of identifying what counts as one's community. I am a citizen of two countries and a permanent resident of a third. Are the moral beliefs of all of these communities—no doubt, mutually contradictory in various ways—true for me, or only of one? Do moral beliefs come to be true as applied to me because I am legally a member of the community, or because I identify with it emotionally, or because I would like to identify with it emotionally? Or is it, perhaps, that every community's beliefs are true for the community, but there is no such thing as being true for the members of the community? (That would nicely solve the problem of contradictions between the various communities I am a part of.) How large does a community have to be to define moral truths? Is a chess club a commmunity that defines moral truths? What if the chess club goes down to one member? Are a pair of friends a community? It is clear that there are many degrees of freedom in a social relativistic theory, and we would have a Mersenne question corresponding to each of them.

7. Hume's objection: Complexity, instinct and nature

Hume saw the complexity of property, inheritance, contract and jurisdiction, and used this complexity to argue for apparently *against* a Natural Law account:

> For when a definition of *property* is required, that relation is found to resolve itself into any possession acquired by occupation, by industry, by prescription, by inheritance, by contract, &c. Can we think that nature, by an original instinct, instructs us in all these methods of acquisition? (??Enquiry::Morals)

To further expand on the complexity, Hume notes the vast variety between these rules in different societies, and analogizes them to the variety of architectures found in housing across societies. A better account, Hume insists, is that we simply engineer rules for the sake of social utility, just as we engineer houses for various ends, and in different environments this results in different solutions, albeit ones with a lot of commonality.[15]

Three responses are possible.

First, we do actually have good empirical reason to think that our moral intuitions and instincts vary quite a bit from case to case. An account of our actual moral instincts is likely to have enormous complexity. Granted, some of the complexity of our actual moral instincts is due to failures. For instance, racist or sexist biases introduce complexity by distinguishing cases that do not morally differ. And it would be *correctly functioning* instincts that we would expect the natural law to be expressed through. Thus, even if Hume's argument fails with respect to our actual instincts, it may work with respect to correctly functioning instinct, since we have reason to think that this would in some respects be simpler than our actual instinct.

However, in fact, it is far from clear that purifying our instinct of malfunctions would make for so much simplicity as would be needed to support Hume's argument. Removal

[15] It is natural to try to solve the problem by adverting to positive law. But Hume notes that there is much complexity with respect to the institutions by which positive laws are produced.

of some biases will indeed remove some complexity. But enough complexity is likely to remain that Hume's argument will not be convincing. Moreover, it is likely that our instincts sometimes fail through not making distinctions that should be made, and hence in some respects our correctly functioning intuitions would likely be more complex.

Second, the account I am defending is one on which our forms set norms. These norms can be arbitrarily complex. Granted, there will be a harmony between these forms and our instincts and intuitions. But this harmony is not a one-to-one mapping. There is such a thing as normal and abnormal food for a type of organism, and we expect the organism to have instincts that tend to direct them to normal consumption and away from abnormal consumpton. But just as correctly functioning sight can still err, so too a correctly functioning feeding instinct can lead organisms to ingest what is abnormal and to refrain from what is normal, especially in environments that have abundances different from the ones present where the organisms evolved. Thus, our natural instinctive preference for high-calorie foods leads humans in affluent Western countries to abnormal consumption, and hence to Reflection can gain additional information as to normative consumption on the basis of high rates of obesity.??refs By reflecting on our nutritive instincts and *other data*—such as the teleology of nutrition and medical facts about us—we can get additional information as to what is normative consumption for an organism, including a human one. This additional information is still fallible, and may fall short of the complexity of the norms involved. And what goes for our nutritive instincts is even more strongly applicable to our moral ones.

Third, recall Hume's own solution to the problem that the complexity of the rules is a result of our social engineering for social utility. Hume's solution is subject to complexity problems of its own. For instance, what groups count as societies and how do we aggregate the benefits to individuals to get a social utility, etc.? But something similar to Hume's solution can be appropriated for the natural law account. We can suppose norms in our nature establishing the goals for certain social institutions, such as property or state authority, and perhaps establishing some constraints on how these goals are to be pursued,

and at the same time requiring us to engineer institutions that satisfactorily pursue these goals within the contraints. These norms will be complex, but will be less complex than the vast complexity in our social institutions. Thus, we have a bootstrapping, from fairly complex norms setting the ends and constraints for social institutions, to the institutions themselves.

CHAPTER III

Ethics and metaethics

1. Metaethics

Metaethics is an account of why the most fundamental ethical truths are true. If we were to make a wish-list for metaethics, it would arguably include the following desiderata for what should follow from the theory:

(6) Ethical truths are objective

(7) Ethical truths are knowable

(8) The explanation of fundamental ethical truths makes them morally compelling to us

(9) The normative implications are plausible.

Recall our π-metaethics on which what made ethical claims true is that they were encoded at some specific position in the digits of π. This gave us objectivity and knowability (at least given the specific position and encoding system).

An individual relativism that says that the right is what agrees with one's belief as to what is right, on the other hand, gives us knowability, and insofar as we find morally compelling the idea that we should obey our conscience it gives us some compellingness, but it lacks objectivity. Furthermore, its normative implications as to what I ought to do are very plausible to me, since obviously I find my own moral views plausible, but the theory's implications for what Hitler should do—namely, that precisely those actions that he believes are right are the ones he ought to do—are implausible to me (and you) in light of the odiousness of his beliefs.

Utilitarianism, on the other hand, considered as a metaethical theory about the nature of the right, yields objectivity, knowability and moral compellingness (the idea that what

we should do is maximize the good is among the *prima facie* most plausible of moral ideas), but it yields a lot of very implausible normative consequences.

A Natural Law metaethics on which for an action to be right is for it to be a proper exercise of the will according to our nature yields a limited objectivity: it makes the right be relative to our kind. But as we saw in Chapter II, this degree of relativity is highly plausible: it is plausible that ethical requirements do vary between different kinds of intelligent beings.

The Natural Law metaethics yields knowability when we accept the Aristotelian harmony theses that things generally function correctly and that the various norms for a thing tend not to conflict. For instance, given such a thesis, the norms for our emotions—including emotions such as moral repugnance or moral admiration or the feeling of obligation—are likely to cohere with our norms for our actions, and by and large our emotions and actions are apt to be correct. This enables us to evaluate normative ethical theories according to the constraint of whether their requirements fit sufficiently with our emotions and require actions that are not too distant from those that people actually perform, especially in the case of people whose lives appear to be harmoniously flourishing. We thus have a rational equilibrium epistemology for our ethics.

The basic idea here is that we ought exercise our will correctly. This is so compelling that it smacks of triviality. Nonetheless, the claim is not trivial, since it provides an analysis of the moral ought in terms of the functional correctness of our wills. We find compelling the idea that we should be true to ourselves. But to be true to ourselves is not just, as is popularly supposed, being true to our changing beliefs and values, but it is to be true to that which makes us be the kinds of things we are: our nature.

The metaethics of right action as the proper functioning of the will is *prima facie* compatible with a very broad variety of normative theories. It seems we can imagine a being whose will's proper function is to will maximal total utilityThus, Aristotelian metaethics is compatible with utilitarian normative ethics, but not with metaethical utilitarianism on which the right is *defined* as what maximizes utility., or to will in accordance with God's

commands, or to will what is universalizable, or to will one's flourishing, or even to cause maximal harm to self. Some of these views will, however, be less plausible given other Aristotelian commitments, such as harmony theses. The harmony theses make it unlikely, for instance, that the right thing be maximal self-harm. Indeed, harmony theses ensure that the normative consequences of the ethical theory be, by and large, fairly intuitive. At the same time, there is a real possibility of error, and of correction of that error.

Natural Law metaethics does justice to the idea that the source of our obligations is in us, rather than in some external fact—such as a divine command—whose moral relevance is questionable. We are our own moral legislators, but because our nature is metaphysically not up to us, we do not have a choice as to what we legislate and we can be wrong about what we have in fact legislated. Natural Law metaethics will thus accept with modifications both the relativist's and the Kantian's insistence on autonomy, but without the ultra-conservative consequences of relativism on which we are always guaranteed to be right and hence never have reason to change our views, and while avoiding the merely formal character of Kantianism which makes it unlikely to yield sufficient normative consequences to guide our lives.

Other metaethical theories may satisfy the four desiderata as well.

2. Internality of morality

An attractive feature of a metaethics is that it grounds moral truths in features of moral subjects. It seems compelling to ask "So what?" about an external ground of moral truth, such as divine or social commands, or Platonic moral facts.

Socrates argued that that virtue is the center of our flourishing on the plausible grounds that moral virtue is the health of the soul.??ref What makes this argument an appealing invitation to moral virtue is that health is an internal good of the person. The analogy opens us to seeing moral virtue as a good grounded internally to the person. But since virtue is a disposition to living by moral truth, this attractive internalism requires moral truth to be grounded internally to us.

Additionally, an internalist intuition is probably a part of the explanation of the attractiveness of subjectivist metaethics: it is harder to say "So what?" to one's own beliefs and values. Aristotelian ethics grounds moral truths in the agent's form, which is a central metaphysical constituent of the agent. It is not exactly the same ground as a subjectivist offers, but it is still an *internal* ground.

In fact, in an important way the agent's form provides a *more* internal ground of moral facts than the agent's own beliefs. Typically, we do not choose our beliefs. At least typically, we catch them, much as we catch flus and colds from those around us.??ref It is largely an accident that they are what they are, especially if there is no metaphysical truth for them to reflect. One of the motivations for subjectivist theories is the intuition that morality is internally grounded. But when the internal ground is something that is itself produced by the accidents of surrounding culture, as is indisputably the case for some of our moral beliefs, then we have a betrayal of the internalist intuition. On the other hand, on an Aristotelian theory, the ground of moral truth is our form, which is a central essential metaphysical constituent of the human individual. Note that a Kantian has a similar advantage: the ground of morality is our rationality, and our rationality is central and, very plausibly, essential to us.

Furthermore, a plausible subjectivism needs to take into account potential disagreement between the agent's beliefs and values. A reasonable solution is to insist that more deep-seated aspects of one's psyche take precedence over more accidental features. If you believe that people should get the benefit of the doubt, but in the heat of the moment you believe that the person you are arguing with should not get the benefit, the subjectivist should probably opt for defining your obligation in terms of the general belief, as it is more likely to be deeply planted in you.[1] We might say that our nature as human beings,

[1] An alternative is a Frankfurt-type??ref idea that higher-order values and ideas should take precedence over lower-order ones. Thus, if Bob believes that Alice should not get the benefit of the doubt, but also believes that he should believe that she should, then it is the second-order belief that trumps. This does not settle all the questions of disagreement between first-order beliefs—there need not be a second-order belief that decides between them. Moreover, although it is psychologically uncommon for a higher-order belief or value to be

understood in the optimistic Aristotelian way as both instituting norms and impelling us to believe and follow them, has a kind of depth that trumps the deepest of our contingent beliefs.

3. What are moral or rational norms?

The idea that norms are species relative suggests that in the space of possibilities—and perhaps in extraterrestrial reality as well—there will be species of beings that are intelligent enough to have advanced science and technology, but whose natural behavior will be quite different from us. This raises a difficult question as to what makes a particular set of natural norms count as a set of moral or even rational norms, and hence makes the species that possesses them a species of moral or rational agents.

Not all natural norms are moral or rational norms. The natural norm behind a properly functioning horse shedding in the spring is neither moral nor rational. Plausibly, a necessary condition for a moral norm is that it govern voluntary behavior. But the question of what behavior counts as voluntary is difficult. It is tempting to say that the behavior of an entity is voluntary if it is subject to reasons. But reasons live in a space made possible by rational norms, and so it seems we need an account of rational norms, at least, to make sense of what behavior is voluntary.

Here is one highly speculative Aristotelian functionalist way to answer the questions. First, we connect reasons and norms with goods considered as such. An (internal) reason for a behavior is an apprehension of the behavior as *good*. A mouse may apprehend cheese as yummy, or maybe even as nutritious, but not as *good*. Of course, being yummy or being nutritious is thereby good, but to apprehend as yummy or nutritious is not to apprehend as good.

---

a mere whim, there is nothing to rule that possibility out. Suppose that I think to myself: "To illustrate the point I am making in this footnote, it would be good if I had the really wacky belief that it's required to insult blue-eyed people on Friday the 13th of January in a prime-numbered year." Nobody should think that such a whimsical second-order value or belief yields a first-order obligation.

Then a necessary condition for a behavior to be voluntary is that it comes from such a reason. This, of course, raises the infamous problem of in-the-right way. A behavior can be caused by a reason without being voluntary. The famous case is the belayer who intends to murder a climber by dropping the rope, and then his hands start shaking at what he has intended to do, which results in an involuntary dropping. Whatever reason he had for the murder is the cause of the dropping, but the dropping is involuntary. Aristotelian metaphysics does, however, seem to have a tool for solving this problem. Causation can be seen to be teleological in nature, and we might say that it is a primitive fact that sometimes the effect *is* a fulfilment the teleology of the cause, in which case we can say that the effect is caused in the right way. A voluntary behavior is one which is a fulfillment of the teleology of the cause.

Finally, the will can then be functionally defined as the system by which reasons lead to behavior that promotes the goods apprehended by these reasons. A rational norm is a norm of behavior of the will favoring some or all reasons, and a moral norm is a norm of behavior of the will favoring some or all reasons that themselves are focused on a good not considered primarily as a good to self.

This is not, of course, the only way to define which norms are moral or rational. And it is quite possible that the question of which norms are moral or rational is largely a verbal question. Go back to the characterization of reasons in terms of goods. The mouse takes the cheese to be yummy, and that is not taking the cheese to be good. But humans often represent good things in thicker ways: as beautiful, courageous, or even divine. Couldn't we imagine a continuum of animals where at one end the cheese is represented as yummy and on the other as having gustatory beauty? Somewhere in the continuum we have moved to representing the cheese as having a thick form of goodness. But where this happens is unclear.

We have a certain set of norms for the will. We can call these rational, and a subset of them moral. What hangs on whether a different set of norms governing the behavior of a

different class of beings counts as rational or moral? Couldn't this be like the question of which animals' hard projections count as horns?

But perhaps the question of which things are moral agents matters for first order moral questions about interspecies relations, such as which organisms we are permitted to eat, whose lives we should save, etc. However, it is not clear that the answers to these questions will neatly line up with determinations of the boundaries of moral agency. Consider intelligent whales that are fine philosophers and scientists in their epistemic life, and that even enjoy contemplating the good, but do so purely non-practically, without the good being any motivator for their actions, which are all instinctive. It would seem wrong to each such beings, even if they turn out not to be moral agents. On the other hand, imagine a shrimp that has the normal behavioral complement of a shrimp, with one exception: it represents the ingestion of algae as good, and voluntarily pursues this good as such. And its intellectual abilities are the minimum needed for an ordinary shrimp's life as combined with the most minimal possible concept of the good. It may well be wrong to eat such shrimp, but given a choice whether to save the life of one of these minimally moral shrimp or the life of one of the intelligent but amoral whales (of course, we should not be biased here by size!)

4. Flourishing

A substance flourishes to the extent that it functions in accordance with its norms. Acting morally rightly is a case of functioning in accordance with the norms for the functioning of the will. Thus, acting rightly morally is an aspect of flourishing for those substances that have a will. At the same time, unless we should deal with a substance that consists of nothing but will, there will be other aspects of flourishing. Because of this, conflict between moral rightness and self-interest is in principle possible.

Admittedly, Aristotelian harmony tends to limit such conflict. Right action tends to promote other aspects of a substance's well-being. But nonetheless just as jogging sometimes promotes cardiac wellbeing at the expense of joint wellbeing, so too right action promotes volitional or moral wellbeing at the expense of life and other goods.

Starting with Socrates, Western ethical reflection has often insisted on moral wellbeing being the most important aspect of a human's well-being. This may seem to be necessary for preserving the idea that one should do the right thing even when this costs one heavily, by allowing one to insist that the cost of doing wrong is always greater than the benefits. But the thesis that moral wellbeing trumps other forms of wellbeing is neither necessary nor sufficient for preserving the need to act rightly.

It's not sufficient since even if moral wellbeing trumps other forms of wellbeing, there are imaginable situations where doing the right thing will on balance very likely harm one's wellbeing. For instance, suppose I am a bank employee of mediocre morals and the best empirical evidence available to me shows that taking an evening ethics class from Professor Kowalska would be deeply inspiring and turn me into a vastly better person. Unfortunately, the only way I could afford the tuition is to embezzle a thousand dollars from a billionaire's account. This embezzlement is wrong, but I can reasonably expect to be a much better off morally from it.

And it's not necessary that moral wellbeing trump other forms of wellbeing, because if morally right action just is action in accordance with the norms for the will, then it is clear apart from any trumping thesis why morally wrong action is defective: it is defective because it fails to be an instance of the proper functioning of the will. One may be the better off if one does the wrong action, but one will still have acted defectively.

Moreover, it is unlikely to be true that moral wellbeing always trumps other forms of wellbeing. Suppose the best science shows that on average there is on the whole a moral improvement—perhaps in the area of compassion—from suffering severe headaches, but this improvement is tiny. A parent who knew this should still relieve a child's severe headache, and wouldn't be acting contrary to benevolence in relieving it.

Nonetheless, it is plausible that moral wellbeing is typically the most important aspect of our wellbeing and that typically other forms of our wellbeing are appropriately sacrificed to it. This gradation is itself encoded in the human form which specifies what is good for us and the ordering between the goods.

Traditionally, Aristotelian action theory has insisted that we always act for our happiness. This happiness thesis is compatible with a metaethics on which right action is the proper functioning of the will, but is neither entailed by it nor particularly plausible. While proper functioning is always good for a substance, a substance when functioning properly in some way need not be doing so *in order to* function properly in that respect. When a flower opens up in the right season, its opening up plausibly has as its end the good of reproduction rather than the good of opening up. Similarly, when you make dinner for your child, your right action is good for you, but you are doing it for the sake of your child and not for the sake of the action itself.

5. Supererogation

It is initially plausible that one should perform an action best supported by moral reasons. But this principle leads to a very demanding ethics with no room for supererogation. Consider a heroic sacrifice of one's life for the sake of others—say, jumping on a grenade to save innocent lives. In typical cases, this is more praiseworthy than refusing to make the sacrifice. If it is more praiseworthy, it is surely better supported by moral reasons. But nonetheless, someone who refuses to make the sacrifice is typically not to be blamed. ??refs

It is very puzzling how to make sense of the permissibility of doing an action less well supported by moral reasons. But a Natural Law account has a way by making a distinction between flourishing and languishing that applies very broadly. A cheetah capable of reaching 120 km/h flourishes less than a cheetah capable of reaching 125 km/h, but neither languishes, while a cheetah only capable of 12 km/h does languish in respect of running. Many areas of functioning have norms allowing both a binary distinction between failing

to meet the norm and meeting the norm, and a distinction between degrees of flourishing in respect of the norm.

We can say that an individual's state or activity S is supernormal in respect R provided that the individual flourishes in S with respect to R more than they would in the case of some alternative state or action S' that still would not be an instance of languishing with respect to R. And now we can say that an action is supererogatory provided that it is supernormal with respect to the exercise of the will. ??connect with later discussion

The mystery of supererogation was how one can count as permissibly acting if on balance one has moral reason to do something better. But once we see that this is just a special case of a phenomenon that extends far beyond morality, and indeed beyond the life of rational beings, the mystery should be significantly decreased.

6. Equality and human dignity

We have an intuition that all humans, or at least all non-disabled adult humans, are in some important sense equal. It is difficult, however, to figure out what the relevant sense of equality is.

A utilitarian has a neat account of equality: we maximize total utility, and total utility is the sum of individual utilities *with equal weights*. But as I have argued, there are serious problems with utilitarianism.??ref And indeed, the equal weight part of the view is itself implausible. Suppose that Alice has kidnapped Bob and taken him to a distant uninhabited planet. On her way back to earth, Alice had an accident and was marooned on a different uninhabited planet. Alice and Bob have enough supplies to survive a year. Carl knows the above, and is himself marooned on a third uninhabited planet, but has enough supplies for two lifetimes. He can send a drone with half of his supplies to Alice or to Bob, but not to both. No one besides him and his recipient will know. It seems clear that Carl should send supplies to Bob rather than to Alice. But on an equal weight utility view, this is false.

Perhaps a deontologist can do better, saying that people have equal rights. Now, rights are correlate to duties, so presumably the equality of rights is correlate to equality of duties

of non-infringement. But while it is equally true of everyone that they have the same basic rights, and it is equally true that it is wrong to infringe on these rights, this is not actually enough for equality of rights. For instance, if E is your right not to be deprived of your eyes and T is your right not to be deprived of your toes, it is equally true that you have E and you have T, and it is equally true that it is wrong to infringe on your E as that it is wrong to infringe on your T. But E and T are not equal rights, since E is more stringent than T. For instance, if one attacker is about to infringe your E and another is about to infringe your T, we would expect the police to prioritize your right to your eyes over your right to your toes. Moreover, while it is equally true that both attacks are morally wrong, they are not equally morally wrong (similarily, it is equally true that elephants and whales are large compared to mice, but whales are larger, both absolutely and as as compared to mice): both assaults deserve significant penalty, but the assault on eyes deserves a greater. Within a single individual there are many unequal rights but that are, nonetheless, equally truly possessed by one. Thus equal truth of possession of rights is insufficient for equality of rights.

Do we in fact have equal rights? Well if we compare rights by stringency, as we did E and T, this is not clear. We have even stronger reason to stop the murder of a complete innocent than to stop the murder of a someone who just falls short of deserving the death penalty, and murder of a complete innocent deserves a greater punishment—if not in law, then in public opinion (??cf.Mill). Similarly, it seems clear that we should make a greater effort to stop the murder of a person with a decade of life ahead of them than the murder of someone who has five minutes of life left anyway.[2]

Or perhaps equality is a political matter. Each adult non-disabled human, perhaps, should have equal input in social decisions. But while this may be the best practical way of running things, it is not clear that it has an imperative beyond the practical. Suppose that we had an indisputably correct way of measuring wisdom and virtue. Would it be

[2]We can assume for the sake of the example that this is not a case of euthanasia.

clearly wrong for a society to weigh votes according to the wisdom and virtue of the voter? Besides, it seems that our intuitions about equality go beyond the political sphere.

If we look at the range of human abilities, we see significant inequalities. But the Aristotelian can at least say this. Possessing the human form is not a function of possessing multiple valuable properties that some have to a greater and some to a lesser degree, and each human has the *human form* to exactly the same degree. The possession of this form is what makes us human, and it is a valuable thing that grounds our dignity.

But does not everyone, Aristotelian or not, agree that there is such a thing as being human, and that we are all equally human? Yes, of course. But on non-Aristotelian views, we differ from one another in regard to the properties that make us human.

Let's say that we define humanity in terms of having DNA sufficiently close to some standard H. Then, first, we likely differ in how closely we hew to that standard. Second, even when we are equally close to the standard, we are close to it for different reasons: I may be close to H because I am very close to H in portion A of my DNA and somewhat close to H in portion B, while you are close to H because you are very close to H in portion B while being somewhat close in portion A. Third, if we take humanity to be the valuable thing that we are equal in, then it seems we are taking closeness to H to be valuable. But closeness to H seems valuable in virtue of the fact that H is a way of coding for various valuable properties, such as a variety of aspects of intelligence (insofar as these are heritable) and a variety of valuable physical features. It is indeed valuable to have DNA that codes for these good things, but nonetheless we differ in this valuable thing—for some of us have DNA that better codes for, say, musical intelligence, and others (very likely including the author of this book) have DNA that codes less well for musical intelligence. Thus our being human on a DNA standard is constituted by a variety of independently valuable genetic features, and we differ in this value. This does not seem to be a good account of human equality.

Or suppose that we define humanity in the way biologists define species, as groups with significantly more genetic interchange within the group than outside it. We are members of the human species, then, because of patterns of genetic interchange in our ancestry. Again, this is something that comes in degrees: there are human populations that are more or less isolated from the bulk of humanity.??ref And, second, it is far from clear that being a part of this genetically interchanging population as such has the kind of immense value that our equality qua human beings needs to have in order to do justice to our intuitions about human equality.

But how does being human ground a deep fundamentally equal value of human beings? We can say that we differ with regard to many valuable features we have, but a particularly valuable, perhaps the most valuable one, feature is our form—it is that which makes us human, which sets the norms that define our dignity, that gives us our high calling of pursuing intellectual and moral virtue, and that is the ground of the possibility of our flourishing. Items used in worship—holy books, vestments, thuribles, altars, and locations—are considered sacred in significant part due to their ends. Humans have high ends, and there is a deep value in that we all have in virtue of these ends. Granted, there is a further value when these ends are fulfilled. But there is an equality in the baseline—in the having of the ends—and we can plausibly say that it is this equal possession of high ends that constitutes our dignity.

We can thus distinguish our valuable features into two classes: dignity, which we have in virtue of the teleological structure specified by our form, and what one might call accidental value, which consists in flourishing according to that teleological structure. We can then say that we are all equal in dignity but vary in accidental value.

At the same time, we talk of "dignity" in the case of roles that one can accidentally have, such as the dignity of a monarch (which can be infringed with *lèse majest´*). Here we can follow Aristotle's idea that philosophically important terms have a focal sense and non-focal derivative senses, so that the focal sense of "health" is the proper function of an organism's body, while derivative senses apply to that which indicates focal health (e.g.,

healthy urine) or produces it (e.g., healthy food) or is in some other relevant way related to focal health. Focal dignity is the value we have in virtue of having the role *human being*. Non-focal dignity is value we have in virtue of having some accidental role, like monarch or parent. And just as we have a relationship between focal dignity and accidental value, where the accidental value is subordinate to focal dignity and fulfills the teleological in it, liekewise subordinate to non-focal dignity (which itself is a kind of accidental value) there will be accidental values of fulfilling the teleological structure in the non-focal dignity (e.g., being a just monarch or a loving parent). It may not be a coincidence that it is natural to talk of our humanity as imposing a high *calling* on us, and we call many accidental roles *callings*.

7. Supervenience

It is widely held that moral facts supervene on non-moral facts: if two possible worlds differ with respect to moral facts, they must differ with respect to at least one non-moral fact. Similarly, it is held that normative facts in general supervene on non-normative facts. The difficulty is then to explain the supervenience relations.

The nature-first theorist has a complex relationship to both supervenience claims. Let us begin with the supervenience of the normative on the non-normative, and first consider general normative claims such as that every sheep should have four legs and every human should refrain from torturing the innocent. Such normative claims are necessary truths, since their truth is a part of what makes a sheep a sheep and a human and a human. Necessary truths vacuously supervene on any basis we might choose, since there are no possible worlds that differ with respect to necessary truths.

But what about *particular* normative claims, such as that Sally ought to have four legs or that Biden ought to discharge the duties of the President of the United States? If we are interested in the supervenience of the normative on the non-normative, we face a serious problem on Aristotelian metaphysics: there are very few non-normative facts. It seems that every natural kind is defined in part by normative properties. That Sally is a sheep is

itself a normative thesis, since a part of what it is to be a sheep is to be such that one ought to have four legs. That Sally is an animal is also normative. And Sally is *essentially* a sheep, and hence being a sheep is central to Sally's identity in such a way that it may even be the case that even the claim that Sally exists may count as a normative claim. Aristotelian metaphysics likely reaches even down to the fundamental particles. Electrons not only do but should repel other electrons. A non-normative fact, thus, will not make reference to any natural kinds, and not even to any particulars falling under natural kinds.

On Aristotelianism, every particular, with the exception of God if there is a God, falls under a natural kind. But facts about God are through-and-through normative, since God is essentially perfectly good. Thus, on Aristotelianism, all facts about particulars are normative. Moreover, on Aristotelianism, a non-normative fact cannot include anything existential. For to be is to be a substance or to be appropriately related to a substance (say, by being its accident). And a part of what it is to be a substance is to have a form that specifies how one should behave. Thus, what it is to be is in part to have norms or to be related to something that has norms. If so, then every existentially quantified claim is normative. The denial of a normative claim is normative as well, and since a universally quantified claim is the denial of an existentially quantified claim (everything is F if and only if there does not exist an object that is not F), universally quantified claims will be normative as well.

But if all facts about particulars and all quantified facts are normative, it seems that *all* facts are normative on Aristotelianism. If this is right, then to say that two worlds are the same in non-normative terms is to say literally nothing about them. And if all facts are normative, then any two worlds that are the same in normative terms are altogether the same. Thus, the thesis of the supervenience of the normative on the non-normative becomes the thesis of modal fatalism: that there is only one possible world! And we have good reason to reject this thesis.

But while this goes against mainstream views of normativity, it is arguably an advantage of the view. For by eliminating non-normative facts, we no longer have any puzzling

phenomenon of the relationship between the normative and the non-normative to be explained.

What about the moral supervening on the non-moral? Moral facts are normative facts about the will. Again, general moral facts, such as that humans should refrain from torturing the innocent or should discharge the duties of the President of the United States if they have voluntarily sworn the relevant oath of office, are necessary truths and hence trivially supervene on whatever facts we want, including non-moral or even non-normative ones. But there are many particular normative facts, such as that Biden should discharge presidential duties, that depend on facts about human wills, such as that Biden *voluntarily* swore the oath of office, and since it is the very nature of the human will to be such that various moral facts about it hold, facts about human wills are not going to be among the non-moral facts. Thus the Aristotelian will also reject the supervenience of the moral on the non-moral.

But this seems problematic. Imagine a world w_1 that *looks like* our world w_0. It has bipeds that look just like us. The history of these bipeds is empirically indistinguishable from our history. There is, for instance, a biped that is empirically indistinguishable from Napoleon and one empirically indistinguishable from Marie Curie. If the moral facts about these bipeds could be other than the facts about us, then it seems that moral skepticism follows. For these beings will have the same collective moral beliefs and intuitions as we do, but presumably these beliefs and intuitions will be wrong. If so, then how can we be confident in our collective beliefs and intuitions?

First, however, we should note that very likely a number of our collective moral beliefs and intuitions *are* wrong. Certainly this was true for many of our ancestors—take, for instance, the acceptance of slavery across large swathes of the world in times past—and it would be very implausible to think the same is not true of us. But absent reason to think that the *majority* of our central moral beliefs and intuitions is wrong, skepticism is not forced on us by the observation that some of them are wrong.

Now, let us go back to the hypothetical bipeds that are empirically indistinguishable from us but have different moral facts. This possibility will not imply skepticism if the moral facts for these bipeds are *somewhat* different than for us. And even if there is no supervenience, it could be that there are metaphysically necessary limits on how far the non-normative and normative can depart from each other. Such limits could, for instance, come from the goodness of a necessarily existing divine creator.??forwardref,compare

But even without metaphysically necessary limits, the mere metaphysical possibility of a radical skeptical scenario of beings non-normatively just like us but radically different morally does not force moral skepticism on us any more than the widely-acknowledged possibility of brains in vats forces external-world skepticism on us. To yield skepticism, the truth of a skeptical hypothesis needs more than mere possibility—it needs to be not unlikely. And whether a world like w_1 is unlikely depends on metaphysical questions about the ultimate origins of the world, just as whether a world where everyone is a brain in a vat depends on such questions.??forward-ref,add

8. Outlandish paradoxes

It is easy to generate paradoxes in ethics and decision theory by invoking outlandish situations. Many, but not all, such situations involves infinities. I will give three representative examples.

First, we have the Satan's Apple paradox about infinite sequences of choices on which something further depends:

> Satan has cut a delicious apple into infinitely many pieces, labeled by the natural numbers. Eve may take whichever pieces she chooses. If she takes merely finitely many of the pieces, then she suffers no penalty. But if she takes infinitely many of the pieces, then she is expelled from the Garden for her greed. Either way, she gets to eat whatever pieces she has taken. ??ref

The puzzle is that for each piece, Eve has conclusive reason to take the piece, but if she acts on all these reasons, something terrible happens. As presented, this is a paradox about self-interest, but we can turn it into an ethical one by supposing that the rewards and penalties of Eve's choices devolve on someone else, say Adam. In that case, we can say that Eve should accept each piece and yet that's the worst option.

Another kind of paradox involves infinite numbers of beneficiaries. Imagine that there is an infinite number of complete strangers, numbered with the integers (negative, zero and positive), as well as two cats, all facing a deadly danger, and you have a choice between one of three equally convenient options:

(10) Save the strangers numbered $0, 1, 2, ...$

(11) Save the strangers numbered $-1, -2, -3, ...$ and one cat.

(12) Save the strangers numbered $1, 2, 3, ...$ and two cats.

Now, you have no reason to prefer the stranger numbered 0 over the stranger numbered -1, the stranger numbered 1 over the stranger numbered -2, and so on. So as far as the saving of people, (10) and (11) are a wash, but it's better to save a cat than not to, so (11) is morally preferable.[3]

But likewise there is no reason to prefer saving the people numbered with negative integers over the people numbered with positive integers, so as far as the saving of people goes, (11) and (12) are balanced. However, saving two cats is better than saving one, so (12) is better than (11).

But now, (10) is clearly better than (12): for in (10), you save stranger 0 instead of the two cats, and wonderful as cats are, it is much better to save that one human over two cats.

So we have a moral preferability circle, and whatever you do, there is something better you could have done at no greater cost. It seems plausible that you have a duty better if you can do so at no greater cost, and yet whatever you do, you violate that duty. And so it

[3]If the reader thinks that cats do not fall in our moral purview, just replace the saving of a cat with saving a human from some minor harm.

seems that you cannot act as you ought, thereby violating the plausible maxim that ought implies can.

Third, consider particularly extreme apparent counterexamples to deontology. We have the intuition that it is wrong to kill innocent people, but even deontologists find it difficult to maintain that intuition in the face of cases where killing an innocent person would save a vast number of lives. In those cases we are apt to be uncomfortable both with killing the innocent and with letting the vast number of others die lest we get our hands dirty. In such cases there appears to be a conflict between the principle that innocent blood is not to be shed and the reason for that principle, the sacredness of life.

There have been various attempts to defuse such paradoxes,???refs and a defender of human nature as the foundation of ethics can accept any of them. However, there is also a simple and highly intuitive alternative to these defusions. A horse's nature may ground facts about the appropriate gait when browsing on grass and the appropriate gait when fleeing a predator through water. But equine nature is simply silent on a horse's gait when fleeing aliens in a zero-gravity environment. Similarly, our human nature could be silent on how we should act in outlandish situations, and our principles just need not extend to such cases. This fits very well with the ordinary person's disdain for philosophers (like me) who spend a lot of time thinking about such cases.

There is a second, and similar, solution. It could be that our ordinary moral rules *do* extend to outlandish cases. Thus, the moral reasoning by which we generated the moral paradoxes in Satan's Apple and the infinite saving case may be correctly grounded in norms in our nature. It may well be that, say, (11) is morally better than (10), that (12) is morally better than (11), that (10) is morally better than (12), and that you ought to do the morally best (or one of the morally best, if there is a tie) between the three options. It's just that these specifications of our nature are impossible to fulfill under these circumstances. In other words, it is very plausible to say that ought implies can in situations that are a part of humans' natural environment, but there may be logically possible outlandish situations that go far beyond this environment where ought no longer implies can. Insofar as our

nature gives us norms fitted to our human environment, we should not be surprised if these norms have counterintuitive implications, such as violating ought implies can, in situations far outside that environment.

We might think of these cases as ones where morality "glitches out". This kind of glitching could have a variety of forms. We might have genuine moral dilemmas where our moral requirements outright contradict each other. Or more mildly we might have cases where true moral requirements conflict with some of our intuitions, or with what one might think of as a *reason* for the moral requirements. For moral requirements often come with a reason in terms of a good that is typically promoted by the requirement, even if all the details of the moral rule cannot be logically derived from that reason. (Compare how in the American constitutional order, copyright law is justified by the value of progress of "Science and useful Arts"??ref.) Thus a prohibition on killing the innocent promotes respect for the sacredness of life, but there may be other ways of respecting that sacredness as well. But in an extreme case where the human race would die out if we refuse to kill an innocent, a prohibition on killing is in tension with the reason for it. In these cases, we would expect to find ourselves pulled in different ways, as we indeed are. And it is unsurprising if norms for the governance of a particular species should work strangely or poorly in situations far from what one might think of as the natural environment of this species. Nor is this some sort of a defect in the norm, any more than it is a defect in a phone that it does not function in a blast furnace.

Similar solutions might well be available on at least two other ethical theories where the laws may be customized to humanity: contractarianism and divine command theory. But, on the other hand, such solutions will be implausible on theories that purport to apply to any kind of rational being at all, theories such as utilitarianism or Kantianism.

A similar point can be made about outlandish epistemological paradoxes. ??refs-and-examples On a natural law epistemology, we should not expect our nature to give us guidance, or at least satisfactory guidance, in situations too far out of the human environment. And while in ethics there are at least two common anthropocentric alternatives to natural

law, contractarianism and divine command, in epistemology anthropocentric alternatives are harder to find.??Hawthorne? Thus, we have perhaps an even stronger consideration in favor of a normativity based on human nature on the epistemological side.

More will be said in ??forward about outlandish scenarios.

9. Agent-centrism

9.1. The egoism objection. According to Natural Law metaethics, an action is right provided that its performance constitutes the will's flourishing, and is wrong provided that its performance constitutes the will's languishing. This seems objectionably egoistic. Paradigm cases of moral wrongness involve harm to others, and are wrong because of that harm. The thought that the action makes the agent languish is a thought too many.??refs Therefore, the argument goes, we should opt for an other-centered metaethics.

My response will be two-pronged. First, I will argue that the very features criticized in Natural Law provide a significant advantage in a number of cases. Second, however, I will argue that the argument against Natural Law's agent-centric character only works against some normative developments of the Natural Law metaethics, rather than against the metaethics.

9.2. The normative advantages of agent-centrism. Other-centered theories nicely account for what is wrong with murder: it gravely harms the victim. They account slightly less well for what is wrong with typical cases of attempted murder: it is an attempt to harm to harm the victim. But they do not account for atypical cases of attempted murder where the victim simply does not exist. Suppose Alice thinks that she has an identical twin living somewhere in Toronto, and sets out to kill her, to avoid the twin's claiming an inheritance. Alice has an extremely rare genetic disorder which an identical twin would share, but which is very unlikely to be otherwise exhibited even in a city as large as Toronto. She adds a poison to Toronto's water supply that targets only people with this genetic disorder, and then takes care to avoid drinking Toronto's tap water. But in fact Alice never had a twin.

There is thus no one that Alice is attempting to kill. Yet morally speaking, she is just as guilty as in an attempted murder case where her twin exists, and depending on one's views on moral luck maybe even as guilty as in the case where she succeeds in killing her twin. It is worth noting that in Anglo American jurisprudence, Alice might get away under the doctrine of impossible attempts, on which an attempt has to have some feasibility, and trying to kill a non-existent person has none. (Of course, Alice is likely to be convincted for pollution and for reckless endangerment of people with this disorder, but these are lesser evils.) However, it is clear that notwithstanding the law of impossible attempts, it makes no difference to Alice's guilt whether in fact she has a twin or not.

We may, of course, try to save the doctrine that wrongs are always wrongs to another by trying to identify other victims, such as society or God. We can try to tweak the Alice case to exclude the society solution. Perhaps Alice is trying to kill her twin sister in a world where she thinks they are the only survivors of a disaster, but in fact Alice is the only survivor and she's never had a twin. That won't help with the God case, at least not within classical theism, since God is traditionally thought of as a necessary being??Refs. Furthermore, the intuition that I am responding to is that there is something particularly centered on the ordinary direct human victim of a wrongdoing that contributes much of the wrong. And if God or society is what we count as the victim in the Alice case, then it seems that we have to say that in the Alice case there is less wrong than in the more ordinary case of attempted murder where the victim actually exists. And yet it seems that how wrong Alice's action is does not depend on whether she has a twin.

The above focused on patient-centric wrongs. We can also think about patient-centric duties. Again, it seems that our duties go beyond these. Plausibly, we have a duty not to deliberately produce a human being who is so genetically constituted as to be practically guaranteed to have a life of unremitting suffering. Now suppose that Bob and Carl both know that they and their spouses have genes such that if they reproduce, the child will have a life of unremitting suffering. In light of this knowledge, Bob refrains from reproduction, while Carl's sadistic tendencies impel him to reproduce in light of this fact. Bob

and Carl both have a duty. In Carl's case, we can identify the individual to whom he has this duty, an individual that that he has wronged. But in Bob's case, the analogous individual does not exist—precisely because Bob has fulfilled his duty. Again, we can try to identify others to whom Bob owes not having a child—society, God and likely Bob's wife. But since there is one less individual here than in Carl's case, Carl has a somewhat more stringent duty, since the unfortunate child exists. However, it does not seem that Carl has a more stringent duty than Bob.

Finally, moving away from cases, it is obvious that some degree of agent-centrism is needed for any plausible story about wrongs and duties. It is *agents* that do wrongs and have duties. We have a duty not to eat humans. Lions, pigs and horses have no such duty. Therefore, an account of what makes it wrong for me to eat other humans has to involve some facts about *me*. It cannot be wholly other-based.

One might respond that on the Natural Law account, what makes it be wrong for me to eat other humans involves *only* facts about me, while it should also involve facts about the prospective victims. But how we understand this objection depends on how we read questio of "what makes it wrong for me to eat other humans".

First, we can understand it as a question about the grounds of the general moral rule that it is wrong for humans to eat other humans. On Natural Law the grounds of that general moral rule are entirely within the agent. However, that is how it should be. For the general moral rule would also hold even if there were no other human beings in the world. And in fact the general moral rule would hold *non-trivially* even if there were no other humans, since even if there were in fact no other human beings, I should avoid actions that are likely to constitute the eating of a fellow human being (e.g., shooting and eating an animal that has a significant epistemic probability of being human). An account of the wrongness of eating other human beings that requires other humans to exist is unsatisfactory.

Second, we can understand the question as asking about particular cases: What is it that makes it wrong for me to eat, say, Carl? But then the Natural Law story is going to

include a fact about Carl: I am the sort of thing that shouldn't eat other humans and Carl is another human.

On neither reading do we have an argument against Natural Law metaethics.

9.3. Avoiding agent-centrism in normative Natural Law ethics. Here let me start with a personal confession. For many years I objected to the eudaimonism I took to be at the heart of Natural Law, which one might take to consist of the twin theses:

(13) What makes an action right is its promotion of the agent's flourishing.

(14) An agent's right actions are aimed at the agent's flourishing.

And these theses seemed objectionably egoistic.

However, while there are ways of pairing the Natural Law metaethics that I have been developing with a normative ethics that embraces (13) and (14), they are both dispensable.

Indeed, (13) is so clearly wrong that it is unlikely that many Natural Law theorists accept it, given the well-known anti-consequentialism of the Natural Law community. It is wrong to rob a bank in order to pay for the tuition of an ethics class even if there is strong empirical evidence that this class will be so transformative that on the whole one's flourishing will be promoted, even if one takes into account the temporary harm done to it by the robbery. Similarly, one may have a duty to continue working in a job that is just barely moral and empirically likely to be destructive of one's flourishing as a moral agent in order to pay the medical bills for a child.

Now, on the metaethics that I am defending, what makes an action right is that it *constitutes* the agent's flourishing with respect to the will. If we add to this the thesis that whatever constitutes the agent's flourishing with respect to the will constitutes the agent's flourishing as a whole, then we get a version of (13) with "promotion" replaced by "constitution". However we should not think that what constitutes the agent's flourishing with respect to the will constitutes the agent's flourishing as a whole. If an agent is flourishing with respect to the will, but is full of ignorance, in great pain, and lying in a bed of vomit??Vlastos-ref, the agent is not flourishing on the whole.

And the thesis that what makes an action right is its constitution of the agent as flourishing in respect of the will seems to be simply a thesis about proper function, and does not imply any selfishness. Consider, for instance, that a bee's defending the hive at the expense of its life fulfills the bee with respect to whatever we call the driver of the bee's activity (we may not wish to call it a "will"). But this does not make the bee in any real way selfish. And certainly a guided missile is not selfish just because it fulfills its nature by exploding.

It is tempting to think that in the case of an agent who is driven by a rational will, if what makes the action right is its constituting the agent as flourishing (in one respect), then by willing the action under the description "right action", the agent aims at flourishing, in a way in which neither the bee nor the guided missile's actions are aimed at flourishing. If this line of thought is correct, then the characterization of rightness in terms of flourishing implies (14), and that seems more objectionably egoistic.

However, a rational agent's intentions are hyperintensional. It is possible to aim at heating up a room without aiming at increasing the kinetic energy of the molecules in the room, even though what makes there be heat in a room is the kinetic energy of molecules, and necessarily one is present if and only if the other is. Indeed, during the millenia before the relationship between heat and kinetic energy was known, no one aimed at increasing the kinetic energy of molecules while heating a room, and even now when the relationship is well-known, few people's intentions in turning a thermostat make reference to molecular motion. Similarly, even if the rightness of an action is grounded in, constituted by or even identical with the action's being an instance of the agent's flourishing as a willer, the rightness can be aimed at without aiming at the flourishing.

Furthermore, as has often been pointed out in the literature??(on moral fetishism), virtuous agents rarely aim at rightness as such. Instead, they aim at thicker right-making features of an action, such as its being an expression of loyalty to a friend, its fulfilling a stranger's need, or its having been promised. It is because the action has such thick features that its performance is an instance of volitional flourishing.

There are, of course, times when a human agent aims at rightness as such. One set of cases is provided by agents who cannot figure out on their own what is right and have to take the rightness of an action on the authority of another, without understanding what makes the action right. Such agents include small children, but also sometimes ordinary well-functioning adults who find themselves in such situations of such moral complexity that they turn to a professional ethicist or a trustworthy friend for advice. Furthermore, in cases where an agent is not sure which action is right but decides on probabilistic grounds, it seems plausible that they are acting for the sake of rightness as such.[4]

Is this objectionably egoistic, assuming the rightness is constituted by the agent's volitional flourishing? There are at least four reasons to doubt this. The first was already mentioned: the hyperintensionality in intentions.

To see the second reason, observe that we actually have a *three layer* story:

(15) the rightness of the action

(16) the action's being such as to constitutive volitional flourishing

(17) the thick features of the action because of which the action constitutes volitional flourishing.

The egoism objection in the case of agents aiming at right as such insists that the willing of (15) inherits an egoistic character from the agent-centrism of (16). But note that (16) is not the end of the story. Just as (15) is grounded in (16), so likewise (16) is grounded in (17). And in paradigmatic cases, the thick features in (17) are other-centric features. If we think that willing the rightness of the action inherits egocentrism from the flourishing, we should even more think that it inherits other-centrism from the thick features, since the thick features are a yet more ultimate ground of rightness than the flourishing is.

[4] In fact, some authors??refs-in-https://www.jstor.org/stable/44122234 have used this to argue against deciding on probabilistic grounds, but it is so plausible that in cases of uncertainty one should decide on some kind of probabilistic ground that it seems better to use this kind of a case to argue that there is no objectionable moral fetishism here.

Finally, recall that we already noticed that an action can be right and constitute volitional flourishing but hamper one's flourishing as a whole, as in the case of working a soul-destroying job to pay family medical bills or refusing to rob a bank in order to pay for a morally transformative class. In such cases, it is absurd to say that by aiming at volitional flourishing one is being selfish, since the action does not, in fact, contribute to one's good overall.

Indeed, a metaethics that grounds rightness in flourishing as a willer is compatible with one's never being required to intentionally pursuing one's own good. We can (perhaps with some difficulty) imagine an alien species whose members pair off in such a way that the proper functioning of each one's will is just to will the good of the other member of the pair. Perhaps this is a species so physically constituted that they are always more effective at benefiting others than at self. In such a species, the Natural Law metaethics would require utter unselfishness—and yet what would *ground* the rightness of an action would be that the action is proper to one's will and hence constitutes the will as flourishing.

We could imagine two versions of such aliens. They might be less reflective than ourselves, and never act on higher-order reasons like rightness as such. Or they might be reflective, and might even come to a Natural Law metaethics on which an action is made right by its constituting the agent as flourishing. In such a case, they might aim at an action under the description "right", but only because they know that an action's rightness is ultimately grounded in their species in the action's benefiting the other member of the pair, though mediately in its constituting the agent's flourishing. In such a case, there is no more objectionable egoism than in a case where an eccentric rich person says that they will donate lots of money to charity if you do something that is good for you, so you eat a healthy and delicious salad.

In fact, for our final response, it is worth comparing the case to certain somewhat odd cases of intentional activity. Suppose that I want to test a device that detects nerve signals between between the brain and the arm, and so I wiggle my fingers. My action aims at finger movement, but in a sense I don't actually care about finger movement. My end

is triggering the nerve signal detector. The movement of fingers not only is not my end, but it is not even a means to triggering the nerve signal detector. To make this point clear, we might suppose that the detector is triggered before the nerve signal reaches the muscles controlling the fingers, so that the finger movement comes after the triggering of the detector. Or, for a more common case, consider the practice of follow-through in racquet sports. Players are counseled to follow-through on their hits, i.e., to keep their racquet moving after it has made contact with the ball or shuttle. This movement makes no noticeable causal difference to the flight of the ball or shuttle, but aiming at a longer movement makes the racquet movement at the time of impact stronger. If one weren't aiming to continue the motion after the hit, one would likely begin slowing down the motion before the hit. In a game, the post-impact motion itself is not a useful end, nor need it be a means to anything one cares about in a game.[5]

We might say in cases like this that one aims at "unnecessary event", the finger wiggling or the extra movement, in order to better calibrate one's action with regard to the ends that one cares about. We can call this "calibrational" aiming or intention. This calibrational aiming has something in common with the case of a hunter who knows that their gunsight is off, and hence aims a meter to the right of the deer, though the difference is that in the finger wiggling and follow-through cases one really does intend the fingers to move and the follow-through to occur, while in the crooked sight case the hunter does not intend the bullet to go to the right of the deer. But in all three cases one merely uses the aim in order to calibrate the aspects of the action that one really cares about.

It is compatible with the Natural Law story that the agent aims at rightness, and hence at the flourishing of the will, only calibrationally. And merely calibrational aiming at one's

[5]Maybe it saves one from the coach's disapprobation or makes one appear a stronger player which may have a psychological effect on the other player, but we can suppose the coach isn't looking and that the psychological effect does not happen in this case.

good is in no way selfish—it does not imply any care, not even instrumental, about getting one's good. This is something that our altruistic aliens could do.[6]

That said, we are not such aliens. The goods that our will naturally aims at include our own goods. The fact that an action results in or constitutes our own flourishing counts in favor of the action about as much as the fact that it results in or constitutes the flourishing of another individual. Thus in our own case we can have two sets of reasons for doing the right thing with regard to other people. First, we have the thick reasons that render the action right, reasons that in typical cases are other-concerning. In cases where we act on another's counsel, we may not be aware of these reasons, and we may be aiming calibrationally at rightness. But, second, acting rightly is good for us. One would be failing to have the proper attitude to oneself if one didn't take the fact that the action is good for one into account. Thus there is an egocentric reason available whenever we know we act rightly. However, this is as it should be. That acting well is good for us is one of the great philosophical discoveries of all time, going back to Socrates, Plato and Aristotle, and every moral theory should include this fact, whether the theory uses this fact in grounding the right, as on Natural Law, or whether this fact is just an additional observation independent of the grounding of the right.

[6] ??cf:Howard: https://jesp.org/index.php/jesp/article/view/1249/332

CHAPTER IV

Applied ethics

1. Introduction

Thinking that ethical duties are grounded in norms innate to human nature does not by itself logically entail answers to controversial questions of applied ethics. One can think that our nature requires us to kill those whose suffering we cannot stop, and hence that euthanasia is required, one can think that our nature prohibits the killing of the innocent even if that killing would be in their interest, and one can have an in-between view.

But the nature-based approach provides at least two benefits for applied ethics. First, because of the Aristotelian harmony principle, it allows facts about our natural behaviors and needs as the kinds of organisms we are to to provide us with defeasible but often strong evidence about what we should do. Second, it makes it more plausible than it would be on a number of competing theories that the answers to applied ethics questions might be irreducibly intricate—not reducible to a small number of simple principles—and might include domain-specific ethical rules for the various areas of our natural lives, such as family relationships or sexuality or (if it's natural) property rights.

We will thus explore some things that we can say on nature-based ethics. The plausibility of what we will say will serve as indirect evidence for the underlying Aristotelian metaphysics.

2. Natural relationships

2.1. Siblings and cousins. An interesting test case for an ethical theory is whether it can make good sense of our duties to our siblings and cousins. Duties to friends and spouses plausibly arise from commitments we make. Duties to parents have traditionally been grounded in our obligation of gratitude for our life. Duties to children can typically

be grounded in the decision to perform actions that have a non-negligible probability of producing a person dependent on us. Duties to strangers might be grounded in our shared rationality. But we owe more to our siblings and cousins than we do to strangers, even though typically we had no say in whether we were to have siblings and cousins, and even when we have no favors to return.

