Heuristics for clueless agents: how to get away with ignoring what matters most in ordinary decision-making

David Thorstad & Andreas Mogensen

Global Priorities Institute | June 2020

GPI Working Paper 2-2020

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Heuristics for clueless agents: how to get away with ignoring what matters most in ordinary decision-making

David A. Thorstad and Andreas L. Mogensen

Abstract: Even our most mundane decisions have the potential to significantly impact the long-term future, but we are often clueless about what this impact may be. In this paper, we aim to characterize and solve two problems raised by recent discussions of cluelessness, which we term the Problems of Decision Paralysis and the Problem of Decision-Making Demandingness. After reviewing and rejecting existing solutions to both problems, we argue that the way forward is to be found in the distinction between *procedural* and *substantive rationality*. Clueless agents have access to a variety of heuristic decision-making procedures which are often rational responses to the decision problems that they face. By simplifying or even ignoring information about potential long-term impacts, heuristics produce effective decisions without demanding too much of ordinary decision-makers. We outline two classes of problem features bearing on the rationality of decision-making procedures for clueless agents, and show how these features can be used to shed light on our motivating problems.

1. Introduction

Recent finds in the Jebel Irhoud cave in Morocco indicate that *Homo sapiens* has been on Earth for at least 300,000 years (Hublin et al. 2017). If we play our cards right, we could be around for many more. This planet will continue to be hospitable to complex life for around another billion years, at which point the increasingly brighter Sun will drive a catastrophic runaway greenhouse effect. If we are lucky, humanity will survive throughout this time and spread to other worlds, giving us 100 trillion years before the last stars burn out (Adams 2008). Countless lives could be lived, filled with

flourishing, suffering, freedom, and oppression on a scale unparalleled in human history.

Suppose there were something you could do to significantly impact humanity's long-term future. Perhaps you could lower the probability of existential catastrophe by working against risks such as nuclear proliferation that threaten to bring our future to an early close. There are so many people yet to live that anything which improves their chances of leading flourishing lives appears to have tremendous moral significance. The expected value associated with actions of this kind seems to dwarf the expected value of just about anything else you could do (Beckstead 2013; Bostrom 2003, 2013; Greaves and MacAskill ms). Assuming a total utilitarian axiology, Bostrom (2013: 18-19) argues that a conservative projection of the total future population yields an estimate of the expected moral value of reducing extinction risk by one millionth of one percentage point that is at least the value of a hundred million human lives. Giving a mere one percent credence to less conservative estimates that take into account the potential for (post-) humanity to spread to the stars and for future minds to be implemented in computational hardware, Bostrom calculates the expected value of reducing the risk of extinction by as little as one billionth of one billionth of one percentage point to be one hundred billion times the value of a billion human lives.

Notice, however, that most of your actions have *some* probability of impacting the long-term future. Whether you sleep in today or get up early determines what you will eat, who you will interact with, and what you will accomplish today, all of which have myriad effects on others, carrying far into the future. If you get up early, you might be more productive. You might get in more reading and more writing. There is some very slim probability that this boost to your productivity will result in a work of philosophy that will be studied by future readers for as long as Plato's *Republic* has been studied today. If nothing else, it might influence the thinking of students who will one day be in positions of political power and whose decisions will impact generations to come.

Recent theorists have taken this to suggest that the expected values of most options available to us are dominated by their possible long-term impacts (Beckstead 2013; Bostrom 2003, 2013; Greaves and MacAskill ms):

> **Ex Ante Axiological Longtermism (EAAL):** In most cases, the vast majority of an option's ex ante (expected) value is determined by its effects on the long-term future.

While EAAL has plausible consequences for decision-making in what we intuitively take to be high-stakes contexts, it raises a pair of puzzles for decision-making in more mundane contexts. We are often clueless about the long-term effects of our actions. We do not know whether we will change the future for the better by getting up early instead of sleeping in. Decision paralysis threatens. It is unclear if and how rational agents can ever be justified in acting. Worse still, longtermist decision-making can be highly demanding. To correctly evaluate the long-term effects of our actions we must consider a huge number of future contingencies. Perhaps by getting up early and going to work we will speed up the rate of technological progress, accelerating the interstellar expansion of humanity in the 24th century. But in most decision-making contexts we cannot spare enough time and cognitive resources to consider even a handful of the relevant future contingencies. Does this mean that we are often doomed to choose irrationally?

In this paper, we aim to sharpen and solve these challenges. In Section 2, we give precise statements of each challenge. In Sections 3 and 4, we review apparent solutions that ultimately do not work. In Section 5, we introduce a number of constraints on a successful solution. In Section 6, we suggest that the problems can be solved by turning from substantive to procedural rationality. In Sections 7 and 8 we develop a procedural solution to both problems.