On utilitarianism, our duties to siblings and cousins come mainly from the contingent fact that we tend to be better positioned to do good to them, say because we know their needs better, are likely to be physically closer, and help from us is likely to be more welcome. But if such contingencies are all that is involved, then we also have to accept an error theory about our intuitions when they go beyond these contingencies. If a sibling or a stranger is drowning, other things being equal one should try to rescue the sibling, even if the stranger is slightly easier to pull out, or is likely to have a slightly better future life. If one finds out that a local homeless person is a cousin one has not seen since early childhood, it is more vicious to ignore their needs than to ignore similar need in a random stranger. Murder of a stranger is evil, but fratricide is worse.

In general, utilitarianism, contractarianism and Kantianism focus on the agent's rationality, taking the details of the agent's humanity to provide no direct normative input into ethical decisions. The fact that most humans hate eating mud gives one reason not to feed mud to them, and the fact that we are unable to instantly teleport ensures we do not have the same obligation to those on other continents as to those nearby. But these are non-normative facts, and the normativity of the conclusions here comes from general normative considerations applicable to all rational beings. There is some *prima facie* plausibility to the idea that the non-normative facts about the relationships between parents and children, together with normative facts applicable to all rational beings, could explain distinctively filial and parental duties. But this is not plausible for the cases of siblings and cousins.

However, if we see ethics as based on the norms written into our *human* nature, given a harmony between the rational and animal aspects of this humanity, will very plausibly

allow for distinctive ethical norms tied to particular kinds of natural human relationships, including perhaps in the first instance familial ones. There is no need on our Natural Law ethics to derive the duties to cousins from non-normative facts about cousinhood and norms for all rational beings: such rules can be fundamental. And the laws can, in principle, be at any level of precision, be it to simply consider one's siblings at a higher weight in one's moral calculus than more distant relatives (we are all relatives, after all, as we learn from evolutionary theory) or to prefer one's siblings over one's cousins to such-and-such a specific degree. The laws could even have social construction built in: they could require us to respect our relatives in the ways that our society prescribes, and require us to establish societies that institute ways for us to respect our relatives.

Divine command theory has a similar advantage: God's commands can be at any level of generality or precision, be it to love one's neighbor or to telephone one's cousins at least twice a year if one can. In principle, rule utilitarianism can do this as well: it is plausible that having rules concerning special relationships like fraternal ones could maximize utility. But Natural Law arguably gives a better explanation of the duties tied to these special relationships. For the nature of these special relationships is very plausibly tied to our humanity, and hence it makes sense that the special obligations attached to them should flow from that humanity rather than the commands of a God or the results of an abstract hypothetical optimization procedure.

Indeed, on Aristotelian natural law, we can say that having these kinds of special obligations is an important aspect of what makes us human—for it is an important aspect of our form, which is precisely what makes us human.

2.2. Less natural relationships. We have a broad variety of socially-instituted and culturally-variable relationships which are very unlikely to have norms encoded for them in human nature. In English-speaking countries the relationship to the parent of one's godchild or the godparent of one's child tends not to have sufficient importance to even have a name, while in other cultures it is important and specifically named. The relationship between an employer and employee varies so broadly with legal and social structures

that it is probably best seen as an umbrella for a number of different relationships, none of which is likely to be encoded in human nature.

Admittedly, a relationship could fail to be culturally widespread and yet could have norms encoded for it in human nature, but there is a more elegant approach to analyzing such relationship: we can see them as cultural determinations of a more fundamental relationship type, with some of the norms coming from human nature's rules for the more fundamental relationship and others from the culture. Moreover, human nature may prescribe the scope for cultures to establish the rules. Such relationships can be thought of as "less natural". At the same time, the difference between these relationships and the "more natural" ones like siblinghood are likely to be largely of degree. For while there may be a fundamental normative relationship of siblinghood, it has further culturally-determined norms.

2.3. Marriage. A particularly interesting question, of significant relevance to controversies in our society over the past century, is where *marriage* lies on the naturalness spectrum. I shall argue that it is likely to be quite natural, with a number of fundamental norms grounded in our human nature by arguing against two main alternative theories and combinations of them.

The first theory holds that marriage is an institution defined by many human societies. Like other such social institutions, such as judgeship, parliament membership, monarchical sovereignty, exchequer chancellorship, and presidency, it is defined by the rights and obligations conferred by society on those who enter into the institution. While we use the same words "judge" and "monarch" across societies, there is only a family resemblance between the institutions these terms refer to, since the actual rights and obligations defining the institutions are often very different indeed. The resemblance may be very weak: the rights and obligations of the monarch of England in the 13th century are about as different from the rights and obligations of the current monarch as the rights and obligations of modern day judge are from a modern day executioner. Nonetheless, for historical

reasons we may use the same word "monarch", sometimes clarifying with adjectives like "absolute" or "constitutional".

The second theory has it that couples choose to undertake certain obligations with respect to each other, which obligations give rise to rights, and this complex of rights and obligations defines the marriage. In more traditional societies, a couple may not choose the obligations specifically but rather will simply opt for the "customary" obligations and their consequent rights. In modern Western society, many couples write their own wedding vows, specifying general obligations. But even in those cases, it is likely that these vows are not typically thought of as a precise and exhaustive legal contract, but rather as a way of customizing one of the prevalent packages of obligations. Again, on the individual theory, we use the same word "marriage" for all these different packages of rights of obligations due to some sort of vague family resemblance between them.

A more sophisticated theory??refs may combine aspects of the social and individual theories, holding that not only do couples undertake obligations to each other and gain rights with respect to each other, they also undertake obligations to society and gain rights with respect to society.

But the individual and social theories are unsatisfactory for multiple reasons.

2.3.1. *Discovery.* People in good marriages come to discover new normative aspects to marriage as they go through life together. ??add-specifics? If the norms of marriage were simply whatever it was that the parties to the marriage chose, there would be nothing to discover. And if the norms of marriage were simply set by society, it would be odd to think that it is particularly by living the married life that one discovers the norms. Rather, the norms would be discovered by study of the history of the social institution of marriage, the laws surrounding it, the intentions of the legislators, and so on.

We might, admittedly, in individual and social institution cases discover new normative facts by logical derivation from previously known ones, but that is not actually the primary way in which we learn about marriage: we learn about it by observing it from the inside. And we discover new facts, including normative ones, about natural kinds of

entities precisely by observing these entities. By observing water, we come to see that it is H_2O and by observing mammals, we come to see that their middle ear should have the malleus, incus and stapes bones??check. And it is in our own case that we are best positioned to observe marriage at work, so it is unsurprising that such observation produces knowledge of normative aspects of the relationship.

Central to this growth is the Aristotelian harmony between different norms. Living according to the norms of marriage tends to fulfill us in other respects, while living contrary to the norms of marriage tends to be bad for us in other respects, and these are things we can often see. A happy marriage makes for happy spouses and an unhappy marriage for unhappy spouses.

2.3.2. *Travel.* Generally speaking, when a married couple emigrates to or visits another society, they are deemed married in their new place, unless there is some general reason that precludes them from counting as married, such as when they are of the same sex and move to a jurisdiction that does not recognize same-sex marriage.

Moreover, this recognition of them as married is not just an honorific indexed to their country of origin. When the Queen of Denmark visited the United States in 1991, she was referred to as a "queen"??check, but obviously she did not have rights and obligations of a monarch with respect to the United States, and so "queen" here was indexed with respect to the Kingdom of Denmark, and similarly for the title "prince" held by her husband. However, if someone referred to Henrik as Margrethe's *husband* or to Margreth as a *married woman* during the visit, these terms would not be merely indexed to Denmark. Rather, they would have the rights and obligations of an American married couple, as modified by their special immunity to persecution, and an ordinary non-diplomatic visitor from Denmark would not even have that modification.

Should we say that by the mere fact of entering a country, a couple that was married in their country of origin enters into a new marriage institution? On the social theory that is exactly what happens: the couple receives a new package of rights and obligations, definitive of marriage in a new society. But this would be quite surprising: it would mean that

a couple going for a honeymoon in another country would have had two weddings (one might tongue-in-cheek wonder if theyn they shouldn't then be entitled to a second honeymoon?), and globetrotting couples would rack up marriage after marriage. Moreover, relinquishing one's citizenship in a country one no longer lives in would be tantamount to a divorce.

Or perhaps instead of the new institution being entered into upon entry into a country, a couple by marrying enters into the marriage institution of every jurisdiction that is willing to recognize them as married, but the rights and obligations of these institutions are merely conditional on their being in those countries. While this would alleviate the problem of multiple weddings, such automatic entry into institutions in states that have no jurisdiction over one seems implausible. Moreover, the problem of multiple weddings is not solved. When Margrethe and Henrik married in 1967, there was no state of East Timor. Then in 1975 it declared independence. By that declaration, did they impose a new marriage institution on Margrethe and Henrik, a marriage institution that disappeared next year when East Timor was annexed by Indonesia, and then reappeared in 2002?

On the purely individual theory, the travel problem disappears. Different states may add rights and obligations, but what defines the marriage is the complex of rights and obligations that the couple entered into on their own, and it counts as a "marriage" in their travel destination because of the family resemblance between these rights and obligations and those that members of that society take on when they enter into an analogous relationship.

2.3.3. *Cross-cultural criticism.* Andronia is an especially sexist society, and Bob and Alice is an Andronian married couple. You've never interacted with Bob in a context that made his sexist views clear, but one day you find out that Alice is sick, and Bob is not showing any consideration for Alice besides the minimum needed to be shown to any human being. You call Bob out on this, and he tells you that in Andronia it is the wife's job always to show consideration for her husband while the husband need only keep the wife alive and show her the kind of consideration one owes every human being when she is

sick. He adds: "This is how my parents behaved, how Alice's parents behaved, and Alice knew that this is what she was getting into when we got married."

If marriage were a natural relationship, we could say that Bob and the rest of Andronian society is just wrong about what marriage requires, and we could say to Bob: "That may all be, but it's not how husbands *should* behave!" We could then show Bob examples of virtuous, caring egalitarian couples in the hope that these examples would open his mind to what marriage really entails. Or we might say to Bob: "If that's all you've committed to, then you're not really married, and so you are reaping the benefits of marriage from Alice under false pretenses."

But if the complex of obligations in marriage is either socially or individually defined, and if neither Andronian society nor the couple included any special obligation of husband to wife in sickness beyond that which we owe any other human being, then Bob could well be simply right in his understanding of his marital duties. This is an unattractive position.

Granted, if Bob and Alice are now living in a less sexist society, we could tell Bob that by moving to this society they have accepted the additional duties of husbands to wives. This is, however, dubious. It may be that be immigrating to a country we take on the legal obligations of that country, but it could well be the case that Bob is meeting these legal obligations, as they tend to be fairly minimal. What Bob is failing to do is to meet the customary obligations attached to marriage in less sexist societies, but it is implausible that by moving to a country one becomes obligated by the customs of the country. No moral criticism would necessarily attach to an American couple if after moving to Canada they failed to celebrate Thanksgiving in October. Furthermore, nothing of significance is changed in the above story if we specify that Bob and Alice are *still* living in Andronia. Be they in Andronia or elsewhere, a husband owes more to his wife than Bob thinks.

Perhaps we could tell Bob: "If that's all you committed to, then you aren't married in *our* sense of the word." But that isn't a criticism of Bob's behavior with regard to Alice. Bob could just say: "So what?" At most it is a criticism of his misuse of words if Bob claimed to be married. Moreover, even as a linguistic criticism it is unlikely to hold water. For we

do in fact use "marriage" and related words for relationships in a vast array of historical and present societies, many of which are quite sexist indeed.

Admittedly, on both the social and the individual view, we could criticize Bob for the relationship that he is in. We could tell him: "If that's what marriage in your context is, it's a corrupt institution, and you shouldn't be married to Alice." But this is unacceptably weak tea. It allows that Bob is married to Alice but does not owe her consideration in sickness beyond that owed a stranger.

2.3.4. *Fulfillment of a natural desire.* Plausibly, apart from reasonable moral and practical restrictions, people should be able to marry those whom they wish to. A society that did not make this possible would be failing its members.

Now, society has no obligation to make possible the fulfillment of every desire people have. Rather, it is reasonable to make a distinction between natural desires and more contingent desires, and hold that society should support the fulfillment of natural desires, such as for food, drink, shelter, useful employment, and knoweldge. Given the plausibility that marriage is one of those things society ought to make available to its members, it is plausible that the desire for marriage is a natural human desire. But if it is a natural human desire, then it is plausible that marriage itself is natural rather than constructed.

This is perhaps the weakest of the arguments for marriage being a natural relationship, however. First, not everyone shares the intuition that a society ought to make marriage possible. Second, it is not clear that we couldn't have a natural desire to construct—individually or socially—an institution of a certain type.

2.3.5. *Same-sex marriage.*

2.3.5.1. An argument for liberals. Let us assume that egalitarian justice requires one to advocate for same-sex marriage in jurisdictions where same-sex marriage is not available.

But suppose that marriage is socially constructed, and that we are in a locality in which one of the norms of marriage is that it be a relationship between a man and a woman. Then, if we understand "marriage" as the word is locally understood, it makes no more sense to

advocate for same-sex marriage than to advocate for chess without pawns: these are simply contradictions in terms. Granted, we may choose to advocate for social recognition of another, more egalitarian institution than marriage. But that will be a different institution.

If we advocate for this different institution, we have two choices. Either, we propose to maintain the institution currently called marriage, whether for everyone who wishes to enter into it or just for those grandfathered into it, or not. If we propose to maintain the current non-egalitarian institution, then we are not really advocating for same-sex marriage. We are advocating for a two-institution model, closely akin to marriage plus civil unions compromises that have generally been seen as unacceptable by advocates of same-sex marriage.

On the other hand, if it is proposed not to maintain the current institution of marriage, then the common and plausible arguments that extending marriage to same-sex couples does no harm to currently married opposite-sex will ring hollow. For it is a part of the proposal that the institution they are a part of be annihilated. Furthermore, in practice, in jurisdictions where marriage has been extended to same-sex couples, generally those who were previously married still count as married. Therefore, on the assumption that marriage is socially constructed, not only is there annihilation of the institution that couples used to be a part of, but these couples are, without their express consent, inducted into the new institution. Such automatic induction into a new relationship does not seem consistent with the ideals of a free society, and yet generally defenders of same-sex marriage have not been bothered by this.

If, however, one holds that marriage is a natural human relationship, then one can argue for marriage equality without arguing for a two-institution model or for the annihilation of the existing institution. Instead, one can hold that marriage is a natural human relationship which non-defectively can be instantiated by couples of the same sex as well as couples of the opposite sex. Given that marriage, understood univocally, can be entered in by both same-sex and opposite-sex couples, it is clear why it is discriminatory for a state to limit recognition of it to opposite-sex couples. And in advocating the end of

this inequality, one isn't advocating for an end of an existing institution, but simply for the state's recognition of the fact that this natural institution can equally well include same-sex and opposite-sex couples.

It is worth noting that defenders of marriage equality who hold that marriage is constructed by individual couples can also avoid the above problematic consequences of social construction. On the individual construction view, in recognizing a marriage, the state is recognizing is a certain type of contract, where the type is defined by a kind of family resemblance. But recognition of opposite-sex contracts of a certain type without recognition of same-sex contracts of a relevantly similar type would be unreasonable. Imagine if one could only sell a house to someone of the opposite sex, after all. On the individual construction view, the claim that no harm is done to opposite-sex couples by state recognition of same-sex marriage is easily defensible. So, the above argument from same-sex marriage advocacy supports the natural relationship view and the individual construction view, but not the social construction view.

2.3.5.2. An argument for conservatives. Here is a plausible principle: If we limit access to an institution on the grounds of gender or sex, absent very strong reason we should strive to make an equivalent available.??ref For instance, perhaps there is some reason for colleges to limit certain sports to one gender, but then they should make other sports available to the other gender. But many conservatives have not only object to same-sex marriage but also to the availability of civil-union institutions for same-sex couples. I will argue that such conservatives should embrace a view of marriage as a natural relationship.

For if marriage is constructed, either individually or socially, then even if the norms of that construction limit marriage to persons of the opposite sex, an equivalent institution without that limitation could be constructed, and by the principle at the top of this argument, it ought to be. In fact, it seems that the best way to resist this argument would be for the conservative to hold that marriage is a natural relationship, and that this relationship is only possible or only normatively possible for opposite-sex couples, while any

superficially similar relationship between persons of the same sex is not a natural relationship. Because no merely social institution would be a natural relationship, it would not be an equivalent to marriage. Therefore, the conservative can respond to the original argument by saying that there is very strong reason not strive to make an equivalent available, namely that no equivalent is possible.

In response, as per our previous argument, the defender of same-sex marriage should say that marriage is a natural relationship that *can* legitimately hold between persons of the same sex. So this conservative response does not close the debate. But it provides the conservative with a way forward. Indeed, it seems that both sides on the same-sex marriage debate will be better served by moving to a natural relationship view of marriage, and then discussing whether this natural relationship has norms that make it possible and permissible for persons of the same sex to instantiate it.

3. Double Effect

Consider these two cases, where all the people other than the dictator are assumed to be innocent.

(18) TROLLEY: A runaway trolley is heading for a fork in the tracks. If nothing is done, it will turn left and kill five people who are on the track. You can redirect the trolley onto the right track, where there is one person on the track, who will be killed by the trolley.

(19) DICTATOR: A dictator tells you to kill one person. If you fail to do so, the dictator will kill five others.

Assume there are no further relevant consequences. Deontologists tend to have the intuition that redirecting the trolley is permissible but obeying the dictator's orders is murder??refs, while consequentialists assume that both are permissible, and indeed obligatory. Let us assume that deontologists are right.

The usual explanation of the difference is that in TROLLEY one merely foresees, without intending, the harm to the person on the right track if one redirects, while in DICTATOR

one intends the death of the person if one obeys the order, even if one does so merely as a means to saving the five.

Often, this difference is formalized into a Principle of Double Effect, which says that an action with a foreseen evil effect and an intended good effect is not ruled out simply on account of the evil effect just in case the evil effect is intended neither as an end nor as a means, and the bad consequences are not disproportionate to the good ones.

But now consider the following variant:

(20) GUNSHOT: If you do not cause the death of one person, the dictator will kill five others. The one person is tied up, and a loaded gun is permanently affixed pointing at that person. The dictator does not care about your intention. You are curious what a gunshot close-up sounds like, so you consider pulling the trigger.

In GUNSHOT, if you pull the trigger out of curiosity, you don't intend to kill the one. Moreover, the consequences on balance are slightly better than in TROLLEY: five are saved and one dies, but also your curiosity about gunshots is satisfied. But if we allow GUNSHOT, then with a bit of cleverness one can come up with a way of non-intentionally killing the one person in DICTATOR. Perhaps one is curious what size of hole a bullet makes in a shirt pocket.

How can one rule out the non-intentional killing in GUNSHOT without allowing one to accede to the dictator's demand?

Here are four lines of response. The first is that proportionality needs to hold not just between the good consequences and the bad ones, but between the *intended* good consequences and the foreseen bad consequences. In GUNSHOT, while the foreseen goods are proportionate to the death of the person at the other end of the gun, the the intended good of the satisfaction of curiosity is not proportionate to that evil. Effectively, this approach forbids one from counting goods that are caused by the foreseen evil towards proportionality, since if one were to intend these goods, one would be intending the evil as a means to them.

But there are other cases where counting goods causally downstream of evils towards proportionality seems exactly right.

(21) BEAR: You have a leg caught in a trap. If you open the trap now, you can save the leg. If you wait to open the trap later, your leg will become infected and will need to be amputated. However, nearby two other people are trapped in a way that they cannot escape (even with your help) until rescue comes tomorrow. A hungry bear is about to eat one of them, Alice. You foresee that if you open your leg-trap now, the noise of the opening will distract the bear from Alice so that it will eat Bob instead. Once the bear has eaten one person, it won't eat until after rescue comes.

The problem here is that a foreseeable consequence of your opening your trap is Bob's getting eaten, and saving a leg is not proportionate to Bob's death. Now, it is true that Bob's getting eaten keeps Alice from getting eaten. But if we do not count goods causally downstream of evils towards proportionality, then we cannot count Alice's being saved among the goods in our proportionality calculation, and all we get to compare is your leg and Bob's life, and hence there is no proportionality. But, intuitively, it is permissible to escape the trap even if this leads to a different person getting eaten.

A second line of response is that intentionally firing a gun pointed in someone's direction is very "close" to intentionally killing them, and we should forbid not only intentionally killing innocent people, but also actions that are very close to such killing.??refs But it is difficult to see why intentionally firing a gun that happens to be pointed in someone's direction should count as too close to killing, but redirecting a trolley at someone does not. After all, nothing of moral significance should hang on the means by which one redirects the trolley. Now imagine that the trolley is heading by inertia towards the left track, but a carefully placed barrel of gunpowder near the junction can push the trolley in the direciton of the right track. And why should it matter whether it is a trolley or a bullet that is being powered by the gunpowder? (We can imagine the trolley is shaped like a bullet train!)

That said, there is something intuitive about closeness theories. However, on such theories we have a seemingly arbitrary parameter defining the forbidden degrees of closeness to intentional killing, and we have a Mersenne question about what grounds the parameter's value.

??ref(Koons) have, however, offered a view of Double Effect on which it *does* matter whether we are dealing with a trolley or a bullet, and where Mersenne questions are less apparent. The reason is that trolleys and bullets are artifacts with different socially-assigned ends, and when one employs an artifact according to its usual operating instructions, the usual ends count morally. The end of a bullet is killing, and when one propels a bullet in that direction, we might say that one's action counts as close enough to intentional killing.[1] One difficulty with this view is that it claims a perhaps implausible moral difference between firing a gun and detonating explosives packed in a steel pipe sealed at one end and with a ball-bearing inserted on the side of the explosive facing the opening, when both are pointed at the same innocent person.

Additionally, the ??ref account does not entirely escape Mersenne questions. An artifact has a particular end and particular set of operating instructions in virtue of patterns of social behavior. However, there are infinitely many functions from social behavior to ends and operating instructions that roughly match our intuitions but disagree on edge cases, such as exactly how often must forks be used to scratch one's back for them to acquire the end of scratching the back. Again, we need an explanation of why this function rather than another defines the artifact for moral purposes.

A third line of response depends on contingencies of psychology. It seems unlikely that one could fire the gun pointed at Bob's heart without intending Bob's death, unless one has done something to protect Bob (say, putting a shield between the gun and Bob).

[1] ??refs taken literally claim that one *intends* killing in such a case. This seems wrong. After all, after the bullet lands, one can rationally, and with no change of one's ends, try to save the person hit by it. But it is not rational to act against one's former ends if one has not repudiated them. But one does not need the implausible claim about intention: all one needs is that the action is close enough to intentional killing for moral purposes.

But note that a really callous person who does not care about human life probably *could* fire a gun that is pointed at Bob simply in order to hear the gunshot. Our actions have plenty of consequences we know about but that we are completely indifferent to, and hence do not intend. I know that moving my fingers to type this sentence disturbs air molecules. But I have no intention to disturb air molecules. If it turned out that due to some random quantum oddity the air molecules were unmoved by my fingers, my prediction that I will disturb the air would be falsified, but my action would not be unsuccessful in any way. But an action that fails to achieve one of its ends is at least partly unsuccessful.

We might, however, make an Aristotelian move here. Perhaps it is abnormal to fire a gun pointed at Bob's heart without intending death, and we are morally responsible for the *normal* ends of an action in addition to its actual ends. And perhaps our human nature defines what ends an action *should* have, given the agent's knowledge of the circumstances. This is a version of the closeness view, perhaps similar to the ??ref:Koons version.

A fourth response notes that there is something callous about aiming at a minor end when there are evident great evils at stake. Suppose a variant of the original trolley problem where there are equal numbers of people on each track, but when you press the button to redirect the trolley, a candy pops out. It is callous to redirect the trolley for the sake of the candy. Similarly, it is callous to fire a shot in the direction of a person in order to see the hole the bullet makes in their clothing. The intended end of the action is absurdly small in comparison to the foreseen harm. The only way for the action not to be callous would be if one intended to save the five other people as in DICTATOR, but if one intended to save the five other people, one would have to intend the means to that, namely the death of the one at whom the gun is pointed, and I have assumed that it is impermissible to do that.

To expand on the fourth response, imagine that Carl is the agent in TROLLEY, but Carl is someone who does not have much in the way of moral feelings, though he does intellectually desire to the right thing. Carl also enjoys redirecting trolleys, though like most people he rarely has a chance to do so. So when he finds himself in TROLLEY, he rejoices, and he redirects the trolley solely for the fun of it, while reasoning that by Double

Effect he is acting permissibly, because the death of the one person on the left track is not a means to the pleasure of redirection, and proportionality holds, because on balance the effects are good.

What is wrong with Carl's action? It seems that he is being callous: the one person on the left track dies for Carl's pleasure. If Carl were intending to save the five on the right track, in the way that we normally expect someone in a trolley case to do, there would be no callousness. This suggests that in checking proportionality, we need to compare the *intended* good against the *foreseen* evils.

We could simply specify that the intended goods are proportionate to the foreseen evils. However, BEAR points away from that. In BEAR, your intended good, the saving of the leg, is not proportionate to the foreseen evil of Bob's death. Moreover, let's modify the case of Carl and the trolley. Suppose instead that the agent is Dave, and he is pinned down in such a way that his leg lies across the left track, behind the fork, and yet he can redirect the trolley. Dave is terrified of his leg getting cut off and is about to redirect the trolley when he sees that there is a person on the right track. Despite his fear, he now thinks it is wrong to redirect. But then he notices that on the left track there are five people. At this point, he concludes that it is permissible to redirect because the overall consequences are positive. But his end in redirecting the trolley is simply to save his leg. The overall consequences function in Bob's reasoning as a defeater to the observation that that redirecting the trolley will result in the death of the person on the right track, rather than as an end.

While Carl is callous and acting wrongly in having a person die as a side-effect of a trivial end, Dave is acting for a quite serious end. It would be better if Dave adopted the lives of the five people on the track as an end as well, but given his terror at the amputation, it is not reasonable to *require* that (though it would be wrong for him to redirect the trolley if the only other relevant effect he saw was the death of the person on the right track).

Our fourth response thus suggests a conjunctive proportionality condition in the Principle of Double Effect:

(22) First, the intended good is not trivial in comparison to the foreseen evil, and, second, the foreseen goods are proportionate to the foreseen evils.

Now, the triviality condition clearly involves an apparently arbitrary parameter measuring relative triviality, and raises serious Mersenne questions.

We have seen that of the available solutions to the problem presented by GUNSHOT, the two most tenable ones involve parameters calibrating closeness or triviality, and hence raise Mersenne questions.

Additionally, we should think a little about the condition that the goods—whether specifies as foreseen or as intended—are proportionate to the foreseen evils. To a first approximation, one might opt for a utilitarian analysis: the overall utility is positive. We have already seen serious problems with this in ??backref in regard to uncertainty, expected value and risk, as well as incommensurability. Further, there is probably a significant overlap between deontology and a friendliness to partiality in ethics. Imagine a trolley problem where on the right track there is a stranger and a cat and on the left track there is one's child. The utilitarian consequences of redirection are negative, but the fact that the person on the left track is one's child seems to make it permissible (some will even say required) to redirect.

But we have also seen that taking relationships into account has many apparently arbitrary free parameters. So the complexity only multiplies.

4. The task of medicine

The realism about teleology and normalcy provided by the Aristotelian framework allows for an elegant solution to the problem of what the task of medicine is.??ref:Lennox

The medical professional is a *professional*. Of course, everyone should refuse to act immorally on behalf of a client. But a professional has norms and pursues goals that go beyond general morality, and has reason to refuse to further the client's aims even when there is nothing generally immoral about these aims but the aims nonetheless violate the

professional goals. Thus, while it is not immoral to create kitsch, a professional artist nonetheless has to refuse a commission that would be unavoidably kitschy.

In the case of some professions, the goals are very much socially defined, and apart from legal minutiae, the delineation of these goals is of relatively small importance. For instance, we have at least three professions that deal with the directing of water: gutter installers, sewer maintainers and plumbers. All three professions are important, but the division of labor between them is not of great importance. It would do little harm to society if we had a single profession for all three tasks, or if we divided up the tasks in some other way, say in terms of dealing with potable and non-potable water, or incoming and outgoing water relative to a house.

However, the division of labor between the medical professions and other professions does seem to cut nature at its joints. The medical professions directly aim at the goods of bodily health, a very natural subdivision of the space of human goods.

Moreover, there is a special value in the medical professional having a very sharp focus on health. Medical considerations are of great importance to everyone's life. But in the end, the patient (or their representative; I will simplify by talking just of patients) needs to be able to make a prudent decision about the recommendations from a medical professional, weighing this recommendation against non-medical considerations such as ones of economics, interpersonal relationships, personal pleasure and convenience, and so on. The patient is typically not an expert in biological matters, but tends to have a good grasp of other relevant goods: for instance, they will know what effect giving up alcohol would have on their social life, or what goods their children would have to give up if a medical procedure is to be paid for. It is important, however, that a medical recommendation be primarily concerned with the good of the patient's health, so that the weighing between medical goals and other goods be delegated to the patient as much as possible, and that the non-medical goods not be double-counted (once by the medical professional and again by the patient) in figuring out the prudent course of action.

At the same time, it is also important for guiding patients to prudent decisions that medical professionals understand health holistically, rather than narrowly thinking only of the kidneys or the feet. Thus, a focus on health in general is important for the medical professional, or at least the medical professional who has an advising relationship with a patient.

But what is health? Health is not *simply* the good of the body. There are many goods of the body besides health, such as athletic prowess, beauty, and reproduction. These goods depend on health, but are not a part of health: for instance, a relatively healthy reproductive system is needed for reproduction, but one can have such a system without using it.

Apparently, physicians see their task as the return of the body to normal function, and then further claim to understand normalcy in a statistical way, as average function.??ref Tying health to normal function seems quite plausible indeed. But the normal cannot be understood merely statistically. ??xref? If it were merely statistical, then the adult who can deadlift 400 kilograms would be as abnormal as the adult who cannot deadlift one kilogram. Rather, normalcy often has a directionality that it inherits from a teleology towards some good. Adults who can deadlift 400 kilograms exceeds the norm, and might be said to be supernormal, but are not thereby abnormal, nor do they need medical treatment to reduce their strength.??Vonnegut

Moreover, among our goods, it is perhaps health that is most clearly species-relative. As a result, an Aristotelian metaphysics of forms is perfectly fitted to grounding the norms of health as the norms of sufficient capacity to function bodily in accordance with our human teleology.????more, better definition

5. Our animal nature

5.1. The moral significance of our animal nature. On the theory being defended, we have a will whose proper function defines the right for us. At the same time, we are biological organisms, and our will plays a certain functional role in driving our activities

and organizing our lives. Among those higher animals, say cats, that are not agents, that functional role is still filled by *something*—some kind of activity driver which we may or may not wish to call a will. It is plausible when we reflect on these higher animals that their activity driver will be malfunctioning if they are not driven to live the kind of animal lives that are proper to their kind. By analogy, it is plausible that if we are not driven to live the kind of animal lives that are proper to our kind, our activity driver is malfunctioning. But our activity driver is the will, and on the Natural Law metaethics I am defending, right action is defined by the proper functioning of the will. Thus the argument from analogy combined with the metaethics yields a case that we *ought* to pursue lives proper to the kinds of animals we are. Indeed, we expect a certain degree of harmony between our animal lives and our lives as agents.

This harmony has some plausible implications for environmental ethics and our relationship to other animals.

5.2. Living naturally.

5.2.1. *Transhumanism.* Various animal activities, such as breathing, drinking, eating and walking are part of a flourishing life for the kind of mammal that we are. To the extent that we should be seeking our own flourishing, we have reason to pursue animal activities. However, this may not yield a very strong norm of pursuit of such activities for two reasons. First, our own flourishing does not appear to be the only thing the norms of our will call us to—the flourishing of others seems just as important.??add-earlier-to-egoism-discussion Second, flourishing as animals is likely only a small part of our flourishing—the bulk of our flourishing consists in specifically human forms of flourishing such as understanding and love.

However, in addition to seeing the animal activities as part of a flourishing life, there is another consideration in favor of living a life consistent with our animal nature. Aristotelian optimism gives us reason to think there is a harmony between the norms of the will and the norms of the rest of a rational animal's organism, and so a rational animal typically ought to act in accord with, and not contrary to, its broader animal nature.

This rules out more radical forms of transhumanism. It would not accord with our broader animal nature to upload ourselves to a computer, even if (which is dubious) we could survive such an upload. The virtual life would not include activities such as breathing, drinking, eating and walking that are a part of a natural human life.

Exactly how much modification of the human body is compatible with the norms of our will is not immediately clear. But the above lines of thought suggest a significant degree of caution. Again, this is an area where we expect many ethical parameters specifying permissible and impermissible modifications.

??what are some relevant considerations?

5.2.2. *Ecology.* It is in the nature of any organism to interact in certain ways with its natural environment. While that interaction may involve out-competing some other organisms, it is plausible to think that the natural behavior of an organism tends to harmonize with its environment to a significant degree (from an evolutionary point of view, organisms and their environment tend to co-evolve). Given a harmony between the norms of the will and the norms of the rest of the rational animal, we would expect a rational animal to have norms that aim it at a certain degree of integration with a natural environment. These norms will then constitute moral norms to harmonize with the environment.

It could be that the reason why an intelligent organism has such norms is because harm to the environment harms the organism. But even if so, it is likely that the norms are not simply derivative from the organism's need to do well in its own niche. Imagine that biological warfare has wiped out all humans beyond you and a handful of friends, leaving the ecosystem intact but lethal to humans. You prepare to flee earth for another solar system, expecting humanity never to return. The spaceship has a good library, but the last novel in a series of fantasy novels that you like has not been released. Instead, it is stored in a computer system so set up by a crazy copyright owner that if you download the novel, all the nuclear weapons on earth will be launched. Your safety is assured if you wait to download the novel once your spaceship has left the earth's atmosphere, but the ecosystem will never recover from the damage then. While under more normal circumstances, severe

damage to the ecosystem has evident repercussions for humans, in this case it does not (except maybe for psychological ones—but we can suppose that these are outweighed by the delights of the novel).

The evident wrongfulness of destroying the ecosystem for the sake of one enjoyable novel is not simply grounded in our duties to protect our species. It may, nonetheless, be *explained* by the need to protect our species. For it might be better for us to have coarser-grained duties to the environment, that forbid large-scale destruction for the sake of minor goods, than to have the finer-grained duties that would allow such destruction in extremely rare circumstances. After all, we know that it is all too easy to fool ourselves as to where the boundaries of the "rare" circumstances lie. The rule utilitarian has room for such an account, but so does an Aristotelian who thinks that there is an inner harmony in our nature such that the various aspects of our nature are good for us in other respects. This pattern of explanation will be discussed further in Section 5.1 of Chapter **??**.

6. The definition of life

??move?? Here is an intuition that until fairly recently would have been widely shared: There are deep metaphysical divides between non-living and living things, and between merely living things and persons, and these divides mark a hierarchy of value, a chain of being. If we could defend such a divide, it would dovetail with the idea that persons are in an important way *sacred*, having rights while other things have mere interests, if that.

I want to offer a highly speculative Aristotelian reconstruction of this intuition. To introduce the reconstruction, start with a puzzle for Aristotelian views. It seems that on such views:

(23) Each thing naturally strives for its own perfections.

(24) The natural activity of a thing is a perfection of it.

But this generates a regress. Let's say that reproduction is an oak tree's perfection. Then by (23), the oak tree naturally strives for reproduction. This natural activity of striving for reproduction, by (24), is then itself a perfection of the oak tree. Therefore, by (23), the

oak tree must naturally strive for it: hence the oak tree naturally strives for striving for reproduction. And so on, *ad infinitum*. But surely an oak tree does not pursue infinitely many things. And even after a few level of meta-striving we exhaust plausibility.

I suggest that we can deny (23). Some perfections of a thing are not actually naturally striven for by the thing.[2] The oak tree does strive for reproduction with its reproductive organs. Moreover, it has a second order striving: it strives to strive for reproduction, by growing the reproductive organs with which it strives for reproduction. There may be one or two more meta-levels, but at some level we can say: it just does this, without striving to do it.

Non-living things, on an Aristotelian metaphysics, also have form and also strive for ends. But plausibly they don't strive to strive: they just strive. We thus have a hierarchical division between inorganic things which do not strive to strive and living things which have second order teleological strivings.

The problem of the definition of life is a thorny conceptual problem in biology or its philosophy. Different authors give different lists of features such as homeostasis, growth and reproduction as part of the definition of life. The multiplicity of features listed makes the concept of life seem arbitrary. Moreover, it is philosophically problematic to tie the the concept of life too tightly to the physical forms of life around us. For it is very plausible that if there are immaterial agents such as deities, spirits or angels, they should also count as alive.[3] After all, those who believe in such beings sometimes hold them to be immortal. But if they were not alive, their immortality would be a trivial claim: a being that is not alive in the first place cannot die. However, these beings are conceptualized as alive, even when

[2]An interesting theological example may be the idea in the Thomistic tradition that both the beatific vision and our striving for it are gifts of God's grace, rather than natural for us, even though the beatific vision perfects us.??

[3]It is worth noting that not everyone who believes in deities, spirits or angels believes them to be immaterial. The ancient Greeks did not think their deities immaterial. And a minority opinion among Christian theologians held angels to be made of "subtle matter".??ref But the argument only needs that some do believe them to be immaterial.

they cannot engage in homeostasis, growth or reproduction. And yet while a particular existence claim about the existence of immortal immaterial agents might be false, it does not seem to be fundamentally conceptually confused. Thus, a good account of life should include the kind of life that is attributed to immaterial agents, and none??check of the accounts in the philosophy of biology do that.

Furthermore, it is a merit of a definition that when applied to cases where we do not know how to classify a thing, the definition does not trivially decide the issue, but it points to the question we need to answer if we are to decide the issue. To that end, consider two borderline cases: viruses and sophisticated robots, like Star Trek's Data. In neither case are we confident whether we have life. Viruses are famously a borderline case. And while Data is described as a "synthetic life-form"??ref, and the Star Trek canon clearly favors his being actually alive, the question is not so philosophically clear. Data obviously fails typical biological definitions of life: while he engages in self-maintenance, he doesn't grow or reproduce in the biological sense of the word (though he does make other androids), in a way that does not match typical viewers' intuitions.[4] And whether a virus qualifies as alive varies from definition to definition??ref in a way that makes it sound like the question of viruses being alive is merely verbal. Yet given the strong intuition that there is something of great value about life, even something sacred, the question of what is and is not alive should not be merely verbal.

On the other hand, an account on which what it is to be alive is to have a second order teleological striving—to strive to strive for a perfection—will nicely include any immaterial agents. It will include any entity that prepares itself for future teleological activity, say by growth, and hence will include all the physical forms of life we know about. It will exclude elementary particles. And whether it includes viruses or sophisticated robots is unclear—as it should be. For it is unclear whether viruses and sophisticated robots have form at all. If viruses have form, then it is likely that their activity of attaching to hosts

[4]Though, admittedly, there may be some static due to the show confusing the question of consciousness with that of life.??check

for purposes of future replication is a striving for replicative striving, and hence they are alive. But it is not clear whether they have form. If sophisticated robots have form, they also exhibit meta-striving, and hence are alive. But in both cases we do not know whether there is form, or whether we are dealing with a mere agglomeration of particles. Aristotle himself seems to have thought that artifacts only had form in the analogical sense of a blueprint in the mind of the designer??ref, but he could have been wrong in the case of artifacts like Data. (For more on the epistemic issues here, see Section **??** in Chapter X.)

We thus have two levels in a chain of being: things that strive but don't meta-strive, and things that meta-strive. Now, among the things that meta-strive, we can describe a higher kind of thing: a thing that strives for all of its perfections. The premises of the regress argument with which we started this section apply to such a being. Thus, this is a being that strives for striving for ... for perfection, at any number of levels. While this is implausible for an oak tree or even a dog, we do actually know of one kind of being that does that: humans. Human beings not only conceptualize particular perfections, such as friendship or striving for striving for health, but they conceptual perfection as such, and strive for it as such. If a trustworthy being offered you to increase some perfection or other, and assured you that you would in no way be harmed, it would be rational for you to accept the offer, because perfection as such is one of the things you and I pursue.

At the same time, in a minded being, the infinite chain that results from striving for all one's perfections need not be a chain of separate desires and hence does not require a being that is actually infinite. Rather, all that's needed is for the being to be such that it has or teleologically strives to have the concept of a perfection as such and a desire for perfection as such. This desire then can manifest in a striving to figure out what the perfections are—a striving that is central to the search for happiness (*eudaimonia*) that was so characteristic of Socratic and post-Socratic Greek philosophy—and a striving to be ready to accept whatever one finds. In fact, it might be that for reasons having to do with the nature of infinity *only* a minded being can pursue an infinite number of ends—for any non-minded being that did that would need to have infinitely many distinct causal sources of

its activity in a way that might well violate causal finitism, the thesis that it is impossible for an infinite number of causes to work together (for a defense of causal finitism, see ??ref). And among minded beings, perhaps it is definitive of *persons* that they pursue all good.

We thus have a qualitative hierarchy of being between the mere strivers, the mere meta-strivers and the universal strivers. The first division in the hierarchy may well correspond to that between the non-living and the living, and the second might—depending on speculative questions about infinity—align with the division between mere life and personhood. And it is very natural to see qualitative divisions of value here as well.

7. Infinity

We saw in ??backref that population ethics raises Mersenne questions. But *infinite* population ethics not only raises questions, but creates serious paradoxes. For instance, suppose there is an infinite line stretching to infinity both to the left and the right, with tickmarks every meter labeled by an integer (bigger numbers being to the right), and one person standing at each tickmark. All the people are on par. Suppose you now have two choices:

(25) Benefit the people at $2, 4, 6, ...$

(26) Benefit the people at $1, 3, 5, ...$

where all the benefits are the same.

Intuitively, we should be indifferent between these. It makes no difference whether we should benefit the people at the positive even- or positive odd-numbered locations. The options are on par.

But now add a new option:

(27) Benefit the people at $3, 5, 7, ...$

with the very same benefits. Observe now that (27) benefits the people standing immediately to the right of the beneficiaries of (25), while (25) benefits the people standing immedaitely to the right of the beneficiaries of (**??**). Thus, the moral relationship between (27)

and (25) should be the same as that between (25) and (**??**). But the latter two, as already noted, are intuitively on par. Thus, likewise, (27) and (25) are on par.

We can argue for the parity of (27) and (25) as follows. If we re-label tickmark n as $(n-1)^*$, then options (25) and (27) are equivalent to:

??* Benefit the people at $1^*, 3^*, 5^*, ...$

??* Benefit the people at $2^*, 4^*, 6^*, ...$

If benefiting those at positive odd-numbered locations is on par with benefiting those at positive even-numbered locations, then surely this should not depend on whether we used the original or the asterisked numbering. Thus (**??***) and (**??***) are on par. But they are logically equivalent to (25) and (27) respectively, so these are on par as well.

But being on par morally is transitive. So, if (27) and (25) are on par, and (25) and (**??**) are on par, it follows that (27) and (26) are on par. But that conclusion is clearly false, since if we can benefit a person without anybody else losing anything, we have moral reason to do so barring some deontological consideration, and if we are set to do (27) then switching to (26) benefits the holder of ticket 1 without anybody losing anything.

Perhaps, however, we should deny that (25) and (26) are on par. There are two ways of doing that. One is to say that the two cases are incomparable. The other is to say that (26) is better than (25).[5]

Neither option is particularly appealing. Suppose that the benefit is the saving of a life. Then if (26) is better than (25), then by the above reasoning (25) will be better than (27). So we will have this preference ordering:

(28) (26) $>$ (25) $>$ (27).

Now (26) is better than (27) by exactly one life saved. So it seems that (26) will have to be better than (25) by less than saving a life—presumably, by half a life-saving—and (25)

[5]Saying that (25) is superior to (26) is not tenable. The relationship of (25) to (26) is the same as that of (27) to (25), and if we say that (27) is superior to (25), we will then by transitivity have to say that (27) is superior to (26), which is absurd.

will have to be better than (27) by less than saving a life—again, presumably by half a life-saving. But this is very implausible. When the scenarios differ in whose lives are saved, and there are no probabilities involved, surely any two scenarios that differ must do so by one or more lives.

On the other hand, suppose that we have incomparability between (25) and (26) and by the same token between (27) and (25). Now suppose that you have a button pressing which saves the lives of the people in positions $1, 3, 5, ...$ and a button that saves the lives of the people in positions $2, 4, 6, ...$ but where the first button has a side-effect: it causes a migraine to a perfect stranger, Alice. It seems very plausible that pressing the second button is the morally better choice. But given the incomparability claims, there is a conclusive argument against this moral preference. For saving the lives of the people in positions $1, 3, 5, ...$ while triggering a migraine for Alice as a side-effect is better than saving the lives of the people in positions $3, 5, 7, ...$, since saving the life of the person in position 1 is well-worth the migraine to Alice. If saving the people at $2, 4, 6, ...$ were better than saving the people at $1, 3, 5, ...$ plus triggering a migraine, then saving the people at $2, 4, 6, ...$ would be even better than saving the people $3, 5, 7, ...$ But saving the people at $2, 4, 6, ...$ is not better than saving the people at $3, 5, 7, ...$, since (27) and (25) are incomparable.

Here is another variant of the above problem. Let L_n be the action of benefiting all the people to the left of position n, and let R_n be the actions of benefiting all the people to the right of position n, and for simplicitly consider only such left- and right-benefit actions. Write $A \leq B$ to say that action B is at least as good morally as action A, and $A < B$ to say that $A \leq B$ but not $B \leq A$. Say that A and B are comparable provided that $A \leq B$ or $B \leq A$. Here are some assumptions about the moral preferability relation:

(29) Transitivity: If $A \leq B$ and $B \leq C$ then $A \leq C$.

(30) Strict monotonicity: For any m, we have $L_m < L_{m+1}$ and $R_m < R_{m-1}$.

(31) Weak translation invariance: For any m and n, we have $L_m \leq R_n$ if and only if $L_{m+1} \leq R_{n+1}$, and $L_m \geq R_n$ if and only if $L_{m+1} \geq R_{n+1}$.

Transitivity is very plausible. Next, by switching from L_m to L_{m+1} or from R_m to R_{m-1}, one benefits the person at location m, without taking benefits away from anyone, and this is surely better, thereby yielding strict monotonicity. Finally, weak translation invariance is based on the observation that the relationship between L_m and R_n is exactly the same as that between L_{m+1} and R_{n+1}.[6]

In Appendix??forwardref, I prove that given (29), (30) and (31), exactly one of the following conditions holds:

(32) For all n and m, actions L_n and R_m are incomparable with each other.

(33) For all n and m, we have $L_n < R_m$.

(34) For all n and m, we have $L_n > R_m$.

In other words, we either have complete incomparability between any left- and any right-benefit action, or else we have a radical skew where all the right-benefit actions beat all the left-benefit actions or all the left-benefit actions beat all the right-benefit actions.

We could imagine a kind of agents whose morality exhibits the radical directional preference of (33) or (34). Perhaps this would be a kind that lives in a spacetime without the symmetries that our spacetime exhibits, or it is a kind without requirements of egalitarianism. But we are not that kind. This kind of radical skew seems deeply implausible *to us*, and so it seems we would need to have the radical incomparability of (32).

But the radical incomparability of option (32) is also implausible. One could adapt the Alice argument given before. Intuitively, L_0 is morally preferable to R_0-plus-migraine for Alice: you shouldn't cause a migraine to a stranger to make sure that the people you save are to the left of zero. But R_0-plus-migraine-for-Alice clearly beats R_{1000}, since R_0 saves a thousand additional people that are not saved by R_{1000} (namely the people at locations 1 through 1000), and a side-effect of a migraine to a stranger is definitely worth

[6]Strong translation invariance would be the thesis that all the L_m are morally equivalent and that all the R_m are morally equivalent, since the L_m are all translations of one another and the R_m are all translations of one another. But strong translation invariance would be incompatible with strict monotonicity. For a discussion of weak and strong invariance conditions, see ??Pruss:nonclassical.

tolerating to save a thousand lives. But R_{1000} is not worse than L_0 by (32), and and hence R_0-plus-migraine-for-Alice cannot be worse than L_0. So the incomparability view undercuts a plausible judgment about avoiding side-effects.

It seems clear that something morally paradoxical happens in these kinds of infinite cases. But an Aristotelian has a neat way out. These kinds of choices are outside the human ecological niche. If morality were kind-independent and necessary, morality would have to extend to such cases. But it is quite reasonable to suppose the human nature either does not contain principles that apply to such cases, or contains principles that do apply to such cases, but end up contradicting each other in those cases—with morality glitching (cf. ??forwardref to ch 10)—or end up applying to such cases but generating conclusions that don't fit with some of the moral intuitions built-into that nature.[7]

A different kind of being could have different norms from us and, if Aristotelian optimism applied to them as well, correspondingly different intuitions. As already mentioned, they might be less egalitarian than us and tolerate more arbitrariness in preferences between infinite groups. Or they might have different norms regarding side-effects and thus be able to morally embrace a greater degree of incomparability than us.

It should be noted that the paradoxes of infinity here only scratch the surface of the range of oddities imaginable.??refs

[7]It is worth noting that we cannot entirely escape the need to address such cases by saying that they are outside our sphere of activity. For, adapting the Pascal's Mugger story, imagine you are approached by a strange person who tells you that she is a magician from another universe where there are infinitely many people arranged a meter apart on a line, and they are all drowning, and she can by a spell effect, say, L_0, R_{1000} or R_0-plus-migraine-for-Alice. She can't decide which to do and wants your advice. Obviously, you wouldn't *believe* her story. But if you are a good Bayesian, you would assign it a non-zero probability, and the question would indeed become one of moral relevance. That said, it is not surprising if morality behaves strangely once you are in an odd epistemic state. What should you do, we might ask, if you come to think that dialethism is true and you should do the wrong thing? Or what should you do if you become convinced of solipsism or its opposite, alterism (the view that you don't exist but other people do). We should not be surprised if either there are no answers to such questions or the answers are strange.

*Appendix: Skew in benefiting infinitely many people

In ??backref, it was claimed that a preference ordering on certain actions that benefit an infinite number of people satisfying certain axioms either suffers from massive incomparability or has a massive left-right bias. That result follows from the following.

??number??

THEOREM. *Let L_n be the set of integers less than n and R_n the set of integers greater than n. Let $\mathcal{A}$ be the set of all the L_n and R_n. Suppose $\leq$ is a transitive relation on $\mathcal{A}$ such that $A < B$ whenever $A \subset B$ and $n + A \leq n + B$ if and only if $A \leq B$ for any integer n, for all A and B in $\mathcal{A}$. Then exactly one of the following holds:*

(i) for all m and n, we have neither $L_m \leq R_n$ nor $R_n \leq L_m$,

(ii) for all m and n, we have $L_m < R_n$

(iii) for all m and n, we have $R_n < L_m$.

Here, $A < B$ provided that $A \leq B$ but not $B \leq A$, and $n + A = \{n + m : m \in A\}$ is the translation of A by n.

For we can identify the actions in ??backref with the sets of people benefited by them. Then note that if $\leq$ is transitive, so is $<$.[8] The condition that $A < B$ whenever $A \subset B$ follows by induction and transitivity of $<$ from (30) since the only way $A \subset B$ can hold is if $A = L_m$ and $B = L_n$ with $m < n$ or $A = R_m$ and $B = L_n$ with $m > n$. The translation invariance condition then follows by induction from (31) since $n + L_m = L_{m+n}$ and $n + R_m = R_{m+n}$.

PROOF OF THEOREM. Suppose (i) does not hold, so for some m and n we have $L_m \leq R_n$ or $R_n \leq L_m$.

[8]More generally, if $A \leq B \leq C$, and at least one of the inequalities is strict, it follows that $A < C$. For by transitivity of $\leq$ we have $A \leq C$, and if we don't have $A < C$, then it must be because $C \leq A$. Then $A \leq B \leq C \leq A \leq B$. Hence $B \leq A$ and $C \leq B$ by transitivity of $\leq$, which contradicts the claim that $A < B$ or $B < C$.

First suppose $L_m \geq R_n$. We will now prove (iii). We have two cases. First suppose $m < n$. Then $L_n > L_m \geq R_n$, so $L_n > R_n$.[9] By our translation invariance condition, we have $L_k > R_k$ for all k. Now fix any j and k. If $j \geq k$, then $L_k > R_k \geq R_j$ by monotonicity so $L_k > R_j$. if $j < k$, then $L_k > L_j > R_j$. So we have (iii).

Next suppose $R_n \geq L_m$. Let $-A = \{-x : x \in A\}$. Define $A \leq^* B$ provided $-A \leq -B$. It is easy to see that $\leq^*$ also satisfies all of the assumptions of the Theorem. Moreover, since we have $R_n \geq L_m$, we have $-R_n \geq^* -L_m$. But $-R_n = L_{-n}$ and $-L_m = L_{-m}$. Thus $L_{-n} \geq^* R_{-m}$. Applying the previous paragraph to $\leq^*$ with $-n$ and $-m$ in place of m and n, we get, for all j and k, the inequality $L_k >^* R_j$. Hence $-L_k > -R_j$, and so $R_{-k} > R_{-j}$ for all j and k, which implies (ii). □

[9] See note 8.

CHAPTER V

Epistemology

1. Balancing doxastic desiderata

I observe one raven, and it's black. I observe another and it's black, too. The story goes on. Every raven I observe is black. After a certain number of ravens, in a sufficiently broad number of settings, it becomes reasonable to believe that all ravens are black. But when?[1]

William James famously identified two incommensurable doxastic desiderata: attainment of truth and avoidance of falsehood. The larger the number of black ravens that are needed for me to believe that all ravens are black, the more surely I avoid falsehood, but the more slowly I attain truth. Intuitively, there is room for differences between reasonable people: some tend to jump to conclusions more quickly, while others are more apt to suspend judgment. But on either extreme, eventually we reach unreasonableness. Both someone who concludes that all ravens are black based on one observation and someone who continues to suspend judgment after a million broadly spread observations are unreasonable.

There is, thus, a range of reasonable levels of evidence for an inductive belief. And, as in the myriad of ethical cases of Chapter II, this raises the Mersenne question: What grounds facts about the minimum amount of evidence required for an inductive inference and the maximum amount at which suspending judgment is still rational? Of course, the "minimum" and "maximum" may may depend on the subject matter, on higher-order evidence such as about how well previous inductive generalizations have fared, and even on pragmatic factors(??ref). But that added complexity does nothing to make the Mersenne

[1] I am grateful to Sherif Girgis for raising the issue of incommensurable desiderata in connection with these issues.

question easier to answer. And, as we discussed in ??backref, invoking vagueness does not solve the problem, but multiplies the complexity even further.

And, of course, my contention will be that conformity to the human form is what grounds the answers for us. The rational way to reason is the way conforms to our form's specification of the proper functioning of our intellect.

It appears to be quite plausible that different answers to the rationality questions would be appropriate for species of rational animals adapted to different environments. First, some possible worlds as a whole have laws of nature implying a greater uniformity than that found in other worlds, and hence make it appropriate to make inductive inferences more quickly. Second, the environments that the rational animals evolved in may have greater or lesser uniformity, despite the same laws of nature. Third, the ecological niche occupied by the rational animals may punish falsehood more or may reward truth more. ??explain with examples Because of this, the Aristotelian species-relative answer to the Mersenne questions is particularly appealing.

2. Logics of induction

Attempts have been made to give precise answers to the questions about the reasonableness of inductive inferences using a rigorously formulated logics of induction.??refs Let us suppose, first, that some such logic, call it L_{12}, does indeed embody the correct answers. Nonetheless, we will have a Mersenne question as to why L_{12}, rather than one of the many alternatives, is the logic by which we ought to reason inductively.

In the truthfunctional deductive case, there is a system that appears to be both particularly natural and matches our intuitions so well that it has gained a nearly universal following among philosophers, logicians, mathematicians and computer scientists: two-valued boolean logic. It is a sociological fact that no logic of induction has anything like this following, and a plausible explanation of this sociological fact is that no logic of induction has the kind of naturalness and fit with intuition that would privilege it over the

others to a degree where it would seem non-arbitrary to say that it is *the* logic we should reason with.

Further, observe that logics of induction can be divided into two categories: those with parameters (say, parameters controlling the speed of inductive inference–??refs) and those without.

A logic of induction with parameters raises immediate Mersenne problems about what grounds the fact about which parameters, or ranges of parameters, are in fact rationally correct.

A parameter-free logic of induction, however, is not likely to do justice to the fact that different ways of balancing rational goods are appropriate in different epistemic and pragmatic contexts. Moreover, it is unlikely to do justice to the intuition that the balancing should be different in different species of rational agents.

3. Goodman's new riddle of induction

All the emeralds we've observed are green, and it's reasonable to infer that all emeralds are green. But Goodman's famous riddle notes that all the emeralds we've observed are also grue, but it's not reasonable infer that all emeralds are grue. Here, an emerald is grue if it is observed before the year 2100 and green, or if it is blue and unobserved. According to Goodman, the predicate "is green" is *projectible*, i.e., amenable to inductive inference, while the predicate "is grue" is not. But how do the two differ?

As Goodman notes, the fact that "grue" is defined in terms of "green" and "blue" does not help answer the question. For if we specify that something is bleen if it is observed before 2100 and blue, or it is never observed and blue, then we can define something to be green provided it is observed before 2100 and grue or never observed and yet bleen, and similarly for "blue" with "grue" and "bleen" swapped.

Whatever the *justification* may be, it is clear that induction with "green" is reasonable, but not so with "grue". Notwithstanding Goodman's symmetry observations, "grue" is

a gerrymandered predicate, as can be seen in accounting for it in terms of more fundamental physical vocabulary. But now observe that "green" is also gerrymandered. An object is green provided that the wavelength profile of its reflected, transmitted and/or emitted light is predominantly concentrated somewhere around 500 to 570 nm. The actual boundaries of that region are messy and appear vague, the measure of predominant concentration is difficult to specify, and accounting for reflective, transmittive and emissive spectra is a challenge. The full account in terms of more fundamental scientific terms will be complex and rather messy, though not as badly as in the case of "grue", which is more than twice as complex since it needs to account for blueness and the rather messy date of "2100", which is quite a messy date in more fundamental physics units (perhaps Planck times since the beginning of the universe?). Where the boundary between non-projectible and projectible lies—what counts as too gerrymandered for projectibility—is an excellent Mersenne question.

There is a very plausible way to measure the degree of gerrymandering of a predicate. We take a language the content of whose symbols are terms for fundamental physical concepts, or more generally concepts corresponding to fundamental joints in reality, and we look for the shortest possible formula logically equivalent to the predicate, and say that the predicate is gerrymandered in proportion to the length of this formula. It is indeed likely that by that measure "is grue" is more than twice as complex "is green".??ref:Lewis

But now notice something odd. Say something is "pogatively charged" if it is positively charged and observed before 5×10^{60} Planck times or never observed and negatively charged. All the protons we have seen are pogatively charged. But we should not conclude that all protons are pogatively charged. It seems that "is pogatively charged" is just as unprojectible as "is grue". However, notice that by the formula length account, "is green" is more gerrymandered than "is pogatively charged". Pogative charge is much closer to the fundamental than colors. It seems, thus, that our Mersenne question about the boundary between the non-projectible and projectible is not merely defined by a single

number—a threshold such that predicates definable with a length below that number are projectible.

Perhaps, however, what is going on here is this. The hypothesis that all emeralds are grue cannot overcome the hypothesis that all emeralds are green, even though both fit with observation. Similarly, the hypothesis that all protons are pogatively charged cannot overcome the hypothesis that all protons are positively. So perhaps rather than an absolute concept of projectibility, we have a relation of relative projectibility: "is green" is projectible relative to "is grue" and "is grue" is non-projectible relative to "is green".

We can once again try to account for this in terms of the complexity of formulae. But now we need to compare the complexity of two formulae. And where previously we had a single numerical threshold as our parameter of projectibility, we now have a threshold and a new non-numerical parameter that specifies the mathematical way in which the complexities of the two terms are to be compared. This parameter specifies how we test against the threshold: the ratio of complexities, the difference in complexities, or some other mathematical function of the two complexities?

Furthermore, while the idea of a language all of whose terms reflect fundamental joints in reality can be defended, the grammar of the language will make a difference to the precise complexity measurements. For instance, if we have the fundamental predicates Cx, Dx and Ex, then the complex formula expressing the predicate "is C as well as either D or E" will be

$$Cx \,\&\, (Dx \vee Ex)$$

in infix notation, and hence five times longer than the formula Cx represening "is C", but in Polish notation will be

$$KCxADxEx$$

and hence only four times longer than Cx.

For a relative projectibility relation defined in terms of linguistic complexity, we thus have at least three free parameters, each a fit subject for a Mersenne question: a threshold, a comparison function, and a grammar for the basic language.

But in fact we probably should not think of a binary projectible / non-projectible distinction, whether relational or absolute. As Goodman himself observed??ref-in-https://www.jstor.org/stable/pdf/686416.pdf, what we have instead is a range of predicates that are more or less projectible. We have "is green" and "is grue". But we can also say that x is grue* provided that x is green and observed by a French speaker before 2100 or by a non-speaker of French before 2107, or not observed, and "grue*" will be less projectible than "is grue". On the basis of our observations, the probability that all emeralds is green is very high, and the probability that they are all grue or grue* is very low. But nonetheless, the probability that they are grue is somewhat higher than that they are grue*. After all, an alien conspiracy to recolor emeralds upon observation with a sharp cut-off in one year seems a little bit less unlikely than one where the cut-off depends on whether the observer speaks French. Similarly, it makes sense to think of "is green" as less projectible than "is positively charged", and of "is cute" as even less projectible.

Projectibility now becomes a matter of degree. An advantage of this is that perhaps we no longer need to make it relational. The reason for the superiority of the green-hypothesis to the grue-hypothesis and for the positive-charge-hypothesis to the pogative-charge-hypothesis can be given in terms of the relationship between the degrees of projectibility. However, the cost is that now we need a function from predicates to degrees of projectibility, and the choice of that function will have infinitely many degrees of freedom.

4. Epistemic value

4.1. Epistemic value on its own. Plausibly, the more sure you are of a truth, the better off epistemically you are, and similarly the more sure you are of a falsehood, the worse off you are.

But what exactly is the dependence of value on the degree of certainty? Fix some hypothesis H and let $T(p)$ be the epistemic value of having degree of belief or credence p (where $0 \leq p \leq 1$) in H if H is in fact true and let $F(p)$ be the value of credence p in H if H is in fact false. The pair T and F is called an accuracy scoring rule in the literature.??ref

We can put some plausible constraints on T and F. First, $T(p)$ cannot decrease if p increases, and $F(p)$ cannot increase if p decreases.[2] But that still leaves infinitely many degrees of freedom for the selection of T and F.

We can, however, make some progress if we reflect on expected values. If your current credence in H is p, then by your lights there is a probability p of your having epistemic score $T(p)$ and a probability $1 - p$ of your epistemic score being $F(p)$, so your expected score is:

$$pT(p) + (1 - p)F(p).$$

Suppose now you consider doing something odd: without any evidence, brainwashing yourself to switch your credence from p to some other value p'. By your current lights, the expected epistemic value of this switch is:

$$pT(p') + (1 - p)F(p').$$

And this shouldn't be higher that the expected epistemic value of your actual credence p. For surely by the lights of your assignment of p to H, no other credence assignment should be expected to do better. Indeed, if another credence assignment p' were expected to do better by the lights of p, then p would be some kind of a "cursed probability", one such that if you assign it to H, then immediately expected value reasoning pushes you to replace it with p'. This is not rational. So, it is very plausible indeed that:

$$pT(p) + (1 - p)F(p) \geq pT(p') + (1 - p)F(p').$$

[2]We might more strongly specify that $T(p)$ always strictly increases with p, and $T(p)$ strictly decreases. That is plausible, but one might also have a view on which there is a finite number of discrete thresholds at which increase/decrease happens.

If T and F satisfy this inequality for all p and p', we say that the pair T and F is a *proper* scoring rule. And if by the lights of the assignment of p to H, that assignment has better expectation than any other, i.e., if the inequality above is strict whenever $p \neq p'$, we say that the rule is *strictly proper*.

Propriety reduces the degrees of freedom in the choice of scoring rule. Given any non-decreasing function T, there is a function F that is unique up to an additive constant such that the parir T and F is a proper scoring rule, and conversely given any non-increasing function F, there is a T unique up to an additive constant such that T and F is a proper scoring rule.??? Hence, once we have one of the two functions, the other is almost determined. However, at the same time, this result shows what a profusion of proper scoring rules there is: for every non-decreasing function, there is a proper scoring rule that has that as its T component.

The question of epistemic value assignment may seem purely theoretical. However, it has real-world ramifications. Suppose a scientist has attained a credence p in a hypothesis H, and is considering which of two experiments to perform. One experiment will very likely have a minor but real effect on the credence in H (think here of a case where you've gathered 1000 data points, and you now have a chance of gathering 100 more). The other will most likely be turn out to be irrelevant to H, but there is a small chance that it will nearly conclusively establish H or its negation. For each experiment, the scientist can use their present credence assignments to estimate the probabilities of the various epistemic outcomes, and can then estimate expected epistemic values of the outcomes.

It is well-known??ref that if the scoring rule is strictly proper, for each experiment that has potential relevance to H (i.e., there is at least one outcome that has non-zero probability by the scientist's current lights and learning which would affect the credence in H), the expected epistemic value of performing the experiment is higher than the expected epistemic value of the *status quo*. Thus if the experiments are cost-free, it is always worth performing more experiments, as long as we agree that the appropriate scoring rule is strictly proper, and it does not matter which strictly proper scoring rule we choose. But if in addition to

deciding whether to perform another experiment, the decision to be made is *which* experiment to perform, then the choice of scoring rule will indeed be important, with different strictly proper scoring rules yielding different decisions.??ref:fill-in

There are a number of mathematically elegant strictly proper scoring rules, such as the Brier quadratic score, the spherical score and the logarithmic score. Of these, the logarithmic score is the only that is a serious candidate for being *the* correct scoring rule, in the light of information-theoretic and other arguments (??ref:phil of sci paper). In our setting where we are evaluating the value of a credence in a single proposition H, the logarithmic score is $T(r) = \log r$ and $F(r) = \log(1 - r)$.

However, there are also reasons to doubt that the logarithmic score is the One True Score. First, there is an immediate intuitive problem. If you are certain of a falsehood, your logarithmic score is $\log 0 = -\infty$, while if you are certain of a truth, your score is $\log 1 = 0$. Now, while there is good reason to think that the disvalue of being sure of a falsehood exceeds the value of being sure of a truth, it is somewhat implausible that it infinitely exceeds it.

For the next two problems, note that logarithmic scores and the arguments for them only really come into their own when we are dealing with more than two propositions (in our above setting, we had H and $\sim H$ are the only relevant possibilities). Suppose we are dealing with n primitive possibilities or "cells", $\omega_1, \ldots, \omega_n$ (say, the sides of an n-sided die), and that our agent has assigned credence p_i to ω_i. If in fact ω_i eventuates, the logarithmic score yields epistemic value $\log p_i$.

One of the merits touted for the logarithmic score is that ???for how many cells??? (up to multiplicative and additive constants) it is the only proper score where the epistemic value depends only on the credence assigned to the cell that eventuates. But this is also a serious demerit. Suppose that you and I are trying to figure out how many jelly beans there are in a jar. Let's say that our range of possibilities is between 1 and 1000. I look very quickly and assign equal probability $1/1000$ to each number. You count very carefully and arrive at 390. But then you think that although you are really good at counting, you might

be off by one. So you assign 998/1000 to 390, and 1/1000 to each of 389 and 391. It turns out that the number is 391. We both have the same logarithmic score, $\log(1/1000)$, since we both assigned the same probability $1/1000$ to cell 391. But intuitively your assignment is much better than mine: you are better off epistemically than I.

Finally, observe that in real life, credences are not consistent—do not satisfy the axioms of probability. And the logarithmic score allows one to have extremely inconsistent credences and still do well. If I assign credence 1 to *every* possible outcome, I am guaranteed to max out the logarithmic score no matter what. Thus one of the least rational credence assignments results in the best possible score.

We now have two different approaches to the Mersenne questions about epistemic value and scoring rules. First, we could suppose that there is such a thing as *the* One True Score. Since only the logarithmic score seems significantly mathematically privileged over all the other scores, and the logarithmic score is not the One True Score, there will be an appearance of contingency about the One True Score even if there is one.

Second, we might suppose that just as rational people can differ in prudential preferences, they can differ in epistemic preferences. Some may, for instance, have a strong sharpish preference for gaining near-certainty in truths, while being fairly indifferent whether their credence in a truth is 0.6 or 0.8, as neither is that close to certainty. Others, on the other hand, may value increased certainty in a gradual way, like the logarithmic rule does.

However, it is important to note that while there may be room for rational people to differ in epistemic preferences, there is reason to think that there are rational constraints on epistemic preferences that go beyond formal conditions such as strict propriety, continuity or symmetry—where the last is the condition that $T(p) = F(1-p)$.

Let $T_0(x) = 1000$ if $x \geq 0.999$, $T_0(x) = -1000000$ if $x \leq 0.001$, and $T_0(x) = 0$ otherwise. Let $F_0(x) = T_0(1-x)$. Then the pair T_0 and F_0 is a symmetric and proper scoring rule.??ref

Consider now a scientist who adopts this scoring rule for some hypothesis H of minor importance about some chemicals in her lab that she initially assigns credence $1/2$ to. She

has a choice between two methods. She can use clunky machine A that she has in her lab, which is guaranteed to give an answer to the question of whether H is true, but for either answer there is a 0.11% chance that the answer is wrong. Or she can use spiffy new machine B which has the slightly lower 0.09% chance of error either way. The only problem is that her lab doesn't own machine B and her grant can't offer the price. Her only hope for using machine B is to go and buy a scratch-off lottery ticket which has a one in a million chance of yielding a prize exactly sufficient to purchase machine B. However, because some chemicals involved in the experiment are expiring exactly in a week, and machine A is slower than machine B and takes exactly a week to run, if she is to use machine A, she needs to start right now and doesn't have time to buy the lottery ticket. And once she starts up machine A, she can't transfer the experiment to machine B.

In other words, her choice is between using machine A, and then learning whether H is true with a credence of 0.9989, or buying a lottery ticket, which gives her a one in a million chance of learning whether H is true with a credence of 0.9991 and a 999,999 out of a million chance of being no further ahead. Going for the second option seems irrational if all that is at stake is epistemic value: the difference between 0.9989 and 0.9991 is just not worth the fact that most likely going with the lottery route one won't learn anything about H. (If what was at stake wasn't epistemic value but something pragmatic, then things could be different. We could imagine a law where some life-saving medication can be administered to a patient only if we have 0.9991 confidence that it'll work, and then there will be no practical difference between 1/2 and 0.9989, but a big one between 0.9989 and 0.9991.)

But a scoring rule like the one described above prefers the lottery option. For the epistemic value of using machine A is guaranteed to be zero since after using machine A, the scientist will have credence 0.9989 or 0.0011, depending on whether the result favors H or not.

On the lottery option, however, conditionally on winning the lottery, the expected epistemic value will be:

$$(1/2)(0.9991 \cdot T(0.9991) + 0.0009 \cdot F(0.9991)) + (1/2)(0.9991 \cdot F(0.0009) + 0.0009 \cdot T(0.0009)) =$$

since it is equally likely given the scientist's priors that the machine will return a verdict for or against H, which will result in a credence of 0.9991 or 0.0009, respectively, and in either case there will be a 0.0009 chance that the verdict is erroneous. Since $F(0.0009) = T(0.9991) = 1000$ and $T(0.0009) = F(0.9991) = -1000000$, it follows that the expected epistemic value, conditionally on winning the lottery, will be:

$$(1/2)(0.9991 \cdot 1000 + 0.0009 \cdot (-1000000) + 0.9991 \cdot 1000 + 0.0009 \cdot (-1000000)) = 99.1 > 0.$$

And if we multiply this by the 1/1000000 chance of winning the lottery, we still have something positive, so the expected epistemic value of playing the lottery with the plan of using machine B is positive, while that of using machine A is zero.

Thus, by considerations of epistemic value, the scientist with this scoring rule will prefer a 1/1000000 chance of gaining credence 0.9991 as to whether H is true to a certainty of gaining the slightly lower credence 0.9989. This is not rational.

Now, in the above example, our scoring rule while proper, symmetric and finite, was neither continuous nor strictly proper. However, we will show in the Appendix??ref that there is a sequence of continuous, strictly proper, finite and symmetric scoring rules T_n and F_n such that $T(x) = \lim_{n\to\infty} T_n(x)$ and $F(x) = \lim_{n\to\infty} F_n(x)$ for all x. If n is large enough, then the pair T_n and F_n will require exactly the same decision from our scientist as T and F did, since the expected value of the expected (T_n, F_n)-scores of the two courses of action will converge to the expected vlue of the expected (T, F)-scores.

Hence not all epistemic valuations that satisfy the plausible formal axioms are rationally acceptable, then we will have Mersenne questions about what grounds the further

constraints on the epistemic valuations. These constraints are likely to include messy prohibitions, with multiple degrees of freedom, on the kinds of sharp jumps that our pathological scoring rule above exhibited.

Furthermore, things become more complicated when we consider that the epistemic value of a credence in a truth will differ depending on the importance of that truth. Getting right whether mathematical entities exist or whether we are material or how life on earth started has much more epistemic value than getting right Napoleon's shoe size. Epistemic value will thus not only be a function of credence and truth, but also of subject matter. Moreover, we will have further degrees of freedom concerning the operation of combining epistemic values for different propositions—addition may seem a plausible operation, but the logarithmic and spherical rules are not combined additively across propositions.

We thus have multiple indicators of a contingency about epistemic value assignments. And there is good reason to think that different forms of life are more suited to different epistemic value assignments. The most obvious aspect of this is that once we move away from the toy case of assigning a value to one's epistemic attitude to a single proposition and consider that attitudes to a large number of propositions need to be considered, it is obvious that the subject matter of the propositions will affect how great a weight we give credences about them in the overall evaluation. And the importance of subject matter obviously depends on the form of life. It is plausible that for intelligent agents whose natural environment is more hostile it would be more fitting to have a greater epistemic value assigned to practical matters, while agents that have few natural enemies and can get food easily might more fittingly have a greater epistemic value assigned to theoretical matters. One imagines here that intelligent antelope might be properly expected to be less philosophical than intelligent elephants.

4.2. Connection with other values. A scientist who is deciding whether to do another experiment at the end of a long day, an experiment that cannot be done on another date (maybe the subjects will be unavailable or the chemicals will go bad), or go home to be with family is weighing the epistemic goods arising from the experiment with the value of

interacting with family. What is the right thing to do depends here on many unspecified factors. How much of a delay in returning home would the experiment create? What is the current state of the scientist's relationship with the family and what is their current level of need? What practical benefits, if any, would accrue to the scientist, the scientist's family, or humankind from the experiment? And, finally, how important is the hypothesis being tested and how much can the experiment be expected to contribute to confirming or disconfirming the hypothesis?

Only the last question adverts to epistemic value. It could be that the answers to the preceding questions suffice to determine whether scientist should go home. Perhaps the scientist's progress towards tenure is crucial to the family's livelihood. Or perhaps there is a sufficiently high chance that the experiment's answer will contribute to a cure for cancer that apart from any epistemic goods it is worth staying. Or, alternately, perhaps the scientist has a commitment to family to go home on time this evening, and regardless of the epistemic goods involved, that commitment needs to be honored.

But of particular present interest is what should be done if the non-epistemic goods do not answer the question of what the scientist should do. In that case, one needs to weigh the epistemic goods against the non-epistemic ones. There is no automatic priority of one over the other. Some purely epistemic goods are great enough that significant sacrifice of non-epistemic goods is worthwhile. Thus, much sacrifice would make sense if the outcome would be to know whether there is life in other galaxies, even though that knowledge seems to have no practical value to us. On the other hand, very little sacrifice would be worthwhile to find out whether the lab's largest rat has an even or odd number of freckles.

If we could find the right scoring rule, that would let us compare different epistemic values. But we would still need a way to weigh the outputs of the scoring rule with non-epistemic values. And then we will have our old familiar appearance of contingency. For whatever is the right way of comparing these values, we can imagine beings slightly different from us for whom a different comparison is appropriate: beings that ought to be

more contemplative or more practical. And we have the familiar Mersenne question about why it is that for us the weighing goes as it does, rather than in some other way.

5. Bayesianism

5.1. Introduction. Bayesianism is the best developed picture of what a precise and rigorous account of epistemic rationality would be like. It is thus worth looking carefully at what kind of answers the Bayesian could give to the questions we have been asking.

On Bayesianism, the interest is not in beliefs as such but in the agent's credences, which come in degrees. The most familiar model is of numerical credences ranging from 0 to 1, and this is the model that will be assumed here. Other models include qualitative probabilities where instead of a specific value being assigned, one has comparisons between probabilities, and interval-valued probabilities where instead of a specific numerical probability, a range of numerical probabilities are assigned.??refs Most of the issues raised will apply *mutatis mutandis* to the other models.

If a perfectly rational agent assigns credence $C(p)$ to a proposition p, then these credences will satisfy plausible axioms for probabilities, such as:

(i) Non-negativity: $0 \leq C(p)$

(ii) Normalization: $C(p) = 1$ if p is necessarily true

(iii) Finite Additivity: $C(p \vee q) = C(p) + C(q)$ if p and q are mutually exclusive.

Furthermore, perfectly rational agents update their credences on receipt of evidence by conditionalization. In other words, upon receipt of evidence E, the "prior" credence of p goes from $C(p)$ to the "posterior" conditional credence $C(p \mid E)$. If $C(E) > 0$, then we define $C(p \mid E) = C(p \,\&\, E)/C(E)$.[3] Essentially, this conditionalization is designed to ensure the plausible rule that if p and q each entail the evidence E, then the ratio of the credences of p and q does not change upon learning E.

[3]There are some technical issues with handling cases where $C(E) = 0$, such as if one finds out that a spinner has landed in a precise location.????

5.2. Induction and priors. From a Bayesian point of view, how induction works is determined by the probabilities prior to all evidence, the ur-priors. Suppose, for instance, that I assign equal prior probability to every logically possible color sequence of observed ravens. For simplicity, suppose that there are only two colors, white and black. I find out that there are a million ravens, and I observe a thousand of them, and find them all black. I am about to observe another raven. The probability that the next raven will be black will be $1/2$. For the sequence $B, ..., B, W$ (with 1000 Bs) is just as likely as the sequence $B, ..., B, B$ (with 1001 Bs), and both sequences fit equally well with our observations.

On the other hand, suppose I assigned probability $1/3$ to the hypothesis that all ravens are white, $1/3$ to all black, and split the remaining $1/3$ equally among the $2^{1000000} - 2$ multicolor sequences. My observation of the first 1000 ravens then rules out the all-white hypothesis. And it rules out most of the multicolor sequences: there are $2^{999000} - 1$ multicolor sequences that start with 1000 black ravens, which is a tiny fraction of the original $2^{1000000} - 2$. Since as a good Bayesian I keep the ratios between the probabilities unchanged, each of the remaining multicolor sequences has $1/(2^{1000000} - 2)$ of the probability of the all-black sequence, and since there are only $2^{999000} - 1$ multicolor sequences remaining compatible with the evidence, the ratio between the multicolor probability and the all-black probability is $(2^{999000} - 1)/(2^{1000000} - 2)$ to 1, or approximately 1 to 2^{1000}. Thus, we have overwhelming confirmation of the all-black probability, and hence an even more overwhelming confirmation of the hypothesis that the next raven will be black.

Other ways of dividing the probabilities between the hypotheses yield other results. Carnap??ref, for instance, had a division that worked as follows. For each number n between zero and a million we have the hypothesis H_n that there are exactly n black ravens, and Carnap proposed that all million-and-one of these hypotheses should have equal probability, and then each hypothesis H_n is divided into equally likely subhypotheses specifying all the subhypotheses that make there be n ravens. Thus, H_0 and $H_{1000000}$ have only one subhypothesis: there is only one way to have no-black or all-black. But H_1 and H_{999999} have a million subhypotheses each: there are a million options for which is the raven with

the outlying color. Using the same constant-ratio technique as before, after observing 1000 black ravens, the chance that the next one is black will turn out to be approximately 0.999, but the chance that all million are black will only be 0.001. More generally, if there are N ravens, and the first m of them have been observed to be black, and $n \geq m$, then the probability that the first n will be black will be $(1+m)/(1+n)$.[4] Hence we have very reason to think that the *next* raven is black, but unless we have observed the bulk of the ravens, we won't have reason to think that all the ravens are black.

Intuitively, while Carnapian probabilities support induction, they result in induction being too slow—it is only when we have observed the bulk of the cases being a certain way that we get to conclude that they are all like that. My 1/3–1/3–1/3 division is too fast. Even with 1000 black ravens having been observed, the probability of a white raven shouldn't be *astronomically* small in the way that $1/2^{1000}$ is. Reasonable priors, thus, yield a speed of induction somewhere between these.

We can presumably come up with a formula for the priors which will fit with our intuitions of how fast induction should work. For instance, we could take Carnap's setup, but increase the prior probability of the all-white and all-black raven hypotheses. But such an increase would be apt to involve one or more parameters. If the specific assignment of priors were rationally required of us, then we would have the Mersenne question of why it is these and not some other very similar priors that are required. And if there is a range of priors rationally permitted to us, then we would have Mersenne questions about the boundaries of this range.

Further, imagine beings other than us that inhabit a more Carnapian world than we do. While in our world, we have a significant number of natural kinds that exhibit or fail to exhibit some basic property exceptionlessly—for instance, every electron is charged, and no photon has mass–in that world there are few such natural kinds. Instead, if we were

[4]Let B_m be the claim that the first m ravens are black. Then $P(B_m) = \sum_{n=0}^{N-m} \binom{N-m}{n} / ((N+1)\binom{N}{m+n}) = \frac{1}{1+m}$.??why:Mathematica The probability that the first n are black given that the first m are black where $n \geq m$ will then $P(B_n \mid B_m) = (1+m)/(1+n)$.

to tabulate the frequencies of basic binary properties in various natural populations—say, tabulating the frequency of blackness among ravens, charge among electrons, mass among photons—we would find the frequencies to be distributed uniformly between 0 and 1. In that world, Carnapian priors would lead to the truth faster than the more induction-friendly priors that we have. And let us imagine that in that world we have intelligent beings who reason according to Carnapian priors. Even if we happily grant that Carnapian priors are irrational for us, it seems plausible to think that they could be rational for those beings. To insist that these Carnapians are irrational, because it would be irrational for us to have these priors, seems akin to saying that bigamy would immoral for aliens who need three individuals to reproduce, or that there is something wrong with fish because they lack lungs.

Consideration of the rationality of induction thus once again reveals an appearance of contingency in the normative realm, which once again yield an argument for an Aristotelian picture of human nature, where the rationally required priors or ranges of priors are those that we are impelled to by our human nature.

But before we embrace this conclusion fully, we should consider two Bayesian challenges, from two opposed points of view. The algorithmic Bayesian thinks that considerations of coding can yield a reasonable set of priors, while the the subjective Bayesian says that there are no constraints on the priors.

5.3. Algorithmic priors. Suppose, first, we have a language L of finite sequences of symbols chosen from some finite alphabet of basic symbols, with some of the sequences representing a member of some set S of situations.

For instance, S could be arrangements of chess pieces on a board[5], and L could be a declarative first-order language with no quantifiers, twelve piece predicates (specifying both the color and piece type) and sixty-four names of squares. We could then say that a

[5]The arrangements contain less information than chess positions, since a chess position includes other information, such as whether a given king or rook has already moved, whose turn it is, as well as historical information needed for adjudicating draws.

symbol sequence represents an arrangement a provided that the sequence is a syntactically valid sentence that is true of a and of no other arrangement. However, in general L need not be a declarative language. It could, for instance, be an imperative computer language for an abstract Turing machine or a physical computer, and the situations could be possible outputs of that machine. Then we might say that a symbol sequence s represents a possible output a just in case s is a program that, when run, halts with the output being a. Or if we like we might add an additional layer of representation between the outputs of the machine and the situations—for instance, the outputs of an abstract Turing machine might represent the physical arrangement of particles in a universe. We can even chain languages. For instance, we could have a computer language L_1, with the outputs being sequences of symbols in some declarative language L_2, whose sentences in turn represent members of a set S of situations.

Next, consider a natural way of choosing at random a finite sequence of symbols of L. Here is one. Add to L's finite alphabet a new "end" symbol. Then randomly and independently, with each symbol being equally likely (i.e., having probability $1/(n+1)$, where n is the number of non-end symbols), choose a sequence of symbols until you hit the end symbol. The sequence preceding the "end" symbol will then count as the randomly selected sequence in L. Every sequence of length k has probability $1/(n+1)^k$, so the probabilities decrease exponentially with the length of the sqeuence. We repeat the random selection process until we get a sequence that is both syntactically correct and represents a situation in S.[6] We now stipulate that the prior probability of a situation s is equal to the probability that the above process will generate a sequence that represents s.

Alternately, we can formulate this as follows. If a is a sequence of symbols of L and s is a situation in S, write $R(a,s)$ if a is syntactically correct and represents s, and let $R(a)$ be shorthand for the claim that a syntactically correctly represents S, i.e., $\exists s(s \in a \;\&\; R(a,s))$.

[6] If at least one finite sequence is syntactically correct and represents a situation in S, then with probability one, we will eventually get to a sequence that syntactically correctly represents some sequence.

Then if A is a randomly chosen sequence of symbols of L, we can define the prior probability $Q(s)$ of s as the conditional probability that A syntactically correct represents s on the supposition it syntactically correctly represents something, i.e.,

$$Q(s) = P(R(A,s) \mid R(A)) = \frac{P(R(A,s))}{P(R(A))}.$$

We can call these L-Solomonoff priors.

These priors favor situations that can be can be more briefly represented in L over ones whose representations are long. The effect of these priors depends heavily on the choice of language L and how well it can compress some situations over others. For instance, in our chess case, if we have no quantifiers, it is easy to see that any two piece arrangements with the same number of pieces will have equal prior probability, because each square's contents have to be separately specified. Thus, if we know that every square contains a pawn, and we have observed the first 63 of these pawns and found them all to be black, the probability that the 64th square will be black is still $1/2$. On the other hand, if quantifiers and identity are allowed into our language, then the all-black-pawn situation can be briefly represented by

(1) $\forall x(\text{BlackPawn}(x))$

(where the domain is squares on the board), while the situation where squares $1, ..., 63$ have black pawns and square 64 has a white pawn is harder to represent. We might, for instance, use a sentence like:

(2) $\forall x(\sim(x = 64) \rightarrow \text{BlackPawn}(x))$ & $\text{WhitePawn}(x)$.

Since the probability of generating a given sequence of symbols decreases exponentially with the number of symbols,

(3) i

s much less likely to be randomly generated than

(4) ,

and it is intuitively very likely (though proving this rigorously would be quite difficult???refs) that in general the probability of generating a sentence representing 64 black pawns is higher than that of generating a sentence representing 63 black pawns followed by one white pawn. We can thus expect that the conditional probability of the 64th pawn being black on the first 63 being black to be very high. (Maybe even too high? It is very difficult to get good estimates here, because there are many ways that a single situation can be represented. ??refs)

Just as in the case of using linguistic complexity to quantify projectibility, the choice of language here provides many Mersenne questions. If we opt for the algorithmic version of the theory, we need to choose some computer language for a real or abstract computer, and then we need to choose a representation map between outputs and situations in the external world, with infinitely many possible candidates. And on the more descriptive versions, we still need to choose a language, with many decision points as to syntax and vocabulary. It is very unlikely that there is a privileged language. And not every will fit with our intuitions about induction. For instance, we can easily create a language, whether algorithmic or descriptive, where 63 squares of black pawns followed by one square with a white pawn are much more briefly describable than 64 squares with black pawns. For instance, on the descriptive side, we might use a BlitePawn(x) predicate, where something is a blite pawn provided it is a black pawn and on one of the first 63 squares or a white pawn and on the 64th, and an analogous WhackPawn(x).

In the unlikely case that there is a privileged language L such that L-Solomonoff priors are rationally required for us, we will have a vast number of Mersenne questions about the various parameters of the language and its representation relation. In the more plausible case that there is a set of languages such that we are required to have L-Solomonoff priors for some L in the set, we will have a vast number of Mersenne questions about the parameters that control the range of languages. All of this gives rise to a significant degree of appearance of contingency.

Consider, too, the following observation. It is implausible that the languages defining rational priors for us should be ones that are completely beyond our ken. But, on the other hand, it is plausible that there are possible languages that to the smartest human are as incomprehensible as one of the less intuitive computer languages like Haskell or Verilog or one of the creations of logicians further from natural language like lambda-calculus is to a typical six-year-old. Imagine now beings to whom such these languages beyond human ken are easy. It *could* be the case that the norms for rational priors for them are formulated in terms of L-Solomonoff priors for one of the "baby languages" that humans can understand, but this does not seem a particularly plausible thesis. It seems more likely that for those beings, the algorithmic rational priors would be different than for us.

The standard way??ref to defend algorithmic measures of complexity from the problems presented by a plurality of languages is to observe that sufficiently sophisticated languages have translational resources. Thus, one can write a Haskell interpreter in Javascript, and so anything that can be expressed in Haskell can be expressed in Javascript by including the code for a Haskell interpreter, and then using a string constant that contains the Haskell code. The result is that the difference in the length of code needed to generate a given output in different computer language will typically not be more than an additive constant: if one can produce the output with Haskell code in n bytes, then one can produce it in approximately[7] $n + k_{H,J}$ bytes in Javascript, where $k_{H,J}$ is the length of the Haskell interpreter in Javascript. For large enough n, the additive constant will be unimportant. If we are to measure the complexity of a one-hour broadcast-quality video by the length of code needed to compress the video, the addition of $k_{H,J}$ will likely be negligible: a Haskell interpreter is about half a megabyte, while an hour of video compressed losslessly can be reasonably expected to be in the gigabytes.

However, two points need to be made in our epistemological context. First, even if two different languages give very similar sets of priors, if they give even slightly different

[7]The approximation is due to complications due to having to embed code in a string constant, which may involve various escape characters.

priors, we either have the Mersenne question of what makes one of these sets of priors be the objectively correct one, or we have the Mersenne question about the boundaries of the range of permissible languages. Second, unlike in the case where we are measuring the complexity of a large set of data, such as a video file, in the inductive cases we need to look at ways of expressing relatively simple statements, such as "All electrons are negatively charged" or Schrödinger's equation or "The first 63 squares have a black pawn and the last square has a white pawn." But for such statements, a translation manual will dwarf the length of the translated text. To say "All electrons are negatively charged" in French by first describing how English works, and then saying that this description should be applied to the English sentence, will produce a French sentence that is many orders of magnitude longer than the English one, and hence not a sentence that is relevant to measuring the prior probability that all electrons are negatively charged.

Finally, while there is something elegant and natural about randomly choosing items in L by randomly choosing within the set of symbols with an end marker added, there are other ways to proceed. For instance, instead of making the end marker equally likely as each of the ordinary symbols, one could at each step of generation flip a fair coin. On heads, one is done generating. On tails, one then uniformly randomly chooses one of the n symbols.[8] Or one might first randomly choose a positive integer specifying the length of the sequence of symbols according to some probability distribution on the positive integers, and then make all the sequences of that specified length be equally likely. Or one might randomly choose a positive integer, and then choose the nth sequence of symbols in some ordering (e.g., alphabetical). While the initial symbol-by-symbol method with an end-symbol may seem more elegant, it is hard to say that it is rationally privileged to the point that the priors generated with it are rationally required. But if it's not thus privileged, then the range of random choice methods will provide more Mersenne questions.

[8]This will actually not make a difference in those languages where the syntax already determines where a syntactically valid sequence ends. This will be the case with some Polish notation languages, where a valid sequence ends when the main operator is filled out with arguments.

For not every random choice method yields priors that are plausible candidates for rational permissibility. There will be random choice methods where the sentence "Birds are a government-run drones" is many orders of magnitude more likely than all other sentences taken together, and so a boundary would need to be posited between the admissible and inadmissible random choice methods.

On a final note, ne might think that the intuitively most natural way of choosing a linguistic item at random is to make them all be equally likely. Unfortunately, this presents serious mathematical and philosophical difficulties. For a language based on finite sequences taken from a finite (or countable) alphabet, there are countably infinitely many sentences: we can enumerate them $s_1, s_2, s_3, ...$ in some arbitrary way. But if each one is equally likely, with probability some real number α, then we have a problem. In classical probability, we will have to have:

$$1 = P(\{s_1\}) + P(\{s_2\}) + P(\{s_3\}) \cdots = \alpha + \alpha + \alpha + \ldots.$$

But if $\alpha > 0$, then the right-hand-side is infinite, while if $\alpha = 0$, it is zero, and in neither case is it 1. There are technical ways of escaping this by departing from classical probability. They all require restricting the additivity axiom of probability that says that if $A_1, A_2, \ldots$ are countably many disjoint events then the probability of the union of the events is equal to the sum of the probabilities of the events to the case where there are only finitely many events. After that, one either takes $\alpha = 0$ or takes α to be a positive infinitesimal—something that is bigger than zero and smaller than any positive real number.

But whatever one does on the technical side, there will be philosophical difficulties. Emblematic of them is this paradox. Suppose you and I play a game where we each randomly pick a sentence with all sentences equally likely, and without seeing the other's sentence. When I see my sentence, whatever it turns out to be, I inevitably become nearly sure that your sentence comes after mine in the sequence, because there are only finitely many sentences that come before mine and infinitely many that come after. And you come

to be convinced of the same thing. This leads to paradoxical decision-theoretic conclusions. For instance if we are playing a game where one wins if one has a sentence further down in the sequence, you will then be willing to pay me any amount short of the prize to swap sentences with me, no matter what sentence you got. ?????

In any case, the equal probability approach itself does not appear to be a good way to generate induction-friendly priors, because it lacks the preference for shorter descriptions that is central to the functioning of algorithmic or linguistic priors.

Once we abandon, as we should, the equal probability approach to choosing a linguistic item, anything else is a matter of choosing from among infinitely many ways to be biased in favor of shorter expressions. Some of these are mathematically more elegant than others, but none is decisively so.

5.4. Anti-skepticism. There are skeptical hypotheses that predict pretty much the same thing as our best non-skeptical theories about the world. Indeed, for any reasonable theory T of the world, there is a variety of skeptical hypotheses about how the world merely looks as if T, say due to an evil demon, or us being in a simulation, or random quantum fluctuations in a Boltzmann brain. If an as-if-T skeptical theory T' starts off with prior probabilities in the ballpark of T's probabilities, because all of the evidence fitting with T fits equally well with T', no matter how much evidence we gather, T' will still be in the ballpark of T's probabilities. For instance, if T' starts off at half of T's probability, then it will remain at half of T's probability. In particular, it follows that T can never have more than $2/3$ probability, no matter how much evidence we gather for it, since we will always have $1 \geq P(T') + P(T) = (1/2)P(T) + P(T) = (3/2)P(T)$.

A sane epistemology thus requires low priors for skeptical hypotheses. How is it going to achieve this? Perhaps the best initial candidate would be a strong preference for simplicity. We might then expect to assign a lower probability to the hypothesis T' on which an evil demon makes things seems as if theory T were true on the grounds that the evil demon complicates the story. But on reflection, an evil demon need not *greatly* complicate the story, and may even simplify it in some respects.

For instance, if T is the true philosophical and scientific story about the world (assuming that in fact the skeptical hypotheses are false), then T will need to have a complex account of the mind-body problem—either giving an account of how consciousness and intentionality can be grounded in a physical system or account for the interaction of non-physical mental states and physical states. But the skeptical hypothesis can be entirely Berkeleian, supposing the evil demon to impose phenomenal states on a purely non-physical being. Likewise, a real physical world that fits with our best theories has vast amounts of complexity beyond our observational abilities: details of the behavior of the world too small or too far away for us to observe them. But an evil demon world need not include any of that complexity. Only what's empirically relevant to us needs to be included in the evil demon's deception. If the evil demon is just deceiving *me*, the evil demon's deception at most needs to encompass a small pretend-universe of radius of about a hundred light-years, since anything outside that radius is irrelevant to my observations. Nor need the evil demon think about the exact positions of all the particles in that sphere: the evil demon need only work with an approximate physics good enough to fool me.

Similar points apply to a number of other skeptical hypotheses. A theory on which I am a Boltzmann brain that's been floating for five minutes in a bubble of oxygen in an otherwise empty universe, about to die once the oxygen dissipates, involves much less in the way of informational complexity than the world of our best theories which posits vastly many more degrees of freedom due to positing many orders of magnitude more particles.??appendix:calculation

One might think a simulation hypothesis—that I inhabit a computer simulation—would involve significant complexity, because not only would I need to be simulated, but we would have all the complexity of the physical cosmos in which the simulating computer lives. But that physical cosmos need not be nearly as complex as the universe of our best theories. It could, for instance, be a cosmos optimized for computing. Whereas in our world, the logic gates, memory cells, oscillators and wires of computers are built from many atoms each (the number steadily decreasing, however), and the atoms themselves

are made of multiple particles with very significant informational complexity due to their apparently analogue (or at least very high resolution discrete nature), we could suppose a cosmos whose basic particles *are* logic gates, memory cells, oscillators and connectors, and which have only the degrees of freedom relevant to their computational role. Such a cosmos could be vastly simpler than the cosmos of our best physics.

Complexity does not appear to let us escape from skeptical hypotheses. What else can we do? Well, one option is to just go with common sense: assign low priors to skeptical hypotheses because they are "crazy". But it seems like an anthropocentric cheat to build such an anti-skeptical bias into the conditions for our priors. It would be odd to think that our priors should be such as to please our intuitions.

But this is not odd on Aristotelian optimism. If our world is as Aristotelian optimism has it, we would expect the structure of our priors to match our world to some degree, and hence to reject skeptical hypotheses. Assuming that in fact skeptical hypotheses are false, and that without presupposing their falsity it is difficult to engage in any sort of reasoning, it is unsurprising that our human nature simply tells us to assign them low priors. It's both epistemically and holistically good for us to do that, common-sense agrees, and on optimistic Aristotelianism that some practice is good for us is evidence that the practice is normative.

The above discussion was formulated in Bayesian terms. But similar points apply to any reasonable epistemological framework. We need some way to choose a non-skeptical over a skeptical hypothesis, even if the skeptical one fits our observations just as well. And an Aristotelian story on which our anthropocentric intuitions of what is and is not a "crazy" hypothesis are normative for us seems like a plausible account of why this kind of anthropocentric intuition is appropriate.

Outside of a natural law metaepistemology, however, it is difficult to see how one could have an account on which skeptical hypotheses consistently should have very low priors. Moreover, this kind of anthropocentric approach allows for the possibility that beings of a different sort, whose natural niche is a different kind of environment—say, non-material

beings whose proper environment is a Berkeleian world of spirits—would have different norms of reasoning, and what to us is a skeptical hypothesis might be to them the most common-sense thing.

5.5. Subjective Bayesianism. Subjective Bayesians avoid all the difficulties of specifying permissible priors by merely requiring the priors to satisfy some formal properties. These are taken to include the axioms of probability and, sometimes, the regularity constraint that all contingent propositions have non-zero probability. Rationality then constrains transitions from one set of probabilities to another: these must follow the Bayesian update rule that upon receiving evidence E, one's probability in a hypothesis H goes from $P(H)$ to $P(H \mid E)$. But the initial choice of priors is up to the individual, subject to the formal constraints.

The resulting picture of rationality does not match common sense. Take the most ridiculous set of conspiracy theories that we would all agree is unsupported by our evidence, but where nonetheless the conjunction of the theories is logically consistent with the evidence. Then there is a possible assignment of priors such that updating in the Bayesian way on our actual evidence strongly confirms the conjunction of these theories, both in the incremental sense of greatly increasing that probability and in the absolute sense of making that probability high. (All we need is that the prior probability of the conjunction of theories be low, but the conditional probability of that conjunction on the conjunction of our evidence be high.) Yet to reason that our evidence strongly supports these theories is paradigmatic of irrationality. It shouldn't be the case that a rational person could come to exactly the same conclusions on exactly the same evidence as are paradigmatic of irrationality.

Or consider the implausible asymmetry between the freedom in choosing initial priors and the rigid constraint in updating. Suppose I don't like my current set C of posteriors, and I would feel better if I had some cheerier alternative set A of posteriors. There is some set of priors Q that, given the evidence E that I had received over my lifetime, would have yielded A. According to the subjective Bayesian there would have been nothing irrational

in having adopted those priors in the first place, and thus having ended up at the cheerier posteriors. Why should I be tied to the priors that I actually had?

The picture of priors here is like a rationally unbreakable vow to live one's life as an evolution of these priors under Bayesian application of evidence. But it is a vow made without any rational ground, indeed without any choice, and likely in childhood. We would think it unsupportable that someone be held committed for life to a promise made early in childhood. Why then should we be bound to our priors and the posteriors coming from them?

5.6. Indifference. An intuitive way to solve the problem of priors is to place a constraint on one's priors that epistemically equivalent situations get equal credences. This is essentially a version of the principle of indifference: For each side of a perfect die, the proposition that the die will land with that side up is epistemically equivalent, so our prior credences should be indifferent between the sides. Assuming that the logical features of the situation necessitate that exactly one of the six sides eventuates, the finite additivity of probabilities ensures that each side has probability $1/6$.

There are many well-known problems with this approach. Some of these are technical??refs, but I will focus here on two non-technical ones.

First, in many cases it is difficult to see how we should carve up the space of possibilities into equivalent options. Should we, for instance, divide grand metaphysical views into naturalism and non-naturalism, with the two getting equal probability $1/2$, or should we have a finer-grained division into naturalism, anti-naturalism (nothing is natural) and dualism (there are both natural and non-natural things), with each getting probability $1/3$? Or should we have a longer list, with dualism being split into Platonism, monotheism, polytheism, etc., but naturalism left as one item, now of rather low prior probability?

Second, as we basically already saw in ??backref, if we are not careful, we will undercut induction. If every assignment of black and white coloration to a set of ravens has equal probability, then until they have observed *all* the ravens, the Bayesian's credence that all

the ravens are black will be no more than $1/2$.[9] Now, perhaps we should not consider every sequence of ravens on par with every other one. But now the question of which sequences are on par with which does not seem to have any canonical answer. As we discussed in ??backref, Carnap thought that lumping together sequences according to the *frequency* of blackness was the way to go. But if not all sequences are on par, why should we think all *sequences* are on par? Intuitively having exactly one out of N being black is equivalent to having exactly one out of N being white, but why should either of these be equivalent to having exactly half of them be black, or all of them be white, say? But if we have no good way to divide the raven color assignments into equivalent events, indifferentism does not help us.

5.7. Basic probabilities, norms and explanationism. Recently, Climenhaga??ref:`https://link.springer.com/article/10.1007/s11098-019-01367-0?utm_source=toc` introduced the very helpful notion of *basic* epistemic probabilities, which determine other epistemic probabilities. For instance, the algorithmic approach tends to take unconditional probabilities as the basic ones, and then conditional ones are defined by the rule $P(A \mid B) = P(A \,\&\, B)/P(B)$. Or on an indifferentist approach to a die roll, the basic probabilities are the equal probability $1/6$ of each of the six sides, and the other probabilities are determined by these and by logical relations. Thus, the probability of the die showing an even number will be $P(2) + P(4) + P(6) = 3/6$, because the three options are logically mutually exclusive.

An Aristotelian meta-epistemology can accommodate a picture of the grounding of epistemic probabilities that takes some of these probabilities to be more fundamental than others. For it is open to the Aristotelian to have a distinction between basic and derived norms. Thus, the prohibition of torturing blue-eyed persons is derived from the prohibition of torturing persons. The latter norm might be basic, or might derive from more basic

[9] Specifically, if there are N ravens in total, and n have been observed to be black, there are 2^{N-n} sequences of raven coloration compatible with the observation, only one of which has all the ravens black, the probability that they are all black will be $1/2^{N-n}$, which is less than $1/2$ if $n < N$.

ones. The question of the grounding structure of norms is a substantive question worthy of investigation, and indeed comprises a significant amount of the work in normative ethics.

It is worth noting that the grounding structure of basic norms is not exhausted by norms of the form: "Assign x as the epistemic probability of A (or of A given B)", nor does Climenhaga claim it is. There surely are basic norms like: "Make your epistemic probabilities satisfy the axioms of probability." But there could also be substantive constraining norms, like: "Make your epistemic probabilities for A-type events independent of your epistemic probabilities for B-type events." Furthermore, it is quite possible that some of the basic norms will be range-based, such as "Assign A an epistemic probability between x and y", or qualitatively or quantitatively comparative: "Assign A a higher epistemic probability than B" or "Assign A half the epistemic probability of B."

Bracketing some important technical detail, Climenhaga's own approach is that the basic probabilities are conditional ones like $P(A \mid H)$, where H is a hypothesis putatively immediately explanatory of A, and the basic probability values are determined by the explanatory relation between H and A. We could, for instance, suppose that H is a thorough description of the chemistry and physics of a match being struck and A is the event of the match igniting. The probability value then would be determined by the chemical and physical facts in H.

There is reason to be pessimistic about this as a complete solution, however. One set of difficulties comes from the thought that such a system of basic probabilities might work for an ideal agent, but humans are not ideal. The explanatory hypotheses we formulate, especially in the pre-scientific era or in ordinary life, are not formulated with the kind of precision that allows for precise probabilities to be read off. If Alice feels grave resentment against Bob, that could explain her insulting Bob, but the hypothesis of grave resentment does not yield a specific probability. We might think that the hypothesis of grave resentment is a disjunction of a large number of more scientifically precise hypotheses, ones that most people would not even be able to formulate. But if our basic probabilities only involve conditioning on these hypotheses, we will need norms for what probabilities a real

human should assign in light of a very, very vague sense of what the range of the more precise hypotheses is, what their probabilities are, and so on. We will need norms, coming along with many seemingly arbitrary parameters and consequent Mersenne questions, on how a real human being needs to approximate the ideal agent.