2. Two problems for Ex Ante Axiological Longtermism

The first order of business is to sharpen the challenges facing EAAL. Throughout this section we assume an ex ante form of consequentialism: the right action is the ex ante best action, considered impartially. Similar challenges could be raised using a number of non-consequentialist theories, since most non-consequentialists are like Rawls (1999) in accepting that the consequences of our actions matter and that pure time discounting is morally unacceptable.

The first problem begins with the observation that we are typically clueless about the long-term effects of our actions (Lenman 2000). In the first instance, we use 'cluelessness' as an opaque label for the sense of deep uncertainty that we characteristically experience when reflecting on the long-term impact of our actions. We leave its exact nature unspecified, at least for now. However, in order to fix ideas, we note that a necessary condition for cluelessness is that we do not know the long-term impacts of what we choose today, nor the relevant chances. Thus, we do not know whether getting up earlier or later will change the long-term future for better or for worse, nor how strong these effects will be. What's more, for any relevant contingency associated with the long-term future of humanity, such as the occurrence of a bioengineered pandemic killing at least 50% of all living people, we either do not know whether getting up earlier or later raises or lowers the chance of that outcome or at least fail to know with any suitable degree of precision by how much the chances are raised or lowered. There are too many contingencies to consider and we have far too little evidence to make any headway. Whether cluelessness involves something more than this is a question that we leave unresolved, for the time being.

We summarize these ideas as follows

> **Cluelessness About Long-Term Effects:** Decision-makers are typically clueless about the direction and magnitude of the long-term effects of the options available to them.

We are not always clueless about the long-term effects of our actions. For example, although we are clueless about the long-term effects of getting out of bed this morning, we may be reasonably certain that nuclear disarmament will change the future for the better. But in most contexts, we are hopelessly lost.

Given EAAL, it seems to follow from Cluelessness About Long-Term Effects that we are clueless about the values of most options that we face.

> **Cluelessness about Ex Ante Value:** Decision-makers are typically clueless about the direction and magnitude of the ex ante (expected) values of options they face.

In other words, it seems to be a feature of the deep uncertainty characteristic of cluelessness that we do not know the expected values of the options available to us, nor are we able to narrow these down with a suitable level of precision. Under consequentialism, this means we don't know the first thing about what we ought to do. Decision paralysis threatens.

> **Problem of Decision Paralysis:** In typical choice situations, it is unclear if and how a rational decisionmaker should act.

Decision paralysis would be a troubling consequence of EAAL. In important life decisions or large-scale interventions aimed at affecting the long-term future, some degree of decision paralysis may be appropriate. Someone who had an easy time making those decisions would be suspected of oversimplifying things. But in most ordinary decision-making contexts, rational decision-makers are not paralyzed with indecision.

In addition to the Problem of Decision Paralysis, there is a separate problem for EAAL. This problem draws on the difficulty of evaluating long-term consequences. We summarize this idea as follows

> **Long-Term Evaluative Difficulty:** To correctly determine the direction and magnitude of all relevant potential long-term effects, agents must consider a great number of future contingencies.

Just imagine the detailed modelling required to accurately assess the expected impact of sleeping in or waking up early on the long-term future of Earth-originating civilization. By the time you got round to completing your model, you would be several years late for work. This generates a second problem for EAAL.

> **Problem of Decision-Making Demandingness:** In typical choice situations, decision-makers must expend a great deal of time and effort to choose rationally.

As before, the Problem of Decision-making Demandingness is not always a problem. Large-scale decision-making about the long-term future should be difficult and demanding. But ordinary decision-making about matters like what time to set the alarm for should not be especially demanding.

We have seen that despite the plausibility of EAAL, it generates a pair of puzzles. Longtermist decision-making threatens to be unduly demanding, and it often appears to make it completely unknowable how a rational longtermist decision-maker is permitted to act. What can be said in response to these puzzles?

3. What won't work

In this section, we consider and reject four quick solutions to the Problem of Decision Paralysis and the Problem of Decision-Making Demandingness. Then in Section 4, we consider an imprecise Bayesian solution, before proposing our own response to these problems.

As a first try, we might accept Decision Paralysis and argue that clueless decision-makers should not act, at least not in the conventional sense. Instead, they should gather evidence to reduce their cluelessness. Only then should they act.