A second set of difficulties comes from the difficulty of applying the explanationist approach to explanatorily fundamental hypotheses, such as that God exists, or that the beginning of the world is entirely a natural event satisfying some precise unexplained physical theory. Here it seems we either need one fundamental hypothesis to be *a priori* certain or we need basic unconditional probabilities for the various fundamental hypotheses. The option of basic unconditional probabilities raises Mersenne questions. As for the possibility of an *a priori* certain fundamental explanatory theory, there is perhaps only one candidate for such a theory in the philosophical literature: theism, either as supported by classical deductive arguments from premises that are themselves certain (such as one's own existence, and various metaphysical principles), or supported as a basic belief required of all human beings by their nature.??cf.Plantinga

Such an *a priori* theism is worth considering, but it raises its own difficulties for the explanationist project. For while the hypothesis that God exists would yield immediate explanations for various claims about the world, it does not seem theologically plausible to think that God is bound by precise probabilities. Thus even with an *a priori* theism, we would have Mersenne questions about priors for conditional probabilities about what God would do if God existed.

5.8. Non-Bayesian update. Bayesianism is committed to updating by conditionalization being the only permissible way to update credences. If my current credence in a hypothesis H is $P(H)$, and I receive evidence E, my credence should move to $P(H \mid E) = P(H \,\&\, E)/P(E)$. No other changes of credence are permitted. There is an elegance and non-arbitrariness to this, of course modulo the above-discussed issues about the choice of the priors $P(H)$ and $P(H \mid E)$.

But as is the case for many other formally simple philosophical theories, this is too simple. Consider several cases.

PILL: A trustworthy oracle offers you a pill which will shift your credences closer to truth and does not introduce any incoherence.

MISTAKE: An hour ago, I made an arithmetical evidence when updating on evidence. I moved from credence 0.4 in H to credence 0.6, even though in fact $P(H \ \& \ E)/P(E)$ equalled 0.7. I have just discovered my mistake. Surely it would be good for me to go back and correct my credences, essentially rewinding my epistemic life from the mistake and re-conditionalizing on all the subsequent evidence.

STUPIDITY: My original priors had extremely high probabilities for some conspiracy theory involving members of a certain minority group, so high that despite the fact that all my life I have been receiving strong evidence against the theory, nonetheless I was quite convinced of the theory. I reflect on my original priors, and note that my priors of the conspiracy theory in question were quite out of proportion to my priors for similar theories involving other minority groups. I conclude that my original priors were stupid and racist. I go back and change them to treat the various groups more equally, and then as best I can I fix up my posteriors to match the evidence I recall having had.

In all three cases, my update of credences seems quite rational, but is not a case of conditionalization. We should thus suppose that while it may be a good general rule that we should update by conditionalization, we should at times depart from it. But the question of how we should depart from conditionalization now introduces complexity in the theory that raises significant Mersenne questions. We need more than just simple exceptions for each case: there will be parameters to set.

The pill case as given is straightforward. We can say that you should depart from Bayesian update when doing so would move some of your credences closer to truth and none further away from it, where a credence's "distance from truth" is the distance of the credence from 0 or 1 depending respectively on whether the proposition the credence is in is false or true. However, there are variants of the pill case. Suppose that the pill has

a 0.9 chance of moving you significantly closer to truth and a 0.1 chance of moving you slightly away from truth. It seems like it could be worth it. Or suppose that the pill moves your credences in important propositions closer to truth at the expense of moving your credences in some unimportant propositions further from the truth.

This suggests that we will need to quantify epistemic value, much as we did in Section??backref, and then say something like this: You are epistemically required (respectively, permitted) to opt for a pragmatically costless change of credence when doing so increases (does not decrease) expected epistemic value. If we quantify epistemic value in terms of strictly proper scoring rules, a delightful result of this move is that we do not need to handle the ordinary case of Bayesian update any differently. For it can be proved that in the kinds of cases where one simply receives evidence and sets one's credences according to it, conditionalization is the unique best policy for maximizing expected epistemic value.??ref:(Also Isaacs and Russell??)

However, as we also saw in Section??backref, the choice of a scoring rule or measure of epistemic value involves infinitely many free parameters, and at the same time not every scoring rule that satisfies the formal constraints is rationally plausible. In our present context of evaluating non-conditionalizing updates the point is particularly clear. For when evaluating credence-changing pills, we need to take into account the *importance* of the affected credences, and that is not a formal criterion.

Note that pill-type cases are not as outlandish as they may initially seem. We do in fact modify people's thinking with psychiatric medication as well as with psychotherapy.

Next, consider the case of the mistake in updating. We don't in fact remember all our evidence: we update our credences on the more important things and forget the less important, which often includes some or all of the evidence. I know that Beijing is the capital of China, but I don't know where I learned it from. I know that there was a cat in the yard earlier this morning, but much of the rich sensory evidence is now gone from my mind. An attempt to go back and correct a past update error is bound to be a messy affair. Sometimes when the mistake is minor it might be better to let it go rather than miss out on

some of the evidence gathered since. And the situation is often too messy and too complex for any sort of an expected epistemic utility calculation.

Yet the fact that a situation is messy does not get us off the hook. There will still be a distinction between right and wrong ways to proceed, even if they cannot be formalized. We have here something very similar to the kinds of messy everyday ethics cases that are grist for the situationist's mill.??backref-or-add It is unlikely that the answer is given by any elegant principle without the kinds of parameters that raise Mersenne questions, but at the same time there is an answer, and there must be something to ground it.

And what goes for the messiness in correcting a calculational mistake applies even more to the messiness involved in the kind of intellectual conversion that occurs when one realizes that all of one's thinking for years has been based on irrational prejudice, and one attempts to dig oneself out of the resulting heap of epistemic defects. Again, there are right and wrong ways to proceed, but the likelihood of a clear and elegant principle that solves the problem is nearly nil.

There is a final set of issues with update. Presumably, there is something to the maxim that ought implies can. And typically we can't do Bayesian update. We don't have the time, don't have the mathematical skills, or perhaps most importantly aren't able to quantify our priors in such a way as to make them amenable to precise mathematics. We need an epistemology that works in this all too human predicament, a *human* epistemology, and we need an account of what grounds that epistemology.

One may insist that under these imperfect conditions, we should simply say that we are stuck with irrationality. But nonetheless there are cases where it is clear that something should be done under the unhappy circumstances. Suppose that the completely right credence on my priors in some proposition p is 0.9874, but I can only calculate to two significant figures. Then I should take the credence to be 0.99, not 0.98 or 0.01. But things will be less clear if I need to form credences in both p and q, which a perfect Bayesian with my priors and evidence would take to be 0.9874 and 0.0332, but due to limitations of time or energy level, I can only get a total of four significant figures. Are the right credences

for me 0.99 and 0.03, or 0.987 and 0.0, or 1 and 0.033? It depends, surely, on the relative epistemic importance of p and q. And once importance needs to be taken into account, it is very unlikely that we will have a precise and elegant account with no free parameters. Instead, we have human messiness.

6. Why be epistemically rational?

One of the main questions in ethics for the past century has been why one should bother being moral. The analogous question for epistemic rationality has received comparatively little attention??ref:exceptions. One reason for the low amount of attention could be that there is an obvious answer: It's good (both instrumentally and not) to have the truth, and following the norms of epistemic rationality helps one get there.

Indeed, there are proposed norms which can be justified in this way. If one measures the utility of a doxastic state by a strictly proper scoring rule, then by the agent's own lights Bayesian update is an algorithm that maximizes epistemic utility.??refs Similarly, it seems reasonable to think that the epistemic rationality involved in choosing which evidence to pursue, i.e., which experiments to perform??backref, can be accounted for by such a maximization. And finally the norm that one's credences should be (probabilistically) consistent can be justified by score maximization, because of a famous theorem that if one's credences are not consistent, there is a set of credences that are guaranteed to yield a better score no matter what the truth turns out to be. If Bayesian update, choice of experiments and consistency are all there is to epistemic rationality, then this kind of reasoning has some hope of answering the question of why one should bother being epistemically rational, assuming we take it for granted that the good is worth having.

But this hope may not be borne out. As argued in ??backref, there is good reason to think that prudential rationality does not always involve maximizing utility—risk needs to be taken into account. Once we have seen that expected utility maximization is not the only reasonable position to take with respect to prudential goods, why should we assume that it is so for epistemic goods?

But more seriously, Bayesian update, consistency and the norms governing choice of experiment do not exhaust the scope of epistemic norms. For a Bayesian should admit there are substantive objective norms constraining rational priors going beyond consistency. It is not rational, for instance, to have a very high prior credence for the hypothesis that one is a brain in a vat hooked up to a simulation of precisely the physics of our universe (cf. ??backref). And then we need an account of why one should bother to have consistent priors that are rational rather than some others.

It won't do to say that the rational priors best conduces to epistemic utility. For they don't. It would have been irrational for human beings to have a high prior for such hypotheses as that water is made of hydrogen and oxygen atoms, that 2020 would be a year of a global pandemic, or that the fine-structure constant is approximately $1/137$. And yet epistemic utility would have been increased by adopting such priors, because all these hypotheses are in fact true. The priors that uniquely best conduce to epistemic utility are a prior of 1 for every actual truth and a prior of 0 for every actual falsehood. But such priors would not be rational for human beings.

An Aristotelian has a story. Although truth is partly constitutive of our cognitive flourishing, so is conformity to the norms of rationality. It is thus non-instrumentally good for us that we reason in accordance with our nature. A human being who thought *a priori* that the fine-structure constant is approximately $1/137$ would be right, and would flourish to that extent, but would also have crazy priors, which is contrary to cognitive flourishing. Typically, we have more control over whether we follow the norms of epistemic rationality than whether we get to the truth, and so from the point of view of pursuit of one's own flourishing it makes sense to follow these norms in typical situations.

That said, one can imagine cases where tradeoffs are reasonable. Suppose a trustworthy but eccentric alien offers to inform you of the correct theory of stellar formation in exchange for the very small sacrifice of brainwashing yourself into thinking that you have an even number of hairs on your head. Irrationally coming to a belief about the parity of

your hair count is contrary to our cognitive flourishing, but in terms of epistemic utility, knowing the correct theory of stellar formation seems worth it.

That said, there is another possible answer to the "Why bother?" question. It could be that to the extent that our cognitive life is under our voluntary control, there are moral requirements that we follow epistemic norms. After all, on the account defended here, morality extends throughout the sphere of the voluntary: the moral is defined by the proper functioning of the will. Thus it is possible that one have a deontological view on which even if voluntarily brainwashing yourself to believe something you have insufficient evidence for is beneficial to your epistemic flourishing, nonetheless it is morally wrong to do so. Much of our cognitive life is not under our direct voluntary control. But we may nonetheless have a *prima facie* moral duty to do what we reasonably can to develop our cognition in such a way that even the involuntary parts cohere with the norms, just as we may have a *prima facie* moral duty to promote the health of our bodies. It does, after all, seem to be a part of the proper function of our will to direct us to both physical and cognitive health.

And we should not take the moral account too far. We can have two people who act equally well morally, but one of them does cognitively better than the other, due to innate talent, opportunities for cognitive self-improvement, and simple luck at getting to the truth. Epistemic value goes beyond moral excellence, if only because it includes the value of epistemic success, much as health goes beyond moral excellence, even though there are *prima facie* duties to promote our own health.

7. Intellectual limitations

Limitations in recognizing logical implications and contradictions are a clear case where rational norms are at least species-relative. It is epistemically irrational for one to conclude that it's raining from the fact that it's neither raining nor snowing. But it is not irrational for one to take $22828 \times 2219 = 50645332$ (which is in fact false) to follow from the

axioms of arithmetic due to an arithmetical slip. Of course, someone with exceptional calculational skills may immediately that see the latter is mistaken, but the ordinary person's failure to see it is not an instance of irrationality.

Ethics also contains intellectual limitation cases. Analogically to the multiplication case, we would not consider a person to be less than virtuous because they are unable to see an extremely complex medical ethics case rightly. But if someone doesn't realize that it's wrong to ambush strangers in order to sell their organs, there seems to be something morally wrong.

But the ethical cases may also be different. Consider an adult who strives to follow their conscience as best they can, but nonetheless fails to see that it's wrong to kill strangers for their organs. This person presumably has a severe intellectual and/or emotional disability. We might judge them to be a good person, or at least not a vicious one. However, the logical case seems different. The person who, due to intellectual disability, concludes that it's raining from the claim that it's neither raining nor snowing *is* failing at rationality, though of course they are inculpable. One can defensibly define moral worth in terms of the seriousness of their attempt to do what seems right, but it is difficult to define someone's degree of rationality in terms of the seriousness of their attempt to think as seems right. For it is paradigmatic of irrational people that they take themselves to be acting quite rationally—if one derives that it's raining from its neither raining nor snowing, this is precisely because the conclusion seems to follow. It is, indeed, unclear whether it is even possible to conclude p from q without its seeming that p follows from q. But it is paradigmatic of immoral people that they lack integrity and violate their own conscience.

The line to be drawn in the epistemic rationality cases (and in the ethical ones if they turn out to be similar), is a line that we are unlikely to be able to draw in a species-independent way. It seems plausible that for beings that as a species are more adept at logic than we are, a failure to see the truth of Fermat's Last Theorem as following from the axioms of arithmetic is a failure of rationality, but it is not so for us.

Recently, Jeffrey??ref has argued that ethical norms may be relative to a stage in life. Whether this is so, it is very plausible that some epistemic norms are so. There may be logical facts failure to notice which constitutes a failure of an adult's rationality, but would not constitute a failure of a child's rationality. If this is right, then we have all the more reason to think the norms of epistemic rationality to be species-relative, since surely what the stages in life are, and when they occur, is something that is species-relative.

7.1. Innate beliefs and testimony. Humans are said to have much less in the way of instinct than other earth animals. Furthermore, we do not seem to have any innate beliefs. But we can imagine a species of intelligent animals which do more by instinct than we do, and which have evolved to be born with some unshakeable and true beliefs, such as that purple winged things are to be avoided and electrically charged spiky fruit is good to eat. For these beings, there is nothing irrational about having such beliefs without any evidence. For us, there is. In a Bayesian mode, we might say that for these beings very high priors in these empirical claims are appropriate, but not so for us.

Experiments in machine learning suggest that for certain kinds of problems it is useful for a system to be pretrained and hence have priors that embody some substantive information. Pretraining presumably comes with a sacrifice of flexibility. Thus whether pretraining is appropriate depends on the subject area and the kinds of situations the system will face. And of course the content of any pretraining is highly contingent. It would be surprising if our evolutionary process did not involve any pretraining. We have good reason to speculate that it is normal for us to have certain substantive priors, and that not to have them would be abnormal. Some evidence for this is given by infants' neural responses to snake pictures.[10] But what those priors should be is surely highly species-dependent: we would not expect alien babies evolved for an environment without any snakelike objects to naturally react to snakes. And even if humans, surprisingly, turn out not to have any substantial information encoded in their priors (if that is even logically

[10]??ref:https://www.nature.com/articles/s41598-020-63619-y

possible), that fact is surely species-dependent. In such a case, for us to have substantively informative priors would be abnormal, but for intelligent cats, say, it might be quite normal.

We can also imagine a species where memories are inherited. A member of this species could, then, be born with a large number of beliefs that they had no evidence for. For, like we often do, the members of this species could forget the evidence that led to a conclusion. But while in our case, we once had the evidence, in this species it was some ancestor of theirs that had the evidence. It is plausible that for us it is generally irrational to have beliefs without ever having had evidence. But this is just a normative fact about our species, not a fact about species-independent rationality.

Now, let us return to our species. Perhaps we should not think the species where memories are inherited to be all that different from us. For do not our parents bequeath much knowledge to us? It is true that this is mediated by soundwaves and inkmarks rather than by gametes, but does that make a significant difference? We should, thus, take seriously the idea that just as for members of a species where memories are inherited it is fundamentally rational to believe these inherited memories, and to question them without special reason is irrational, for us it may be fundamentally rational to believe at least some testimony, and to question it without special reason is irrational. If so, then we have another example where the Bayesian way of looking at update may not be correct. ???fill out

7.2. Epistemic self- and other-concern. Suppose it turns out that I have a near-identical twin on a planet just like Earth, all of whose mental life has been exactly the same as mine, with one exception: the first bird I saw today was a bluejay, while the first bird he saw today was a robin. Neither of us is an avid birder, and so this has no significant effect on me. We both learn that next time we fall asleep, we will be transported to a third planet, and my twin's memories will be changed so that his robin memory will be replaced by my bluejay memory, with corresponding tweaks for anything else affected by the memory. Furthermore, I am given a choice of epistemic policies: I can ensure that

tomorrow both I and my twin will firmly believe that the bluejay memory is correct, and hence will believe ourselves not to have had our memories changed, or I can ensure that we will both assign credence $1/2$ to the bluejay memory's being correct.

Insofar as my epistemic concern is solely for myself, it makes sense to take the policy of sticking to the bluejay memory. But doing so harms my twin epistemically. What about the total epistemic value? As before, let $T(r)$ and $F(r)$ be the accuracy scores or epistemic utilities for assigning credence r to a proposition that is true and false respectively. Then on the policy of sticking to the memory, the total epistemic value is $T(1) + F(1)$: I will have credence one in a truth and my twin in a falsehood. On the policy of assigning credence $1/2$, the total epistemic value is $T(1/2) + F(1/2)$. If the pair T and F is strictly proper, then $(1/2)T(1/2) + (1/2)F(1/2) > (1/2)T(1) + (1/2)F(1)$, and so $T(1/2) + F(1/2) > T(1) + F(1)$. Thus total epistemic utility is higher when the policy is of total uncertainty. However, the total epistemic utility *for me* is higher when the policy is of certainty: obviously, $T(1) > F(1)$.

When evaluating epistemic policies, then, we have two different points of view: a self-concerned and an other-concerned point of view. And we are led to different policies by the different points of view.

The above is a strange case. But there are everyday cases to be considered. It might well turn out that as a community we get to the truth more effectively if investigators have a stronger commitment to their theories than the evidence warrants, because then they will be more motivated to search for evidence and will have "skin in the game". However, such a commitment can be expected to be epistemically harmful to the individuals. Thus the epistemic goods of the community may conflict with individual epistemic goods.

For completely asocial rational agents, we would expect norms of epistemic rationality favoring individual epistemic goods. But for social rational agents, we would expect norms of epistemic rationality favoring social epistemic goods. But sociality comes in degrees. We would expect that in more social kinds of rational animals, there would be a

stronger favoring of social epistemic goods, and in less social ones, a weaker such favoring. We are neither asocial nor maximally social (we are not a hive mind!). For us, we would expect some kind of an intermediate between favoring individual epistemic goods and favoring social epistemic goods. But that intermediate normative position will raise Mersenne questions as to why it lies where it does. For it is not plausible that there is some metaphysically necessary rule that takes the degree of sociality of a rational agent and spits out the degree to which social epistemic goods should be favority. It is much more plausible to suppose that the degree and manner of preference for social epistemic goods is a part of the kind norms.

Nor is it likely to be a single numerical degree of preference per species. Surely the appropriate degree of preference for social epistemic goods depends not just on the species, but also on one's role in society. So we need a ground not just for single human numerical degree of preference for social epistemic goods, but a function from social role, and probably area of knowledge, to degrees of preference for social epistemic goods.

7.3. ***Imprecision.** Suppose that a spinner is randomly spun, and it comes to a stop at some angle between 0° and 360°, inclusive, though of course 0° and 360° label the same outcome. To make the logic of the situation a little simpler, let's label that outcome "360°", and call the spinner angle X. Thus $0^\circ < X \leq 360^\circ$. Independently of the spinner, a fair coin is tossed. If the coin is tails, the spinner is left unchanged. If the coin is heads, the spinner is adjusted so that it now points to half of the angle it started with. Call the final angle Y. All of the above happens out of your sight.

Note that if the coin lands heads, then we are guaranteed to have the final spinner angle $Y \leq 180^\circ$. Thus, if you were to learn that $Y > 180^\circ$, you would thereby have learned that the coin landed tails. It is a standard result in Bayesian epistemology that if something is evidence for a hypothesis, then its negation is evidence against the hypothesis. Hence, if you were to learn that $Y \leq 180^\circ$, that would be evidence against the tails hypothesis, and hence in favor of the heads one. More specifically, the probability of heads given

that $Y \leq 180^\circ$ will be $2/3$. For there are initially four equally likely options given your evidence:

(i) heads and $X \leq 180^\circ$

(ii) heads and $X > 180^\circ$

(iii) tails and $X \leq 180^\circ$

(iv) tails and $X > 180^\circ$.

Learning that $Y \leq 180^\circ$ rules out (iv), since on tails no adjustment is made and so $X = Y$. However, learning that $Y \leq 180^\circ$ does not rule out any of the ways of getting any of the options (i)–(iii). Thus, (i)–(iii) remain equally likely. Since heads occurs on two of these three options, thus the probability of heads is $2/3$ after learning that $Y \leq 180^\circ$.

Suppose, now, that we have a three-step process. At t_0, the spin, toss and any adjustment is done out of your sight, but you know the process that was followed and will be followed. At t_1, you will be informed whether $Y \leq 180^\circ$. At t_2, you will be informed of the exact value of Y.

Imagine that at t_1 you learn that indeed $Y \leq 180^\circ$ and at t_2 you learn that Y is exactly 22.44°. At t_0, your credence in heads was $1/2$. At t_1, it went to $2/3$. What about at t_2? Well, here is the problem. At t_2, the total information gained since t_0 can be summed up as $Y = 22.44^\circ$ (this logically entails the information from the t_1 stage that $Y \leq 180^\circ$). But the information that $Y = 22.44^\circ$ is equivalent to:

(v) Either tails and $X = 22.44^\circ$, or heads and $X = 44.88^\circ$.

But note that (v) tells you nothing about whether you got heads or tails, in light of the fact that all the spinner angles are equally likely, the coin is fair, and the spinner and the coin are independent. Both paired coin-and-spinner outcomes are equally likely, and when we learn that exactly one of them happened, they remain equally likely.

In light of the fact that your total information at t_2 is that given in t_0 plus (v), at t_2 your probability for heads seems to be $1/2$.

But now observe that this argument would work no matter what angle you learned Y to have at t_2. Thus at t_1, you would already know for sure that given the additional information you will receive at t_2, your final credence in heads will be $1/2$.

Famously, Bas van Fraassen has posited the plausible reflection principle that if a rational agent knows for sure that their future credence will be some value, then that should *already* be their credence. This principle needs some qualifications to avoid easy counterexamples. The agent needs to know for sure that they are will maintain rationality, and that they will not lose any information.??refs We can suppose that all of these qualifications are met in our case, and given these qualifications, the reflection principle is very plausible. Yet our case appears to be a counterexample to the principle.

It is worth noting the pragmatic effect of this failure: you are subject to a Dutch Book, a sequence of bets with a guaranteed loss. At t_1, you will be willing to accept a wager where you get \$45 on heads and pay \$55 on tails, since $(2/3) \cdot \$45 - (1/3) \cdot \$55 = \$11.67$. But at t_2, you will be regretting this wager and expecting an outcome of $(1/2) \cdot \$45 - (1/2) \cdot \$55 = \$(-5)$. Given your expectation at t_2 of a five dollar loss, you will be willing to pay the bookie four dollars if they offer to let you change your mind. As a result, you have a guaranteed loss of four dollars. Nor is this some kind of a rare situation: there is nothing that special about the case where $Y = 22.44^\circ$: any angle less than or equal to 180^circ has the same property. Thus half of the time that the whole game is played you will end up Dutch-Bookable at t_1 and t_2.

Note that there is a technical problem in the above reasoning. I argued that updating on the information in (v) yields equal probability of heads and tails. However, the information in (v) has probability *zero*. To see this, note that for any small positive number ε, the probability that X, when measured in degrees, is between $22.44 - \varepsilon$ and $22.44 + \varepsilon$ is $2\varepsilon/360$. Thus the probability that X is *exactly* 22.44° is no bigger than $2\varepsilon/360$ for *any* positive real number ε. But the only way a non-negative number p (and probabilities are

non-negative) can be less than or equal to $2\varepsilon/360$ for any positive number ε is if p is zero.[11] Standard Bayesian reasoning updates the probability of a hypothesis H on evience E to $P(H \mid E) = P(H \;\&\; E)/P(E)$. But if $P(E) = 0$, then this formula yields $0/0$, which is undefined.

The literature contains various ways out of the difficulty of updating our information on an event that seems to have zero probability. We might say that it's not zero but an infinitesimal, or we might have a mathematical framework like Popper functions that allows for such conditionalization, or we might work with probability densities. Whatever we do, a criterion of adequacy on the method will be that we preserve the symmetries in the situation by taking the two disjuncts in (v) to be probabilistically on par, and hence making the update yield equal probabilities for heads and tails.[12] And that will lead to our paradox.

The solution I want to propose instead is this. Stop the search for general ways of updating on zero probability events. Except in special cases (more on that soon???), such updates should not be done. Instead, embrace imprecision. In practice, we never exactly measure the position of a spinner. Suppose that instead of measuring the spinner position at 22.44° at t_2, we measure it to be 22.44 ± 0.01 degrees. Let E be the evidence, then, that the spinner finally ended between 22.43 and 22.45. Then $P(E) = (1/2)(0.02/360) + (1/2)(0.04/360)$ (on tails, to get E we need need the spinner to initially land between 22.43 and 22.45; on heads, we need it to initially land between 44.86 and 44.90). Then by Bayes' theorem:

$$P(\text{heads} \mid E) = \frac{P(E \mid \text{heads})}{P(E)} P(\text{heads}) = \frac{(1/2)(0.04/360)}{(1/2)(0.02/360) + (1/2)(0.04/360)} = frac23.$$

[11] Suppose that $p \geq 0$ and $p \leq 2\varepsilon/360$ for every $\varepsilon > 0$. For a *reductio ad absurdum* suppose $p > 0$. Then let $\varepsilon = p$. This is a positive number, so $p \leq 2\varepsilon/360 = p/180$. But we cannot have $p \leq p/180$ for a positive p.

[12] There are technical problems with maintaining symmetry in general in contexts with infinitesimal probabilities. But the amount of symmetry needed for the above argument—namely, equal probability of 22.44° and 11.22°—can indeed be maintained. ??ref

Now, if we were to measure things will full precision, we would indeed have to update on zero-probability events, and we would sometimes end up in a bind. But our rationality, I propose, is not made for full-precision measurements. Our rationality is made for imprecisionand is kind-relative. Beings who made full-precision measurements of real numbered quantities would have to reason differently from how we do. Maybe they would have some other method than conditionalization for dealing with these cases. Maybe violating the reflection principle and being Dutch-Booked would be rationally acceptable for them. This all fits very well with our Aristotelian picture. We have our nature and the rationality that our nature requires of us.

8. *A priori* intuitions

We have a large store of what one might call substantive *a priori* intuitions, such as that nothing can cause itself, that there is contingency, that nothing is a proper part of one of its proper parts, that a material object couldn't have been always immaterial, that if it is not necessary that there exist green objects then neither is it necessary that there exist blue objects, that every thought has a thinker, that every number has a successor, that the Peano axioms of arithmetic are consistent, and so on. These intuitions vary in strength, though all the above examples are very plausible.

An Aristotelian account of epistemological normativity can hold that among the norms of our nature are requirements that we have a high credence, maybe in some cases even certainty, in such propositions, and add optimistically that we are so constituted that we tend to follow this requirement of our nature. We can then have a basic justification in such claims simply in virtue of our nature requiring us to assign them a high credence.

9. Going beyond the applicability of human epistemology

Suppose that you are certain you are one of infinitely many people have each rolled a fair die, none of whom have seen the result. It is then announced by a perfectly reliable

angel that all but finitely many of the dice show six. What should be your credence that your die shows six?

On the one hand, it seems very likely that your die shows six. After all, the vast majority of the dice show six, and you have no reason to think yourself exceptional.

On the other hand, prior to the announcement, your credence in six was 1/6. And the announcement told you nothing about your die. For the following two statements are logically equivalent:

(1) All but finitely many of the dice show six.

(2) All but finitely many of other people's dice show six.

For your die's state makes no difference to whether there are finitely or infinitely many non-sixes. But then given your certainty that the dice are fair, (2) gives no information about your die's result. And since (1) is logically equivalent, it too gives no information. Having received no information, you should stick to 1/6.

This reasoning seems convincing. Yet if everyone sticks to 1/6, then all but finitely many people are quite far from the truth. Furthermore, if everyone is playing a game where you are asked to guess if you have a six or not, and you are rewarded for getting it right and penalized for getting it wrong, then if everyone sticks to 1/6 for having a six, everyone will guess that they don't have six, and all but finitely many people will be penalized, which is surely not the right result.

In ??ref, I argued that a good solution to epistemological paradoxes like this is to reject the metaphysical possibility of an event depending causally on infinitely many events, in the way that the angel's accouncement depends on the infinite number of die rolls. But there is another possible solution: we can deny that our epistemic norms extend to far-fetched situations, just as in ??ref it was suggested that our ethical norms do not apply to far-fetched situations. Whether this is a completely satisfactory resolution to the paradox is not clear. For there is some plausibility in thinking that if predicaments like the above were metaphysically possible, there could be rational beings who could reason in them. But it seems there couldn't be any. However, no matter what we say about the metaphysical

possibility of such beings, there is something right about the thought that *we* are not made to reason about such things.

There are many other examples where epistemology seems to break down in radical cases. What should you think if you came to be convinced that an evil demon is trying to get you to believe as many falsehoods as possible? Any thought you might have ends up undercut. Should your credences in everything be at 1/2, then? But that is incoherent: for then you have credence 1/2 that the last die you rolled is 1, and credence 1/2 that it was 2, and credence 1/2 that it was 3, and credence 1/2 that you never tossed a die, and so on. Perhaps suspension of judgment is something other than assigning probability 1/2. But if so, should you suspend judgment about everything, including about the norm of fitting your beliefs to you evidence? However, if you suspend your judgment about that, then why bother with suspending your judgment about other things, given that you don't think you need to fit your beliefs to your evidence? Or what should you think if you come to be convinced that you are a computer simulated non-player character in a sophisticated video game? What kind of a world should you think you are in?

Note now that there is such a thing as realizing that when you were using certain words, you had no concept behind them, and the words were mere meaningless words. For instance, take a word that we as laypeople defer to experts on the meaning of, such as "gluon". But now imagine that physicists have been pulling our collective legs about gluons all this time: it was just a made-up word without any meaning behind it. We can imagine discovering that. And words whose meaning we get by deference are not the only words like that. The phenomenon of a person who doesn't know what they are talking about is not uncommon. My metaphysics includes accidents. But I could imagine finding myself in a position where I come to be convinced that I never had a concept of an accident—that "accident" is a meaningless word. Now imagine a radical hypothesis: I come to be convinced that I have no concepts at all. What should I think now?

There is something plausible about the idea that in certain extreme situations there is no right way for us to think, but beings other than us could have moral norms that provide

additional specifications—ones that may appear arbitrary to us, say—as what one should think in some of these extreme situations.

10. What is epistemic rationality?

We have an Aristotelian answer as to what epistemic rationality is among humans: it is the proper functioning of our cognitive faculties. That answer nicely generalizes to yield an account of epistemic rationality for a variety of kinds of beings, as long as we can make sense of what cognitive faculties are.

In the case of morality, a similar functional approach defined morality in terms of the proper function of the will, and the will as a system directing behavior in the light of practical good-directed reasons. We might try here try for a similar account but this time in terms of reasons directed at epistemic goods, such as correct and informative (whether in the sense of yielding understanding or some other sense) representation of the world. This approach is more problematic, however, than on the moral side, especially if we are open to a wide range of possible epistemologies for different kinds of rational beings.

First, we could imagine beings that directly contemplate the truth, like Aristotle's gods or Kant's intellectual intuiters. It is difficult to identify reasons in such a case, and yet the contemplation appears to be a rational activity. And while humans do not normally directly intuit reality, it is possible that we do so in special cases, such as Platonic contemplation of the Forms or the Catholic of human flourishing as bound up with a direct beatific vision of God.

Second, reliabilist epistemology can involve processes that reliably directly generate beliefs in us rather than doing by the intermediary of reasons. And even if reliabilism is not a correct description of human epistemology, it seems imaginable that there be rational beings of whom it is normatively true.

Third, objective Bayesianism has three places for epistemic normativity: updating on evidence, choosing which experiments to perform, and having rational ur??-priors. The update and experiment-selection are indeed based on reasons. However, the priors are

not typically attained on the basis of reasons, but appear to be simply innately had by an agent (though in cases where one changes one's priors reasons may be involved). Again, even if objective Bayesianism is not the right normative theory of humans, it is plausible that it could be the normative theory for a different kind of rational being. Moreover, there are other phenomena that could be governed by rational normativity that are similar to that of priors, such as innate beliefs, or beliefs passed on genetically in imaginable alien species.

Without adverting to reasons, we might simply define an epistemically rational system as one directed at epistemic goods, such as correct and informative representation of reality, and epistemic rationality as the normativity of the proper functioning of such a system. However, if we did that, then epistemic rationality would go very far down in the animal kingdom: even worms have systems directed at correct representation of reality. To rule worms as not rational, we might want to add an element of self-reflection, where the agent is such that they ought to explicitly embrace the epistemic goods as such, and ensure that their epistemic functioning is correctly aimed at this goods. This last condition would contemplation, reliable belief formation and prior-generation back into the fold of rationality: we might say that it is precisely when the agent reflectively monitors these processes in their directedness at epistemic goods that the agent is epistemically rational in engaging in them.

It is worth noting that on this reflective account, the high forms of cognition, such as the Aristotelian gods' direct vision, might well be rational in a somewhat different way from the rationality of lower forms of cognition. For the high forms of cognition are often seen as essentially correct, neither in need of correction nor repair, and hence the self-monitoring could be limited to realizing that necessarily everything is working correctly. This is akin to the way that morality would work differently for a perfect being that does not need to engage in self-correction, but that knows that it always functions morally correctly.

11. Metaepistemology

The primary question of metaethics is whether ethically normative claims have truth value, and if so, what grounds their truth value. The Natural Law answer that I have been defending is that ethically normative claims have truth value grounded in facts about the flourishing of the will, which facts in turn are grounded in our form (??add this!). Analogically, we would expect epistemologically normative facts to have truth value grounded in facts about the flourishing of the intellect, which facts in turn would be grounded in our form.

Natural Law metaethics as it stands is compatible with a wide variety of normative ethical theories, though some combinations are less appealing philosophically. Thus, while one could suppose that the proper function of the will is to maximize utility, normative utilitarianism fits more elegantly with metaethical utilitarianism according to which facts about the right and wrong just are facts about utility maximization.

Similarly, Natural Law metaepistemology fits with a variety of normative epistemological theories. Reflection on ethical cases shows that flourishing can have internal and external components (e.g., one's friends' well-being is partly constitutive of one's well-being), and so Natural Law metaepistemology is compatible with internalist and externalist norms. It is compatible with full-blown Bayesianism, though just as normative utilitarianism fits more neatly with metaethical utilitarianism, so too those versions of Bayesianism on which the norms all have simple formal statements—say, subjective Bayesianism or Carnapian objective Bayesianism—fit more simply with an epistemology on which the norms are grounded in formal constraints. On the other hand, an objective Bayesianism with non-formal constraints on priors fits very well with a Natural Law metaepistemology.

However, over the last century or so there has been a disciplinary difference between ethics and epistemology. While the ethicists have had a primary focus on the question of how one should act, the epistemologists' primary focus has not been the question of how we should think, but what knowledge is. One might take this question of the nature of knowledge to be akin to many ethicists' interest in the nature of virtue. In both cases,

there is a sense in which the question does not carry normativity on its sleeve, in that there are possible answers to the question that do not involve any normative components. Thus, Descartes identified knowledge with the clear and distinct, and a naive Aristotelian (though of course not Aristotle himself) might identify virtue with the mathematical center between the humanly possible extremes. Of course, in both cases it is highly plausible that the item is worth having, but that does not make the item essentially normative—water is also worth having, but is not essentially normative. However, likewise, in both cases more plausible theories of the item have a significant normative component. Thus, the real Aristotle would identify virtue with a *reasonable* mean, while many have tried for accounts of knowledge where one of the components is *justification*.

Natural Law metaepistemology more easily yields an account of the grounding of more obviously normative concepts such as justification. For instance, a justified inference could be an inference performing which constitutes intellectual flourishing (or the intellect *qua* engine of inference). Again, this is compatible with many theories as to what kinds of inferences are in fact justified. But it is more difficult to see exactly how to ground knowledge. Nonetheless, the Natural Law metaepistemologist can simply say that facts about knowledge are grounded in facts about intellectual flourishing with specifying how the grounding goes. Or they might punt and say that the concept of knowledge is not essentially normative, and hence explaining how knowledge is grounded is beyond the scope of the normative parts of metaepistemology—and perhaps it is only to the normative parts that the Natural Lawyer has a distinctive contribution.

I think there is a plausible story about something in the vicinity of knowledge, however. We can think of two aspects of the intellect, namely process and accomplishment, and both aspects have their distinctive flourishing. Justification will be a matter of process flourishing. But knowledge seems more like accomplishment. Thus, we might try to identify an accomplishment of the intellect as knowledge provided that this output constitutes intellectual flourishing. However, this is not exactly right. For flourishing intellectually in an accomplishment will include not just the aspects of knowledge acknowledged in

the analytic tradition, like true belief and justification, but also *understanding*, structural connections between different things found in the intellect that enlightens us about the things that are important. Aquinas famously distinguishes the vice of *curiositas* from the virtue of studiousness, and the distinction lies in curiositas' trivial pursuit of mere bits of knowledge while the virtue pursues a holistic structured understanding of the world's explanatory conenctions. Thus flourishing with respect to intellectual accomplisment is more like what the medievals called *scientia* than what we call knowledge.

So a Natural Law metaepistemology elegantly gives us a theory about the grounding of *scientia* rather than of knowledge. Is this a disadvantage? Or is it a hint that in fact many of the blind alleys of post-Gettier epistemology have been due to the fact that the more natural concept in the vicinity is the intellectual accomplishment of *scientia* rather than mere knowledge? There is room for significant further exploration here, as well as room to search for a purely normative characterization of knowledge that might end up being grounded in human nature. How such a search might go will depend on controversial questions about the role of knowledge in our rational life.

While my own view suggests we should minimize that role in favor of understanding and credence, someone attracted to knowledge having a significant role might have room to give a normative-functional account of knowledge that has a significant role for human nature. Thus, if one thinks that the knowledge is the norm of assertion, then one might say that knowledge is that state which makes assertion non-defective, and the standards of defectiveness of assertion are given by our nature. Or one might try to characterize knowledge as whatever it is that needs to be added to beliefs about explanatory relations in order to yield understanding, with understanding now having a normative characterization as the object of a human's intellect. Given an account of knowledge that goes back to the flourishing of the intellect in this kind of an Aristotelian way, we would have an explanation for why there are so many epicycles in post-Gettier accounts of knowledge. For what we have seen of other areas of the normative life suggests that we should not expect a great deal of simplicity in the norms governing us.

That said, it is not the task of metaepistemology to provide a detailed account of knowledge—that is the task of epistemology—but simply to provide an account of the types of facts that ground knowledge. We might roughly break up these facts into three categories: (i) mental facts constituting one's belief and the inferences that led to it; (ii) the facts (in typical cases extra-mental) making the belief true and entering into the warrant or justification; and, finally, (iii) facts because of which the first two sets of facts count justifying the belief and yielding what additional condition is needed to get out of the Gettier condition. It is only the third class of facts that seems to fall under metaepistemology, and it is plausible that facts from that category are normative facts about the flourishing of the intellect.

12. Epistemic supererogation

In ??backref, we saw that a Natural Law ethic can easily accommodate the concept of supererogation by allowing for a case where one flourishes more in willing to ϕ than in willing to ψ, but nonetheless willing to ψ would not be a case of languishing. This was a special case of a general phenomenon where one state can constitute greater flourishing than another, even though neither is a case of languishing.

Does this phenomenon appear in epistemology as well? ??refs Intuitively, it does. Often, when Dr. Watson fails to see the connections that Sherlock Holmes sees, we cannot say that Watson is epistemically defect—rather, Holmes is, at times, superlatively insightful in weighing the evidence. This distinction, often specific to a particular epistemic area, between insufficient insight, acceptable insight and superb insight is familiar to us, and an Aristotelian metaepistemology simply makes it a part of a general phenomenon. The possibility of epistemic supererogation thus provides some support for the present account.

If the correct epistemology is Bayesian, however, one might think there is no room for supererogation. Either one updates correctly on the evidence or one updates incorrectly. There is, no doubt, a gradation among failures of correct update, but there is only way

to update correctly. Thus, it seems, on a Bayesian epistemology, there is no room for supererogation, and an Aristotelian's ability to account for epistemic supererogation is no advantage.

This is not correct. There are at least three areas within a reasonable Bayesianism where there is room for supererogation. First, if Bayesianism is to be at all a candidate for our epistemology, and assuming some version of an ought-implies-can doctrine, Bayesianism has to accommodate imprecise probabilities, because we humans are only capable of having probabilities that are both precise and rational in very limited situations such as coin flips. Maybe I could make myself have a credence of *exactly* 0.9997 that my copy of *Being and Time* is in my office, but that number wouldn't be rational. But we are required to have some level of precision. I would be failing in rationality if I vaguely judged that the probability of *Being and Time* being in my office is "at least a little more likely than not". If I simply judged it to be "very likely", I would not be irrational. But it could be that a person more excellent in precision could have a finer grained probabilities than my "very likely", whether partly numerical ("somewhere around 0.9997" or maybe "between 0.9997 and 0.9999") or just with more refined non-numerical levels. Such a person's judgment could well count as supererogatorily rational.

Second, we are simply incapable of updating fully on all of our evidence—there is too much of it. When I consider the evidence that it's getting to be time when I will go to the gym, I don't update on the fact that there are smudges on the living room window. Our world is deeply interrelated. There is likely some kind of a correlation. (E.g., smudges may indicate something about the weather, which may indicate something about whether I am going to exercise outdoors or indoors.) But I ignore it. And that's part of being an acceptable human reasoner. Similarly, when Watson ignored the dog's famous failure to bark??ref in his holistic evaluation of the evidence, he wasn't irrational. Holmes, on the other hand, in focusing on this piece of evidence was more than merely rational: he was an excellent reasoner.

Third, and more speculatively, if rational constraints on priors involve ranges of acceptable asisgnments, it could well be that sometimes assignments near the edges of a range are less rational than ones further from the edge. For instance, plausibly, we should have anti-skeptical priors that assign high credence to such things as the uniformity of nature or the reliability of our senses, and our nature requires us to keep these credences between the bounds of two epistemic vices: undue skepticism (setting the priors too low) and excessive credulity (too high). Then someone whose credences are just slightly above undue skepticism or slightly below excessive credulity would be rational, but perhaps only barely so. It would be better to have a more balanced credence, away from the two boundaries of rationality.

Thus, even the Bayesian should be glad for the possibility of accounting for epistemic supererogation.

Appendix: *Approximating the pathological scoring rule with continuous ones

We need to show that the stepwise scoring rule (T, F) from ??backref can be written as a limit of symmetric, strictly proper, finite and continuous scoring rules.

First note that any symmetric continuous proper scoring rule (t, f) can be written as the limit of symmetric continuous strictly proper scoring rules by letting $t_n(x) = t(x) - (1-x)^2/n$ and $f_n(x) = f(x) - x^2/n$, since the Brier scoring rule defined by the functions $-(1-x)^2$ an $-x^2$ is strictly proper, and the sum of a proper and a strictly proper scoring rules is strictly proper.

Thus, all we need to show is that is that we can approximate (T, F) with symmetric, finite and continuous scoring rules. Furthermore, we can drop the symmetry requirement. For write $f^*(x) = f(1-x)$. Then (t, f) is a proper scoring rule if and only if (f^*, t^*) is a proper scoring rule. Now if $T(x) = \lim_n t_n(x)$ for all x and $F(x) = \lim_n f_n(x)$, then

$$(T(x) + F^*(x))/2 = \lim_n (t_n(x) + f_n^*(x))/2$$

and

$$(F(x) + T^*(x))/2 = \lim_n (f_n(x) + t_n^*(x))/2.$$

But $T = F^*$ and $F = T^*$, so the left-hand sides are just $T(x)$ and $F(x)$, respectively. Moreover, $(t_n + f_n^*, f_n + t_n^*)$ is will be a continuous symmetric finite proper scoring rule if (t_n, f_n) is a continuous finite proper scoring rule.

Fix $\varepsilon > 0$. Let ϕ_ε be a continuous non-negative finite function that is zero except on the set $U_\varepsilon = [0.999 - \varepsilon, 0.999) \cup (0.001, 0.001 + \varepsilon]$, and is such that $\int_{0.999-\varepsilon}^{0.999} (1-x)\phi_\varepsilon(x) = 1000$ and $\int_{0.001}^{0.001+\varepsilon} x\phi_\varepsilon(x) = 1000000$. Define

$$T_n(x) = \int_{1/2}^{x} (1-u) phi_{1/n}(u)\, du$$

24 and

$$F_n(x) = \int_{1/2}^{x} u\phi_{1/n}(u)\, du.$$

By ??SchervishThm4.2, (T_n, F_n) is proper. It is clearly continuous and finite. It is, further, easy to calculate that $T_n(x)$ and $F_n(x)$ equal $T(x)$ and $F(x)$ except perhaps on $U_{1/n}$.??check For any x in $[0, 1]$, there is an N such that x is not in $U_{1/n}$ for any $n \geq N$. It follows that $T(x) = T_n(x)$ and $F(x) = F_n(x)$ if $n \geq N$, and we have the limiting condition we wanted.

CHAPTER VI

Mind

1. Multiple realizability

Some conscious beings have brains. Start with the hypothesis that it is a necessary truth that all conscious beings have brains.

First, this hypothesis is just implausible: it seems quite plausible that we could have conscious beings with a very different body plans.

Second, observe that brains are a specific type of organ in DNA-based animals. To have a brain, thus, you need to have DNA. To have DNA, you need to have hydrogen atoms. To have hydrogen atoms, you need to have electrons. A particle with a different electric charge would not be an electron, and the charge of the electron is definable in terms of the fine structure constant $e^2/(2\varepsilon_0 hc)$. If the fine structure constant were different, we wouldn't have electrons. We might have shmelectrons that behave almost exactly like electrons, but they wouldn't be electrons. If we didn't have electrons, we wouldn't have hydrogen, but at best shmydrogen. And if we didn't have hydrogen, we wouldn't have DNA, but at best shmDNA.

But now imagine a world extremely so similar to ours that no instruments of a sort humans ever have a hope of constructing could ever tell the difference, but where, nonetheless, the fine structure constant has a slightly different value. In that world we have beings that behave, as far as any of us could ever tell by external and internal examination, just as we do. But they not only would *be* unconscious zombies, they would *have to be* zombies—no beings with shmelectrons in place of electrons could be conscious on the hypothesis we are considering. That such a slight difference in physical constitution would make the difference is extremely implausible.

Third, if the hypothesis is true, we should be quite surprised at the existence of consciousness. The argument just given shows that consciousness requires the precise value of the fine structure constant that we have. How likely is that? Well, there are infinitely many possible values that agree with our world's fine structure constant to within a thousand significant figures. Unless our fine structure constant turns out to be some very special distinguished value (for a while, some physicists thought it was exactly 1/137??refs, but later measurements disproved that, and a recent estimate is 1/137.03599921), the chances of getting the exact value randomly we have is zero or at best infinitesimal. Given the fact that consciousness has great value significance (??shvalue??), if consciousness depends on brains, and hence on electrons, then the fact of consciousness would loudly cry out for explanation.

The line of thought above is akin to fine-tuning arguments, where narrow ranges of fundamental constants are claimed to be needed for life, and call out for explanation, with two options being typically offered: a multiverse (unlikely things will happen if dice are rolled enough times) and an intelligent designer. But there are some relevant differences in our present case.

First, our range is much narrower—only one exact value is compatible with consciousness on the hypothesis we are exploring—which means that objections from the rescaling of ranges do not apply as they do in the case of the fine-tuning argument.??coarse-stuff

Second, plausibly an intelligent designer would be conscious, and if consciousness requires brains as we are hypothesizing, a designer will be of no help here, on pain of circularity.

Third, because the consciousness-permitting range has only one point on it, and there are uncountably infinitely many possible other values of the fine structure constant, hitting this value will not automatically be probable even given a multiverse. If you spin a continuous fair spinner once, your chance of hitting a particular value is zero or infinitesimal. But the same is true for any finite number of independent spins. Moreover, in classical probability theory, this is also true for a countably infinite number of spins. And for an

uncountably infinite number of spins, the probability is simply undefined. In light of this, the multiverse hypothesis only really solves the problem of consciousness in our context if it is a Lewisian or Tegmarkian hypothesis that *every* possible cosmic arrangement is realized in reality. But such a hypothesis only solves the problem at the expense of introducing serious sceptical problems, since there will be cosmoses, just as real as ours, where every coherent sceptical hypothesis hold, and it does not appear reasonable to think that we got so lucky as to escape them all.??refs

Tying consciousness to brains thus links consciousness to the precise laws of nature we have. That is not only intuitively implausible but leads to serious problems. We should think that there is some flexibility in what kinds of bodies conscious beings can have.

Perhaps instead of supposing that consciousness is tied to brains, we could suppose that consciousness is tied to a range of brain-like organs. Thus, consciousness would be compatible with having somewhat different laws of nature, resulting in fundamental particles slightly different from the ones we have, and behavior somewhat different from the one we have, but not *very* different. But now consider the Mersenne questions about the boundaries of physical constitution compatible with consciousness. These questions cannot be settled by invoking human nature, since they are questions that transcend the nature of any one species. Nor can they be settled the way Mersenne settled his original questions, by invoking God's creative decision, because we are supposing that the connection between consciousness and brain-like organs is necessary. We should avoid Mersenne questions that do not seem to have a plausible answer.

Furthermore, the issue of worlds practically indistinguishable to our instruments but where one has consciousness and the other does not returns on the range view. Suppose that the upper cut-off for the fine structure constant to be compatible with consciousness is 1/100 (recall that our world's fine structure constant is about 1/137). Then either 1/100 is the highest value compatible with consciousness or the lowest value incompatible with consciousness. If it is the highest value compatible with consciousness, there should be a

world w_c with consciousness and fine-structure constant $1/100$ and a world w_z that is practically indistinguishable from w_c but where the fine-structure constant is slightly more than $1/100$ and hence where there are only zombies. If, on the other hand, $1/100$ is the lowest value incompatible with consciousness, then for a value of the fine-structure constant $(1/100) - \varepsilon$ for some positive ε less than one divided by a googolplex there will be a world w_c with consciousness. Then we should expect there to be a possible world w_z with fine-structure constant $1/100$ that is practically indistinguishable from w_c (a difference of one in a googolplex should not affect anything observable), but w_z will be a zombie world, since we have assumed that a fine-structure constant of $1/100$ is incompatible with consciousness. So in either case there will be a world with consciousness and a world with zombies which are physically indistinguishable to humans. But it is implausible that consciousness should depend on physical features that are so insignificant.

This line of thought pushes one to a very liberal view about what kinds of physical constitutions are compatible with consciousness. It does not appear, in particular, that consciousness should depend on having a physical constitution that includes brains or anything similar to brains. We thus have very significant multiple realizability.??check-mr-book

2. Functionalism

2.1. Introduction. Full-blown dualism, of course, yields significant multiple realizability. Indeed, a minded being's body could be an oak tree or even a rock, as long as it had the right kind of non-physical mind on dualism. We will discuss the interaction of dualism with Aristotelian forms in Section 5. In the meanwhile, however, let us continue to consider broadly naturalistic accounts of mind.

We have seen that there is good reason to be very liberal about the type of physical aspect that a minded thing can have. But if we are to remain in a broadly naturalistic theory, we need to put some limits on the kinds of physical constitutions that minds can be based on. We saw earlier that limits based on particular natural kinds—DNA, brains,

electrons, etc.—are highly implausible. The most plausible remaining option is functionalism: to have a mind is to have a certain kind of functional structure, so that, necessarily, if there is a functional isomorphism between two entities with their respective functionally-specifiable causal histories, if one of these entities has a mind, so does the other. Moreover, the isomorphism between causal histories implies a significant degree of identity between the mental histories.

On what we may call strong functionalism, their purely internal mental histories will be the same, and in particular they will have qualitatively the same states in their histories—whenever one felt hot, so did the other, and whenever one had a perception as of red, so did the other. The restriction to purely internal mental histories allows for some externalism. Thus, an individual on Earth may be thinking about water, while the analogous thought in an isomorphic individual on Twin Earth, may be thinking about XYZ, where XYZ fulfills the same causal role on Twin Earth as H_2O does on Earth.

On weak functionalism, the non-qualitative purely internal mental histories will be the same, and whenever one has a conscious state, the other has an analogous conscious state, but the exact qualitative phenomenal character of the conscious states may depend on the precise physical substrate underlying the two conscious states. On weak functionalism, a silicon-based isomorph of a human being, will have some sensation in a functional state isomorphic to a human's eating sugar, but that sensation's qualitative character may be different from the taste of sweet.

I will now argue that functionalism, whether weak or strong, has serious problems which can be solved by combining it with an Aristotelian hylomorphism.[1]

2.2. Interpretation. Begin with the well-known observation that simple causal systems, like the electrons buzzing inside a rock, can be re-interpreted as emulating the functioning of our brains, simply because they have such a vast number of states.??refs If this

[1]The arguments based on the possibility of malfunction will be based on the ones in ??ref:Koons-Pruss.

is right, then functionalism appears to lead to the absurd thesis that rocks not only think, but think like we do. One version of this argument will be given in ??forward:appendix.

One might try to get out of this difficulty by insisting that gerrymandered functional systems do not count: only simple causal systems count as implementing the functions. However, it is very likely that complex evolved brains like ours do have some significantly gerrymandered states. One might try to draw a distinction between more and less gerrymandered systems, however. The functional states that need to be attributed to a rock to re-interpret it as thinking our thoughts are doubtless many orders of magnitude more gerrymandered than our functional states. But now we have a nasty Mersenne question again: what makes it be the case that the transition between the degrees of gerrymandering compatible with having a mind those incompatible with having a mind lies where it does?

2.3. Reliability. Next, consider the question of the reliability of functional systems. Whether our universe is deterministic or not, functional systems are imperfectly reliable. How reliable do they need to be to count as the functional systems they are? Consider a subsystem that given two inputs representing numbers puts out their sum 99.9% of the time, and 0.1% of the time puts out the product. Obviously, it is more reasonable to interpret it as a reliable addition system than an extremely unreliable multiplication system. But suppose the system puts out sums 50.1% of the time and products 49.9% of the time. What then?

There are two natural cut-offs. We could require that a system is defined by how it behaves 100% of the time or by how it behaves more than half of the time. Requiring perfect functioning would have the empirically false consequence that humans don't think. A 50% cut-off, however, may be problematically low. If every subsystem of a complex functional system had a 49% failure rate, then the typical outputs of the system would be largely random, because any output of the system is the result of a causal chain of many subsystem states, and any such chain would likely contain multiple failures. Moreover, while 50% reliability seems to be a natural and well-defined cut-off, the actual reliability

of a system cannot be captured by a single number, if only because reliability depends on environmental conditions. If we say that a system is an adder provided that the output is the sum of the inputs more than half of the time, we have to specify the temperature, background radiation, and other conditions under which that reliability is to be defined. And now we lose the neat elegance of specifying the reliability with the single number 1/2.

We might try to define the reliability with respect to the actual environmental conditions the system was in. But suppose that Alice finds herself in extremely harmful conditions—say, great heat or toxic fumes—but by a fluke survives with what is intuitively full brain function for a few seconds longer than we would expect, screaming seemingly in pain. Given that under those extreme conditions the functioning is extremely unreliable, we would have to say that Alice is in fact a mindless zombie and feels no pain. This is implausible.

Additionally, note that the reliability of a system varies over time, perhaps increasing over an initial burn-in period, and then eventually decreasing. We can then define the reliability of a system instantaneously or via a time-average. If we proceed via a time-average, then we will have the highly counterintuitive consequence that an individual who died at a hundred was actually conscious through their life, but had they lived a decade longer, they would *never even have been* conscious, because some crucial subsystem's reliability average over a 110-year lifespan would have been below the cut-off, but over the hundred-year lifespan was above the cut-off. So we should define the reliability instantaneously.

In any case, the very idea of a sharp cut-off in reliability for mindedness seems counterintuitive. Imagine two humanoids whose brains are nearly identical, with the exception that one brain structure crucial for consciousness in one of the humanoids has 50.000001% reliability and the other has merely 49.999999% reliability. Suppose that both brains *in fact* function exactly the same way, so that the sequences of internal states are exactly the same—it's just that one is slightly more likely to fail than the other. It does not seem very plausible to think that one would be conscious and the other not.

Finally observe that a functional system can retain its function even when highly defective and unreliable. A car that starts only on one of three mornings is still a car.

2.4. Damage. Plausibly, a functionalist account of (nociceptive) pain will include two essential aspects of its causal role: the pain is caused by bodily damage and the pain causes a tendency to aversive behavior.[2] But imagine that the brain pathway leading to the tendency towards aversive behavior is broken. David Lewis??ref thought that in such cases the subject would still feel pain. The phenomenon of pain asymbolia where pain is felt without aversive behavior—and apparently even without unpleasantness!—provides empirical evidence for this plausible hypothesis.??ref Similarly, if a patient's nerves are severed short of the tissue whose damage they are supposed to signal, surely the patient can feel pain originating in the nerves themselves ("neuropathic pain"??refs) rather than in damaged tissue—the case of phantom limb pain may be like that.?? Presumably, if *all* of an unfortunate patient's nerves were so severed, the patient would still be able to feel pain given stimulation of the nerves themselves, even though the patient's pain would no longer actively play a part in a damage-detection system.

To account for this, it seems we need to define the functions of functionalist theories in terms of normal or typical behavior. David Lewis's own proposal is population-based: the function of a state is defined by what the state does in the relevant population.??ref This faces the notorious reference class problem: should we take the relevant population to be the biological species or the more narrow subgroup of people who have the particular condition that disrupts the causal network? If we take the relevant population to be the species, then we can make the empirically supported judgments about pain asymbolia and the intuitive judgment about total severing of nerves, namely that there is pain present in these cases. If we took the narrower population as the reference class, we would have to say that there is no pain in these cases.

[2]This may be necessary, but is unlikely to be sufficient for the causal role of pain, or else a computer that monitors the state of a storage drive and moves data from damaged to undamaged portions feels pain when it detects damage and acts on the detection by averting putting data in the damaged location.

So suppose we take the species to be the right reference class. But now imagine that a functionalist tyrant has all pain-signaling nerves severed, but does not wish to feel any pain originating from these nerves, and hence kills every organism that does not share her unfortunate condition. Then there seems to be no reference class according to which her states homologous with pain have the causal role of pain, and so it seems that on Lewisian functionalism, the genocidal tyrant's plan has succeeded—by killing others, she has killed her pain. But it is obviously false that one can end one's pain by killing others.

One might object that the relevant reference class should be counted diachronically, and so the many generations of past humans outnumber the survivors of the tyrant's genocidal plan, and hence ensure that the state still has the functional role of pain. But on such a diachronic interpretation, we need only change the story. Move the story into the realm of science fiction, and suppose that the tyrant institutes a massive cloning program which within a few years produces more human beings—living doubtless in appalling conditions—who outnumber all the humans who lived previously, and the tyrant ensures that each such human's nerves are severed just as her own are. Now on a diachronic species-based account, the function of the tyrant's states has changed, and hence the tyrant ceases to be capable of feeling pain. This, too, seems patently false.

An alternative to population-based approaches is evolutionary approaches. The function of a trait is the causal role that it evolved for—the causal role such that its historical fulfillment led to the present existence of organisms with that trait. There are many details to fill out to make this a tenable account, but they won't matter for the critique that I will offer, based on Koons and Pruss??ref.

Imagine a world w_1 where super-powerful aliens secretly snatch each organism from earth just before it dies and transport it in some form of stasis to another galaxy, while replacing it with a fake corpse. There they cure whatever was ailing the organism and ensure that the organism succeeds in reproducing many times over. All the descendants of that organism are in the care of the aliens, who ensure that no genetic line goes extinct—that every organism in the care of the aliens succeeds in reproducing with its offspring also

being taken care of. Moreover, in w_1 life on earth is an exact duplicate of how it is in the actual world, w_0—nobody ever catches on to idea that the organisms we have get snatched before death.

Without extinction, with every organism successfully reproducing, there is no evolution—just exponential population growth. It is essential to an evolutionary explanation of a phenotype that the phenotype contributes to the the organism's probability of passing on its genotype. In w_1, there are no evolutionary explanations of any earth organisms, and hence on a functionalism where function is defined evolutionarily, no organisms have minds. This already seems somewhat counterintuitive given physicalism: the physical causal history of every actual organism on earth is the same in w_1 as in w_0, and yet there are no minds in w_1.

To make the counterintuitive aspect of this story even more problematic, now imagine a world w_2 which is just like w_1, except that once the organisms arrive in the other galaxy, they are immediately slaughtered. In w_2, we do have evolution just as in w_0, the only difference being the exact location of death. So in w_2, on evolutionary functionalism, there are minded beings on earth. But it is highly counterintuitive that whether the aliens choose to slaughter the abducted organisms once they take them to another galaxy affects whether earth primates are conscious.

One might object that in w_1 there is still natural selection, but one centered around location rather than existence. While the reason there are giraffes has nothing to do with the selective benefits of their necks in reaching food—without the necks, they might come close to dying of hunger but would still reproduce in the second galaxy—the benefits of their neck explain why there are giraffes on earth. But locational selection is highly implausible as an account of the function underlying minds. Imagine a world where there is a galaxy full of planets with randomly generated ecosystems, all of them in their first generation of macrofauna, some of which are phenotypically just like us. Suppose next that aliens pick out those macrofauna whose brains function in a way congenial to the aliens'

plans, these macrofauna including organisms physiologically just like ours, with the reason for their selection by the aliens being that the brains function in the way required for mindedness. The aliens then take all these macrofauna and transport them to a new planet. On a locational selection version of evolutionary functionalism, it seems that as soon as the macrofauna are all in a new location because of the mental-like functioning of their brains, they become conscious. But it is absurd to suppose that organisms can become conscious simply due to a change of location.

Granted, on some evolutionary accounts of selection,??ref selected function is only exhibited in the generation subsequent to the selection. If so, then only the immediate children of the transported macrofauna have minds. But that is only a little less absurd: it is surely not the case that simply a locational change in the parents makes the children conscious. Moreover, the aliens need not even move the selected-for organisms. They could just move the unselected-for organisms.

Besides, it seems *ad hoc* to allow locational selection in addition to the much more natural notion of Darwinian existential selection. Once we allow locational selection, we should probably allow selection in terms of properties other than location. For instance, the aliens could paint a red circle on the head of those organisms they like, and set up machines that put red circles on the descendants of any organisms with red circles. We would then have red-circle selection, and if we allowed that to count for defining function, then we get the absurdity that simply by painting red circles on some animals one can make them, or their descendants, conscious.

??locational selection

2.5. Many functions. It is not enough that a subsystem always outputs the sum of its inputs for it to be an adder. After all, if the only inputs ever given are pairs of zeroes, then the subsystem could just as well be a multiplier. As is well-known from Wittgenstein and Kripke??refs, no finite amount of data is sufficient to determine the function of the system, since any finite collection of inputs and outputs is consistent with infinitely many possible functions—admittedly, perhaps messy ones.

We might try to define the function of a system or subsystem in terms of counterfactuals. Perhaps an adder is something that *would* output the sum for all inputs in the range of allowable inputs. However, Frankfurt's counterexamples to the Principle of Alternate Possibilities??ref can be adapted to show that this is untenable. Imagine that Bob counts as thinking that $5 + 7 = 12$ in virtue of the fact that his thought involves the operation of an adding subsystem in his brain. But suppose that a neuroscientist has placed a neural scanner and bomb in Bob's vicinity in such a way that if the scanner detects the adding subsystem getting any input other than 5 and 7, the bomb blows up Bob. Now counterfactuals like "If the inputs were 4 and 3, the output would be 7" are false. If the inputs were 4 and 7, there would be no output, just an explosion.

One might try to define functions by asking what the output would be given the inputs *absent external interference*. But what counts as interference with a system, as opposed to, say, a helpful or neutral effect, depends precisely on the system's function. If the purpose of a wood-pulp product is to preserve inscribed information, then fire is an external interference; but if the wood-pulp product's function is to be kindling, then fire activates its the object's function. Furthermore, one can internalize Frankfurt-like cases. Imagine that through science-fictional genetic modification Bob's liver comes to behave just like the scanner-and-bomb system.

2.6. A neo-Aristotelian solution. If functionalists do not have the resources to define functionalism, then this is presumably the most fundamental possible flaw in functionalism. However, the problem of defining functions is exactly one of the one that Aristotelian forms are designed to solve. The proper function of a subsystem in an organism is that the fulfillment of which constitutes the organism's flourishing with respect to that subsystem. This does not require the subsystem to be reliable, though Aristotelian optimism predicts that most systems will be reliable.

A robust view of organisms as having forms that specify normative, and hence functional, features thus solves a central problem with functionalism. And as we saw, there is very good reason for a naturalist to adopt functionalism—it is the best naturalist option for

saving multiple realizability. Thus, we have an argument from naturalism to Aristotelianism. Of course, whether Aristotelianism is compatible with naturalism is not clear. If naturalism is understood to say that there are no fundamental normative properties, then of course there is no compatibility. ???refs

2.7. A spectrum. There is a spectrum of functionalisms, depending on the level at which we have the functions. Think here of the analogy of a computer. We could suppose that proper function is specified at a fairly low level: say, the level of components like transistors, which are specified by their datasheet—the transistor is designed to function in this range of temperatures, voltages, and currents, saturates at this level, has this current gain, has this transition frequency, etc.—and by having the correct data bits in the memory units. On this picture, for mental function it doesn't matter how the components are implemented, say at the level of doped semiconductor materials. Or, at higher level, we could specify the CPU as a whole, rather than its transistors, defining the syntax and semantics of the machine code that it runs (with or without some abstraction, say with respect to edge cases), and then specifying what machine code and data are loaded. Or at a yet higher level we could specify the details of the algorithms, say in a high level programming language or even pseudo-code, for instance specifying the precise sorting algorithm that is being used for operationalizing preference rankings in an autonomous robot. Or we could abstract from details of the algorithms, and specify what the computer accomplishes at a very high level, e.g., simulate n-body Newtonian gravitational interaction, or translate spoken Cantonese into written ecclesiastical Latin. A similar hierarchy can be imagined in the case of minds implemented wholly or partly by brains, though we know much less about the details of how high level function is implemented by our neurons, so it is easier to illustrate the point with computers.

There is thus an important degree of freedom in functionalist theories of mind with regard to the level at which the mind-constituting functions are specifying. Some levels are more plausible than others. The lower the level of specification, the less there will be of multiple realizability. In fact, in the computer analogy, it seems that even the fairly high

level at which details of algorithm choice are specified is too high. Surely an elephant and a dog could equally exhibit preference-driven agency, with some efficiency differences, if one employs a merge sort and the other a bubble sort in generating its preference ordering. Furthermore, the lower the level of specification, the more Mersenne problems we raise, because the more degrees of freedom we have when we require the specification of detail.

There is thus reason to prefer the highest levels of specification. And Neo-Aristotelianism can rise to that challenge, because forms can specify proper function at an arbitrarily high level of abstraction.

3. Supervaluationism about minds

A tempting solution to the problem of the multiplicity of functionalist (or other) theories of mind differing with respect to fine details is to treat each theory as a precisification of concepts like *mind* or *pain*, none of which is privileged over the others. And, if all has gone well, then typical adult humans fall under all the precisifications of "has a mind" and in paradigmatic cases of being in pain they fall under all the precisifications of "is in pain".

One difficulty with this approach has already been discussed in the ethical?? context??backref, namely that similar problems arise at the meta level: What *range* of, say, standards of reliability yields a functionalist concept that is in fact a precification of our concept of mind?

Furthermore, recall the argument??backref that the centrality and overridingness of ethical norms to our lives makes it deeply implausible to think that there is a plurality of closely related concepts, none of which is privileged over the others. But given the deep importance of the mental to ethics, the same concern applies to the mental. That an action causes severe pain to a non-consenting individual with no significant benefit is a conclusive moral reason not to perform the action. But if there are many concepts of pain with none of them privileged, and similarly of consent, then whether one has a conclusive moral reason for an action will depend on the choice of precisification as well. Hence,

whether one has a conclusive moral reason becomes a verbal question in such cases, and that is not compatible with the force of moral reasons.

There is also something counterintuitive about the implications of this kind of vagueness about mind. There is some plausibility in thinking that there can be vagueness about the exact phenomenal character of a mental state. But vagueness about whether there is a phenomenal character at all seems more problematic. For instance, maybe it can be vaguely true that someone is in severe pain, but in such a case it is definitely true that they are at least experiencing a discomfort. This inference, however, is invalid on the precisification approach to solving the problems with functionalism. Imagine a perfectly reliable computational system S_0 which would definitely implement severe pain—a system that, say, emulates what happens in a human brain when one has one's leg amputated without anaesthesia. But then imagine a continuum of systems that in the actual world behave just like S_0, but are less and less reliable. At some point at this continuum we simply have completely random behavior that by chance matches that of S_0. Somewhere in between it is vague whether the system has the same functions as S_0. But we can imagine that everywhere along the continuum it is definitely true that *if* the system is conscious of any discomfort at all, it is conscious of a severe pain. Thus in the vagueness region, it is vague whether the system is experiencing severe pain, *and* whether it is experiencing even any discomfort. This kind of state of vagueness about whether there is severe pain and whether there is discomfort appears impossible.

To make the counterintuitive character of this kind of vagueness clearer, imagine that you start as a conscious human that is continually observing the visual situation around you, a field of colorful wildflowers. But then your neural system gradually becomes less and less reliable, but by luck nothing goes wrong—everything actually behaves as if all the reliability were in place. Moreover, all your subsystems are always equally reliable or unreliable, I suppose. Eventually, according to functionalism, mental states stop, and they all stop together, because the level of reliability is the same for all your subsystems. On the vagueness solution currently under consideration, at any given time in the process it is

definite that you are either mindless or are having normal human perceptions of a field of wildflowers, but it eventually becomes vague which of the two it is. We are apt to imagine this process as a one involving fading mental states—the wildflowers get less colorful, your thoughts get sluggish—but that's incorrect. It is definitely true that if you are having any mental function at all, you are seeing the wildflowers with all their vibrance and are thinking as well as ever, and hence it is definitely true that the wildflowers are not getting less colorful and your thoughts are not deteriorating. Yet, allegedly, at a vague point it all transitions from full vividness to nothing. We can, of course, imagine a transition from full vividness to nothing with no in-between states—a sudden loss of consciousness, as when you receive a massive head injury. But what seems absurd is the idea of a sudden transition from full vividness to nothing with no in-between states that happens at a vague time.

??ref:fading-qualia?

Additionally, the kind of vagueness involved here could generate a situation where it is vague whether something is a person, and hence has the kind of dignity or even infinite value??ref that persons are often thought to have, or whether it is a mindless thing—with its being definite that it is nothing in between. For suppose it is vague whether a certain degree of statistical reliability is enough for proper function, and then suppose that all of the computational systems of an entity have that degree of reliability, so that it is definite that if that degree of reliability is sufficient for function, the system is a person, and if it is insufficient, the system completely lacks computational function, and hence is mindless. One might hold that it is possible to have a system where it is vague whether it has enough mentality to have the value of a person, but our best candidates for imagining such a system are ones that have significant mental functioning, but it is vague whether that amount is sufficient for personhood. But the idea that it is definite that it is either a person or mindless, but vague which one, seems highly counterintuitive.

4. Substances

The functions necessary for human cognition appear to be housed in each of a multiplicity of entities, including the whole human organism, the human organism less hair, the head, the brain, and perhaps just the upper brain. Standard functionalism seems to imply that all of these entities think the same thoughts. But this has some very implausible consequences. If all of the above entities think, then by the same token for any subset S of the set of my hairs, there is an entity A_S that contains all of my organism less hair plus precisely the hairs in S, and on standard functionalism it seems that A_S thinks exactly as I do, for all S. The number of A_S entities is equal to 2^N where N is the number of hairs that I have. If I pull out one of my hairs, half of these entities perish. Since I have hundreds of thousands of hairs, the loss of conscious beings is vast—more than 2^{100000} beings. Moreover, pain had by a more hirsute person is shared by a much greater number of entities, all other things being equal, so it follows that other things being equal (such as the mass of the non-hair parts), given a choice between relieving the pain of a bald and a non-bald person, absurdly we should help the non-bald person. A similar argument applied to atoms instead of hairs shows that more massive people should typically have vastly greater preference in our moral calculus, and that weight-loss is akin to murder.

We might try to escape these ethical worries by insisting that what counts for moral purposes are not the tokens of thought but the types. This too is counterintuitive. Suppose disease A is had by a million people that causes a severe pain, and careful functional analysis shows that every sufferer from the disease feels *exactly* the same pain. On the other hand, disease B is had by a two people, and causes in each of them a different qualitatively unique pain but equal in severity to that of the sufferers from A. If we count types of thought, then it seems that it is intrinsically twice as good to distribute pain killers to the B sufferers than to the A sufferers. And that seems quite wrong.

Perhaps, though, what we should count for moral purposes are not types of thought, types of lifetime streams of thought. Even though each A-sufferer has the same pain, that pain is embedded in a different type of lifetime of thought. On the other hand, the

thoughts of each of the A_S entities are embedded int he same type of lifetime of thought. The lifetime stream type approach, however, is not all that plausible. I think most of us will have the intuition that if disease A is found among a million mentally duplicate twins, it's still better to relieve the pain of A than of B. Furthermore, the lifetime stream approach implies that how much moral consideration you now deserve can depend on what *will* happen to you, since what type of lifetime stream of thought you have depends on what will happen in the future.

Alternately, we might modify the functionalism to rule out all but one of the nearly-colocated entities as candidates for thinking. The most natural approach here is to say that what thinks a thought is a minimal or a maximal system exhibiting the relevant type of functioning. Neither account seems tenable. The maximal approach seems particulary absurd: my thoughts aren't being had by some giant system that includes my brain and the Orion Nebula. Probably the only defensible version of the maximal approach is a holism on which it is the universe as a whole that thinks—but then the moral counting problems reappear, because all the thoughts are had by the same entity, and so there does not seem to be a reason to relieve the pain of disease A rather than of B.

The minimal approach is problematic in a different way. There need not be a unique minimal functional system. Consider a bridge made of planks, whose function is to allow a one ton load to be rolled over the bridge. It might turn out that there is some unique subset S of n planks such that the planks in S would suffice to bear the load, and such that any other subset sufficient to carry the load would have $n + 1$ planks. But this also need not be true. It could well be that the bridge is such that there are two or more different subsets with a minimal number of planks that can carry the load.

The question of how many minimal functional systems there are in a given human organism is a potentially interesting empirical question, but it is intuitively irrelevant to the question of how much moral consideration should be given to this human's suffering. Furthermore, there seems to be a rather arbitrary decision to be made as to how one defines the minimality of a system. One might try to do so by minimalizing the number of

particles, the number of atoms, the number of molecules, or the mass.[3] All of these could result in different determinations.

The Aristotelian, of course, has a natural solution: there are only as many substances as substantial forms, and substantial forms unify some but not all arrangements of material parts. Something like this can be said, however, by anyone who believes in non-vague restricted composition, i.e., that only some precise arrangements of material parts make up a whole, for instance a variant of van Inwagen's theory those only arrangements of multiple material parts that are caught up in a life make up a whole but without van Inwagen's belief in the vagueness of life. However, the Aristotelian has a further advantage over other restricted compositionalists with respect to functionalism. It is not a coincidence, on the Aristotelian view, that those systems of material parts that have a function also have a substantial form, because the substantial form defines function. On a non-Aristotelian restricted compositionalism, on the other hand, where function is not defined in terms of form but, say, in terms of actual causal interactions or population statistics(??backref,add?,Lewis), both brains and organisms will be apt to have the kinds of functions that define minds, and yet only organisms really exist—are real composites of parts. ?????so what?

5. Dualism

The neo-Aristotelian teleological twist on functionalism allows one to remain to a significant degree a naturalist (more on that in ??forward). The account reduces mental properties to functional properties, but the functional properties are not reducible to the kinds of properties that physics studies. A more traditional Aristotelian approach, however, is to refuse to reduce the reduction of mental properties to non-mental ones. This is certainly also compatible with the robust view of human nature that has been defended in this book.

[3]The last is a little tricky, because we learn from Einstein that energy is a component in mass, and the mass of a part then depends on the energy it has as a part of a larger whole.

Whether the more dualist or the more functionalist view is more plausible depends on how satisfying the reduction of mental properties to teleologically laden functionalist properties is, and one way to evaluate this reduction is to consider whether the arguments against the reduction of mental to physical properties apply to this neo-Aristotelian functionalism. These arguments can be divided into three categories, depending on which aspect of mental life they object to the reduction of: content, consciousness, or freedom of will.

As has been noted??backref, the neo-Aristotelian teleology offers hope for a reductive account of mental content. Teleological properties can be hyperintensional—it is the purpose of eyes to see rather than to see or be such that $2 + 2 = 5$, even though necessarily everything that sees is such that it sees or is such that $2 + 2 = 5$. Thus concerns that there is no way to reach the hyperintensionality of mental content from the extensional or at best intensional content of the physical world do not transfer to the neo-Aristotelian account. Similarly, concerns about the need to account for purpose in the mind—beliefs are states that are there *in order to* mirror the world—and that the only source of purpose for the physicalist, namely evolution, is inadequate do not apply to the neo-Aristotelian theory. Thus, content-based arguments do not tell against a neo-Aristotelian functionalism.

The case of consciousness is less clear. We can try to run standard knowledge arguments transposed to the neo-Aristotelian context. Suppose Mary is raised in a black and white environment, and knows all the physical facts as well as all the normative facts about human beings. In particular, she knows that certain states of the human being are such that they should occur precisely when the human being is looking at a red object, and that these states typically occur when people see a ripe tomato. She, further, knows all the normative interconnections between these states and other states, including all the inferential connections. Will she learn anything further by seeing the tomato? Here intuitions may differ. It is at least easier to hold out for a negative answer to the learning question here than in the case where Mary has a purely physical state.

Similarly, we can try for imaginability arguments. These come in two versions: zombie arguments that one can have our physical constitution without consciousness and afterlife arguments that our consciousness could survive the destruction of the body. On these arguments, the imaginability of a scenario provides defeasible evidence of its metaphysical possibility.

Could there be beings that have isomorphic physical and normative properties to ours but that are unconscious zombies? Again, this is not completely clear. Among our normative properties are moral duties. Could one have something isomorphic to moral duties without consciousness? If not, then zombie arguments against neo-Aristotelian functionalism do not work as they stand.

Alternately, what can the neo-Aristotelian functionalist say about surviving the destruction of the body? Those "fainthearted" neo-Aristotelians who hold that the form is nothing but a kind of arrangement of matter, of course, will find it troubling to suppose the form to survive the destruction of the matter. Though even there, there is a potential precedent for a view that would allow the form to survive. Aquinas's account of transsubstantiation holds that in the Eucharist, the accidents of bread and wine continue to exist after the cessation of the existence of the substance. Among the accidents there will be *shape*. If a shape can exist without that of which it is the shape, then why not an arrangement as well? Similarly, some contemporary trope theories hold that there are "unaffiliated" tropes, tropes that exist without their substance. If a trope can exist without a substance, why can't an arrangement exist without matter?

But of course the view defended in the book is more robust: forms are not mere arrangements. The more reality the form itself has, the more plausible that there is no metaphysical impossibility about the form existing without the matter.

One may, however, wonder whether *we* could continue to exist without a body. ???????

Finally, we have freedom of will. Those convinced that an action that comes solely from the causal powers posited by a completed physics??ref would not be free will not be moved by being told that these causal powers have a normative organization. Such

??Doesn't functionalism have the same problems?

??vagueness of consciousness

6. Teleology and representation

7. Teleology and mental causation

8. Soul and body ethics

Appendix: Functionalism gone too far

Consider a deterministic functional system Q consisting of a finite number of possible computational subsystems $S_1, ..., S_N$, where S_k is always in exactly one state from the finite set $\mathcal{A}_k$ of possible states at each of the discrete "significant" moments of time $t_1, ..., t_m$ over its finite lifetime[4], and a finite number $I_1, ..., I_M$ of sensors, where I_k is always in exactly one state from the finite set $\mathcal{I}_k$ of possible input states.

We can think of Q as a finite digital computer. The total state of the system at any given time can be represented as the $(N+M)$-tuple $(a_1, ..., a_N, b_1, ..., b_M)$ where a_k is a state in $\mathcal{A}_k$ and b_k is a state in $\mathcal{I}_k$. We can designate some of the subsystems as outputs, connected to external effectors (muscles, motors, lights, etc.) Furthermore, we suppose there are functional laws which provide a deterministic mapping f from the total state of the system at time t_i to the total state at time t_{i+1}. Thus, $f(a_1, ..., a_N, b_1, ..., B_M) = (a_1, ..., a_N, b_1, ..., b_M)$ provided that the system would transition from state $(a_1, ..., a_N)$ to state $(b_1, ..., b_N)$.[5] Presumably any deterministic analog system can be approximated by such a system Q to aribitrarily high precision. We suppose for convenience that there is some fixed initial state of Q, with fixed initial input and computational states.

[4] A standard modern digital computer only has defined computational states at ticks of its internal clock. Between ticks of the internal clocks, it is in an analogue state that is not computationally defined. A lot of careful engineering goes into ensuring that the states become properly "digital" at the clock ticks.

[5] The mapping is independent of time. But we can always suppose there is a finite clock, e.g., given by a subsystem S_k such that the set of states $\mathcal{A}_k$ is the set of times during the system's lifetime, and with the transition rule that the time always gets incremented. Or we can suppose an input from an external clock.

Now consider a different system P consisting of a single particle moving in the xy-plane in space, with a constant (sublight) non-zero x-velocity v, starting at x-coordinate 0 at time t_1. For ease of visualization, suppose the x-axis runs left to right, and the y-axis runs down to up. Let $\mathcal{J}$ be the set of all possible single-time input state vectors $(b_1, ..., b_M)$ where $b_k \in \mathcal{I}_k$ for each k, and let K be a positive integer such that $\mathcal{J}$ has at most 10^K members.

We now recode Q's inputs into P's inputs as follows. Suppose our particle is at y-coordinate 0 at a time $t_0 < t_1$. For $1 \leq n \leq N$, let ψ_n be a one-to-one function from $\mathcal{J}$ to integers between 0 and $10^K - 1$, both inclusive. The analogue to inputting the sensor state vector $(b_1, ..., b_M)$ into Q at significant time t_n will now be this. We take the y-coordinate value y_{n-1} at t_{n-1}, and add to it the value $10^{-Kn}\psi_n(b_1, ..., b_M)$, shifting the particle upward along the y-axis as needed to ensure it reaches that by time t_n.[6] Thus, the first set of inputs of Q will be encoded in the first K digits after the decimal point, the second set of inputs will be encoded in the second K digits, and so on.

Next, suppose Q's computational states begin with the fixed initial state vector $(a_{1,1}, ..., a_{1,N})$ at time t_1. Given the determinism of the system, there is a mathematical function f that takes a sequence s of length n of members of $\mathcal{I}$ and returns the computational state $f(s)$ that Q would be in at time t_n if it were to have received the sequence of inputs s at the times $t_1, ..., t_n$. Let $n(x)$ be the integer n such that $x = vt_n$, if there such an integer, where we recall that v is the velocity of our particle along the x-axis. Let $s(x, y)$ be the sequence of $n(x)$ sensor state vectors encoded by the y-coordinate value y, assuming $n(x)$ is defined. Thus, the first member of $s(x, y)$ will be the sensor state vector s_1 such that $\psi_1(s_1)$ is equal to the decimal number given by the first K digits after the decimal point in y, and so on. We now deem P to be in a computational state corresponding to $f(s(x, y))$ when P is at (x, y). This is defined at all the significant times $t_1, ..., t_n$. We have, thus, an isomorphism between the computational states and sensor inputs of Q and P.

[6] We may need to ensure the units in which these are measured are such that the particle can be shifted in the requisite time without exceeding the speed of light.

We now go for one final twist. Suppose that in the actual world, the system Q gets the sensor input vectors $s_1, ..., s_N$ at times $t_1, ..., t_N$, respectively. We can now choose the function ψ_n such that $\psi_n(s_n) = 0$ for all i. Then a single particle moving with constant velocity v along the x-axis, with y-coordinate always equal to 0 is an isomorph to the actual functioning of Q. But the very same particle will be an isomorph of any other system Q' with the same significant times. Thus, the particle will think your thoughts *and* my thoughts!

??do we need a clock?

CHAPTER VII

Semantics

1. Communication and norms

1.1. A problem about cooperation. ??cut:squeaking is better?

There are scenarios, such as the Prisoner's Dilemma or the Tragedy of the Commons??Refs, where it is difficult to see how to rationally secure cooperation between agents. The following should not be one of these. You have two agents who will each in a separate booth choose whether to press a red button or a blue button. If they both press the same button, they each get a reward, say a chocolate bar. If they press different buttons, they each get a penalty, say a nasty electric shock, with the penalty outweighing the award by a significant factor, so it's better to get neither than to get both. If either player omits to press a button, neither gets anything, and the buttons are so set up that one cannot press both. Moreover, the players are allowed to confer ahead of time.

Obviously, when conferring ahead of time, they will need to decide which button to press, by rolling a fair die or flipping a fair coin if necessary, and then they need to go into their booths and press that button. Neither has any incentive to defect to pressing the other button, and there is no risk in pressing a button should the other defect fail to press anything. This is a really easy win-win game.

But now suppose our two players, Alice and Bob, are perfect expected utility maximizers who break ties with fair coinflips, and the only relevant utilities are the rewards and penalties of the game. There are no further games that will be played. Nobody outside the game is in any way affected by the results (e.g., nobody will be disappointed if one of them breaks a promise). And because each player gets the same payoff, it won't matter whether Alice and Bob maximize collective utility or their own personal utility. Finally, the above

information is completely luminous to both players. I claim that at this point the obvious strategy—to decide on a button and then both press it—is no longer rationally available.

For concreteness, let's suppose that Alice and Bob have agreed to press the red button. They go into their booths. What will Alice do? She is a perfect expected utility maximizer. She will only press the red button if the expected utility of doing so is at least as big as that of all the alternatives (these being being pressing the blue button or pressing no button). Now the expected utility of pressing the red button is only going to be at least as big as the expected utility of pressing neither button if Alice takes it to be significantly more likely that Bob will press red the button than that Bob will press the blue button.[1]

But why should Alice take it to be significantly more likes that Bob presses the red button than the blue button? Ordinary human beings take themselves to be beholden to norms of promise-keeping, and tend to abide by those norms, especially when there is no obvious benefit to failing to do so. But Bob is a pure expected utility maximizer. Whatever normative force he takes promises to have has to be derivable from the norm of expected utility maximization. In ordinary contexts, dealing with ordinary human beings, keeping promises certainly does maximize utility, because ordinary human beings believe in norms of promise-keeping and punish those who break those norms (if only by castigating or refusing to enter on joint projects with promise-breakers). But Bob is not dealing with an ordinary human being. He is dealing with an expected utility maximizer.

Here is one way to see the difficulty. Imagine that Alice and Bob are perfect utility maximizers with a perverse value theory that in addition to common-sensical value assignments to chocolate bars and electric shocks assigns non-instrumental negative value to keeping promises and non-instrumental positive value to doing the very opposite of

[1]Suppose Alice maximizes only her own utility (if she maximizes collective utility, just double all the utilities). Suppose $x > 0$ is the reward and $-y < 0$ is the penalty with y bigger than x by a significant factor. Then the expected utility in pressing the red button will be $\alpha x - \beta y$. Alice will only press the button if $\alpha x - \beta y \geq 0$, i.e., if $\alpha/\beta \geq y/x$. Since y is bigger than x by a significant factor, this requires α to be bigger than β by a significant factor.

what one has promised. I will stipulate that pressing the button of the other color counts as the "very opposite". In this case, if there is joint knowledge of the perverse value theory, it is more reasonable to expect Alice and Bob to press the button other than the one they promised. And now imagine that they are perfect utility maximizers with a value theory that assigns positive value to keeping promises and non-instrumental negative value to doing the opposite. In this case, it will be more reasonable to expect Alice and Bob to press the button they promised. But then the in-between case, where Alice and Bob are perfect utility maximizers and assign zero value to promise-keeping and promise-breaking, should be one where the probabilities of pressing the promised button and pressing the opposite button are equal.

What if we suppose that the solution here is that as a matter of contingent fact people have a preference for promise-keeping over promise-breaking: we feel bad when we break promises and good when we have fulfilled them. Preferences enter into utilities, and so if Alice and Bob have the standard preferences, they will have a bias in favor of promise-keeping, and if each knows the other to have the preference, then each can take the other's preference into account, and hence each can expect the other to keep the promise.

First, it is not clear if this solves the problem if we imagine the penalty for mismatched button presses increased so that a preference for avoidig the penalty is an order of magnitude stronger than the preference for promise-keeping. In that case, unless Alice and Bob are going to be very confident in the other's choice, a preference for promise-keeping will not do the job.

Second, and more importantly, if a preference for promise-keeping is needed to solve the problem, we now have an argument that some norm encumbent on humans requires such a preference. For if two human beings are stuck in a suboptimal solution in the button-pressing game, they are clearly falling short of what humans should be able to achieve. The argument thus shows that there must be norms on human beings that go beyond utility maximization, whether collective or individual.

1.2. Arresting the regress of meaning. Some communicative actions—speech acts or gestures—have their significance assigned through earlier communicative actions. Thus, sometimes one coins a word and stipulates its meaning in terms of other words, sometimes one uses gestures to introduce a new word, and sometimes one just hopes that use in a rich enough communicative context will clarify the meaning. But barring outlandish hypotheses such as that humans got their language from aliens, who got theirs from an infinite regress of angels, we cannot suppose an infinite regress. There must be ur-communicative actions, ones which did not get their significance from earlier communicative actions.[2]

At the same time, there is a contigency here. While it feels natural to us to use an extended index finger to indicate the nearest salient object approximately along the ray extending from the knuckle in the direction of the fingertip, it would be possible to have rational beings that use this gesture to indicate the third-nearest salient object along the ray extending from the index finger's tip to the knuckle. There is no necessary connection between the physical behavior and its significance. We thus have a Mersenne question here: What explains the correlation between physical behavior and significance in ur-communicative actions?

We might try to explain the correlation in terms of the actual contingent behavior of individuals and communities which arises by natural or social selection. Suppose, for instance, that some social animals evolve to squeak at a certain pitch when observing a predator, thereby warning other members of their group. Eventually, their descendants develop rationality, but the squeaking behavior is maintained, and remains correlated with the presence of a predator, even though it is now under voluntary and rational control. It is plausible to say that the squeaking is now a communicative activity whose significance is "Predator!"

[2]There is a Sellarsian objection to this. Perhaps there are behaviors prior to the advent of rationality that count as having communicative significance in virtue of *later* behaviors, in a kind of virtuous significance-conferring circle.??ref If so, however, then we can just count the whole circle of behaviors as the first ur-communicative community action.

But this plausible claim deserves more careful examination. Rationality complicates things. Suppose Alice sees a predator and is considering to squeak to trigger her groupmates' defensive behavior. If Alice's squeaking and her fellows' defensive behavior is to be rationally chosen, the agents need reasons. The fact that their ancestors used to squeak when predators were present and start defensive behavior upon such squeaking is an interesting bit of pre-history, but there does not appear to be a reason for them to imitate this quaint custom. Even if we add the fact that they find themselves with a desire to squeak in the presence of the predator and to initiate defenses upon hearing a squeak, these desires at most generate the very weak kinds of reasons one has to fulfill miscellaneous subrational desires rather than the strong kinds of reasons that one has to warn one's fellows and protect oneself and one's cmmunity.

There is no difficulty here if Alice is the only rational one, and hence the others will act on instinct. Alice can then rationally squeak to trigger the instinct. Similarly, if Alice acts on instinct and the others are rational, they can infer from her instinctive behavior that there is a predator present, as one infers fire from smoke. But when both are acting purely rationally, we have a difficulty. Likewise, if Alice thinks there is a fairly high chance that her fellows will follow their habit and prepare themselves, or if she knows that her fellows think there is a fairly high chance that Alice would find herself squeaking in the throes of instinct, there is no difficulty. The difficulty shows up when we have nothing but rationality at play.

Perhaps we can solve the problem by positing a non-rational preference for squeaking when seeing a predator and for preparing a defense when hearing a squeak? After all, arguably, even a perfectly rational being will act in accordance with preferences when other things are equal.

But the non-rational preference is insufficient here unless it is implausibly strong. For even if one finds oneself with an urge to squeak in the presence of a predator, the squeak itself endangers one. Because of this, for a rational being, the brute preference for squeaking is not sufficient to motivate the squeak. It is only when one thinks the squeaking will

trigger defensive behavior among one's fellows that it's worth squeaking. Similarly, we may suppose that defensive behavior is costly and inconvenient, and is only worth engaging in, not withstanding the non-rational preference, when there is reason to think there is an actual predator.

Instead of a brute preference, perhaps convenient priors will do. Thus, suppose that members of the community simply find themselves with a high prior conditional probability that a member of the community rationally squeaks when presented with a predator and does not squeak when not presented with a predator.[3] Knowing about that this prior is wide-spread in the community, Alice can squeak in order to get her fellows to update their credences in favor of a predator. The priors seem pleasantly self-confirming: if community members have the priors, then it will become public that they do so, and the squeaking behavior will match the priors.

But now suppose Bob is reflecting on his convictions. Bob finds himself accepting a correlation between Alice's rational squeaks and the presence of predators. But since the squeaks are rational, there must be a rational explanation of these squeaks. Since we are no longer attempting a preference-based story, presumably the explanation is that Alice accepts a correlation between community members hearing squeaking and their rationally coming to think there is a predator. In other words, Bob finds himself having a brute prior concerning a contingent and empirical matter—namely, another community member having a certain credential state. However, when we find out that our conviction about a contingent and empirical matter is simply a brute prior, that tends to undermine the conviction. Suppose that I find myself believing there is vast treasure buried under my house. I search for the source of my belief, and find it's just a prior. Absent a story such as that angels put that prior in my head to encourage me to dig out the treasure, finding out that this was *just* a prior should undermine the confidence, contrary to what subjective Bayesians think.

---

[3] I am grateful to ?? for the suggestion that priors might do the job.

Couldn't natural selection play the angel, though? There is an adaptive advantage to correlated priors in Alice and Bob that make communication possible, and these priors then end up automatically matching reality. ???

1.3. Reason-generating mechanisms. The transition from non-rational signalling to rational communication is thus difficult to analyze. We need some sort of a reason-generation mechanism.

Can we suppose that the reason-generation mechanism here is a necessary one? Perhaps it is a necessary truth that when there is a pre-rational behavior that tends to be triggered by circumstances C, then that behavior when done rationally *signifies* C? But there would be multiple Mersenne questions that would be raised by such a necessary truth. First, we need to select one item C in the causes rather than another—does the squeak signify the predator's, or the light in the air between the predator and the observer's eyes, or the immediate cause of the predator's presence? There are multiple selection rules, no one of them significantly more natural than the others. Second, what reliability does the tendency have to have in order to yield a signification fact? ??more The parameters in the connection between behavior and significance point to something contingent. We can imagine different species of rational beings where the parameters are different from what they are in us.

Reasons are normative entities. Thus a contingent reason-generation mechanism will, plausibly, be a mechanism for generating norms. Aristotelian form fits well here. The form could directly specify that squeaking properly occurs only when there is a predator present, or it could specify a general rule for connecting pre-rational behavior with norms of significance.

But even if there is such a norm-generating process, what makes the norms be norms of *communication*? A cat's nature requires it to turn its ears towards relevant sounds. When we see a cat turn its ears in some direction, that provides us with evidence that there was some sound relevant to it. But the cat is not communicating that there is a sound relevant to it by turning its ears. What, then, makes it be the case that a norm in the nature of

a communicative animal is a communication-constituting norm? Do we not need some further primitives besides norms of proper function to make it be *communicative* proper function?

We can speculatively sketch a part of an answer in terms of the *content* of norms. A toy story could be that some norms come in pairs, where one norm posits that a certain overt behavior is only proper when some fact p is known to a community member to obtain and another community member is known to be present and capable of observing the behavior, and another norm posits that when that overt behavior is observed in another member of the community, there is a tendency to form a belief in p. In that case, the toy story says that a behavior that is a fulfillment of the first norm counts as a communication of p and a behavior that is a fulfillment of the second norm counts as a reception of p. Of course, the full story would need to be much more complicated.

And all that said, it is not clear that we need a full story as to which exact behaviors are in fact communications. What matters for figuring out what to do is the content and force of the norms, not what kind of norms it is. It is a tautology that if the force of a norm is kept fixed, the norm has the same reason-giving impact on us, whether it be a norm of semantics, prudence, etiquette or morality.

2. Content and indeterminacy of reference

Wittgenstein, Kripke, Quine and Putnam??refs have problematized reference and content in light of the fact that different content attributions to our locutions can be made to fit with our behavior. The Wittgenstein-Kripke line of thought notes that any finite number of cases of behavior can be made to fit with infinitely many rules. Any finite number of utterances of "$a + b = c$" that fit with our "usual" interpretation of "$+$" will also fit with infinitely many rules, including, say, the rule that "$a + b = c$" means that c is identical with a plus b when a and b are less than or equal to x and means that c is identical with a times b when at least one of a and b is bigger than x, where x is the largest number we have ever discussed in the context of "$+$". The Quinean line of thought observes that the

same word, "Gavagai", can be interpreted to mean a rabbit or an undetached rabbit part, with both interpretations fitting equally well with the community's practices. And finally the Putnamian line of thought observes that a remapping of the truth conditions can make "The cat is on the mat" mean any other true proposition, as well as noting that the identities of mathematical objects—such as the integers—would be underdetermined even by a countably infinite number of statements about them.??ref ??expand-and-exposit

In all of these cases, we have the initial intuition that there is a well-defined meaning to the locutions, an intuition that is destabilized by the arguments. These cases, thus, can be seen as the opposite of the cases of vague terms like "bald", where our initial intuition is that there is no well-defined meaning.

We thus have two families of arguments. One family of arguments pushes in the direction of indeterminacy. And it does so not just in the cases where indeterminacy is intuitive, as for "bald" and "heap", but alas also in cases where we expect determinacy, as with the question whether we are referring to rabbits or undetached rabbit parts. Another family of arguments, mainly those based on insistence on classical logic??backref, push in favor of determinacy, but alas also in cases where we expected indeterminacy. If we want to maintain the determinacy of pretty much any term, then we will need to hold to something somewhat problematic—we will need to bite the bullet by denying a premise of a relevant indeterminacy argument (plausibly, some variant of the Quinean argument applies to all terms). But similary if we want to maintain the indeterminacy of any term, we will have to wrestle with classical logic.

With regard to these arguments, it would be simpler either to embrace determinacy in all cases or to embrace indeterminacy in all cases. For then we would only need to bite the bullet on one set of arguments. If we are to do this, then embracing determinacy in all cases seems preferable—embracing indeterminacy about all of language seems like it could undercut too much of our practices. However, by treating all the cases alike, we go against common sense which distinguishes "bald" from "rabbit".

The Aristotelian has a particularly good hope of having a metaphysical answer to the arguments for indeterminacy. This can be embraced in all cases, thereby resulting a picture of a sharp world that we will discuss below, or only in some cases, which fits with common sense.

The arguments for indeterminacy are all based on an assumption that the correct semantic theory will make semantic facts supervene on facts about our actual behavior and the world around us. But we can reject this assumption, and add normative facts about humans to the facts about our actual behavior and the world around us as part of what the semantic facts supervene on. These further facts could be hyperintensional normative facts, such as that it is only appropriate to say "Gavagai!" in the presence of a rabbit. Granted, necessarily, one is in the presence of a rabbit if and only if one is in the presence of an undetached rabbit part. But there can still be a difference between the norm of its being appropriate to say "Gavagai!" in the presence of a rabbit and a norm of its being appropriate to say it in the presence of an undetached rabbit part.

For norms are hyperintensional: ϕing and ψing might be such that necessarily one does one if and only if one does the other, but it is still a different thing to be required to ϕ than to be required to ψ. One way to see this is that if one is required to do something, one is required to try to do it. But trying, like intending and believing, is clearly hyperintensional. It is a different thing to try to bisect or trisect an angle with ruler and compass than to try to bisect an angle with ruler and compass, even though, necessarily, one bisects or trisects if and only if one bisects, since trisection is impossible. If I do not know that trisection is impossible and I have promised a friend to show them a trisection or bisection, what I am obligated to try is different than had I promised to demonstrate a bisection. It is one thing to have a reason to bisect and another to have a reason to bisect-or-trisect. Or, to adapt an example of Faroldi's (??ref:p137, Hyperintensionality and Normativity), you might be obligated to drive to the hospital, without being obligated to either drive to the hospital or drive to the hospital while drunk.

The above examples all involve moral normativity. But plausibly the same is true of other kinds of normativity. The function of the $\times$ key on a calculator is to multiply quantities, not to calculate the exponential of the sum of their logarithms. It is the proper function of a duck embryo to develop two feet, but it is not the proper function of a duck embryo to grow a number of legs that God would believe to be the smallest prime number.

Similarly, reasons are hyperintensional, and norms give rise to reasons, which makes it likely that the norms themselves are hyperintensional. To see that reasons are hyperintensional, note that reasons sometimes provide explanations of actions, and explanations are always hyperintensional.[4] That explanations are hyperintensional is easiest to see in the special case of entailing explanations, where the explanans entails the explanandum (e.g., that Bucephalus is a horse and all horses are mammals explains and entails that Bucephalus is a mammal). For if p explains and entails q, then p is equivalent to p & q, but a conjunction does not explain its own conjunct. But if explanation were hyperintensional, then anything equivalent to an explanation would also be an explanation. But even without entailment it is easy to see the hyperintensionality of explanation. For p is equivalent to $(p \ \& \ q) \vee (p \ \& \sim q)$. But this complex disjunction does not explain q. For the second disjunct, namely p & $\sim q$, does nothing to contribute to explaining q, and so if the complex disjunction explained q, it would do so by means of its first disjunct, p & q, and that does not explain q.

A similar argument may irectly shows the hyperintensionality of reasons. Suppose that p is a reason for x to ϕ. Then $(p \ \& \ \phi(x)) \vee (p \ \& \sim \phi(x))$ does not seem to be a reason for x to ϕ. For, plausibly, if a disjunction is a reason to ϕ, then at least one disjunct is a reason to ϕ. But p & $\sim \phi(x)$ is not a reason to ϕ: that I promised to call you by noon and failed to do so is a reason to apologize, not a reason to call. And that I ϕ is not part of a reason for me to ϕ, so p & $\phi(x)$ is not a reason for me to ϕ either.

[4]Cf. Faroldi (p. 139)??ref. Faroldi restricts this to non-causal explanations, but causal explanations are also hyperintensional by our argument below.

??add connective text?? Note that on our normative account we don't need any causal connection to the objects we speak about. Mathematical objects are no more problematic than physical objects. Even if an infinite number of sets of mathematical objects satisfy the Peano axioms, it is open to the normative semanticist to say that there is a pair (N, s) that make it be the case that according to the norms of our nature saying "There are infinitely many primes" is appropriate just in case there are infinitely many members of N that are prime with respect to the successor function s, so that the members of N are *the* natural numbers. At the same time, the normative semantics could allow that there is no privileged system (N, s) but instead all mathematical statements are conditional. Settling the question of which of these is true is difficult to task for the philosopher of mathematics, but neither presents a special semantic difficulty.

2.1. Illocutionary force. Typically, one asks someone for something that one wants. But asking is not the same as communicating one's desire. First, sometimes one asks for something one doesn't want. For instance, a security specialist could conduct a phishing call where they ask a fellow employee for their password, hoping that few if any will give it. Or a middle manager might be tasked by upper management with requesting something from staff that the middle manager thinks is actually bad for the company, and hence hope that no one will agree to the request. Conversely, one may want something but not ask for it for moral reasons. To adapt a situation that occurs twice in P. G. Wodehouse stories??ref, one may own an ugly heirloom that one cannot give away because of one's relationship with the person from whom one received it, but one would be glad if it were taken away. One could imagine a frank conversation where one happens to slip that one wouldn't mind the heirloom taken away. In the Wodehouse cases, the communication *is* a surreptitious request—and that, of course, is illegitimate, much as a king's exclamation "Would that someone rid me of this troublesome priest" is an invitation to murder. But one could also imagine a case where the slip is not a request, but simply a frank statement to a friend, followed by sincere emphasis that one isn't requesting removal. In that case, removal of the heirloom would be theft, even if it were desired by the owner. Or, for a

different case, one might have a moral objection to a particular life-saving medical procedure, and hence one's conscience would forbid one from requesting it, but nonetheless wish that the procedure were done to one against one's will, say by a medical mistake, and one could in a frank conversation communicate that wish *without* that constituting an underhanded request.

In making a request, one creates a reason for the other party to provide one with someone. And not just any reason, but a special kind of reason in light of one's own request.???

But now consider the first time anybody ever requested anything. In requesting, they created a moral reason for their interlocutor. This was a power they already had, and the meaningfulness of the communicative act of requesting must have already been in place. How? How could that communicative act not only had its illocutionary force but been *understood* to have that illocutionary force given that no one had ever requested anything? The meaning of a request is largely defined by the kind of reasons it gives rise to. But how can one grasp these reasons if one has never encountered them before?

3. Sharpness and levels

3.1. Sharpness at the second-level.

3.1.1. *Declarative practices and reasons.* We should all agree that it would be possible to have a communicative practice on which an declarative utterance instead of expressing a specific proposition subject to classical logic, expresses, say, a *family* of propositions, and the utterance has the following axiology: it is bad if none of the propositions is true, it is good if all of them are true, and it is neutral if some but not others are true. We take someone to be engaging the communicative practice well (badly) to the extent that their utterances are good (bad).

We can complicate things in various ways. We might, for instance, make the family of propositions itself be fuzzy, in the technical sense that we assign a degree (say between zero and one) to which a given proposition is in the family. Then we might say that an utterance is maximally good provided every proposition even partly in the family is true,

and maximally bad provided every such proposition is false. But for utterances on which there is no such unanimity in the fuzzy family, we find a way of measuring an intermediate value, say by a count weighted according to the degree to which the proposition is in the family of the true propositions in the family minus the false ones. Thus, the family of propositions corresponding to "Alice is rich" may contain to a high degree the proposition that she is at least a millionaire, to a low degree the proposition that she is at least a billionaire, and to zero degree the proposition that the sky is blue.

And such practices are perfectly comprehensible within an Aristotelian framework: we can suppose that human nature makes some declarative utterances fulfill us, others be contrary to our fulfillment, and others be indifferent to us. I am not yet claiming that our practices are like that, just that practices like that are perfectly comprehensible.

However, it is very plausible that some of our everyday utterances are in fact not merely epistemically vague. An inquiry into whether three rocks can make a heap or exactly how much money one needs to have to be rich seems a waste of time, not only because the answers are useless to us, but because there is no answer. In particular, it is very natural to say with the supervaluationists that there are infinitely many ways to make "heap" and "rich" precise, for each of which there is a well-defined answer to the question, but without such precisification there is no answer. This fits well with the hypothesis that our everyday communicative practices are like the ones described above.

Our description of the practices presupposed that there are propositions that are simply true or false, and that given the truth values of the propositions, the values of declarative utterances—whether just bad, neutral and good, or more fine-grained ones—are thereby determined. However are they determined sharply, definitely? Isn't it plausible that not only is it vague whether some group of four rocks is a heap, but one can have a grouping of rocks—say, five of them—where it's vague whether it's vague or whether it's definite that it is a heap, so that it's vague whether the statement "These rocks are are a heap" is neutral or good (or it's vague what exact value it gets).

Now the value of an utterance gives rise to reasons. If an utterance has positive value, that constitutes a reason to make the utterance. If it has negative value, that constitutes a reason to refrain from making it. And if it has no value, then we neither get a reason to make it nor a reason not to make it. But now recall the argument from ??backref(and make cohere with this!) that moral evaluation is merely epistemically vague. If moral evaluation were more than epistemically vague, and yet we had classical logic, then the right account of that vagueness would be that there are many moral concepts that fit with our usage of terms like "is right" and "is wrong", no one of which is privileged. But the central importance of morality to our rational lives requires privileging. Nothing can ultimately compete with moral wrongness as a reason against an action.

But what was true of moral concepts is *a fortiori* true of reason concepts. Nothing can ultimately compete with *reasons* in guiding our lives. Just as we might say that the overridingness of morality is the central discovery of ethics—explicit in Socrates—the centrality of reasons is the central discovery of action theory. It is difficult to say more here than to bang one's fist on the table. The concept of a reason is not an unprivileged one among many closely similar concepts. And this undercuts the possibility of non-epistemic vagueness about reasons. But because of the way the good and the bad immediately give rise to reasons, it follows that they cannot be vague either. And hence in the supervaluationist social practices described above, where declarative utterances express a family of precisifying propositions, there must always be sharp facts about whether the family is such that all, none or merely some of its members are true. In other words, vagueness must stop at the first level.

A reason-giving practice can only generate sharp reasons, since there are no others. Thus if vagueness is somewhere involved in a reason-giving practice, that vagueness must be "flattened out" once reasons are generated. Depending on the details, there may be more than one way of doing that. For instance, plausibly there is first-level vagueness about a person's height: there are multiple ways of measuring wihtout any privileged one,

depending on how much upward stretching is allowed, what gravity regime the measurement is made in (remembering that even on earth, gravitational acceleration varies about half a percent with location), whether skin flaking at the top of the scalp is included, etc. A practice that rewards the "tallest person" (say, a Guinness World Record practice) might generate a reason to bestow the reward on a person who is definitely the tallest, or on one who is definitely or vaguely the tallest, or it could (likely unfairly) flatten the vague data about height into sharp facts about reasons in some more complex way, especially if the family of precisifiers of "x is the tallest person" comes along with a degree-of-membership function.

In principle, any finite number of levels of vagueness could also get so flattened out. Thus, if there is second-level vagueness but no third-level vagueness, one might specify that there is reason to give the reward to someone who is definitely definitely tallest or definitely vaguely tallest or vaguely definitely tallest, and to no one else. And then we have fully sharp reasons, since there will always be a definite fact of the matter whether there is a reason to give the reward absent third-level vagueness. In light of this, one might think that the sharpness of reasons argument would allow declarative practices corresponding to any finite number of levels of vagueness. A limitation of vagueness to a finite number of levels would itself be a significant result.

However, there is reason to hold out for one level of vaguness in our declarative practices. The most plausible account of the flattening in our ordinary declarative practice d is:

(1) There is d-reason to declare p if and only if p is definitely true.

Moreover, the rule (1) itself seems definitely correct. Thus, if the existence of a d-reason is a definite matter, then whether p is definitely true must be a definite matter. And if definiteness is definite, so is vagueness, since, definitely, p is vague if and only if neither p nor $\sim p$ is definite. Thus vagueness stops at the first level.

The main competitor to (1) would be:

(2) There is d-reason to declare p if and only if p is definitely or vaguely true.

Given that (2) would give us reasons to affirm evidently contradictory sentences in cases of vagueness, it is not very a plausible candidate for an account of the relation of d-reasons to truth. However, even if (2) were the correct account, as long as it was definitely the correct account, the sharpness of "There is d-reason to declare p" would imply the sharpness of "p is definitely or vaguely true". Now, it is definitely true that

(3) p is definitely or vaguely true if and only if $\sim p$ is not definitely true.

Hence, we would have sharpness of "$\sim p$ is not definitely true", and hence we would have sharpness of "$\sim p$ is definitely true". But since it's definitely true that q is definitely true if and only if $\sim\sim q$ is definitely true, letting p be $\sim q$, we conclude that we would have sharpness of "q is definitely true". Since this would work for all q, we would again have sharpness of definiteness and of vagueness in general.

One might try for some other flattening of vagueness profiles to reasons. But (1) and (2) seem to be the most plausible two candidates. Thus we have good reason to stop at first level vagueness.

Finally, we might also argue for the claim that vagueness stops at the first level by thinking about the specifics of the morality of promises. If I promise to ϕ, and I do in fact ϕ, then I have done morally well; if I do not in fact ϕ, then I have done morally badly. But what if it's vague whether I have ϕed? Let's say that I have promised to cure your baldness, and because of my treatment you have some meager tufts of hair that you didn't have, not enough to make it definite that you are non-bald and yet not enough to make it definite that you are bald. Did I do well or badly? Well, it is very natural to say: my activity was neutral in respect of the promise. Given a sharpness in attribution of moral goodness, badness and neutrality, and given the above plausible matching of moral value with attributions of definite truth, definite falsehood and vagueness, we have good reason to think that in the case of predicates that can figure in promises at least, we can at most have first-level vagueness—it must be sharp whether someone is definitely bald, vaguely bald or definitely non-bald. But if we have second-level sharpness about baldness, plausibly we have second-level sharpness about everything.

3.1.2. *Logic.* A standard argument for full-blown epistemicism is the Sorites series.??Sorensen Consider this argument, where the if-then statements are material conditionals.

(P_0) Charles wasn't old on his first day.

(P_1) If Charles wasn't old on his first day, he wasn't old on his second day.

(P_2) If Charles wasn't old on his second day, he wasn't old on his third day.

...

(P_{26999}) If Charles wasn't old on his 26999th day, he wasn't old on his 27000th day.

(C) So, Charles wasn't old on his 27000th day.

Clearly C is false: a 73-year-old *is* old. The argument, however, is valid by a sequence of 26999 instances of *modus ponens*. The only way a valid argument can have a false conclusion is by having a false premise. Now, premise P_0 is clearly true. Thus, at least one of the premises P_n, for $n = 1, ..., 26999$, is false.

Now, if P_n is false for $n \geq 1$, then since P_n is a material conditional, its antecedent is true and its consequent is false. Thus, Charles wasn't old on his nth day but became old by his $(n+1)$st day. Hence, we have a one-day transition from young to old. And it is precisely such sharp transitions that non-epistemicist advocates of vagueness reject as absurd.[5] Yet logic forces us to accept them.

On the view I am defending, logic applies to propositions, not to sentences of our declarative practices. If we consistently precisify the premises and conclusion of the argument, the precisification of one premise will turn out to be false.[6]

Sticking to the sentences in the argument, we can say that P_n is definitely true for small n, then becomes vague (maybe somewhere around $n = 22000$), and finally becomes

[5]We should assume that by "old" we mean something like "calendrically old". For we all understand such locutions as "Alice became old the day she found out she had lung cancer."

[6]Only one. For if Charles is old on a day, he's old on all subsequent days. If P_n is false for $n \geq 1$, then the consequent is false, so Charles is old on his $(n+1)$st day, and if P_0 is false, then Charles is old on his first day. Either way he is old on day m whenever $m > n$, and so P_m is true, since it's a material conditional with false antecedent. Thus, as soon as one premise is false, all the subsequent ones must be true.

definitely true again. The points at which P_n becomes vague and then again definitely true are fully precise, which is counterintuitive, but that counterintuitiveness is less evidentially significant than the violation of common sense in Charles becoming old on a specific day.

If we think a sentence is bad to say when definitely false, neutral when vague, and good when definitely true, then none of the P_n are bad to say, but the conclusion *C is* bad to say. We might think that this means that we have a case where something bad to say logically follows from a number of things that are not bad to say, and this may seem absurd. But it's not clear that this is absurd. For we are not here dealing in propositions, where a conjunction of acceptable ones is also acceptable, but with sentences in a vague declarative practice. We should not import logical intuitions that apply to propositions in thinking about the value of a declarative practice. It may also help to see that in the case of *moral* badness there is nothing particularly paradoxical about cases where asserting a conjunction is bad but no conjunct is bad to assert. For instance, consider the case[7] of a racist who writes a series of factually correct articles, each one about a highly immoral member of a minority group. Each article may be such that it is not bad to write, but the oevre as a whole is racist.

We do, however, have a famous difficulty. Like other supervaluationist views, the account being defended has the consequence that definitely:

(4) There exists n between 0 and 27000 such that Charles is not old on day n but is old on day $n+1$.

At the same time, for any specific day m in that range, say $m = 22003$, it is *not* definitely true that:

(5) Charles is not old on day m is but is old on day $m+1$.

In particular, the disjunuction of instances of (5) as m ranges over all the numbers between 0 and 27000 is definitely true, but every disjunct is merely vague. This is, admittedly,

[7] Not hypothetical.??ref

counterintuitive. This counterintuitiveness is, I suspect, tied to the fact that if we have a true disjunction of propositions, at least one disjunct is true....????

??bite bullet on exists n, but ok as logic doesn't apply

3.1.3. *Some objections.* The conclusions above are counterintuitive. Intuitively, just as it sounds silly to think that on such-and-such a day it was true to say that Elizabeth II was (chronologically[8]) old while it wasn't true to say it a day earlier, it sounds silly to say think that there was a precise day on which she was definitely old, while the day before she was merely vaguely old, and similar objections can be made at higher levels of vagueness.

However, our linguistic intuitions are typically more trustworthy than our metalinguistic intuitions. Thus, our intuition that oldness is can be vague is more to be trusted than our intuition that vagueness can be vague. We have good arguments for second-level sharpness, and these arguments undercut the intuition.

A second objection starts like this. Communicative practices of a declarative sort that embody first-level vagueness are indeed possible. All we need to do is to specify the values in the way we did at the beginning of Section 3.1, so that an utterance is good when all the propositions in an associated family are true, bad when they are all false, and neutral otherwise. If we had a community of perfectly sharp speakers, we could even imagine introducing such a practice, either as a sort of pleasant relaxation (it's not fun to always have to be precise) or for practical reasons, to be engaged in at times. Suppose we have such a practice, and it includes terms like "hairs" and "scalp". But then we could engage in this practice to introduce a new non-sharp communicative practice. For instance, we might specify that in the new practice a predication of "is bald" is good when the person has more than 2000 hairs on their scalp, is bad when the person has fewer than 1000, and is neutral otherwise. However, how much filament needs to stick out of a follicle for the follicle to count as hosting a hair itself appears to be fodder for first-level vagueness matter, and it can be first-level vague whether a hair is on the scalp or the upper part of the

[8]It is perfectly comprehensible to say that someone *psychologically* became an old person after some traumatic event.

cheek. Consequently, in the new practice we can have cases where it's vague where Jim's crinal profile falls—whether he has fewer than 1000 hairs on the scalp, more than 2000, or in-between. And in such a case it may well be vague that it's vague whether he is bald.

Now, given the possibility of such a practice, its actuality is likely. Surely we often do extend our communicative practices, using old—and presumably vague—terms to introduce new ones. And we do that with our vocabulary.

However, notice that the quick sketches of communicative of social practices corresponding to vagueness were accounts of declarative practices. One does not introduce a new social practice—a game, say—by an declarative practice, but by an institutive practice: "Let's a play a game with rules $R_1, ..., R_n$."[9]

While the norms of a declarative practice that includes vague utterances is easy to sketch, it is more difficult to sketch those of an institutive practice that involves vagueness. The rules of a practice provide reasons for the practitioners. But if it cannot be vague whether something is a reason for ϕing, a reason against ϕing, or neither, then we cannot make it vague what the rules defining the practice are and how they apply. What we can do is at most something this. We can have a practice of instituting practices, where we specify a collection of ordered "rule" pairs (D, E) where D is a description of a behavior (or maybe situation) and E is an in-practice evaluation, such as "permissible" or "bad", and where the description D is such that $D(b)$ is a declarative utterance of a linguistic practice embodying first-level vagueness when b is a name of a behavior. Our practice of instituting practices then specifies that if b falls under D, where (D, E) is one of the rules then, $D(b)$ is definitely true if and only if b has E. In cases where b does not definitely fall under the description D in any of the rules, then we can simply say that b definitely lacks all of the relevant evaluative properties.

For instance, the rules for an oversimplified race r could be (A, W), (B, T), and (C, L), where W is a win, T is a tie, and L is a loss, and $A(r)$ says that r is a performance that is a run with the upper body of the runner crossing the finish line before the upper body of any

[9]Though of course one might use declarative grammar: "We will play a game with rules $R_1, ..., R_n$."

other runner, $B(r)$ says that r is a performance that is a run with the runner's upper body never crossing the finish line or another runner's crossing earlier, and $C(r)$ is the denial of $A(r)$ conjoined with the denial of $B(r)$. Now suppose that in our first-order declarative language, sentences like "Alice's upper body crossed the finish line at t" are vague when what crossed the finish line at t was a loosely attached scab on the forehead. Then in cases where Alice was ahead only by such a scab, Alice definitely lacks a win, a tie or a loss. If a win and a tie have positive valence, while a loss has negative valences, then we can say that Alice's performance definitely is neutral.

That all our games are in fact perfectly sharp in their values is counterintuitive. But it is hard to avoid this conclusion while holding on to the privileged role that reasons play in our lives that forces reasons to be sharp.

3.1.4. *A sharp world and a fuzzy language.* A broadly-held intuition is that the world is sharp but our language is fuzzy, so all the vagueness is due to our language. There is a well-known objection to this: our language is itself a part of the world, so linguistic vagueness is still vagueness about the world.

The view defended above where vagueness is restricted to the first level does justice to the sharp-world-fuzzy-language intuition. Our declarative sentences express a range of propositions, maybe even a graded range, and that's the fuzziness of the language. However, the propositions in the range are themselves sharp, and the truth of the sentences is derivative from the truth of the propositions, so ultimately the world impacts our language through the sharp end of our practices, and the world itself can be said to be sharp.

At the same time, the view manages to escape the objection that our language is a part of the world. The relevant part of the world is our declarative practice. And the semantics and norms of this practice are fully sharp. There is always a definite answer to the question whether a given proposition is a precisification of a given declarative sentence, to what degree (if the account involves degrees), and what the norms governing the use of the sentence are. It can be vague whether a sentence is true, but it is not vague what the sentence means: it means its set of precisifiers.

Requiring a piece of sports equipment to be between exactly 200 and 250 grams does not render a sports practice vague. Similarly, the fact that the evaluative properties of our declarative language are sometimes defined in terms of ranges of propositions, rather than individual propositins, does not make the practice itself vague. Linguistic vagueness on this view does not mean that linguistic practice as a practice is vague. It just means, very precisely, that some of the sentences of the language express a range of propositions. This kind of vagueness does not make the practice itself vague, and does not introduce any vagueness into the world.

On the other hand, if there were higher-level vagueness, so that sometimes it was vague whether a sentence is, say, definitely true, the linguistic practice as such would be vague. For the practice depends normatively on whether a sentence is definitely true: there is a practice-internal good to declaring a sentence that is definitely true which is not had when one merely declares a sentence that is vaguely true. ??Merricks

A closely-related intuition is that the world as it objectively is is sharp, but the world as it is relative to us is fuzzy. Someone who accepts this intuition may insist that practice-internal values are themselves relative to us, rather than a part of the objective furniture of the world. Again, this is a difficult line of defend given the observation that if something has a practice-internal value for some individual x, then it is an objective fact about the world that it has that practice-internal value for x.

Perhaps, however, one could read this intuition as implying that there is second-level vagueness but no third-level vagueness. Thus, one might say that for our declaratory practice, it is not sharp whether a performance is good, neutral or bad, but there is some objective range of precise interpretations of that practice, and we say that a performance is definitely good provided it is good on all interpretations, vaguely good if good on some but not all, and definitely not good if good on none. What that range is, however, is an objective fact about the world. Within each interpretation, the evaluation of a performance will be sharp. However, there does not seem to be any significant philosophical benefit to placing the sharpness at the third-level rather than the second. We are left with the

problematic idea that the concept of a reason is subject to multiple precisifications, in a way that does violence to the overridingness of rationality, and besides we have a more complex view, while still requiring an account of the third-level sharpness, which is no easier to have than an account of second-level sharpness. And the intuitions supporting second-level vagueness are just not as robust as those supporting first-level vagueness.

3.2. A distinguished semantic theory. ???????? Canonical choice of naturalness???

4. A neo-Aristotelian account

The neo-Aristotelian account allows us to have a completely sharp world with our language being completely sharp, with there being definite facts of the matter about significant and insignificant questions: Is Alice dead yet? Is this pile a heap? Should Alice rebuke Bob publicly? Is the smashed object a car? Is this still the ship that Theseus sailed in? Is Beethoven a better composer than Bach? Should I lower my credence in quantum mechanics after the mildly senile retired physicist told me she just found a contradiction in it? Is a cat a dommal (where "dommal" is a term whose patterns of use are that users are content to call all dogs "dommals" and infer mammality from being a dommal).

Our nature *could* provide us with this sharpness. First, when the propositions in turn are concern normative matters, our form can at least partly ground the truth values. This can yield complete sharpness about all normative matters.

Second, our nature could ground the facts about the proposition expressed by and illocutionary force of each of our utterances by grounding the norms that specify how facts about our symbolic behavior together with facts about the non-symbolic aspects of the world attach propositions and illocutionary force to utterances. These rules could be very complex, and known by us only approximately and in general terms.

But at the same time, instead of grounding the specific proposition expressed by each utterance, our nature could ground facts about ranges of propositions, thereby allowing declarative utterances to be vague with multiple precisifiers. Again, our nature can do

this by fully precisely grounding our semantic norms—for, as we saw in ??backref, it is possible to have fully precise semantic norms and yet genuine first-order vagueness.

??blunt language and sharp world

???complex embeddings and logic?

??backref to indeterminacy

CHAPTER VIII

Metaphysics

1. Composition

Intuitively, your parts compose a whole, namely you. On the other hand, if you choose exactly one atom from every star in every galaxy in the universe, intuitively these scattered atoms do not make up any whole.

Yet it seems that one could produce a continuous series of worlds, adding, subtracting or moving one particle at a time, where the first item in the series consists of the scattered atoms in different stars and the last item consists of you. Somehow, as one moves along the sequence, at some point a new macroscopic object pops in (maybe it's you, or maybe it's something else), despite the variation between successive terms in the sequence being extremely slight. What makes for the transition? ??ref:Sider

???vagueness, nihilism, Sider, etc.

As Tomaszewski has noted??ref, the Aristotelian, however, has a solution, namely the invocation of form. After all, it is not possible to have a sequence where at one end you have one atom from every star and at the other have all of your parts, and the transition is particle-by-particle. For your parts include a human form, while the scattered atoms contain at most the forms of particles or atoms. Adding, substracting or moving particles is not enough: one needs to add a human form somewhere in the sequence.

We can, further, give an account of when a plurality of objects, the xs, compose a substantial whole: namely, there is a form among the xs which informs all the other xs, and everything informed by that form overlaps with at least one of the xs:

(1) $\exists F[F \in xx \;\&\; operatornameForm(F) \;\&\; \forall y((y \in xx \;\&\; \sim y = F) \rightarrow operatornameInforms(F, y))]$,

assuming that the $operatornameForm(x)$ predicate only applies to substantial forms. Here, the informing relation is one where the form gives identity to the thing it informs. [1]

??why is the transition where it is?: Mersenne

2. Identity over time

One of the classic questions of metaphysics is about the grounds of identity over time. A very general way of posing the question is to ask for an explanation or ground of claims of the form:

(2) Object x_1 is identical to object x_2,

where x_1 and x_2 respectively exist at times t_1 and t_2, which are presupposed to be different[2] and where we require the explanation not to involve identity and to involve only purely qualitative properties and relations.[3]

Put in that very general way, it seems unlikely that we will have a solution, absent some extremely controversial metaphysical assumptions, such as Leibniz's Principle of Identity of Indiscernibles (PII).[4]

---

[1]Understood this way, informing is a relation that holds both between an a substantial form and the material parts of the substance and between a substantial form and the accidents or accidental forms of the substance. One may, however, object that these are two different relations. If so, take my $operatornameInforms(x, y)$ relation to be a disjunction of the two.

[2]One may worry that this presupposition cannot be stated without using identity, i.e., without denying that $t_1 = t_2$. However, it is not clear that even if this is true, it affects the significance of the question. Moreover, if time turns out to be linearly ordered, then we can state the presupposition disjunctively: t_1 is earlier than t_2 or t_2 is earlier than t_1.

[3]If we allow the account to involve non-qualitative properties like Socrateity (the property of being Socrates), then we can offer a infinite account: x_1 is identical to x_2 provided that any property had by x_1 is had by x_2.

[4]The PII says that two things are identical just in case they have the same purely qualitative properties. If the PII holds, then we can say that $x_1 = x_2$ if and only if for every purely qualitative property we have $Q(x_1)$ iff $Q(x_2)$.

But there is a somewhat less general way of putting the diachronic identity question. Suppose that at time t_1, some proper plurality of items, the xs, compose an object and at time t_2 the ys compose an object. Then we ask for the grounds of:

(3) An object composed of the xs at t_1 is identical with an object composed of the ys at t_2.[5]

This is not asking for an account of diachronic identity in general, but of diachronic identity of complex objects (note that we only require the xs to be a proper plurality, i.e., for there to be more than one of them).

Here we *can* give an Aristotelian account:

(4) There is a form F such that at t_1, F unites the xs, and at t_2, F unites the ys.

Here we might stipulate that a form F unites the zs just in case F is one of the zs, at least one of the zs is a non-form, every form among the zs informs each non-form among the zs, anything informed by F overlaps at least one of the zs:

(5) $\text{Unites}(F, zz) \equiv [F \in zz \;\&\; operatornameForm(F) and \exists x(x \in zz \;\&\; \sim operatornameForm(x)) \;\&\; \forall G \forall x((operatornameForm(G) \;\&\; \sim operatornameForm(x) \;\&\; G \in zz \;\&\; x \in zz) \rightarrow \text{Informs}(G, x)) \;\&\; \forall x(\text{Informs}(F, x) \rightarrow \exists y(y \in zz \;\&\; O(y, x)))]$.

Here $O(y, z)$ says that y overlaps z, i.e., $\exists w(w \leq y \;\&\; w \leq z)$, where $\leq$ is the parhood relation. There is no identity anywhere in (5).[6]

[5]The reason for the indefinite pronoun is that, first, there might be more than one object composed of the same parts and, second, using the definite pronoun introduces another instance of the identity relation given the Russellian analysis of "The F is G" as saying that some F is G and has the property of being *identical* with every F that is G, whereas we only want an account of a single identity relation.

[6]Note that the right hand side of (5) is much more complex than (**??**), even though we are giving an account of a very similar phenomenon. A simpler formulation than (5), and more in line with (**??**), would say that F unites the zs just in case it is one of them and informs every one of the zs other than itself and everything informed by overlaps at least one of the zs. However, this formulation makes use of identity in talking of z *other than* F. To avoid the identity operator, we quantify over all the substantial forms among the zs and say that they inform all the non-forms. For the present account to work, we cannot have a case where all the things

It may appear that (4) uses the concept of identity in claiming that the *same F* unites the xs as unites the ys. Even if that were so, progress would have been made: an account of identity for complex would be given things in terms of identity for simple things. But in any case, we need not concede the point as we can rewrite (4) in first order logic without any equal signs:

(6) $\exists F(\text{Form}(F)\ \&\ \text{Unites}(F, xx, t_1)\ \&\ \text{Unites}(F, yy, t_2))$.

Granted, (7) is logically equivalent to:

(7) $\exists F \exists G(\text{Form}(F)\ \&\ \text{Form}(G)\,\text{Unites}(F, xx, t_1)\ \&\ \text{Unites}(G, yy, t_2)\ \&\ F = G)$,

but this only show that the identity is eliminable from (**??**), just as the identity can be eliminated from

(8) $\exists x \exists y(\text{Tall}(x)\ \&\ \text{Green}(y)\ \&\ x = y)$

("There is a tall object which is identical to a green object") to yield:

(9) $\exists x(\text{Tall}(x)\ \&\ \text{Green}(x)$

("There is a tall green object").

Thus, (4) gives us an account of the identity of complex objects that does not presuppose identity.[7]

---

informed by a substantial form F are also informed by another substantial form G, i.e., it cannot be that all the non-substantial-form constituents of one substance a are parts of another substance b. For if we had that, then the (5) would make F unite all the constituents of a together with the form of b, which does not seem right. On a natural interpretation of the metaphysics of conjoint twins (??cross-ref,??ref), there are common parts that informed by two forms. Our assumption is compatible with that interpretation, but rejects the odd possibility of conjoint twins where the common parts include *all* the non-substantial-form constituents of one of them. If our assumption is rejected, then we may need to make use of the identity operator in (5). However, the identity operator would only need to be used in the case of where one of the relata is a form, and so we could still have an explanation of identity of substances in terms of identity of form, and metaphysical progress will have been made.

[7]One might also worry that complexity presupposes identity: an object is complex provided it has two distinct parts. But the Aristotelian can make sense of complexity of objects by saying that an object is complex provided that it has a part that is a form and a part that is not a form.

3. Teleological animalism and cerebra

Each premise of the following argument is very plausible.

(10) We are humans.

(11) Humans are mammals.

(12) So, we are mammals.

(13) Mammals are animals.

(14) So, we are animals.

Nonetheless, the conclusion—labaled as "animalism"—is denied by many philosophers. One traditional path to this denial is a Cartesian dualism on which we are immaterial souls that inhabit human animals. The other path is modern colocationist views on which we are a special kind of material object—a person—constituted by a human animal. The third path is brain views on which we are brains, or parts of brains (namely cerebra), which in turn are a proper part of a human animal.

On all three paths, one will want to deny (12). Perhaps the best way to do so would be to distinguish "humans" in (10) and (11) into human animals and human persons, and insist that premise (10) is true only of human persons, while (10) is true only of human animals. Thus the argument is unsound if "humans" is used consistently and otherwise invalid.

In any case, the intermediate step (12) gets denied. Instead, we are *associated* with mammals, by ensoulment, constitution or parthood. Yet (12) is by itself extremely plausible. To deny that we are mammals seems akin to denying that earth is round. Further, both the Cartesian and brain views imply that we rarely if ever see or touch another person. One needs extremely good arguments for such counterintuitive theses.

Probably the main candidate here is a family of arguments about the difficulties of accounting for cerebrum transplants on animalism.??refs The simplest version is that if your cerebrum is removed from your skull and placed in a vat in such a way that it can continue functioning, then intuitively you continue to think and come along with the cerebrum. But

a cerebrum is not an animal. On the contrary, the cerebrumless body appears to be an animal. After all, some animals (e.g., fish??) lack cerebra, and it seems that the destruction of a cerebrum in an animal that normally has one would result in the animal becoming severely disabled rather than ceasing to exist. Thus, animalism points to the cerebrumless body as you—the cerebrumless body is the same animal as you—and the cerebrum in the vat as something or someone else, contrary to our intuitions.

Cartesians, colocationists and brain theorists who identify with the cerebrum have no such problem. They can all say that the cerebrumless body is not you, and instead you inhabit or are colocated with or are just plain identical with the cerebrum in the vat.

Here I want to argue that an Aristotelian about humans can embrace a teleological variety of animalism on which it is natural to say that we go along with our cerebra. Now animalism is highly plausible as a view of the human person that does justice to the intuitions that humans are mammals, weigh about 80 kg and can be seen without surgery, but faces a serious cerebrum transplant problem. If Aristotelianism can help animalists overcome that problem, that is some further evidence for Aristotelianism about humans.

There are two teleological features in the animalism I will sketch. The first teleological feature is the thesis that what defines somethingas an animal (and indeed an animal of a particular type) is its teleology. While it is usual to think of animals as things that nourish themselves, grow, reproduce, and have a certain level of autonomy from the enviroment, the teleological animalist instead insists that animals need not engage in these activities, but need only have a teleological orientation towards them, need to be the sorts of things that *should* engage in these activities.

The second teleological feature is to see organisms, including humans, as having a teleological *hierarchy*, with some telē subordinated to others. Sometimes the subordination is instrumental: our teeth rend food in order to nourish us. But there can also be a value-based subordination, where an activity, while not merely instrumental towards another activity, is less central to the flourishing of the organism and to the organism's identity. On the teleological animalism I am sketching, it is postulated that in humans, activities

common to all animals are subordinated to specifically personal activities, namely rational and moral behavior. ???can we have this in non-theistic versions??? wouldn't we have subordination to reproductive activity???

When there is a hierarchical subordination, it is plausible to think that in cases of splitting, an organism is more apt to come along with the organs supporting the higher-level features. If in humans it is moral and intellectual capacity that is at the top of the teoleological hierarchy, and the cerebrum is much more directly supportive of these capacities than the lower brain, heart, lungs, etc., then we would expect the human organism to come along with the cerebrum. If you cut a worm in such a way that one end contains just the head and the other contains the rest of the body, and both parts behave as if they were alive, it is reasonable to suppose that the form of the original individual may go with the larger piece which contains more in the way of life-supporting organs, rather than with the head, because the head is less teleologically central to the worm than to us. But given human teleology, we would expect us to go along with the head, if the head were given life support, or even with the brain or cerebrum.

What about the remaining cerebrumless human body, which maintains its vital functions? If the original individual goes along with the cerebrum, and if we take the idea of one human form in two bodies to be absurd, there are three possibilities for the cerebrumless body:

(a) the cerebrumless body gains a new human form, or

(b) it gains some non-human form, or

(c) it becomes a non-living substance or a formless heap of matter.

If the cerebrumless body gains a new human form, then we have the counterintuitive consequence that a temporary removal of a cerebrum followed by reimplantation would either result in conjoined twins—two human beings joined at the edges of the cerebrum—or would result in the death of one or more of the two human beings. None of these options seems very plausible, but at the same time, we should not be too surprised if strange things happen when you move cerebra around!

If the cerebrumless body gains a non-human form, this is presumably an organismic form, since the entity is capable of nourishment, reproduction, etc. But we have something moderately puzzling: a non-human organism that would produce a human being if it were to mate with a human or with another organism of the same sort but of the opposite sex. Moreover, it seems plausible that if the cerebrumless body has a teleology at all, that teleology impels it to try to support the cerebrum: oxygen would be directed by the body towards the missing cerebrum, presumably. This suggests that the being is incomplete without the cerebrum. But a being that ought to have a human cerebrum seems to be a human being.

The last option is a formless heap of matter or a non-living substance (such as a body of water might be on some Aristotelian theories??refs). This does not seem particularly puzzling in the transfer case. If the form departs, by going along with the cerebrum, then formlessness would seem to be the obviously expected result, barring some special reason to the contrary.

What if instead of the cerebrum being removed and put on life-support, the cerebrum is simply destroyed? This corresponds to a tragic real-life scenario: upper brain death. Bioethicists disagree on whether upper brain death is death.??refs We have four moderately plausible options for cerebral destruction:

(i) the original individual continues to live, i.e., the cerebrumless body retains the original human form??backref-to-identity, or

(ii) the original individual dies and the cerebrumless body gains a new human form, or

(iii) the original individual dies and the cerebrumless body gains a non-human form, or

(iv) the original individual dies and the cerebrumless body has no organismic form, and is either a non-living substance or a heap.

Between (a)–(c) and (i)–(iv), there are twelve combinations with various intuitive connections between them. However, we can intuitively reduce the number of options quite

significantly. I have assumed that in the transfer case, the original form goes along with the cerebrum in the transfer case, departing from the cerebrum. Suppose that (i) is false. Then the original individual dies and the cerebrumless body is deprived of its original form. It seems very plausible that what happens to the cerebrumless body upon deprivation of its original human form should not depend on what that form does after departing the cerebrumless body—whether it continues to inform a reduced body (the cerebrum), or survives disembodied as many religious people think, or perishes. Thus, if (i) is false, then (a), (b) or (c) holds respectively if and only if (ii), (iii) or (iv) holds.

We thus have three plausible options with (i) false:

(α) (a) and (ii)

(β) (b) and (iii)

(γ) (c) and (iv)

What if (i) is true, so that in the case of cerebral destruction, the original form continues to inform the cerebrumless body? Does this tell us anything about what happens in the transfer case? One might initially think that the falsity of (i) fits poorly with any of (a), (b) and (c). After all, if destruction of the cerebrum results in the form staying with the cerebrumless body, shouldn't we expect the form remain with the cerebrumless body when the cerebrum is transfered?

But this is not clear on teleological animalism. For we might think that in division or partial destruction of the body, the form goes along with the part that is the best candidate for being informed by it, at least when there is a unique best candidate and that best candidate is "good enough". On the teleological account, the quality of candidacy for being informed is measured by how high in the teleological hierarchy are the goals directly promoted by the part. If the part promotes goals at the top of the hierarchy, then the candidate is automatically good enough. Thus, when the cerebrum survives, it is reasonable to think the form goes along with the cerebrum, because the cerebrum is good enough as a candidate, and higher in the hierarchy than the cerebrumless body. But if the cerebrum is destroyed, the cerebrumless body is the unique best candidate, because it is the only

candidate. Whether the cerebrumless body is good enough as a candidate is not clear, but neither is it clear that it is not good enough. It is, after all, on the next step down in the hierarchy after the cerebrum, having among its tasks the full support of the cerebrum's functioning, as well as many important purely animal functions. Thus, all of the following are at least somewhat reasonable epistemic possibilities:

(δ) (a) and (i)

(ε) (b) and (i)

(ζ) (c) and (i)

On teleological animalism we thus have six combinations for making sense of what happens to the form and cerebrum, namely (α)–(ζ), and none of them appear immensely problematic, though (a) and (b) have some moderately counterintuitive consequences. But even if we feel the need to reject these, that still leaves us with (γ) and (ζ): in cerebral transfer cases, the cerebrumless body is not a living thing, while in cerebral destruction cases, the cerebrumless body may (if we have (i)) or may not (if we have (iv)) continue to be a living human being.

One may think (ζ) is implausible, because it implies that whether the body-minus-cerebrum continues to have the original human form depends on what happens to the cerebrum—whether it is merely removed or actually destroyed. But such dependence is, as already noted, to be reasonably expected. For if we think of the cerebrum as a magnet for the original human form, then transfer of that "magnet" might be reasonably be thought to pull the form along, hence depriving the cerebrumless body of it, while destruction of the "magnet" might well leave the form in place.

Teleological animalism, thus, has multiple ways of making sense of what happens in cerebral transfer and destruction cases, while these cases are highly problematic to other types of animalism. Given the plausibility of animalism as such, this gives us another reason to accept teleological animalism.

4. Ill-matched matter, rearrangement, the power to continue existing and immortality

5. Naturalism

Is the Aristotelian hylomorphic account of humans compatible with naturalism? This depends on how naturalism is defined and whether we take the theistic version of the account or not.

Since theism implies the existence of a non-natural causally efficacious substance, namely God, it will be incompatible with most versions of naturalism. And I have argued theat the Aristotelian account is unsatisfactory without theism.

Still, it is an interesting question whether the what the Aristotelian account says about the forms of finite substances is compatible with naturalism. If so, then one could combine the Aristotelian account with a naturalism restricted to finite objects, and save some naturalistic intuitions. Furthermore, it would mean that an Aristotelian not convined by the arguments that theism is needed to make the theory satisfactory could be a full-blown naturalist.

If we take naturalism to say that the only entities that exist are the ones that would figure in a completed science, then it is unlikely that Aristotelian metaphysics would be compatible with naturalism about finite objects. However, such a strong naturalism would likely also conflict with many other metaphysical theories that are rarely taken to contradict naturalism.

For instance, consider theories of time. Of the theories of time, four dimensionalism, on which ordinary objects like ourselves are extended in time as well as space, is what seems to fit best with Relativity Theory. But the most common four-dimensionalist view of changing properties is perdurantism: changing objects are made of temporal parts or slices to which the properties are primarily attributed, so that a tomato that once was green and now is red has a green temporal part and a red temporal part. But slices are unlikely to figure in a completed science if our current science is a good guide to that. Consider that our current physics does not consider particles like electrons and quarks to

be fundamental, and hence not made up of smaller parts. But electrons and quarks change over time (e.g., with respect to spin and flavor), and so they would need tos have temporal parts. But these parts are not found in our physics.

Or consider that our science may quantify over physical objects like particles and fields, and applies predicates to them, but does not quantify over properties. Thus, properties, whether understood as universals or as tropes, go beyond current science. Yet it seems implausible to understand naturalism as implying nominalism.

One may weaken naturalism to say that the only *causally relevant* entities are those of a completed science. But this would still rule out perdurantism, since objects change with respect to causally relevant properties, and hence their temporal parts have these properties, and have them more fundamentally. It would also likely rule out many versions of trope theory, since the causal efficacy of tropes is supposed to explain the causal efficacy of the objects made of them.

Let's step back. Naturalism denies the existence of causally efficacious "supernatural" entities like ghosts, but it is neutral on temporal parts and tropes of "natural" entities like electrons. What is the relevant difference between ghosts and temporal parts of electrons? It seems to be this. While neither entity is posited by science, if ghosts exist, then certain phenomena that it belongs to science to investigate have no scientific explanation, barring systematic overdetermination. If there are ghosts, they are presumably responsible for at least some of the appearances of ghosts, some of the chills people feel in graveyards, etc. These phenomena then either have no physical explanation at all, or by a massive coincidence are overdetermined by a physical and a spectral cause. There is a competition, thus, between scientific and spiritual explanations when we are concerned with ghosts.

On the other hand, there is neither competition nor overdetermination in the case of the objects studied by science and their ontological constituents such as temporal parts or tropes. When a temporal part or a trope of an ice cube makes your hand cold, the ice cube also makes your hand cold. The ice cube's cooling causal influence on your hand does not compete with the ice cube's temporal part's causal influence on you or on the

causal influence of a trope of coldness (if there are temporal parts of ice cubes or tropes of coldness). Nor is this overdetermination, since it is not overdetermination when an object *C* causes an effect *E* by means of *C*'s part. After all, it is not overdetermination when the ice cube cools the hand it is lying on *and* the lower half of the ice cube cools the hand.

We might thus want to say that naturalism holds that any phenomenon that it falls within the purview of science to seek for causal explanations of either has no causal explanation at all, or else has a scientific causal explanation and is not overdetermined by a scientific and a non-scientific one. On this account, theism is incompatible with naturalism if there is a first state of physical reality. For if there is such a state, it belongs to science to seek for its causal explanation. However, there is a theistic explanation of that state and no scientific explanation. We thus have the kind of competition that naturalism rules out.[8] One might object that science knows that it cannot, on pain of vicious circularity, provide a scientific explanation of the first physical state, and hence it does not belong to science to seek that explanation. But this objection confuses the first physical state *qua* first and the first physical state as it intrinsically is. It does not belong to science to seek the explanation of the first physical state *qua* first. But if we just consider it as a specific physical state—particles and fields arranged thus-and-so—then it certainly belongs to science to seek for its explanation, though that search will be rightly abandoned if sufficient evidence is gathered that the state is in fact the first one.

Now, let's go back to Aristotelianism. Bracketing theism, the entity that might trouble the naturalist is form. But the forms of things are components of substances—humans, dogs, oaks, etc.—that are also found in our current science (e.g., biology) and are likely to be found in the completed version of that science. And there is neither competition(??what's that? remove here and earlier?) nor overdetermination between causal explanations provided by forms and their substances. As long as the substances do not have

[8]Matters are a little more complicated if there is no initial physical state, but we can apply the above argument to an initial segment of physical states: science cannot explain that segment but theism claims to do so, and yet it lies in the scope of science to ask for such an explanation.

spooky causal powers beyond the scope of science, adding forms to the ontology no more contradicts a plausibly defined naturalism than perdurantism or trope theory does.

It is, of course, open to the Aristotelian to suppose that some finite substances, whether natural like humans and oaks or supernatural like angels or ghosts, have causal powers of a sort that goes beyond the scope of science, though specifying what those powers would have to be like is difficult. But nothing in the applications given for Aristotelianism posited such powers. Thus as far as the argument of this book goes, there is no contradiction with a carefully specified naturalism with respect to finite things.[9]

[9] I think the most plausible place to look for a tension would be in examining human free will, but that goes beyond the scope of this book.

CHAPTER IX

Laws of nature and causal powers

1. Humean and pushy laws

1.1. Deterministic versions. There are two main views of laws of nature. On Humean views, we have laws simply in virtue of non-causal regularities of the behavior of objects in nature, and causation is then grounded in the laws. On pushy views, we have laws in virtue of metaphysical components of reality that affect or constrain the behavior of objects. In this section, I will argue that pushiness is the right view.

Introduce Lewis??

??causation

??argument from ideological parsimony

1.2. Problems with explanation. The most discussed problem with BSA is explanatory circularity. To introduce the problem, let us begin with an explanatory problem independent of BSA. The old Deductive Nomological (DN) model of explanation??refs in the philosophy of science had it that scientific explanations are deductively valid arguments for the particular fact to be explained, at least one of whose premises is a law. For instance:

(1) All massive objects attract gravitationally.

(2) The sun is a massive object.

(3) So, the sun attracts gravitationally.

Of course, it was soon seen that having this form is not necessary for explanation—statistical explanations do not fit the deductive schema.??ref It was also seen that having this form is not sufficient for explanation, as one needs somehow to account for the explanatory direction between the particular facts stated in the premises and the particular fact in the conclusion. A famous example here is how the laws of nature together with the

length of a flagpole's shadow yield a deductive argument for the position of the sun in the sky, but obviously the position of the sun in the sky is explanatorily prior to the length of the shadow.

But it is worth noting that even the seemingly innocent example (1)–(3) has a crucial problem, which suggests that explanations of the DN sort rarely if ever work. For consider that (1) is a universal generalization, and a universal generalization is true in part in virtue of its instances. That all massive objects attract gravitationally is partly grounded by the Andromeda Galaxy attracting gravitationally *and* partly by the sun attracting gravitationally. But grounding is a form of explanation. Thus, the sun's gravitational attraction is explanatorily prior to the universal generalization about massive objects attracting gravitationally. And yet the universal generalization is, according to the DN model, prior to the sun's gravitational attraction. But loops in explanatory priority are quite implausible.

The most common kind of response to these kinds of arguments is that explanatory loops are permissible when the explanation in the two directions is of a fundamentally different sort. For instance, in the above case, the sun's gravitational attraction is *nomically* or maybe *causally* explained by (1) and (2), while providing a partial *grounding* explanation of (1).??refs Nonetheless, it is not clear that it matters for the anti-loop intuition that the two directions of an explanatory loop are of the same kind. It is the circularity that seems to be the problem.

Furthermore, suppose you light a match in the kitchen. Then the friction applied to the match explains why something is burning in your house. But notice that this explanation consists of a chain of two explanations. The friction applied to the match in your kitchen *causes* or *nomically explains* (in conjunction with some appropriate laws) the match to be on fire in your kitchen, and the match's being on fire in your kitchen *grounds* the fact that something is burning in your house (particular cases ground existential generalizations). The causal/nomic and grounding explanations combine into a generic explanation that is neither just a causal/nomic explanation nor just a grounding explanation, but is nonetheless an explanation. This shows that there is a generic concept of explanation which goes

beyond particular types of explanation, like causal, nomic or grounding. And an explanatory loop with respect to generic explanation seems problematic.

Thus, even paradigmatic cases of DN explanations seem troubling. However, there is a plausible explanation for the initial intuitive appeal of DN explanations like (1)–(3). When we offer scientific explanations in terms of a generalization that is a law of nature, we are implicitly invoking the fact that the generalization *is* a law of nature.

To support this point, suppose that we live in a world with only one massive object, Bob, but where nonetheless it is a law that all massive objects bend spacetime. One might think that this thought experiment will be rejected by the advocate of BSA on the grounds that no universal generalization that has only a single instance can be a law. But this is not clear. First, a single instance can be extremely informative. Thus, many contemporary Humeans accept the "past hypothesis" which says that the initial entropy of the universe is low as a law, because it is so informative when combined with other laws, even though it is a hypothesis that applies to only one time. Second, on standard Lewisian BSA, the laws are the (nontrivial?) logical consequences of the best system. But a universal generalization that has only one instance can be a logical consequence of a law that applies to many other instances. For instance, it follows from the law of gravitation that all massive natural satellites of planets on which a philosopher was executed by being made to drink hemlock attract gravitationally, and hence this complex universal generalization is a law. But as far as we know, there is only one such natural satellite, earth's moon. Similarly, it could be that the fact that all massive objects bend spacetime is a special case of a more general fact about what kinds of objects bend spacetime, and the more general fact could have more than one instance.

Then in this world with one massive object to say that Bob bends spacetime because he's massive and all massive objects bend spacetime is silly. But it is not at all silly to say that Bob bends spacetime because he's massive and *by law* all massive objects bend spacetime, even if Bob is the only massive object. Of course, we do not live in a world where there is only one massive object. But it is plausible that the logical form of the

explanation of a particular massive object bending spacetime is no different in our world than in that one.

Now the hypothesis that nomic explanations invoke a law *as a law*, implicitly or explicitly, can help solve the explanatory circularity model facing DN explanations, by modifying (1)–(3) into:

(4) By law, all massive objects attract gravitationally.

(5) The sun is a massive object.

(6) So, the sun attracts gravitationally.

Whether this remains deductively valid depends on whether the "By law" operator has the property that from its being the case that by law p it follows that p (i.e., whether it is a modal operator satisfying Axiom T??). While this seems a natural assumption, it has been disputed.[1] We need not settle this question. Even if the argument is not deductively valid, it is a plausible ampliative argument form, and it is clearly explanatory.

But now note that the Humean cannot make use of this solution to the circularity problem. For, first, on a Humean analysis, just as the sun's attracting gravitationally was a partial ground of (1), it seems to also be a partial ground of (4), since we may suppose that it is because this universal generalization is true that it makes it into the best system.[2]

And, second, even apart from the circularity problem, explanations of the modified DN sort are problematic for BSA. As Salmon noted, while adding irrelevant facts to an

[1] Thus, van Inwagen thinks that a false universal generalization can still be a law, if the exceptions are miracles.??ref

[2] It is not *always* true on BSA that a universal generalization is a law in part because it's true. It might be that a universal generalization is a law because it is a logical consequence of the more fundamental laws that are axioms of the best system. But if that is the case for (4), then just need to replace the example with a more fundamental one or work in a world where the universal gravitational attractiveness of massive objects is a fundamental law, since for the fundamental laws, their inclusion in the system is explained in part by their truth, as the axioms of the best system must be true.

argument's premises does nothing to damage the argument, irrelevance damages an explanation.??ref But on BSA, the proposition that by law all massive objects attract gravitationally tells us that the proposition that all massive objects attract gravitationally is a logical consequence of the system of truths that best optimizes a balance of informativeness and brevity. But *this* fact does not seem explanatory. That there is a system that best balances informativeness and brevity and has as a logical consequence that all massive objects attract does not seem to contribute to an explanation of why a particular object attracts gravitationally. Facts about brevity of expression just do not seem relevant to explaining the movements of planets.

Imagine a family of worlds where massive objects attract gravitationally, but so do objects with a certain fundamental property Q. Suppose further that there is nearly total overlap beween massive objects and objects with Q, so that both:

(7) All massive objects attract gravitationally

(8) All Q objects attract gravitationally

are highly informative, and equally brief. But now imagine two worlds, w_1 and w_2. In w_M, the number of Q objects is N for some large N, and the number of massive objects is $N+1$, while in w_Q, the number of Q objects is $N+1$ while the number of massive objects is N. Thus, in w_M, it is (7) that is the more informative, while in w_Q, it is (8) that is the more informative. Suppose there are no other regularities that make it into the best system, and suppose that we include (7) in the best system for w_M, but adding (8) would give so little extra information (since there is only one object that has Q but no mass) that only (7) makes it in. On the other hand, for analogous reasons, the laws of w_Q include (8) but not (7).

The proposition (**??**) in w_M is then grounded in part in the fact that there is one more massive object than Q object in w_M (unlike in w_Q). But the fact that there is one more massive object than Q object is not explanatorily relevant to the sun (assuming the sun to exist in w_M) being gravitationally attractive. Yet that fact is explanatorily prior to (**??**), which in turn is explanatorily prior to the sun's attractiveness.

In summary, to avoid circularity, we need to invoke laws *as laws* in our explanations. But lawfulness on BSA consists of facts about truth, informativeness and brevity. The facts about truth ensure that we haven't escaped circularity, and the facts about informativeness and brevity are not explanatory. BSA is inadequate, thus, if laws are supposed to be explanatory.

1.3. Too much power. Now imagine a universe consisting of a single quantum "coin toss" performed a very large number of times. Observing some infinite sequences of these coin tosses, such as $HHH...H$, $TTT...T$, $HTHT...HT$ or $THTH...TH$, would make us confident that the coin tosses are deterministic. But even though this would make us confident of determinism (e.g., $HTHT...HT$ would make us confident that there is a law that each toss is followed by its opposite), any such sequence could also occur without determinism. Not so on our Humean story. On our Humean story, any one of these global regularities *logically guarantees* a deterministic law in a world consisting of a single coin tossed repeatedly. This is highly counterintuitive. Furthermore, suppose that we in fact have a "patternless" sequence that to the Humean yields indeterministic law that says that each coin toss is independent and fair. But while the statement that each coin toss is independent and fair should allow any finite sequence of coin tosses, it turns out that certain sequences, such as our four examples above, are logically incompatible with independence and fairness on our Humeanism.

Or consider magic. Suppose you live in a multiverse consisting of two physical universes, each of which has laws of nature roughly like ours. Now, plausibly, any state of a physical universe like ours can be encoded as a countable sequence of real numbers (i.e., a finite sequence, or one that can be enumerated using the natural numbers: $x_0, x_1, x_2, ...$). For instance, suppose the universe is made up of a countable cardinality of particles, each of which has a finite number of properties naturally expressible as one or a finite number of real numbers (e.g., mass and charge can be expressed as one number, and position can be expressed as three given a coordinate system), which properties either change continuously or have a countably infinite number of times of discontinuity. Then all we need

in order to fully describe the system is to specify the properties at a countable number of times (e.g., the times at which discontinuities happen and all times that can be expressed by a rational number), and the resulting description can be expressed as a countable sequence of real numbers.[3] But a countable sequence of real numbers can, with a well-known trick, be expressed as a single real number.[4] And if instead of a classical particle system one prefers a quantum story, then note that most models of quantum mechanics make the wavefunction be a vector in a separable Hilbert space—and the cardinality of the set of vectors in a separable Hilbert space is the same as the cardinality of the set of all real numbers. Adding a countable number of discontinuities in case of collapse, we can still encode the state of the universe as a single real number.

Now, any real number can be encoded into the tilt angle of a physical rod (e.g., encoding x into a rod tilted—with respect to some axis—at angle $\arctan x$). Thus, we can encode the complete physical state, over all of time of one of the two universes into the tilt angle of a physical rod in the other universe. Now fix some encoding that can be specified in a relatively brief way. Suppose now that you are in one of the universes, and at a precisely specifiable moment (say, one that is a precise number of Planck times since the

[3]*This argument uses the fact that if A and B are countable sets, then the Cartesian product set $A \times B$ of pairs (a, b) with a from A and b from B is also countable. To see this, note that if we can enumerate A as $\{a_0, a_1, a_2, ...\}$ and B as $\{b_0, b_1, b_2, ...\}$, then we can enumerate $A \times B$ as $(a_0, b_0), (a_1, b_0), (a_0, b_1), (a_2, b_0), (a_1, b_1), (a_0, b_2), ...$ The trick here is to first enumerate the one way where the indices add up to 0 ($0 = 0 + 0$), then the two ways where the indices add up to 1 ($1 = 1 + 0 = 0 + 1$), then the three ways where the indices add up to 2 ($2 = 2 + 0 = 1 + 1 = 0 + 2$), and so on.

[4]*Any real number can be remapped to the range from 0 to 1, exclusive, by taking x to $(1/2) + (1/\pi) \arctan x$. Any countably infinite sequence of numbers $x_0, x_1, x_2, ...$ between 0 and 1 exclusive can then be expressed as a sequence of decimal numbers where ith number is of the form $0.x_{i,0}x_{i,1}, x_{i,2}, ...$ (with the convention that an infinite terminal sequence of nines is preferred to an infinite terminal sequence of zeroes, say). One can then encode all these numbers into a single number of the form $0.x_{0,0}x_{1,0}x_{0,1}x_{2,0}, x_{1,1}, x_{0,2}...$ (this is basically the same pattern as in Note 3), so that every digit of every one of the numbers in our initial series occurs at a determinate position in the encoded number. And *a fortiori* if one can encode any countably infinite sequence of reals into a single real, one can encode any finite sequence of reals into a single real.

beginning of your universe) you tilt the rod at angle that happens to match the state of the other universe. Then including the information that the rod angle at this moment matches the complete state of the other universe will indeed provide a vast amount of information about your multiverse in a fairly brief compass, and hence would be included in the optimal Lewis-Ramsey description, and will be a law. Thus, by tilting a rod at a specific angle—and surely any rod tilt angle is physically possible for you (though maybe not possible to induce *intentionally*)—you can create a law correlating the rod angle with the other universe. Thus, by waving a rod, you can make the rod be a vastly informative dowsing rod that by law and not merely by coincidence carries complete information about another universe. This is highly implausible magic!

1.3.1. *A plurality of bestnesses.* A problem that will be familiar from many of our earlier discussions is that there are many free parameters in the account of the bestness of the best system. For instance, given a measure of informativeness $I(T)$ and length $L(T)$ of a system T, we will want to combine them into a measure of quality of theory $f(I(T), L(T))$ with some sort of value function $f(x, y)$ such that $f(x, y) < f(x', y)$ and $f(x, y) > f(x, y')$ when $x < x'$ and $y < y'$. But there are infinitely many functions satisfying these inequalities. Perhaps with some thought we can find some more reasonable constraints, but it is very implausible to think we can reduce the space of reasonable candidates to one.

Moreover, neither $I(T)$ nor $L(T)$ has a unique privileged candidate.

If a world has a phase space—say, defined by values of various natural determinables like charge, position and momentum at various times—it makes sense to think of the informativeness of a theory T as inversely related to the size of the set of trajectories through phase space (i.e., functions from time to phase space) compatible with T. If we could get the set of allowed trajectories down to one, that would be maximal informativeness.

But how do we measure the size of the set of trajectories compatible with the theory? The set-theoretic cardinality of the set of allowed trajectories for many theories that intuitively vary significantly in their informativeness will be the same. For suppose our determinables are position. Now consider theory T_1 according to which there is a single particle

which for all time is found in the same position in three-dimensional Euclidean space, and T_2 according to which the the particle moves through three-dimensional Euclidean space over a continuous trajectory. Clearly, T_1 is much more informative than T_2. But the cardinality of the space of trajectories allowed by T_1 is *the continuum*, the cardinality of the real numbers.[5] However, that is also the same as the cardinality of the space of all continuous trajectories, which is what T_2 allows.[6]

We might, on the other hand, try to define the size of the set of trajectories as a volume rather than a cardinality. An immediate problem may seem to be that of the choice of units of volume. This is not a serious problem for comparing degrees of informativeness, as long as we measure the volumes of the set of allowed trajectories using the same units. Even completely arbitrary units like the length of Charles III's forearm and the time between his mother's and his own coronations can be used for comparative purposes. However, the scaling issue is relevant for the combination problem. For unless our combination function f has some additional special properties, like being linear in the second variable (i.e., $f(x, \alpha y) = \alpha f(x, y)$), the choice of scaling may still be an issue, because it might be that $f(x, y) < f(x, y')$ but $f(x, \alpha y') < f(x, \alpha y)$ for some choice of x, y, y' and α. On the other hand, we might use the scaling worry to justify a linearity constraint on the second variable in f, thereby reducing the arbitrariness of the choice of f.

A more serious problem is cases of two theories that clearly vary in information content will both allow a set of trajectories with zero volume. For instance, suppose a world with only one moment of time, and one determinable: three-dimensional position of a single

[5] A position can be encoded as three real numbers. Any real number can be recoded to be between 0 and 1 (use the function $f(x) = (2/\pi)(\pi/2 + \arctan x)$. Any three numbers between 0 and 1 can be encoded in a single number. For instance, the decimal numbers with the digits $0.x_1x_2x_3...$, $0.y_1y_2y_3...$ and $0.z_1z_2z_3...$ can be encoded as the single number $0.x_1y_1z_1x_2y_2z_2x_3y_3z_3....$??check infinite nines.

[6] To specify a continuous trajectory, one only needs to specify its values at a countable number of times—say, all times that are represented by rational numbers. Thus one needs to specify a countable number of real numbers. But a somewhat more complicated interweaving allows that to be coded in a single real number.??ref and ??backref

particle. A theory that requires the particle to be at the coordinates $(0, 0, 0)$ reduces the space of trajectories to a subset of zero volume, as does a theory that mere constrains the particle to have the first coordinate zero. The first theory requires the particle to be at the origin and the second to lie in the yz-plane. But a point and a plane both have zero volume.

We might find a way of comparing sets of zero volume. For instance, we might try to find the Hausdorff dimension (which may fractional) of the two sets, and deem the set with higher dimension to convey less information. And then we could use Hausdorff measure to compare sets of the same Hausdorff dimension. But what would we do about sets both of whose Hausdorff measures are infinite? Moreover, the above approach leads to a pair of numbers, (α, β), where α is the Hausdorff dimension of the allowed subset of phase space and β is the α-dimensional Hausdorff measure. Thus, instead of combining length with a single number, we need to combine length with *two* numbers. And what do we do in the case of sets whose Hausdorff measure is infinite with respect to the Hausdorff dimension—these may need to be compared as well. Furthermore, Hausdorff dimension itself is not the only way to formalize the concept of dimension.???

Thus we should not expect a canonical measure of informativeness: there will be many free parameters.

Now, length may seem more tractable. Indeed, given a fixed language whose sentences are finite strings of characters from a finite symbol set, there is no difficulty in making sense of the length of the briefest expression of a system of propositions in that language.

But there are many languages. Just think of such decisions as the choice of grouping notation. In recent use in logic, for instance, we have parenthesis notation, Polish notation and dot notation. Surely there are many more reasonable grouping notation. Or think which logical primitives should be included. For truthfunctional connectives, nand is sufficient, as is nor, as is either of the pairs and-not and or-not, but why should we limit ourselves to a minimal sufficient set. Perhaps we should allow all the binary connectives. Or perhaps some but not all. Or perhaps all the ternary ones. Or perhaps just one seven-place nor. It is reasonable to think there is a privileged set of predicates—the perfectly

natural ones. But is it reasonable to think there is a privileged set of logical operators? This is murky. Or a privileged grouping notation? That seems even more dubious.

We thus have little reason to hope in a single distinguished measure of bestness. Now, we might hope that for a wide range R of measures of the quality of a theory in a world, there is sufficient overlap between the best theories to ensure that the things our best science will converge on as the laws will in fact be entailed by the theories that are best according to the measures in R. We could, then, define a law as a proposition L such that for all quality measures Q in R, the Q-best theory entails L. But specifying the boundaries of R will be subject to difficulties very similar to those in defining a single canonical measure Q.

It seems we cannot escape the idea that there is significant vagueness in the concept of a law for the Humean. Is this a problem?

In Section 1.2, I argued that in paradigmatic cases when universal generalizations enter scientific explanations, they do so as laws. In other words, that L is a law is itself a part of the explanation, and not just L itself. Whether this *always* so is not clear, but it seems clearly true that it is sometimes so. Nobody would bat an eye at a scientist saying that planets move in elliptical orbits because by a law of nature objects attract gravitationally in proportion to their mass and the inverse square of distance, while the sun's mass is much larger than the mass of any other objects in the vicinity and the kinetic energy of the planets is not too high, and by theorems proved by Newton an inverse square law of attraction produces elliptical orbits when the kinetic energies are not too high. Here a part of the story is that there is a law of nature.

If I am right about this, and if an account of what it is to be a law of nature is a very complex statement involving a measure of the quality of theories, and various linguistic facts about lengths of expressions, then our scientific explanations in terms of laws are much more complicated than we likely thought. While the concepts figuring *in* the laws may be very natural, the concept of a law is itself rather unnatural given BSA. Moreover, the plurality of bestnesses implies that there is a serious vagueness in the concept of a law,

which makes our scientific explanations, even in the most precise areas of fundamental physics, full of vagueness.

Suppose, on the other hand, that we follow the example of the old deductive nomological model of explanation and deny that laws enter explanations *as laws*. Still, the vagueness of the concept of law would affect what counts as an explanation, assuming that generalizations that are not laws are not allowed in explanations, at least not in the place where we would put a law. We end up damaging the objectivity of the concept of explanation on this approach: it becomes a linguistic question what is and is not an explanation. And if explanation is significantly relevant to justification—for instance, via inference to best explanation—we damage the objectivity of the concept of justification. This is perhaps less problematic than damaging the explanations themselves, as would be the case if the nomicity of the laws were a part of the explanations themselves, but it is still problematic.

Furthermore, our now familiar Aristotelian solution to issues of unacceptable vagueness, which is to ground boundaries in human nature, is not plausible in the case of laws of nature, because doing so would render laws of nature too anthropocentric. (Granted, however, we might be able to use the Aristotelian solution to resolve any infection of vagueness to the concept of justification, since our justification is appropriately anthropocentric.)

??refs on vagueness of laws

1.3.2. *Mind and causation.* It is interesting to note that the full Humean package where causation is grounded in laws and laws are grounded in a best-system analysis has a serious problem with any philosophy of mind on which certain causal relations are essential to mental events. This will include, first and foremost, functionalist account on which the pattern of causal interconnections *defines* mental events. But it will also include any view on which causal interconnections of a particular sort are *essential* to mental events. For instance, while the functionalist may think that what *makes* an event be a pain is that it is typically caused by damage and typically causes motivations to aversive behavior, one need not be a functionalist to think that a causal connection between an event and motivation is necessary for the event to be a pain. Likewise, plausibly, a necessary condition

for an act of inferring q from p is that a thought (say, a believing or an assuming) with content p causes a thought with content q. Finally, it is very plausible that for a wide range of mental events, an essential part of what makes the mental event be *mine* is that it is at least partly caused by me (or by my character, etc.) That I think entails that I cause an act of thinking.

But now notice that best-system laws are defined globally in terms of the four-dimensional arrangement of stuff in the universe. Whether some lawlike generalization does make it into the best system depends, in particular, on what the future is like. One reason is that typical lawlike generalizations tend to non-vacuously apply to events past, present and future, and so if the future were sufficiently different, the generalizations themselves would be false, and hence not in the best system. If some such generalization is needed for a causal relation, say between pain and aversive motivation, then if the future were such that the generalization would fail to hold, there would be no pain now.

The idea that whether there is pain now depends on what will happen in the future is itself highly counterintuitive. Additionally, it allows for implausible deductive predictions about the future. I am in pain now. My being in pain now requires a causal relation between pain and aversive motivation. Such a causal relation requires a law of a certain sort. A law of that sort, on BSA, requires the future to satisfy certain constraints. Hence, the future will satisfy these constraints. But it is highly implausible (and in tension with the historical Hume's scepticism about induction??refs) that we can thus reason *deductively* from present pain to contingent future arrangements of matter. And if introspection counts as *a priori*, this is *a priori* reasoning about contingent future facts.

There is a possible way out of this line of argument. While laws need to hold at all times, their truth need not put a constraint on what the future is like. For a law could be such that it vacuously holds of future times. Thus, we could suppose a law that energy is conserved prior to 2022, or that emeralds are green at all times t such that $t \leq t_1$, where t_1 is in fact the present. Many contemporary Humeans accept one such law, because they hold that it is a law that the initial entropy of the universe is low. Nothing in BSA rules out

the possibility of a law with temporal indexing. Temporal indexing carries a dual price: it reduces informativeness and makes the law less brief. But that price may be worth paying in some cases. Thus, a Humean could say that while my present pain requires laws that suffice to ground the causal connection between my present pain and aversive motivation, for all we know, these laws could be temporally indexed in such a way that they have only vacuous application in the future.

As far as it goes, this point is correct. However, the Humean is not off the hook for making deductive predictions about contingent future facts. For a system that includes such temporal indexing still has to compete for bestness with other systems that make non-vacuous statements about the future. And that the world's best system of laws is such as to ground the claim that my pain causes my aversive motivation seems very plausibly to put some sort of a non-trivial constraint on the future arrangement of matter. For instance, it is plausible to suppose that a certain kind of future arrangement of particles would generate elegant patterns that combine with the behavior of things in the present in such a way as to generate laws that beat out any laws that would ground a causal connection between my pain and my motivation. And so I have a deductive inference to the non-existence of such a future arrangement on the basis of my present pain.

Moreover, plausibly, the same intuitions that disallow deductive inferences about future arrangements of matter from my present phenomenal states should disallow deductive inferences about arrangements of matter outside the solar system from my present phenomenal states. But it seems very plausible that, given BSA, the existence of laws undergirding a causal connection between pain and motivation would have non-trivial implications for how matter is arranged outside the solar system. Laws grounding causal connections shouldn't be such as to be purely local (i.e., to have merely vacuous implications for how things are outside of a local region).

Perhaps our Humean could simply embrace the non-locality of the grounding of our phenomenal states, and allow deductive inferences of facts about temporally and spatially

distant arrangements of matter from our phenomenal states, *pace* the historical Hume. But the counterintuitiveness of this is a definite cost for the theory.

1.4. Indeterministic extensions. BSA was initially formulated for deterministic laws. But what about indeterministic laws, such as that some quantum setup has a chance $1/3$ of resulting in outcome A—or, for homey simplicity, that a certain indeterministic coin toss has chance $1/2$ of heads? The standard move is to go to a Probabilistic BSA (PBSA). In BSA, we required all the claims of the theory to be true. In PBSA, we only require non-probabilistic claims to be true. However, now, in addition to optimizing informativeness and brevity, we also optimize *fit*, where we check how well the outcomes fit probabilitic predictions. A theory which claims there is some large number N of independent fair (i.e., the chances of heads and tails are equal??backref for first use of 'fair') coin tosses will tend to better fit worlds where the number of heads is closer to $N/2$ than worlds where the number of heads is further from $N/2$. Thus, we optimize informativeness, fit and brevity over theories whose non-probabilistic content is true.[7]

1.4.1. *Violations of the Principal Principle.* ??chances vs probabilities

Suppose the world consists of some large number N of independent fair indeterministic coin tosses, where N is large and easily mathematically expressible, e.g., $N = 2^{256}$. As we would expect, very close to half of the coin tosses are heads, and we suppose that a probabilistic best systems view will have laws that correctly assign an independent chance of $1/2$ (or, if one wishes, something close to $1/2$, a complication I will ignore) to each toss. Moreover, because N is easily expressible, and expressing it conveys a significant amount of information about the world, the laws also state that the number of tosses is N.

Now, the following version of van Fraassen's Principal Principle is very plausible:

(9) If from a law U it can be proved that the chance of E is p, then $P(E \mid \mathrm{Law}(U)) = p$,

[7] If we want a bit more elegance in the formulation, we can drop the requirement that the non-probabilistic content is true and instead stipulate that a non-probabilistic statement has good fit when it is true but infinitely bad fit when false, so no theory with false non-probabilistic statements will win the crown of being the best system.

where $\mathrm{Law}(U)$ says that U is a law. Here is another very plausible thesis:

(10) If E and F are logically incompatible and F is logically possible, then $P(E \mid F) = 0$.

(The restriction to the case where F is logically possible is to handle the intuition that perhaps $P(F \mid F) = 1$ even if F is logically impossible.)

Now, let U be our law that says that the world consists of N independent fair indeterministic coin tosses. Let E_0 be the event of not getting any heads. Then it can be proved from U that the chance of E is $1/2^N$, so by (9) we have:

(11) $P(E_0 \mid \mathrm{Law}(U)) = 1/2^N$.

But on our probabilistic best systems analysis, it is logically impossible to have U be a law when no heads have ever occurred. Thus, $\mathrm{Law}(U)$ and E_0 are logically incompatible, and so by (10) we have:

(12) $P(E_0 \mid \mathrm{Law}(U)) = 0$,

a contradiction.

The point here is quite simple. Orthodox probability reasoning shows that it is possible but unlikely that we will have no heads given a large number of independent fair coin tosses, but our probabilistic best-systems analysis must categorically reject such a possibility.

One might think that assigning zero probability to things so incredibly unlikely is unproblematic. But we also have some other strange probabilistic results. In defining the laws, we are maximizing a balance of fit and brevity. If in our world half of the coin tosses are heads, then a law assiging chance $1/2$ to heads is likely the best balance. But suppose that the frequency of heads isn't $1/2$ but something very close to $1/2$. Then the law may still be that the chance is $1/2$, because a slight decrease in fit due to having a chance not quite matching the exact frequency might be more than offset by the gain brevity if chance $1/2$ is significantly more briefly expressible than the actual frequency, say "$1/2 - 3 \cdot 2^{-254}$".

Let's suppose N is very large but easily expressible (e.g., $N = 2^{2^{2^{2^2}}} = 2^{2^{65536}}$), and let $\mathcal{W}_N$ be the set of worlds with nothing but N tosses where there are no briefly expressible

and highly informative patterns other than those conveyed by the independence of the coin tosses, the frequencies and the number of tosses.

Let F_N be the set of all fractions between 0 and 1 with denominator N. Each member of F_N could be the exact frequency of heads in some world in $\mathcal{W}_N$.

Let $A_{N,1/2}$ be the subset of frequencies in F_N such that there is a world w in $\mathcal{W}_N$ in which there is a Humean law specifying the chance to be $1/2$. For simplicity, suppose N is even, so $1/2$ is a member of F_N. For N sufficiently large, we should expect $1/2$ to be a member of $A_{N,1/2}$. But for the reasons given above, we would expect some frequencies in F_N that are very close to $1/2$ to be in $A_{N,1/2}$ as well. If E_α is the event of the actual frequency being α, then on our Humean view we have:

(13) If α is not in $A_{N,1/2}$, we have $P(E_\alpha \mid \text{Law}(U)) = 0$.

On the other hand, very plausibly, if α is in $A_{N,1/2}$, then we have a real possibility of having frequency α on the assumption that the law is U, and so we would expect:

(14) If α is not in $A_{N,1/2}$, we have $P(E_\alpha \mid \text{Law}(U)) > 0$.

Frequencies that are sufficiently far from $1/2$ (such as the 0 in our previous example) are not going to be in $A_{N,1/2}$: the fit of these frequencies is too bad. The case of $\alpha = 0$ has already been discussed.

What is surprising, however, is that very likely there are some members α and β in F_N such that $\alpha < \beta < 1/2$ and yet α is a member of $A_{N,1/2}$ but β is not a member of $A_{N,1/2}$. For we can suppose that both α and β are *extremely* close to each other, and very close to $1/2$, but α is much more briefly expressible than β, so that although the fit of a chance $1/2$ law is slightly poorer for worlds with frequency α than for worlds with frequency β, the greater brevity of an expression for α makes a law giving the chance of heads as α have a better balance of fit and brevity in a world with frequency α than a law giving the chance as $1/2$, while $1/2$ has a better balance of fit and brevity in a world with frequency β than a law giving the chance as $1/2$.

I cannot give precise examples here, because we do not actually have the measures of fit in hand and estimating the brevity of the shortest expression of some number is often a very difficult task. But suppose I am right. Then by (13) and (14), some frequencies in F_N that are further from $1/2$, such as α, will have a higher conditional probability on U than some frequencies that are closer to $1/2$, such as β. I imagined that the counterintuitive result that E_0 has zero conditional probability on U fails might be given the "excuse" that U entails a chance that is so tiny that we might as well take the probability to be zero. But that excuse won't work for E_β getting zero conditional probability on U. For if we say that U entails such a low chance for E_β that E_β deserves conditional probability zero, then E_α should get conditional probability zero as well, since E_α has an even lower chance on U.[8]

??refs

1.4.2. **Chance and propositions.* There is a variety of slightly different accounts of chance that can be given on PBSA. Let "the mosaic" refer to the arrangement of properties that are systematized in the best system. Here are a few options, depending on whether one takes conditional or unconditional chances to be fundamental:

(15) Event E has unconditional chance p if and only if the mosaic's best system entails that E has chance p.

(16) Event E has unconditional chance p if and only if the mosaic's best system conjoined with a complete specification of the mosaic's initial conditions entails that E has chance p.

(17) Event E has conditional chance p on C if and only if the mosaic's best system entails that the conditional chance of E on C is p.

There is room for much technical discussion of the details here, but notice that each of these three accounts is viciously circular: it is attempting to define a chance (conditional or not) by a definition that itself makes use of the concept of chance.

[8]For a large number of tosses, the chance of getting a particular frequency x of heads has an approximately normal distribution centered on $1/2$ if the true chance of each independent toss is $1/2$, and hence drops off as x moves away from $1/2$.

There is a complicated way out of this difficulty that modifies the PBSA.

Suppose that we have a language $\mathcal{L}$ whose vocabulary can express all the fundamental physical concepts acceptable to the Humean, plus which has one more function symbol, Chance(x). This new function symbol is uninterpreted, and I stipulate that sentences using that function symbol are neither true nor false. Let M be some appropriate set of mathematical axioms that include the correct axioms of set theory (perhaps indeed they include all truths of set theory) as well as the right axioms of probability formulated using the chance symbol (so these will be axioms of unconditional or conditional probability, depending on whether the chance symbol is unary or binary). Say that a theory (a set of sentences) T in $\mathcal{L}$ is *non-false* provided that no false sentence can be formally proved from $T \cup M$. This is equivalent to saying that all the provable consequences of $T \cup M$ that do not use the chance symbol are true. We can now specify that the best system is the non-false theory T that optimizes brevity, fit and informativeness.

Defining informativeness is difficult, but we might define it by looking at two kinds of information conveyed by the provable consequences of T. First, we have sentences that do not make use of the chance symbol. The informativeness of that part of the consequences will be measured much as in a non-stochastic BSA. The informativeness of a statement of form Chance$(E) = r$, where E does not make use of the chance symbol, will be null if $r = 1/2$, and while if $r > 1/2$, it will increase in proportion with both $r - 1/2$ and the informativeness of the non-probabilistic claim that E occurs, and if $r < 1/2$, it will increase in proportion with $1/2 - r$ and the informativeness of the non-probabilistic claim that E does not occur if $r < 1/2$. We might for simplicity ignore claims using the chance symbol not of the form Chance$(E) = r$. Of course, we will have to beware of overlap in information between the chancy and non-chancy sentences, and so on.

Finally, we measure fit by looking at all the provable consequences of the form Chance$(E) = r$ where E does not contain another instance of the chance symbol, and saying the fit of each is a measure of closeness between r and the numerical truth value

of E (falsehood being zero and truth being one)—perhaps a measure according to some standard scoring rule.??refs

The details are doubtless very difficult, but that's to be expected.

Given all this, we get a best system T consisting of sentences some of which fail to express a proposition. Given that laws of nature are not linguistic entities in a particular language, and that they are factual, and hence capable of being true, we surely cannot take a set of uninterpreted sentences like T in $\mathcal{L}$ to be a system of laws.

But we can now make a further move. Let B_T be the proposition that all the sentences in T are provable consequences of the best system. Then B_T is indeed a proposition and factual. We can then say that the laws of nature are all the logical consequences of B_T. The laws, then, are not the sentences derivable from the best system, since these include uninterpreted sentences, but rather the laws are the consequences of the claim that these sentences are sentences of the best system.[9]

One technical problem of this approach lies on the probability side. Suppose we are in a world with a fair coin sequentially tossed a large finite number N of times so set up that if T is the best system then $T \cup M$ proves a sentence saying that the coin was indeed tossed sequentially N times and proves the Chance$(H_1) = 1/2$, where H_1 is the event of the first throw being heads.[10] Then B_T will be a law of nature. In the notation of Section 1.4.1, it will be law that the frequency of heads is a member of the set $A_{1/2,N}$. This set contains $1/2$, and contains some frequencies in F_N very close to $1/2$, but also excludes frequencies just as close to $1/2$ as some of the included ones, on the grounds of that some of the excluded ones are so briefly expressible (e.g., $1/2 + 1/2^{256}$ in a world with $N = 2^{256}$) that the brevity cost of the greater fit is worthwhile. This kind of law of nature looks little like the kinds

[9] An alternative would be to consider the logical consequences of the proposition that T *is* the best system. But that would be less satisfactory. For it seems that one could have a world with all the laws we have *and* more. But if it is a law that T is the best system, then it seems there couldn't be a world with additional laws beyond those in T.???

[10] This will require N to be a number that is sufficiently briefly expressible to make it into the best system.

of laws scientists posit. We do not have laws of nature allowing and disallowing events based on the brevity of the expressibility of facts about these events.

1.4.3. *Non-Humean chances.* There is another approach to chances, however, than PBSA. Suppose that we take chances themselves to form part of the Humean mosaic, perhaps being fundamental properties, rather that being something defined via patterns in the mosaic. Then a non-deterministic theory can stick with BSA. The requirement that chancy statements be *true* now is cashed out in terms of the chances in the Humean mosaic. It is now quite possible for the best system to say that the chance of heads is $1/2$ even if the coin never comes up heads. For we no longer optimize fit, but informativeness. And the claim that the chance of heads is $1/2$ might well be quite informative, since it could tell us about the stochastic properties of a vast number of coin tosses—even if it does not give us useful information about the outcomes of these tosses.

From the Humean point of view, the down side of this is that chances are too much like causal propensities, which the typical Humean wants to reduce to patterns in the mosaic. After all, if our account of chance is not based on frequencies, then it seems that our best option for what it means to say that the chance of heads is $1/2$ is that the immediate cause has a propensity of $1/2$ to yield heads.

Thus, if we allow chances into the mosaic, we probably should allow causation as well. The resulting BSA has a richer mosaic, and is harder to refute by counterexample. The problem that coincidental correlations become rigidified into magical laws??backref is less pressing once it is clear that the magical laws are not causal laws—for causation is now a part of the mosaic, and hence mere correlations do not yield causation. However, we still have the problem of the plurality of bestnesses. And because we have so significantly enriched the mosaic, the main reason to believe the BSA is weakened. For that main reason is ideological parsimony: the elegant fewness of fundamental concepts. Once we have introduced causation and causal propensities, one of the great claims to intellectual power on the BSA side—giving an account of causation??—is gone. And since causal facts can be

used for explanation, the explanatory difficulties for BSA are mitigated, since at least some causal explanations become available.

The downside, however, is that at this point we do not have the main motivator for Humean views: an extremely limited ideology, and we still have a number of the difficulties.

2. Aristotelian laws

The most common Aristotelian approach to laws is to ground laws in the essential powers of substances, which in turn are grounded in the forms. If an electron is a substance, then the form of an electron gives it the power to attract positively charged things and repel negatively charged ones, and such facts ground the laws of electron behavior. And if a rabbit is a substance, then the form of a rabbit gives it the power to reproduce with another rabbit, which grounds laws of rabbit behavior.

This makes laws explanatory, without introducing any additional entities beyond the forms that we already need to explain the normative side of nature. Indeed, the integration of the normative and explanatory is central to the Aristotelian optimism on which things typically act rightly (we might even speculate that in the case of some things, such as subatomic particles, they *always* act rightly), and hence we can defeasibly infer ought from is.

But there are some complications. The first is that classical Aristotelianism is committed to the principle that substances do not have substances as parts.(??cross-refs, ??refs) But then if a rabbit is a substance, the electrons in it are not substances. Instead, the electrons in the rabbit may be accidents of the rabbit, or "virtual objects" constituted by bundles of rabbit causal powers. If so, then the laws of electron behavior will be grounded differently for different electrons. The behavior of electrons in humans and those in rabbits will be grounded by the human and the rabbit forms, respectively. Free-floating electrons, on the other hand, would seem to be self-standing substances, and their behavior will

be grounded by electronic substantial forms. On this view, the grounding of the laws of electron behavior appears quite complex.

The easiest way to remove this complexity is to allow substances to be parts of other substances. In this book, however, I am trying to remain as neutral as I can on the details of various robust Aristotelian ontologies, so it is worth seeing what can be done without doing this.

Certainly, forms of different beings can (and presumably do) common patterns of behavior. Nonetheless, we still have a deep puzzle here. Why is it that forms of so many things exhibit such commonalities? Why do oak and rabbit and human electrons behave the same way? To give a partial answer, we might posit a "fine structure" in the forms, whereby the oak, rabbit and human forms (and those of standalone electron substances) have a common component coding for the behavior of electrons. This common component could be found in a genus that the forms are specifications of. Or we might simply suppose that the common efficient cause or causes of all of these substances—say, God, or the particles in the early universe—have so acted as to produce beings with a variety of forms all of which code for common behavior of electrons in them.

There are, however, two kinds of laws that pose a special problem for the Aristotelian approach. One kind is high-level structural laws that depend on all physical things having a common pattern to their behavior. The conservation of energy is a paradigm example. While the forms of electrons and photons (and/or of larger substances containing them) can make it a law that energy is conserved in the interactions of electrons and photons, it is logically possible to have physical objects that interact with, say, photons in a way that does not conserve energy. The law of conservation of energy appears to be partly grounded in the fact that there in fact are no such things—that all physical objects have a form that satisfies the constraints of the law of conservation of energy. If $F_1, ..., F_n$ are all the forms of existing physical objects, then $F_1, ..., F_n$ are not sufficient to ground the conservation of energy without some additional metaphysically contingent fact such as that there are no physical objects with a form not among $F_1, ..., F_n$.

A second kind of problematic law is exemplified by the Second Law of Thermodynamics. The Second Law of Thermodynamics has presented a puzzle given that the physics relevant to the behaviors of the particles whose joint entropy tends to increase is invariant under a time-reversal. The usual contemporary view is that what explains the increase of entropy is a physics that can be time-reversal-invariant combined with low-entropy initial conditions. For given low-entropy initial conditions, entropy, as it were, has nowhere to go but up—there are a lot more possibilities for transitions from low to high entropy states than from low to low entropy states.??ref But, again, this law is not grounded in the natures of the interacting objects, but in those natures as combined with the metaphysically contingent fact that we started in a low-entropy state. We might call laws that depend on special boundary conditions "impure laws".

The two families of laws—the high-level structural laws that transcend the natures of particular objects and the impure laws that depend on boundary conditions—are ideal candidates for BSA. BSA does not care whether the laws depend on boundary conditions or not—just on how elegant and informative they are. But they do create problems for Aristotelian accounts, and more generally for other "pushy" accounts of laws??backref. For it does not seem that any kind of "pushy" law forces the initial conditions to have a low entropy. And while a "pushy" law can move things around in our world, it is somewhat mysterious how it can keep the kinds of things that would have the power to violate the structural laws out of existence.

There are several options for the Aristotelian at this point. The first is to bite the bullet, and hold that the structural and impure lawlike generalizations are not actually laws of nature. This is fairly plausible in the case of impure laws, but less so in the case of structural laws.

The second option is to notice that at least some of our problematic laws can be accounted for as follows. We suppose that in fact none of the physical substances in existence have a causal power to produce a violation of the law directly or indirectly, where indirect

production of a violation would be the production of a causal chain leading to such a violation. In short, there are no physical potential violators of, say, the conservation of energy, and no physical potential producers of such violators. We then say that a lawlike fact such that nothing physical is a potential direct or indirect violator of it is a law. We still need some account of what makes a fact lawlike. One move would be simply to defer to the best version of BSA: a lawlike fact is one that would be a logical consequence of the best system. And, finally, we just bite the bullet on any generalizations that cannot be accounted for in this way: these we just do not account to be laws.

This may seem to involve us in all the difficulties of BSA. But that need not be the case. First, we might only make the move for non-stochastic laws, denying that structural or impure generalizations that are stochastic in nature (such as the Second Law of Thermodynamics) are laws of nature. Second, the explanatory difficulties of BSA disappear in our present context. For laws grounded in this way are explanatory simply in a causal-privative way. We explain, say, why a violation of conservation of energy did not happen by saying that there was nothing physical with the causal power to generate such a violation. This is a genuine explanation independently of whether the conservation of energy is a law. The explanatoriness of the laws here does not fllow from BSA. On this story we might happily embrace the plurality of accounts of bestness, because the laws explain simply in virtue of their content, not in virtue of their being *laws*.

The third option is to embrace theism, and allow for two kinds of laws: laws grounded in the essential powers of physical substances and lawlike generalizations about physical substances intended by God to hold. This will be a partial occasionalism: some explanations in terms of laws will go back to created substances, but others will go back to God (and presumably some cases will be explained by a combination of the two).

CHAPTER X

Evolution, Harmony and God

1. The origin of the forms

1.1. Evolution and forms. We have good empirical reasons to think that the variety of biological structures that fills our planet is largely or completely the product of unguided variation together with natural selection. However, as I have argued, there are good philosophical reasons to think that the organisms with these structures have normatively laden forms which specify how the organisms should behave, endow them with the causal powers that make that behavior possible, and impel them towards that behavior.

It is implausible to think that the forms supervene on the biological structures. For instance, one theory of the evolution of wings for gliding is that small wings are useful for heat dissipation. Larger wings allow for more dissipation of heat, but are also more expensive for the organism to maintain. However, at around size at which the heat-dissipation benefits are outweighed by the maintenance costs, the wings also become useful for gliding. It is plausible that a species A that has the smaller wings has them with the telos of heat dissipation. But a species B that has evolved the larger wings has them with the telos of gliding, either instead of or in addition to heat dissipation.??ref,check But we can now suppose a member of B whose wings are defective and only good for heat dissipation. Such a member's biological structure might be largely indistinguishable from that of a normal member of A, and yet it is normatively different: such wings are defective in B but entirely appropriate in A. If these norms are grounded in forms, it seems there is a different form in members of B than of A.

In general, in the evolutionary process, we expect small transitions in genetically-based biological structure between parents and children, with no change between the parent's

form and the child's form. For if we had constant change between the parent's form and the child's form, our best account would be that the form simply matches the biological structure, which would not allow for genetic defects, and yet genetic defects—deviations of genetically-based biological structure from the kind norms—are clearly possible. Moreover, it is important to our ethics that all human beings, despite a wide variety in physical and mental endowments—including the striking biological difference between male and female—are beings of the same kind.

We thus need an explanation of why it is that at certain apparently relatively rare and discrete points in the evolutionary sequence we have a new form on the scene. This itself yields Mersenne questions: while some transitions of form might happen to coincide with a particularly striking genetic transition, we expect a number of them to come along with only minor genetic transitions, seemingly at arbitrary positions. What explains these transition points?

Hitherto in this book, such questions were answered by invoking the forms themselves. And this can be done in this case as well. We might suppose that the form of species A endows the members of A with a causal power to generate new members of A in some circumstances, together with new instances of the form of A, but also a causal power to generate new members of B in other circumstances, along with new instances of the B form. The difference in circumstances could be determined by the DNA content in the gametes joining together, so that when a descendant is going to have such-and-such DNA contents, the descendant gets the form of A, but with other DNA contents, the descendant gets the form of B.

This story requires the forms to contain intricate specifications of which form is generated when. Granted, the slew of Mersenne questions we have already raised should make us circumspect about balk at mere complexity of form. But now observe that the story as given above requires that the first biological organism on earth—presumably some simple unicellular or maybe even proto-cellular?? organism—contain within it a form that codes for the causal power to produce forms of all possible immediate descendants of it. These

immediate descendant forms then would have to code for the causal power to produce all their possible immediate descendants, and so on. Thus, the form of the first and simplest organism would implicitly code for all the forms of life that would ever actually be found on earth, and indeed all the forms of life that *could* ever descend from it.[1] We thus have here a dizzying complexity.

???few species story!??? no help, still have complexity

But the problem does not stop here. For we can now ask where that immensely sophisticated form of the first organism comes from? If we say that it comes from the causal powers of non-living substances, such as fields or fundamental particles, then we have to posit an even greater complexity in the forms of these non-living substances. The result would be highly counterintuitive, by supposing non-living things to have immense sophistication of form. Further, however, we would need a story of where the first forms arose from. If we take the above account to its logical conclusion, then at the Big Bang we would already have particles or fields whose forms implicitly included the vast formal complexity of all physically possible living organisms. And this in turn yields a powerful design argument. For the idea that such complexity would simply come about for no reason at all is utterly implausible.

Thus, the story that forms contain the rules for the generation of future forms points towards a being whose own power is sufficiently great to generate such forms. And to avoid a vicious regress, such a being would need to be a necessary one.

But note that once we have accepted the existence of a necessary being that is the ultimate source of the varied forms in our world, we can now tweak the story to avoid the implausible idea that unicellular organisms implicitly code for the forms of elephants and unicorns. Instead of supposing that the transitions between forms corresponding to certain selected changes in genetic structure are caused by the parent forms, we can suppose that

[1] It is tempting to say that the number of possible descendant forms is infinite, but that is not clear. After all, there could be some physical limit to the size of the genetic code fo a biological organism given our laws of nature. But in any case, finite or not, the number of possible descendant forms is incredibly large.

the necessary being is directly responsible for the transitions of forms. On such a view, the form of a unicellular organism might only endow its possessor with the ability to generate a descendant of the same kind, and the necessary being would directly produce any new forms when it is appropriate to do so.

1.2. Reasons for creating forms. Of course, this would lead to the question of *why* the necessary being produces new forms when it does so. Here, taking the necessary being to be rational can help. For there can be good value-based reasons for the transitions to fall in some places rather than others.

Consider, first, an odd thought experiment. A horse-like animal comes into existence with an maximally flexible form such that whatever the animal does fulfills the norms in the form. To eat and grow is one proper function, and to starve and produce a corpse is just as proper a function. Whatever our "flexihorse" does or undergoes is equally good for it. But there is something unsatisfactory about the flexihorse as a creature. If whatever the flexihorse does is equally good for it, then the fact that the flexihorse flourishes is just a direct and trivial consequence of its externally imposed form rather than the individual's *accomplishment*.

Reflection on this suggests there is a value in creating organisms that can fail to fulfill their norms. This value might be grounded in the forms themselves: it might be that real horses, unlike flexihorses, have self-achievement of flourishing among the proper functions in their form. And there is a value in creating organisms that have additional types of good written into their form, including such self-achievement. Alternately, one might hold that in addition to kind-relative goods, there may be kind-independent goods—perhaps grounded in imitation of the creator??forwardref?—and self-achievement of flourishing could be one of these.

Either way, a rational being creating organisms has reason to create organisms that can fail to achieve their form, and hence has reason to create beings with less flexible norms. Moreover, there appears to be a comprehensible value—again, either kind-relative or kind-independent—in production of beings of the same kind. As an intuition pump here, think

of the *Symposium*'s idea that the yearning for eternity is exemplified in animal reproduction. Thus, we can give a value-based explanation for why a necessary being would create beings in discrete kinds, with norms that the beings need not live up to.

2. Explaining harmony by natures and evolution

??:won't work for some a priori intuitions??

[2]

2.1. Number of natures.

2.2. Nomic coordination.

2.3. Fit to DNA and niche.

??ethical, organic and epistemological

2.4. Nature zombies. Aristotelian metaphysics allows for a curious hypothesis: nature zombies. Nature zombies are macroscopic entities that have the same physical structure as real organisms—say, oak trees or humans—but have no nature as whole (their physical constituents may have physical natures). A nature zombie would lack mind, and would have no intrinsic normativity (unless there is some in their physical constituents), and above all would not even be a substance, but a mere heap of constituents.

We can ask several questions about nature zombies. First, there is an epistemological one: how can I tell that all the apparent organisms around me aren't nature zombies? (I can tell I am not one, because I can tell that I have a mind.) Second, assuming it is indeed so, why is it that there aren't any nature zombies around? Third, and perhaps most deeply, why isn't it the case that *every* apparent organism is a zombie—i.e., why are there any macroscopic things with natures?

The epistemological question is easily handled in the same way that other skeptical hypotheses are. The Aristotelian can simply say that it is a part of our nature as the specific kind of reasoners we are that we should dismiss skeptical hypotheses. We don't need to

[2]??This section owes much to discussion in my mid-sized objects seminar, and especially to Christopher Tomaszewski's suggestions on the explanatory powers of forms.

(and couldn't) tell the real organisms from the zombies to justifiably think the General Sherman Tree, Seabiscuit and Biden to be (or have been) real organisms. We just need to be able to tell the organisms from the rocks and the like, which in these three cases is easy (though it is appropriately hard if we turn to viruses).

What about the explanatory questions, however? It seems surprising, after all, if a bunch of physical constituents make an organic substance with a rich normatively laden form. Given the fact of abiogenesis—that life arose from non-life about 3.5 billion years ago—we might expect to have a nature zombie world. The zombies could be expected to evolve just as well as the normatively-laden organisms that (assuming the arguments of this book are sound) we have around us.

An Aristotelian move would be to suppose that the physical constituents of the universe themselves have natures, and it could be that their natures have the causal power, when the physical constituents are rightly arranged, to produce a living substance, and lack the causal power to produce a zombie. *Prima facie*, the chemicals in a primordial soup could produce one of three outcomes: a soupy mess, a zombie organism, and a real organism. But it could be that their causal powers are so restricted that a zombie organism is beyond their power—it must be either a soupy mess (the typical outcome of mixing the chemicals) or a real organism with form and all. And the organism, in turn, could be the common ancestor of all the other organism-like entities. ???

2.5. Exoethics.

2.6. Aquinas' Fourth Way and the good. Aquinas' Fourth Way??ref puzzles the modern reader. It begins with a principle that comparisons between degreed properties are grounded in a comparison to a maximal case: one is more F when one is more like the item that is maximally F. Aquinas then illustrates the principle with the case of heat and fire: an object is hotter provided that it is more akin to the hottest thing, namely fire. He then applies the principle to goodness, and concludes that there is a best thing, and this is God.

The fire illustration is not just unhelpful to us, since we know that fire is not the hottest thing (the sun is almost twice as hot as the hottest flame), but it is actually a conclusive counterexample to the degree property principle, since we can easily compare temperatures without reference to an alleged hottest object.[3]

So Aquinas' comparison principle is false. But I contend that there is still something to his argument when applied to the good.

Now, a form-based metaphysics gives a powerful account of the good for a being of a particular kind—an oak, a sheep or a human, say—in terms of its match to the specifications of the form. It also gives a ground to comparisons between the good of different instances of the same kind: a four-legged sheep is, other things equal, better at sheepness than a three-legged sheep, because it more completely fulfills the specification in their ovine nature. In fact, this is itself a counterexample to Aquinas' comparison principle, in that we can compare degrees of success at sheepness without supposing any individual sheep to be perfect.

However, in addition to value comparisons within a kind, there are ones between kinds. When Jesus says that we are "worth more than many sparrows" (Mt. 10:31??ref), what he says is quite uncontroversial. Indeed, even a perfect sparrow seems to have less good than a typical human. While the nature of a sparrow will enable value comparisons between sparrows, and that of a human between humans, we still have the question of what grounds the value the difference between sparrows and humans. Some Aristotelians reject cross-kind value comparisons as nonsense.??refs But given the intuitive plausibility of many such comparisons, this rejection is a costly one.

[3]In any finite universe, presumably there will be a hottest object. However, temperature comparison is not defined by that object, since even if Bob is in fact the hottest object, we would expect it to be physically possible to have a hotter object than Bob. But if degrees of heat were defined by closeness to Bob, it would not be possible to be hotter than Bob, since nothing can be closer to Bob than Bob.

Aquinas' Fourth Way is not infrequently seen as more Platonic than his other arguments for the existence of God, and Plato indeed had a solution to the problem of cross-kind comparisons, by talking of differing degrees of imitation of the Form of the Good, which itself is perectly good. Plato, on the other hand, lacked a satisfactory solution to the problem of intra-kind comparisons. He may well have thought that there was a Form of Humanity??refs, which exemplified humanity perfectly, so that similarity to the Form of Humanity would define how good one is at being human. However, we can see that this solution is clearly unsatisfactory. First, the Forms are immaterial, so the Form of Humanity is immaterial, and hence it lacks fingers. Thus, the fewer fingers a human has, the more they are like the Form of Humanity, and hence, absurdly, the more perfect they are. Second, if somehow the Form of Humanity ends up having body parts, then the Form of Humanity either has an even number of cells or an odd one. But clearly neither option is more perfect than the other.

Central to Plato's solution to cross-kind value comparisons is the self-exemplification of the Form of the Good: the Form of the Good is itself maximally good. But a similar self-exemplifying Form cannot be used to account for intra-kind comparisons. Aristotle, on the other hand, has the non-self-exemplifying forms immanent in things. The Aristotelian form of humanity specifies human perfection, but does not do so by exemplifying it. It has neither fingers nor cells, but it *specifies* that humans should have ten fingers while specifying an age-dependent normal range of cell numbers rather than a specific cell count.

Notwithstanding the general falsity of Aquinas' comparison principle for degreed properties, Aquinas provides us with a plausible extension of the Aristotelian system to allow for comparisons of degrees of good between objects of different kinds in terms of the similarity to or degree of participation in a maximally good being, a divine being that plays the role of a self-exemplifying Form of the Good. The human being participates in God in respect of abstract intellectual activity, Aquinas will contend, while sparrows do not, and in that important respect, at least, humans are more like God. On the other hand, the sparrow's movements approximate divine omnipresence better than the stillness of a

mushroom does, and in that respect at least the sparrow is superior to the mushroom. We have, thus, a ground for something like a great chain of being.

There are still difficulties here. While the human is superior in intellectual activity, the sparrow moves around with greater three-dimensional freedom. How can we say that the human is superior all things considered? Where we previously had a problem of cross-kind comparisons, we now have the problem of cross-attribute comparisons. Intuitively, the human's intellectual superiority to the sparrow trumps the sparrows motive superiority to the human, and enables us to say that the human is more perfect on the whole. This higher level question is difficult indeed.

But there is some hope in thinking that in attributing different divine attributes we sometimes express divinity to different degrees. It may be that there is no meaningful comparison between how well we express divinity by saying that God is all-knowing versus by saying that God is all-powerful, saying that God knows the multiplication table up to 10×10 expresses divinity less well than saying that God can create any possible physical reality. I suggested earlier that motion imitates divine omnipresence. Thus, the sparrow's ability to fly imitates God's presence throughout several kilometers surrounding the surface of the earth, while the human's more limited mobility imitates God's presence in a thin two meter shell of air surrounding that surface. But the degreed difference between the two divine attributes—each a limited special case of omnipresence—imitated here might well be trumped by the fact that the sparrow does not imitate God's abstract intellectual activity *at all* while the human does imitate that activity, and does so in respect of a very wide scope of things (the human can think abstractly about the whole universe, for instance).

In ??backref, we gave a non-theistic Aristotelian sketch of a three-step great chain of being. The account here has a hope of allowing one to fill in more intermediate links. Even if the details in the comparisons between different attributes or respects do not work out, we still have an advantage for the theistic Aristotelian in being able to make cross-kind comparisons under specific respects, like motility or intelligence.

??ref:Jeffrey/Ward

2.7. Epistemology of normativity and form. [Argument: If a guided missile has form, it's alive by the Ch?? account of life. But it's not alive. So it lacks form. Is this a bad argument???]

2.8. Ethics and happiness.

2.9. Modern technology and outlandish scenarios. In ??backref, I argued that an ethics based on human form can simply ignore outlandish scenarios that are far outside of our ecological niche, such as ones involving infinite numbers of beneficiaries. However, there is a danger in this line of reasoning. As Arthur C. Clarke famously said, "Any sufficiently advanced technology is indistinguishable from magic."??ref To human beings 50,000 years ago (or even just 500 years ago!) much of our technology would indeed be magical, and decisions that we routinely need to make, say in bioethics, would be predicated on outlandish assumptions.

We might thus expect an ethics and epistemology grounded in a form possessed by hunter-gatherer primates to be silent on dilemmas of a highly technological society, leaving us to do whatever we wish, or, even worse, to fail to harmonize with the shape of our lives, like that of a fish on land. Yet while there are, as there have always been, difficult and controversial moral and epistemological cases, we do not in fact find ourselves adrift without guidance in the modern world. Virtue continues to contribute to our flourishing, and ancient texts, whether religious or philosophical, continue to point to good ways of living.

This gives us reason to think that if our moral norms are grounded in human nature, human nature was somehow picked out with foresight for what kinds of challenges humans would face in the distant future. Our ethics does may not work in outlandish situations, such as those involving infinities as noted in ??backref, but it works in a broader range of moral environments much broader than that found in early homo sapiens society.

Thus, the theistic version of our natural law theory both accounts for the apparent unsatisfactoriness of our ethics in situations that humans apparently never find themselves with and the applicability in situations across a very wide range of situations, wider and more technologically varied than the natural environment of other animals. This kind of foresight points to a foreseer, indeed a designer, and hence towards a theistic version.

The move I suggested in ??backref for outlandish scenarios, namely that our ethics and epistemology simply does not apply to them, may seem problematic given that we live in a world where many things that our not-too-distant ancestors would have seen as outlandish are real. We fly regularly around the world and irregularly to the moon, speak with people on the other side of the planet, move organs from one person to another, make cats glow by inserting jellyfish genes, program machines to have conversations with us, have bombs that can wipe out most megafauna including humans, and can clone at least embryonic humans. Our capabilities and the situations that we are in are quite different from those we evolved for. And further changes may be facing us. Many think there is a serious possibility that human beings will spread through the galaxy, affecting vast numbers of lives, which may make actual seemingly outlandish questions about where our actions have very slight probabilistic effects on vast outcomes.??ref:Fanaticism,glitchy ethics??backref and discuss:ch4 on infinity

One approach to these modern questions is a principle-conservativism: the questions are settled by moral principles that we accepted for millenia. Because the default for an action is moral permissibility, in the case of qualitatively new kinds of actions, principle-conservativism is apt to lead to a radical expansion of moral possibilities. If we had no principles governing human DNA manipulation in the past, even if this was simply because we had no concept of DNA, now there are no limits, except limits coming from traditional principles of harm and consent. Principle-conservatism in these kinds of cases would paradoxically justify a vast change in the shape of human lives of a sort that arguably is not compatible with human flourishing. Conversely, however, we have cases like the invention of effective therapeutic surgery. Prior to these, any serious degree of

cutting open of the human being would be an instance of grave harm, and generalizing from those cases to therapeutic surgery would have been unfortunate for the human race.

A more naturalistically inclined Aristotelian could despair about ethics in quintessentially modern situations. Absent foresight from a God or an axiarchic principle, we should not expect our natures to provide guidance in these situations, or at least "non-glitchy" guidance??backref. This line of thought could lead the Aristotelian to a dark view on which there just is no answer to a number of contemporary moral questions, or on which the answer conflicts in a glitchy way with our moral intuitions, or perhaps even one on which the true moral norms conflict, and we get hard-to-avoid moral dilemmas.

But a theistic story can restore optimism. God can know what kinds of seemingly outlandish scenarios might actually be relevant to the lives of his creatures, and can wisely choose the forms whose norms that fit with these. The resulting norms may seem *ad hoc*, especially in edge cases: they won't be the elegant principles of classic utilitarianism (though of course classic utilitarianism faces significant difficulties in out-of-our-experience situations, as we saw in ??backref:population-ethics). And an apparent *ad hoc* character in divinely-instituted rules as applying to edge cases should not surprise us—wise legislation does not eschew judgment calls.

An interesting question is whether a theistic Aristotelian should be surprised by having ethics glitch in some actual cases, in one of the three ways discussed in ??backref: (i) real dilemmas, (ii) conflict between moral rules and the reasons for them, and (iii) conflict between moral rules and our intuitions. After all, logical space contains an infinite number of possibilities for a form of a rational being, and a perfect being should be able to combine one with an environment in which there would be no actual glitches.

I grant that it is very plausible that a perfect being *could* do that. But would the perfect being do it? Even for a being whose power is unlimited by anything other than logic, there can be unavoidable costs to options. Intuitively, there is a value to the most important norms of behavior for limited beings[4]—say, norms governing killing—having a significant

[4]Why limited????

simplicity, so that they can be reasoned about more easily, especially under time pressure. At the same time, there is a value to a diverse and rich moral environment. And there is a value to morality lacking glitches. Plausibly, one cannot have all three values to their maximal degree at the same time—there may be logically unavoidable trade-offs. And there does not appear to be strong reason to think that a perfect being would be so enamoured of one of the three value that we would expect that value to be present to the maximal degree. In particular, we should not expect a completely unglitchy ethics.

But we might have reason to hope that glitches are rare in the actual circumstances faced by humanity, or that the worst of the glitches should only occur in the case of agents who have wrongfully produced the circumstances for this glitching.[5]

??hypothetical judgments

2.10. Avoiding radical scepticism. There is a number of sceptical hypotheses that have the property that they cannot be ruled out either on logical grounds or *a posteriori*. These include hypotheses that the world around us is a computer simulation, that our moral intuitions are disconnected from moral reality, that we are Boltzmann brains, i.e., short-lived brains in bubbles of oxygen arising from fluctuations in the vacuum of space, that we live in an infinite multiverse that undercuts all probabilistic reasoning??ref, that simpler scientific theories are more often right other things being equal, and so on. Yet we think these hypotheses false. If we think them false neither on logical nor empirical grounds, it must be because we assign low probabilities to them prior to empirical assessment. Moreover, this assignment is required by our rationality: those who fail to assign low probabilities to them are irrational.

Our Aristotelian account can ground the correctness of this judgment of irrationality in human nature.??cf.backref But there would be something deeply problematic about us if the low *epistemic* probability of the sceptical hypotheses were not matched by a low *objective chance* for them to be true in light of the causal and stochastic structure of the

[5]Compare the theory that real moral dilemmas only occur in the case of agents who have done wrong, say by making contradictory promises.

world. If in fact the most likely way for a being with our rational nature and mental life to arise would be as a Boltzmann brain, then even if we have lucked out and are not a Boltzmann brain, there is a disharmony between the world and our mental life.

In such a lucky case, the connection between our nature-required priors and the world then appears too fragile, chancy and "unsafe"??ref for the beliefs essentially dependent on these priors to count as knowledge. Even a reliabilist should say that if some beings were required by their nature to assign a very high prior probability to the hypothesis that the universe formed an even number of years before life first arose, and it was mere chance that this hypothesis was true with the causal structure of reality not assigning it a higher probability than the hypothesis of an odd number of years between the beginning of the universe and the beginning of life, then that hypothesis is not knowledge. Yet it is very plausible that we *know* the sceptical hypotheses under discussion to be false. (This judgment has admittedly been disputed by a number of epistemologists who admit with G. E. Moore that I know that I have two hands, but will not allow the Moorean inference that I know that I am not a brain in a vat, despite the fact that I have two hands obviously entailing that I am not a brain in a vat. ??ref) And even without considerations of knowledge, we might note that an optimism resting on an assumption of mere luck appears paradigmatically irrational.

The problem is perhaps most pressing in the case of the kinds of highly abstract *a priori* intuitions discussed in ch4:??backref, such as the intuition that the axioms of arithmetic are consistent or that nothing can cause itself. For if there is merely a coincidence between the truth of the intuitions and our possession of a nature that normatively requires us to have these intuitions and causally impels us to them, then even though we may be justified in following the intuitions, they are unlikely to count as knowledge. Aristotelians thus need a theory on which there is the right kind of connection between our rational nature and the metaphysical, causal and stochastic structure of the world.

2.11. Global aesthetic-like features. [6]

2.12. Family.

2.13. Retributive justice.

2.14. Divine authority.

3. Kind-independent goods

Aristotelianism does really well with explaining kind-dependent values. But there also appear to be values that appear to transcend kinds, such as simplicity, diversity, flourishing, achievement, etc. Furthermore, we can compare kinds. The Aristotelian account defended in previous chapters can ground the comparison between a flourishing and a non-flourishing human, or between a flourishing and a non-flourishing chantarelle mushroom. But it is also obvious that a human is a better kind of entity than a mushroom. ????

4. Complexity and explanation

4.1. A problem. A central form of argument for Aristotelianism is based on Mersenne questions and the messy normative complexity of our lives. But do we not normally prefer simpler theories to more complex ones, and hence should we not reject the normative complexity in favor of a simpler theory like utilitarianism in the name of Ockham's razor?

Ockham's razor, however, has always been a defeasible criterion: entities are not to be multiplied *beyond necessity*. But sometimes there is necessity. It would be simple to suppose that all trees of a single species look exactly the same. But that just wouldn't fit with our evidence. In biology, one does not expect individuals of the same sort to be exactly alike, unlike in fundamental particle physics.

Specifying what a rational animal of a particular species ought to be like and how it should behave can be expected to involve a lot of information. How much information?

[6] I am grateful to Nicholas Breiner for drawing my attention, in the context of justice, to this form of explanation of moral features.

Well, we might take the information contained the DNA common to all humans to give us a lower order of magnitude bound, since the common DNA presumably encodes something about what human bodies are supposed to be like. There are 3.2 billion base pairs in human DNA, and 99.1Since each base pair is two bits of information, that means about 6.3 billion bits, or the equivalent of about 500,000 book pages.[7]

Imagine the task of designing the rules of behavior for a rational animal that has a significant complexity in its bodily life, subject to the constraint that the rules lead to a life that elegantly balances moral and epistemic norms, and fits well with the bodily nature of the animal and its niche in the ecosystem. It is plausible to think this will be several orders of magnitude more complex a task than that of designing the rules for a well-balanced and significantly embodied game such as tennis. Generating a game of pleasing elegance and yet compelling complexity, especially an embodied one, takes a fair amount of information, and the official rules for tennis are about forty pages.??ref

We can think of simplicity as an aesthetic criterion in theory choice. But simplicity is not the only factor contributing to beauty! (If it were, the most beautiful art would be no art: you can't get simpler than an installation that can be completely described by $\sim \exists x(x = x)$.) Overall theoretical simplicity is one way of having an elegantly unified theory. But one can also achieve elegant unification in other ways. Consider, for instance, hierarchical organization. Wittgenstein's *Tractatus*??ref achieves a unification by being summed up in seven top-level sentences, with a progressive hierarchical amplification and justification in terms of multiple levels of sentences. Or consider the unification achieved in biology by Linnaean and Darwinian taxonomies.

We could have an ethics that is simply simple: it has a briefly expressible rule that covers everything in full detail. But just as it is unlikely that we would get a compelling racquet sport with a single brief rule, even if we allowed for some vagueness, it is unlikely that we would get a harmonious set of norms for the life of a rational animal out of such

[7]Counting a page at 1800 characters and each character at seven bits: $6.3 \times 10^9 / (1800 \cdot 7) = 500000$ (oddly exactly, by coincidence!).

a rule—that, indeed, is an upshot of the enumeration of the many Mersennian issues of detail in normative phenomena that have been discussed in this book.

Can we have some other kind of theoretical unification? I think we can. As discussed in connection with particularism(??did I??backref), we can suppose suppose a hierarchical structure. We can have a hierarchical ethics, at the top with one or more principles like Aquinas's "Pursue the good and avoid the bad", the Kantian injunction to treat others as ends rather than mere means, or the Biblical "Love your neighbor as yourself". But perhaps unlike the historical Kant, we need not take the top level principle or principles to have all of the normative informational content for morality. Instead, we can think of it as a unifying headline, perhaps to explaining tennis by saying: "Hit the ball back into the other player's side." There will be further rules that are not mere logical derivations, but build on the general principle expressed in the higher level rules by giving more specific rules. The second level rules themselves are not unlikely to need further adumbration.

Consider Hillel's famous response to the request that he explain the Jewish law while standing on one leg:

> That which is hateful to you, do not do to your fellow. That is the whole Torah, all the rest is commentary. Now, go and learn it [the commentary]. ??ref:add scholarly translation

Now it is clear that in fact that the primary Jewish commentaries on the five books of the Torah (i.e., the Mishnah, and the Talmuds which are commentaries on the Mishnah??check,refs) contain normative material not found in these books. Nor should this be a controversial claim, since rabbinical tradition holds that the rabbis have an oral tradition going over and beyond the books of the Torah. Thus we should probably interpret Hillel as saying that the Golden Rule (in his negative formulation) is a kind of summary, rather than as saying that the rest is logically derivable. Similarly, Aquinas, after giving his "first precept"??ref that good is to be done and evil avoided, lists second level laws such as preserving human life, respecting the reproductive life of us a rational animals, knowing

the truth (especially about God), and living in society. It is clear that there we still have not reached the level of normative information needed to resolve all moral questions.

In both the Hillel and Aquinas cases, we have a unification of ethics under one or more general principles that are insufficient for deriving all the specifics. This is akin to the explanatory unification that modern biology receives from evolutionary theory. Besides generalities like that species tend to mutate towards inclusively fitter forms, the basic principles of evolutionary theory—random variation and the survival of the fittest—do not generate specific predictions. However, they do organize the vast sphere of modern biological knowledge.

Famously, Aristotle has observed that in ethics, unlike in geometry, one can only speak in ways that are true for the most part.[8] On the account I am defending, this is not quite right. Instead, many of the higher level ethical claims are what one might call "generalities" that organize ethical reasoning. These claims can be exceptionlessly true, but there are limits to how helpful they are in particular cases. Thus, it may always be true that we should respect human life, but this does not give a clear answer as to what the health care provider should do when the family of a particular patient requests disconnection from life support. The respect claim describes, in general terms, the shape that the finer-grained principles have. And it may well be that the finest grained principles which apply to certain particular cases have a complexity beyond our practical ability to specify, and so we do not have principles that definitively settle a case.

While I have used ethics in the above discussion, the same plausibly applies to epistemic rationality, where we have a very general principle like "Pursue understanding (or knowlege or truth)", with finer-grained specifications such as "Avoiding error is more important than getting at truth", "Prefer elegant theories" and "Direct your attention to more important matters." In the case of semantics, we may, on the other hand, have a high

[8] ??ref. Of course, this claim itself needs to be carefully understood. Aristotle himself says that murder and adultery are always wrong. Perhaps he is thinking that murder and adultery are definitionally wrong—murder being a wrongful killing and adultery being sex contrary to respect for marriage?

level principle that "Meaning follows usage", and then a variety of finer-grained principles about how usage yields meaning. As we get to very fine-grained principles, we have an extremely complex account, but hierarchically organized.

5. Explanation of our normative complex

5.1. A pattern of explanation of norms. Here is a familiar pattern. We have a deeply-seated moral intuition about the general prohibition, call it *g*, of some action, such as incest. It is not clear how to derive the prohibition in its full generality from intuitively more basic principles, such as one of the categorical imperatives. Easy considerations, which I will call the *c*s, show that in *typical* cases the action is wrong, but our moral intuition goes beyond these typical cases. Thus, considerations of the abuse of power, distortion of familial dynamics, and genetic harms show that most cases of incest are wrong, but it is easy to imagine cases of incest to which these considerations do not apply—say, elderly siblings who were raised apart—and yet moral intuition forbids incest in those cases as well.

We can now save the moral intuition by saying that the more general prohibition *g* is simply a fundamental moral rule, not reducible to the *c*s that explain why the action is wrong in typical cases. But if we stop at this, the connection between *g* and the *c*s mere happenstance, and that seems intuitively wrong. The abuse of power, distortion of family dynamics, and genetic harms should be relevant to why incest is wrong.

At this point, often we are in a position to see another fact: it is quite beneficial to have a general moral prohibition beyond the prohibitions arising from the *c*s.

One reason for such a benefit from a general prohition could be that our judgment as to whether the *c*s apply to a given case is fallible, especially given our capacities for self-deceit, and the costs of violating the *c*s are so high that it would be better for us to have a general prohibition than to try to judge things on a case-by-case basis.

Second, in some examples of the pattern, serious deliberation about the forbidden action can itself harm one or more of the goods involved in the *c*s: thus, having to weigh

whether the distortion-of-family-dynamics consideration applies against a particular instance of incest can itself distort the agent's participation in family dynamics.

Third, we could have a tragedy of the commons situation. It could be that the *c*s are actually insufficient to render an instance of the action wrong, but we would be better off as a society if we had general abstention from the action. Thus, perhaps, the genetic harm coming from one more couple's engagement in incest would be insufficiently significant to render the incest wrong, but without a general prohibition, incest would be sufficiently widespread as to cause serious social problems. A general prohibition that is not logically dependent on the *c*s would help avert such social harms.

These considerations are very familiar to us in the case of positive law. Jaywalking involves harms such as disruption of traffic flow and the danger of death of the pedestrian and of trauma to the driver, and the considerations of these make jaywalking wrong in typical cases. There are obvious instances, however, where these considerations do not apply: say, crossing a road where the pedestrian can clearly see that there are no intersections or cars on the road for a significant distance in either direction. However, it may be better for people simply to abstain from jaywalking than judging whether the disruption and safety considerations apply on a case-by-case basis, because there could be so much harm if the judgment were to go wrong. As a result, it is can be reasonable for a state simply to ban jaywalking altogether (or to ban it with some clear and easily adjudicated exceptions). We similarly resolve cases of tragedy of the commons with positive law: think, for instance, about laws against littering.

In the case of positive law we have two different explanations. First, there is an explanation of why the forbidden action is wrong in general: this is because it has been competently forbidden by legitimate authority. This explanation need not make reference to considerations such as disruption of traffic flow or danger of death.[9] Second, there is

[9]Though in some cases *some* such reference may be needed in order to establish that the matter falls within the competence of the authority in question. Thus, a government agency may be permitted to make rules on matters where traffic flow disruption is concerned.

an explanation as to why the action has been forbidden by the authority—and here all the rich considerations are relevant.

5.2. Theism. A theistic version of natural law can have precisely the above pattern. An action is morally forbidden because our nature is opposed to it, an instance of grounding explanation. This explanatory fact does not make reference to the *c*s. But we still have a further question to ask that it is natural to put in the form: "Why does our nature includes this prohibition?" But since our nature is essential to us, the answer to that question could simply be the necessary truth that we couldn't exist without this nature. However, we can put the question in a different way: "Why are there intelligent primates on earth with a nature that includes this prohibition rather than some other kind of intelligent primates with a nature that does not include this prohibition?" And here the theist can answer: Because it would be good, in light of the *c*s and the further considerations in favor of generalizing the prohibition beyond the cases where the *c*s specifically apply, to have intelligent primates with a nature that includes this prohibition, and God acted in light of this good.[10]

The explanation may still appear viciously circular. On an Aristotelian metaphysics of value, what is good for us is grounded in our possession of our form. How could the possession of our form, then, be explained by what is good for us? But it is difficult to see the difficulty in the context of theistic selective explanations. Whatever form will be exemplified will define what is good for its possessors. If God were to choose to exemplify a form that defines one and only one state as good for its possessors but that also makes it nearly impossible to attain that state, the result would be beings that almost universally are in a bad state. There is reason not to do that. Instead, God has good reason to select a form that makes it much easier to attain the good state defined by the form.

Here we should make a distinction between the specific goods grounded in our form—health, friendship and the like—and the good of fulfilling our nature. The specific goods are grounded in our nature. But it is not clear that we should say that the good of fulfilling

[10]A divine command theorist can make the same move, but divine command theory has some liabilities which were discussed in ??backref.

our nature is itself grounded in our nature. It is plausible in the Aristotelian context to say that to be good for x just *is* to fulfill x's nature. This identity is simply reductive. Given this, the circularity in the explanation disappears. What specific things are good for us is grounded in our nature. But it is good for us to fulfill our nature, and that fact is independent of what the nature is. It thus makes sense to explain *which* nature is exemplified by considerations of how apt the possessors of that nature would be to fulfill that nature, and have that good. We thus have a kind of explanation of why rabbit-like beings specifically have reproduction be good for them—having reproduction be good for them is more apt for the fulfillment of their nature than, say, having the discovery of mathematical truth be good for them. It is the general good of fulfillment of any nature that can explain why a nature with such-and-such specific goods is selected.

In fact, we can have explanatory relations running both ways between goods and norms of behavior. If having norm N promotes some specific good S, then that could explain why a being whose form codes for S being good also has norm N of behavior. But conversely, we could explain why a being whose form codes for norm N also codes for S being good for the being. If the norms are selected for exemplification by a perfectly good God, we may expect both forms of explanation to show up, as well as a hybrid model where both N and S are chosen together for their fit.

This kind of divine selection explanation of both ethical norms and goods extends from ethical to prudential, epistemic and semantic norms. As we saw in ??backref, ethical, prudential, epistemic and semantic norms all interact in complex ways with what is good for us, and this interaction can provide God with reasons in favor of some and against other combinations of norms and goods.

We would expect God to have access to the truthmakers of fundamental abstract intuitions. Indeed, on some theological accounts, God himself is the ultimate truthmaker of many of them, including notably mathematical and modal truths.??refs If so, then if God designed our nature so that our intuitions might mirror his knowledge, it is plausible that our justified following of these intuitions does indeed yield knowledge in us.

5.3. Non-theistic alternatives. What could such a selective cause be like? There are three main candidates for selective causes in the philosophical literature: evolution, axiarchic principles, and intelligent designers such as God.

Genetic descent with variation only directly governs the non-normative aspects of organisms. It is not sufficient to explain the form that the organism has, given that the form encodes normative features as well. Nonetheless, it is worth considering the possibility of a law of nature linking DNA to form, a law of nature specifying that when an organism with such-and-such DNA comes into existence, it has such-and-such a form. This law of nature would be immensely complex, with many free parameters raising Mersenne questions. Moreover, if the law of nature is to cohere with the Aristotelian optimism that is crucial to our Aristotelian account, there must be a fit among the various aspects of the form and between the form and the actual physical body plan and physical environment. It is implausible that this fit, in us and presumably in the myriad of other organisms, is just a coincidence. Such a law of nature calls out for an explanation. On pain of vicious regress, the need to explain the law of nature points towards one of the other two explanatory candidates: axiarchic principles and intelligent designers. Moreover, without a value-laden explanation of the linkage between DNA and form, the hierarchical explanations discussed in Section **??**, and the non-deductive hierarchical explanation of normative principles is replaced by a vast coincidence, and hence we do not have a satisfactory answer to the complexity objection.

Our Aristotelian account in order to be intellectually satisfactory requires an explanation that itself delves into some normative domain. This explanation could directly govern the imposition of forms or via some intermediary like a linking law of nature.

Furthermore, the need to explain our fundamental *a priori* intuitions in a way that connects them with their truth is unmet by a purely evolutionary approach. Evolution doesn't respond to the consistency of the Peano axioms of arithmetic and give us a corresponding intuition.

At this point, our choice appears to be between an explanation involving a non-intelligent tendency towards value and an intelligent one, such as the theistic one that we have already discussed.

The main candidate for the non-intelligent tendency are axiarchic principles, such as those defended by Leslie and Rescher.??refs These are fundamental metaphysical principles that require the world to be optimal.

It is worth noting that while I am discussing axiarchism as an alternative to theism, Rescher himself takes his theory to imply theism: it is better for there to be a God, and hence there is a God.??ref And if there is a God, then presumably this God is sovereign and governs the selection of forms, and we can skip forward to the discussion of theism. Leslie's version also involves supernatural beings. But Leslie thinks that what is best is not that there be one infinite all powerful, all knowing and all good God, but infinitely many omniscient observers who enjoy the world thereby adding to its value, though without creating it, since then there would be the possibility of conflict between them.??ref,check

There are four main problems with axiarchic principle explanations.

First, intuitively, a metaphysical principle *constrains* what beings can exist and how they behave rather than somehow explaining the positive existence of beings. But the axiarchic principles are supposed to explain the existence of beings: the beings in reality exist because it is for the best that they do so.

Second, there does not appear to be a unique best world. We could take any good world and add one more happy disembodied mathematician. This might not produce an overall better world. It might, for instance, be aesthetically inferior in some way—say, by having too many mathematicians and thus offending against simplicity, or by having a non-prime number of mathematicians (perhaps the aesthetically best number of mathematicians is a prime of the form $2^n - 1$ for some large n)—but it is superior in at least one significant way, namely by having an additional happy mathematician, and it is not plausible to think that it would be an overall inferior world.

Third, axiarchic principles appear to lead to modal collapse. If metaphysical principles require everything to be for the best, then it seems that everything must be the way it is.

There are at least three potential ways out of the modal collapse objection. The first is Leibniz's solution who distinguished between moral necessity and logical necessity. A proposition is logically necessary, according to Leibniz, provided that there is a finite proof of a contradiction from its negation. It is morally necessary provided that there is a finite *or infinite* proof of a contradiction from its negation. There are infinitely many logically possible worlds (where as usual p is possible just in case its negation is not necessary) but only one morally possible world—the best world. It is logical modality, then, that answers to our intuitions about the broad range of possibilities for reality.

Unfortunately, Leibniz's notion of logical necessity in terms of finite proof does not fit well with much later developments in logic and modal logic. Consider a very weak version of Axiom S4 of modal logic. Axiom S4 says that *any* necessary proposition is necessarily necessary. Weak S4 says that *some* necessary proposition is necessarily necessary. This seems utterly uncontroversial. For instance, surely, that everything is either green or not green is not only necessary, but necessarily necessary.[11] From Weak S4 and uncontroversial axioms of modal logic it follows that some proposition is necessarily possible.[12] But now a proposition p is necessarily possible in Leibniz's sense of logical modality just in case there is a proof that it is possible. And p is possible just in case there is no proof of $\sim p$. Thus, p is necessarily possible just in case there is a proof that there is no proof of $\sim p$. Now, in an inconsistent logical system, there is a proof of *every* proposition. Hence, a proof that there is no proof of $\sim p$ would be a proof that the logical system we are working with is consistent. But as long as the logical system has a recursively enumerable??? set of axioms (and to deny that would not be in the spirit of Leibniz's notion of finite proof),

[11]Weak S4 can be proved to follow from the Necessitation Rule, which says that if p is a theorem, so is $\Box p$, as long as the logical system is such as to have at least one theorem. For if p is a theorem, then $\Box p$ is a theorem by Necessitation, and hence so is $\Box\Box p$.

[12]Suppose $\Box\Box p$. By Axiom T, $\Box p \rightarrow p \rightarrow \Diamond p$ is a theorem. By the Distribution Axiom, it follows that $\Box\Box p \rightarrow \Box\Diamond p$ is a theorem. Since we have $\Box\Box p$, by modus ponens we have $\Box\Diamond p$.

includes the axioms of arithmetic (the idea that the axioms of arithmetic could be false in some possible world seems hard to buy) and is actually consistent, then by Gödel's Second Incompleteness Theorem the system cannot prove its own consistency. And hence it cannot prove p is possible on the Leibnizian account of possibility, and thus does not make p necessarily possible on that account.[13]

The second way out of modal collapse is to limit the scope of axiarchic principles to producing what one might call the best *skeleton* for a world. Say that a skeleton for a possible world consists of all the explanatorily fundamental parts of the world, such as the initial conditions and the laws of nature. As long as the laws of nature and/or causal powers of the initial beings are indeterministic, we could make only one skeleton possible, while yet having a multiplicity of possible worlds differing in how that skeleton evolves indeterministically into a fully fleshed out world. This will save our intuitions about more ordinary possibilities: I might have forgotten to come to class today, you could have found my arguments more convincing than they are, and the French could have emerged from World War II as the dominant world power. But forcing the laws of nature to be necessary is pretty counterintuitive.

The third response is to allow for tied or incommensurable worlds, and say that the axiarchic principle requires *a* best world, but not *the* best one. One might, if one wishes, also combine this with the skeleton move and let the principle require the world to have *an* optimal skeleton. This response would also answer our earlier objection that there is no such thing as the best world. It is mysterious, however, how the axiarchic principle would then go about selecting which precise world or skeleton exists from among the optimal ones. A principle is not a person who can choose between a set of incommensurable options, nor is it an indeterministic cause that has a range of possible effects. ???

The final difficulty for axiarchic views is the problem of evil. Looking at the litany of suffering in human history, our world doesn't look like the best of all possible worlds.

---

[13] ??discuss objection paper

Axiarchic views can make use of many of the responses to the problem of evil given by theists. This vast literature is beyond the scope of this book.??refs

5.4. Theistic choice points. Suppose we are convinced that we need a theistic Aristotelianism. At this point there are metaphysical and theological choice points. One metaphysical choice point is whether there are any uninstantiated forms. If there are, as on a Platonic picture, then we have a theological question: Does God freely choose which ones to create, or do they exist necessarily, say in the mind of God? If there are no uninstantiated forms, as on a more classically Aristotelian picture, then probably the most parsimonious theistic story is that God creates in the act of creating the substances that instantiate them. We thus have three views: Theistic Voluntarist Platonist Aristotelianism, Theistic Involuntarist Platonist Aristotelianism and Theistic Classical Aristotelianism.

On the Platonist versions, the forms have some kind of uninstantiated mode of existence, in addition to the instantiated mode of existence they have in creatures. (I am assuming here that we have already decided in favor of individual forms—??backref. Perhaps, though, on the Platonist versions that choice point should be revisited?)

The Voluntarist Platonism option may seem to have some unnecessary complexity. If God chooses which forms to create, it is puzzling why God would "bother" with the ones that aren't going to get instantiated. There is, however, a possible answer: to open a field of possibilities to creatures. Perhaps the forms need to have some kind of Platonic existence in order for creatures to have the power of producing their instantiations. There is a value to the earth ecosystem "having a choice", with many evolutionary possibilities of what kinds of biological substances should exist, and on a more Platonic version of the metaphysics this could require the pre-existence of these forms in their uninstantiated mode.

On both the Voluntarist Platonist and Classical versions, there is or can be a field of possibilities for other forms than the ones that actually exist. On the Voluntarist Platonist version, these are other forms that God could have created *ex nihilo* independently of

instantiation. On the Classical version, these are other forms that God could have instantiated and thereby brought not existence. Presumably, God knows what these possibilities are, and so they have some kind of existence as ideas in the mind of God. There is much room here for difficult metaphysical exploration of the exact status of these divine ideas.???many-refs Nonetheless, this point shows that there is a commonality between all three versions of theistic Aristotelianism: there is a field of formal possibilities. On the Voluntarist Platonist and Classical theories, this is a field of divine ideas. On the Involuntarist Platonism, this is a field of necessarily existing forms.

On all three views, God selects from that field of possibilities some forms that will be instantiated, and maybe also some forms that creatures can on their own cause to be instantiated. There are significant metaphysical differences between the views, but all three involve a similar kind of divine selection model of which forms are instantiated.

5.5. Participation.

5.5.1. *The account.* But there is also a different way that we could have a theistic explanation of normative features of forms. Classical theism holds that all things are either God or participate in God. In such a setting, it is natural to think of a form as a way for a being to participate in God. But now while God's infinity and otherness may give a wide scope to what sorts of arrangements of features could count as a participation, that scope is plausibly narrower than all logically non-contradictory arrangements of features. Some candidate norms, like a requirement of causing gratuitous pain to others, just may not be included in any metaphysical possible way of participating in a perfectly good God. And some combinations of individually admissible features may also fail to be found in any metaphysically possible mode of participation in a perfectly unified God, such as having conversation with conspecifics as central for one's good while having the essential causal power of deterministically exploding whenever one approaches a conspecific within talking distance.

Such a theistic participatory limitation on forms yields a more metaphysical explanation of some aspects of Aristotelian optimistic harmony than divine selection does. Moreover, this mode of explanation lends itself more easily to supernaturalist stories other than theism, such pantheism or a classical Platonism centered on the Form of the Good.

Nonetheless, a mere limitation on the space of possibilities for forms is insufficient for explaining all the aspects of Aristotelian optimism. First, a limitation of forms by itself does nothing to rule out the possibility of a form being always instantiated in beings that happen to inhabit an environment completely unsuitable for flourishing according to the norm.

Second, unless we think the limitation is really severe, our explanations of norms will be curtailed. For what we will be able to explain is why the complex of norms is minimally acceptable—such as to be minimally capable of participating in God (or the Form of the Good, on a classically Platonic version). But the limitation won't explain cases where norms fit particularly well together, since if they fit less well, the norms could still be found in some possible form. For instance, in humans living by the moral norms is central to flourishing. This ensures that any human that lives by the moral norms automatically has quite a bit of flourishing, and one who does not live by the moral norms cannot be said to flourish overall. Plausibly, a much weaker degree of unity between overall flourishing and the norms governing the will would suffice for a form of a being that participates in God: living by moral norms could be a less central aspect of flourishing. A divine selection explanation can advert to God's having a good reason to produce beings with the greater degree of unity in their normative features, and thereby explain the higher degree of unity, while a participatory limitation explanation would only explain why the degree of integration is at least minimal.

5.5.2. *An objection.* There is, however, a serious problem with the participatory limitation account. One of the main things that led us to grounding ethics in human form was the appearance of contingency implied by the vast number of seemingly arbitrary parameters in ethics. But does not the same thing difficulty apply to the participatory limitation

account? For there seem to be parameters defining the boundaries between participatable combinations of normative features and unparticipatable ones. Just how much unity between the norms is needed in a form that participates in a God who is one? How much normative egoism can be found in a form that participates in a perfectly loving God?

We might try to shift the difficulty onto parameters in the divine nature which determine what is a possible participator in God. These parameters can be necessary, since God is normally thought of as a necessary being.

However, there are shortcomings of this approach. First, it becomes puzzling why we don't simply do a similar thing for the Mersenne questions that led us to forms. We could suppose, for instance, a Platonic account on which there is a vast number of metaphysically necessary ethical ur-norms with metaphysically necessary but still seemingly arbitrary parameters, instead of norms found in natures of particular types of rational beings. Given the plausibility that different possible intelligent species would have different parameters in their norms, the ur-norms would presumably include many conditional ones connecting the non-normative features of a species with the norms governing thier behavior: "If you have such-and-such genetics, then you should prefer parents to strangers to degree x."

One response is that shifting the difficulty from parameters found in human ethics to parameters in divine ethics is still philosophically advantageous. First, as discussed in ??backref-and-add, there is an advantage in norms that are grounded in a nature, whether divine or human, in that these norms are not an alien imposition of dubious binding power, but are the requirements of one's very own nature. Second, there may be a greater unity in a complex set of norms governing a single divine being than in abstract Platonic norms governing all possible beings in conditional form. Third, there is some hope that the parameters governing divine nature are fewer and more unified than any plausible set of parameters governing humans.

On the other hand, the idea of arbitrary parameters in God is theologically and philosophical unattractive. Such parameters fit poorly with classical theism's doctrine of divine

simplicity. Moreover, it is the lack of arbitrary parameters that makes God an attractive explanatory posit. If God is to have a vast number of arbitrary parameters, is it not just as simple to explain things in terms of a brute necessity of a Big Bang, or of a particular set of necessarily selected forms?

Another response would be to hope that a participatory limitation account does not involve parameters. Perhaps there is no degreed limit on norms found in a form that participates in a perfect God, but the rather there are sharp non-arbitrary limits: nothing contrary to divine nature, such as hatred of persons or cruelty or injustice, can be required. If we make this move, we won't be able to account for all of Aristotelian optimism using participatory limitations. For optimism involves more than just belief in the barest minimum of positivity. The optimism that is essential to give us epistemic access to the norms under the Aristotelian synthesis requires a significant degree of union in the form, not just the barest minimum.

One might worry that a similar difficulty applies to the divine choice account. God is more likely to actualize a more unified form. But what are the parameters in the function that governs the relationship between the degree of unity in a form and the chance that God would actualize that form? We might again insist that parameters in the divine nature are not particularly problematic.

But there is another move. The idea that there are numerical chances assigned to divine actions is itself theologically and philosophically problematic. First, numerical chances seem to be a kind of limitation on divine power.??refs:Murphy??? Second, the idea of assigning numerical chances to the vast infinity of possible divine action. This infinity exceeds any infinite cardinality, since for any cardinal number κ, God could create precisely κ angels, while the class of cardinal numbers exceeds any particular cardinal in its size??refs, and so the class of divine actions creating groups of angels is beyond cardinality. Assigning numerical probabilities in such a setting is fraught with mathematical difficulty.

Instead of supposing numerical chances assigned to divine actions, we might simply suppose that there is a non-numerical qualitative rule: what better matches the divine nature is more likely to be instantiated. This rule need not even impose a total ordering on the space of possible divine creative actions, because there may be vast scope for incommensurability between divine actions.??ref:divine-creative-freedom, so that there will be divine actions whose chances are themselves incommensurable.[14] We might then further suppose that the qualitative ordering on the chances still yields epistemic probabilities for humans. For the human form could prescribe specific numerical rations of epistemic probabilities where the objective chances (??explain chances vs probabilities earlier) have a merely comparative or even incommensurable relationship.

The divine selection account, thus, seems to have a greater chance of escaping the arbitrary parameter worry than the participatory limitation account.

??why not divine command theory?? alienness?

5.6. A dual account. Moreover, it is quite reasonable to combine the theistic choice and participatory limitation accounts. Given the considerations in Section 5.5.2, the participatory limitation account is unlikely to suffice on its own, after all.

On a combined account, some aspects of form, especially coarser-grained ones, can be explained by participatory limitation while others, especially finer-grainer ones, can be explained by divine selection. The result is an explanatorily rich account of normativity, which predicts a minimal coherence throughout one's normative complex, and leads one to expect higher degrees of unification in more central aspects.

[14]*What could that mean? Here is a mathematical parable. The Banach-Tarski paradox has it that we can divide a solid mathematical ball into five disjoint pieces, E_1, E_2, E_3, E_4, E_5, and move these pieces to construct two balls of equal size to the original. Imagine randomly choosing a point in the original ball, and asking which of the five pieces it lies in. Paradox ensues if we make too many comparisons between pieces, such as that the point is more likely to lie in E_1 or E_2 than in just E_3, or vice versa, but it is safe at least to say that it is more likely that the point is in E_1 or E_2 than in just E_1. However, it may be reasonable at least to say that if A is a proper subset of B, then the point is more likely to lie in B than in A. There are still technical problems with this (cf. ??ref:Pruss-domination), even though lying in E_1 versus in $E_2 \cup E_3$ have incommensurable chances.

This expectation of unity in turn yields another tool to help us to actually find out what our norms are. We should prefer normative theories that allow for significant integration and harmony between moral, epistemic and other norms. An integrated picture of human flourishing is obviously attractive.

6. Final remarks

??explain how theism grounds the variety of harmonies discussed earlier in the chapter

CHAPTER XI

Eternal Life and Fulfillment

??delete?

??interact with Oderberg on suffering and pain

CHAPTER XII

More aristotelian Details

1. Introduction

2. More on flourishing

2.1. Supernormality. If Bob is facing unjust torture, and Alice finds a way to substitute herself for Bob, Alice has done something morally good. But at the same time, barring special circumstances (such as Alice having made some promise to Bob, or bearing some responsibility for why Bob is facing torture), Alice has no obligation to take Bob's place. Her action is morally excellent, but not obligatory. We call such actions supererogatory.

On the view I am defending, what makes Alice's action morally good is that she flourishes volitionally in her action. If, on the other hand, Alice were to encourage Bob's torturers, her action would be bad, and she would volitionally languish in it. However if Alice merely discourages Bob's torturers, but does not offer to take Bob's place, she flourishes volitionally less than if she offers to take Bob's place, but she flourishes nonetheless.

We find in the will, thus, a distinction (a) between languishing and flourishing, or better malfunction and proper function, as well as (b) between lesser and greater flourishing, or a lower and higher degree of proper function. This distinction lines between the moral concepts of the impermissible and permissible, and the merely permissible and the supererogatory.

It is likely that a similar distinction is found in areas other than morality. A typical physicist presumably has a properly functioning physical intuition. But Albert Einstein's physical intuition was not merely properly functioning: it was uncannily supernormal, so that thinking through thought experiments, such as his famous one about riding a light-beam, led to groundbreaking progress in physics. And, on the other hand, there are people

whose physical intuition is abnormal. We thus have a distinction (a) between the abnormal and the at-least-normal, and then (b) between the merely normal and the supernormal. The supererogatory is then a species of the supernormal.

??medical ethics

2.2. Parts and aspects. When my arm is functioning poorly due to an injury, I am functioning poorly insofar as I have an arm. It it tempting to reduce evaluations of the function or flourishing of a part of a substance to the function or flourishing of the whole in respect of the part. But while this is tempting, there is good reason to resist this reduction.

Of course, everyone agrees that there are cases where the flourishing or languishing of a part or aspect is instrumental to the opposite state of the whole. Xenophon has Socrates give the example of a person who is harmed by their wisdom because a tyrant hears about the wisdom and has the wise person kidnapped to serve as an adviser.??ref And many a person would have escaped a broken leg sustained if they had a minor sprain that kept them from skiing.

But there are more interesting cases where the flourishing state of a part seems not merely instrumental to the opposite flourishing state of the whole, but is constitutive of it. Muscles are torn down by exercise and regrow stronger. The process of tearing down is harmful to the muscles themselves, but is a constitutive part of the proper functioning of the body's system of adaptation to particular activities. And, more generally, the death of cells is part and parcel of the normal self-renewing persistence of a multi-cellular organism.

These cases are perhaps not entirely convincing. One might insist that when cells die as part of the self-renewal of the organism, the death is itself a part of the proper functioning of the cell, and hence both the cell and the whole are flourishing. Historically, Aristotelians have tended to insist that a thing's destruction is bad for it.???refs However, this may be mistaken. If we think of substances as four-dimensional objects—spatiotemporal entities—then a thing is destroyed just in case it has an upper temporal boundary. Now, having *spatial* boundaries is not bad for a thing—indeed, having spatial boundaries of the right sort is constitutive of a thing's having the correct shape and size. A dog so bloated as to

be boundless, taking up the whole universe would not be a healthy dog! Similarly, a thing could have proper temporal boundaries, and if so, it might be harmed not just by living too short a time, but also by living too long a time. Whether human beings are of this sort is a difficult question beyond the scope of this book. I think the way human flourishing seems to always call for more than we have suggests that humans are not like that. But most human *cells* seem to have a proper lifetime.

On the other hand, one might think that a muscle's being torn down is beneficial in the medium term to the organism, but harmful in the short term. The organism does become weaker. Thus the harm to the muscles is matched by the short-term harm to the muscular organism.

But cases of redundancy may provide a more convincing example of where the flourishing of a part comes apart from the partial flourishing of the whole. Suppose that an organism for its basic functioning needs n_1 functioning parts of some sort—say, cells of a particular kind, or legs, or teeth—and for the sake of redundancy it needs some large number n_2. Suppose that having more than n_2 is supernormal for the organism (as per our discussion in ??backref), until we reach some large number n_3 at which point the organism has too much.

For the sake of definiteness, suppose that the parts are teeth of some organism type, while n_1 is 30, n_2 is 35, and n_3 is 40. With fewer than 30 teeth, the organism fails to chew well. With 35 to 39, it has a healthy level of redundancy. And at 40 or more, it has too many teeth. Suppose now that Sally is an organism of this sort and she has 38 teeth. One of the teeth, however, is getting worn down quite a bit. That tooth is no longer fully functional, and hence that tooth is failing to flourish. However, this does not constitute any failure of flourishing in Sally. Even if the tooth stopped functioning entirely, Sally would still have sufficient dentation both for first-order purposes and for redundancy. Sally can still be fully functioning as a whole with respect to her teeth, even though that tooth is not itself fully functioning, and even though she would be fully functioning to a higher degree if that tooth were to fully function. In this case, the direction of flourishing of the part and of

the whole seems to be the same, but nonetheless one can have full flourishing of the whole without every part fully flourishing. In such a case, it would not be correct to say that Sally is languishing with respect to that tooth. For that tooth doesn't make her languish—it just makes flourish less.

Interestingly, redundancy can be between parts or aspects of very different sorts. We might suppose that a flourishing human being has a sufficient number of abilities of various sorts. These abilities can be moral, intellectual, emotional, or physical, and within each category, they can differ quite significantly from each other. Then full flourishing could turn out to be compatible with a severe impairment within certain abilities—for instance, one's mathematical aptitudes might be dysfunctional, and yet the *person* might fully flourish. This would allow an Aristotelian to accept the thesis that some persons with significant disabilities can nonetheless be fully flourishing.??refs For the disability can constitute the failure of a part or aspect of the person to flourish, without thereby constituing a failure of flourishing of the whole.

We could, in principle, suppose that there are some cases where the failure of flourishing of a part might not even make the whole flourish less. Going back to Sally, perhaps 37 teeth is better than 38 (perhaps 37 makes for better fit within the jaw), even though any number from 36 to 39 is fully normal. In that case, if Sally's 38th tooth is starting to deteriorate, this could be moving her to an even better state.

At the same time, there are proper parts and aspects such that the flourishing of the part or aspect always lines up with the flourishing of the whole. In the case of a person, to have one's will flourish in an action is to flourish with respect to the will, and makes one better off with that respect, and hence we do not need to correct looser discussion earlier in this book where no distinction was made between flourishing with respect to the will and having one's will flourish. Plausibly, rationality is similar. One is always better off for being more moral and more rational, and a failure of flourishing of one's morality is a failure of the person as a whole.

How much of the above logically possible differentiation between flourishing of the part and partial flourishing of the whole is realized in humans is a question for further investigation. However, we see that the teleological structure of a substance and its parts and aspects can be quite conceptually complex. If we take the form to be the ground of the essentials of that structure (???what of contingent forms), then this makes for even more work for form.

3. Teleological reductions

3.1. A multiplicity of concepts. The applications above, and the Aristotelian tradition, make use of various normative concepts that are said to be grounded in forms, such as proper function, teleology, and flourishing. It would be good to investigate if these can be further unified, under either one of these concepts, or some further unifying concept.

First, as we have seen in the discussion of the supererogatory and supernormal, we have both binary and comparative normative concepts, often in the same context. Suppose a stranger is about to be hit by a train, and the only four options are:

(1) give them a kick to ensure that they have no chance of survival

(2) make fun of them

(3) stand by idly

(4) jump in and push them out of the way likely at the cost of one's own life.

The binary distinction is that the first two options are impermissible, and the other two are permissible. But there is a comparative distinction as well: (1) is worse than (2), and (4) is better than (3). The binary distinction does not reduce to degree of comparison: while (2) is much worse than (3), (3) is much worse than (4) (though it is more natural to say that (4) is much better than (3)). And the comparisons do not reduce to the binary characterizations.

Proper function and teleology appear to be primarily binary concepts: a thing functions properly or improperly, and a thing either does or does not achieve its end. Flourishing, on the other hand, seems to comprise both the binary and the comparative. The mildly

vicious person unjustly suffering horrendous pain appears not to be flourishing *simpliciter*, but if the pain were increased, they would languish more. Flourishing thus appears the best candidate for a foundational concept for our norms.

But because the notion of ends and teleology has been so important in the Aristotelian tradition, both in the case of voluntary action and involuntary activity, it is worth thinking some more about ends.

3.2. Ends. Many activities seem to occur for an end. The activity then counts as successful provided that the end occurs and occurs as a fulfillment of the activity. An organism produces gametes in order to reproduce. A cat chases birds in order to catch them, and eats them for nourishment. And I put on shoes to keep my feet comfortable when I walk. The end-directedness of much voluntary activity is obvious, but whether there really is teleology in the involuntary cases is more controversial, though highly intuitive.

The Aristotelian tradition tends to analyze voluntary action as always end-directed, but also tends to see involuntary activity as often, if not always, directed at an end. I will argue, starting with the case of voluntary action, that many interesting phenomena would be misclassified as end-directed. The actual structure can be more complex, and while it has a directional structure, it is misleading to think of that structure as teleological in the sense of possessing a *telos*, an end that fulfills it.

Consider a sprinter who is running a hundred meters all out against a clock, rather than against other opponents. The runner has an end, namely to sprint 100 meters. But sprinting 100 meters does not explain the intense effort the runner puts in. Less than half of the effort could have been put in, and 100 meters would still have been sprinted. The bulk of the runner's effort is explained by being directed not at completing the sprint but at completing it in minimum time.

But what state of affairs does the runner's speed-directed effort have as its end? A runner might have a particular target time in mind. However, we are imagining a runner who runs all out. A runner who is just aiming at a particular time could slow down if it

became obvious that a slower run would still achieve the target time, but not so our all-out runner. Our sprinter may have some specific time in mind to motivate themselves, but interpreting their action as merely aiming at that time does not capture all of the directional structure of the performance. Any shortening of the time of the run is welcome given the sprinter's aims.

We would normally describe the runner as trying to run 100 meters "as fast as possible", and that seems to be a coherent description of an end. However, the language of "as fast possible" should not be taken literally. First, we have the question of what the relevant comparison class is. Is the runner trying to run as fast as any human being can on any track? As fast as they themselves can run on this track on this occasion? Or something in between? ??? unlikely success!

Second, suppose we fix a particular sense of "as fast as possible", and then after ten meters the runner realizes that they have been slightly slower than is possible. At this point, it is no longer possible for the runner to achieve the goal of literally running the run as fast as possible. But we do not expect the runner to stop. We expect the runner to resume running all-out, as part of the same directed activity.

There is obviously a teleological structure to the sprinter's run. But it is not a structure of aiming to achieve a *telos*. We can think of this structure that of a preference structure: a faster time is always preferred to a lower time. Such preference structures are common in games, where we ourselves have defined the teleology, but we can also find them in the case of teleologies that are not our own invention. For instance, suppose the human mind aims at understanding. If we understand this in terms of aim at a specific *telos*, the understanding of everything or of everything humanly understandable, then as soon as we realize that we cannot have this (for there are humanly understandable things that I cannot understand because they have faded to much into the past—facts about the taste of dodo birds, for instance), the pursuit would become pointless.

It is better to say that our teleological orientation sets an ordering on our understanding, where understanding a set B of items is better for us than understanding a set A of

items whenever A is a proper subset of B (every item in A is in B, but not *vice versa*) and when the levels of understanding for each item are kept fixed. This ordering then obviously affects our epistemic lives, but also our practical ones (since one needs to make sacrifices in order to understand).

There even seem to be cases where we have a teleological orientation not understood in terms of the pursuit of a specific *telos* outside of rational activity. A pecan tree produces pollen in order to have offspring. The more offspring, the better! Again, the *telos* is not to have infinitely many offspring[1], but rather the teleology seems best understood in terms of a preference ordering: more offspring, keeping the health of the offspring constant, is better.

As an alternative we might suppose that we just have infinitely many *telê* in one organism: to understand the truth or falsity p_1, and to understand the truth and falsity of p_2, and so on, for all the propositions that can be understood by humans, or to have at least one offspring, and to have at least two offspring, and to have at least three offspring, and so on. But while I have argued that our form is complex, to suppose such infinite complexity seems excessive. Furthermore, it is plausible that each *telos* impels the organism causally—e.g., pulling on us in our deliberation. But infinitely many such things causally influencing a single aspect of us would violate the principle of causal finitism defended recently by multiple authors??refs. Furthermore, each organism would always have infinitely many unfulfilled *telê*, and that may seem just too pessimistic to believe.

It may, of course, sound oxymoronic to talk of a teleological structure without a *telos*. But that's a merely verbal point. We can think of teleology as about pointing and directedness, without a *target* as such. Imagine for instance that I am competing in an odd archery competition, where the winner is the one whose arrow hits closer to the center—except that if you hit the exact center, you automatically lose. Then there is an obvious sense in

[1]One might suppose that there is a limit here based on how much offspring it is possible for one pecan tree to have. But it is not clear that there is such a limit in principle. Wouldn't a pecan tree always be better off if it lived an extra year and had more offspring (perhaps transported to another planet, to avoid overcrowding)?

which I am aiming. Indeed, since the chance of hitting the center is infinitesimal, I should even set my bowsight on the center, but it is not correct to say that I am aiming to hit the center. Rather than coining a new phrase, I will talk of teleology not defined by *telê*.

It is, by the way, interesting to note that this way of thinking of teleology damages a Kantian argument for why we should believe in an afterlife and God. The argument observes that we are unable to achieve complete virtue in this life. But since we cannot aim at the impossible, and need to aim at virtue, we should suppose another life, and a being capable of ensuring that in that life virtue is achievable. But following the above understanding of teleology, we might think ???? . ... But might be able to rescue by supposing that a specific degree is a telos of ours that nonetheless cannot be achieved in this life.

4. Individual forms

Recall the long-standing debate whether forms are individual—numerically different ones for different members of the same kind—or shared by all members of the same kind.

In ??backref, we saw that there is some advantage to an individual form account of ethics: individual forms intuitively do a little more justice to the personal nature of ethical obligation.??[but conjoint twins] ??add But are there any other arguments for taking forms to be individual?

I believe so. An initial attempt might be to argue that then the numerically same entity—the form—is present in multiple places at once, since a form counts as present where its matter is. I do not find this argument compelling, however, as I do not think multilocation is absurd.??ref But if you do, that is one argument. Let us consider some others.

4.1. Individual unity. My nose and my heart are parts of the same human, while my nose and your stomach are not. What constitutes the difference? On an individual form view, there is an elegant Aristotelian answer: my nose and my heart are informed by the

same form, while my nose and your heart are not. This answer will obviously not do on a joint-form view.

Nor will it do to say that the difference is constituted by the fact that there is continuous human matter joing my nose with my heart, since if you and I were conjoined twins, but conjoined neither by heart nor nose, my nose would be connected by continuous matter with your heart, even though those would still not be the nose and heart of one human.

One might try for a teleological account of the unity of the human being: my nose and my heart function together (the nose allows the entry of oxygen which is distributed by the heart). However, we are social animals, and teleological cooperation crosses individual boundaries.

It could be that some other account friendly to joint-form Aristotelianism is available. But the simplicity and elegance of the individual-form account of individual human unity is a reason to opt for the joint-form view.

4.2. Distant conspecifics. Suppose a shared form theory is true. Now, imagine that in our galaxy there is only one human being, Adam, and imagine that in a galaxy far, far away, God creates a humanoid comes into existence, with no genetic connection to Adam, but with a form that is just like Adam's: this form unifies matter in the same way as Adam's form does, it imposes exactly the same norms on the form's owner as the human form does on Adam, and it causes the same structure and behavior as the human form does for Adam.

At this point we have a dilemma: either the form of this humanoid must be numerically the same as Adam's or not. Suppose it must be numerically the same as ours. Then somehow simply by creating something in a galaxy far, far away, God causes an entity in *our* galaxy—Adam's form—to become multilocated. This seems counterintuitive.

Suppose that the form does not need to be numerically the same as Adam's. In that case, we have admitted that there can be numerically different forms with the same broadly functional features (including the normative functions). This means that the question of whether you and I have the numerically same form is not settled by noting that the forms

have the same functional features. Indeed, now the question whether your and my form is numerically different or the same becomes a metaphysical question that no empirical data is relevant to the settling of. There is nothing absurd about there being such metaphysical questions. But it is some advantage to a theory if raises fewer such questions, having fewer degrees of freedom. And if one does accept a theory where it is possible but not logically necessary that different individual substances have numerically different forms, then one really shouldn't be accepting that in practice you and I share a form. At best one should be agnostic on this question.

5. Accidental normative forms

If you have promised to ϕ, you should ϕ. Consider two Aristotelian metaphysical explanations of what is going on here. On the conditional-norm explanation, the human form contains the conditional norm that you should ϕ if you have promised to ϕ, which when combined with the fact that you have promised to ϕ grounds an obligation to ϕ. But there is another possible explanation. Perhaps promising to ϕ causes, perhaps in virtue of a power contained in the human form, the normative accident of being such that you ought to ϕ to come into existence. One might think of the second story as a very robust normative power theory: a theory on which normative powers are a type of causal power that brings into existence an irreducible and new normative entity. (Most normative power theories do not consider normative powers to be a type of causal power.??refs)

The conditional-norm account posits fewer entities, and insofar as this is the case, Ockham's razor favors it. And intuitively it seems to be the right account of promissory obligations. If this account holds for all norms, then the human substantial form could be the only normative property of the human being—there would be no normative accidents.

Now, while the conditional-norm account for humans posits fewer entities (i.e., forms), it makes the human form contain more information in the form of conditionals. For instance, on the conditional-norm view, humans without a Y chromosome have a human nature that specifies what range of physical developments that are normal expressions of

the genes that are unique to the Y chromosome. On the normative accident view, it may be that humans with a Y chromosome also have an additional non-physical property governing the expression of Y-based genes.[2]

At the same time, the normative accident view of human beings involves significantly more in the way of *causal* laws presumably grounded in human causal powers, such as that when one makes a promise, a promissory-obligation accident comes into existence, or when one has such-and-such DNA, then such-and-such an accident governing norms of gene expression coms into existence. Moreover, since normative accidents are in large part defined by the norms they embody, the informational complexity of the normative accident is presumably present in the causal power for its production—it is a power to produce an accident of, say, being required to ϕ. Thus while we have reduced the *normative* complexity in human nature, we have done so at the cost of *causal* complexity, *and* a multiplication of entities.

This last point does not apply to normative accidents that have a cause outside of the human being. But it is difficult to think of examples, with the exception of one theological possibility. According to many Christians/, by God's grace human beings can be directed at the "beatific vision", a direct vision of God. This beatific vision exceeds the power of human nature, and it is usually taken that natural human fulfillment does not require it. One way to make sense of this is to suppose that God by grace gives all or some humans

[2]It's worth noting that both views are compatible with a broad variety of views on gender and transsexuality, since choosing between the metaphysics of the two views does not settle the question of what the range of normal physical developments is. One might think that the normative accident view allows for a more conservative theory on which there is a metaphysical component—an accidental form—that determines whether one is really male or female, an accident of maleness or an accident of femaleness. But at the same time, the normative accident view enables one to have a metaphysical basis for the claim that one's real sex and/or gender fails to match one's biological constitution at birth. It is also worth noting that for norms relevant to sex and/or gender, the conditional norm view could make the antecedents of the conditionals be facts about DNA (such as whether one has a Y chromosome) but could also make them involve facts about psychology and society. The metaphysics does not by itself settle the normative questions here.

a normative accident of teleological directedness at the beatific vision, together with the supernatural powers that lead in the direction of fulfillment of this telos.