We are not opposed in principle to evidence-gathering as a strategy for resolving cluelessness. However, this strategy falls short of a general solution to either problem. Note, first, that evidence-gathering only worsens the Problem of Decision-Making Demandingness. If we went along with this approach, we would be required to exert a great deal of effort to improve our evidential situation in everyday decision contexts. That is too much to ask. We do not normally think that agents are required to gather much, if any, evidence before deciding whether to get out of bed in the morning.

Nor will evidence gathering resolve the Problem of Decision Paralysis. Taken in the relevant sense, evidence-gathering *is* an action and we are as clueless about the effects of gathering evidence as we are about the effects of most other actions. So evidence-gathering alone will not solve the longtermist's problems.

As a second try, we might deny Cluelessness About Ex Ante Value. This strategy grants that we are severely uncertain about the long-term future, but holds that our uncertainty is quite manageable. It differs only in degree and not in kind from typical cases of uncertainty. We already have a good theory of rational decision-making under uncertainty, namely precise Bayesianism: rational decision-makers should assign precise probabilities and utilities to the states and outcomes under consideration and act so as to maximize expected utility given these assignments.

We do not think the precise Bayesian approach solves either problem. Note first that precise Bayesianism offers no hope of solving the Problem of Decision-Making Demandingness. Constructing and evaluating detailed probabilistic models of the long-term effects of our actions is extremely challenging and time-consuming.

Furthermore, it is natural to feel that the precise Bayesian approach does not do justice to the severity of our uncertainty about the long-term future. We are not simply uncertain about the long-term future, but clueless. In the jargon of Knight (1921), we seem to be dealing not with *risk*, but *uncertainty*: ignorance of a kind so deep that our evidence does not warrant assigning precise probabilities to all relevant contingencies.

Some may insist that we should nonetheless reason in accordance with precise Bayesianism under conditions of Knightian uncertainty. Broome (2012) advises us: "Stick with expected value theory, since it is very well founded, and do your best with probabilities and values." (129) However, given EAAL, this makes rational decision-making highly sensitive to small, rationally arbitrary variations in probability assignments. If getting up early is assigned a 10^{-16} probability of fending off some future existential catastrophe, then that is presumably what we ought to do, unless getting up late prevents a similar catastrophe with probability 10^{-14}. Humans are ill-equipped to track such minute differences between probabilities. The precise Bayesian approach would thus leave rational decision-making hostage to the whim of decision-makers operating far beyond their discriminatory capacities.

A third try is to deny EAAL. This would solve both problems. But how could this denial be motivated? It could be claimed that, in mundane cases, the options available to us are almost always such that, for any important far-future consequence that might result from the choice of one option rather than another, we have no more reason to expect that this consequence will arise from choosing one option rather than the other. For example, while helping an elderly woman to cross the street might disturb the local traffic and thereby alter various conception events such that a future genocidal dictator is born who would otherwise not have risen to power, we have no

more reason to believe that helping the woman across the street will have this effect than that it will disturb the local traffic and thereby alter various conception events such that a future genocidal dictator is never conceived. If this is true in general, then the long-term effects of most actions may be thought to cancel out in ex ante valuation, and the ranking of available acts in terms of ex ante value will typically be determined entirely by their foreseeable effects on the short-term future (Lenman 2000; Greaves 2016).

However, there are good reasons to doubt that this symmetry claim holds with sufficient generality. Note first that the conclusion of the symmetry argument is too strong. It claims that in typical cases the long-term effects of our actions are entirely irrelevant to their ex ante values. While there might be some plausibility to the claim that long-term effects typically do not dominate ex ante value, it is not plausible that long-term effects are typically irrelevant to ex ante evaluation.

The bigger problem is that widespread symmetry with respect to expected long-term effects would be a striking coincidence that we have no reason to expect and good reason not to. We can readily motivate the symmetry claim for agents who lack evidence *of any kind* bearing on the long-term effects of their actions, as in the case of helping an elderly woman to cross the street. However, that is not the situation that most decision-makers face most of the time. The problem is not that we can say nothing about the potential future effects of our actions. Quite the opposite. There is often simply too much that we can say. Even the simplest among us can list a great number of potential future effects of our actions and produce some considerations bearing on their likelihoods, as in the case of deciding whether to get up earlier or later, where we have good reason to suspect that the former will allow us to be more productive, and some reason to think that greater productivity on our part might influence the long-run in various ways. It would be extremely surprising for perfect symmetry to be a frequent occurrence.

Some readers may suspect that we hereby set ourselves in opposition to a recent discussion of cluelessness due to Greaves (2016). That would be a mistake.

Greaves distinguishes between cases of *simple cluelessness*, exemplified by the case of helping an elderly woman to cross the street, and *complex cluelessness*, exemplified in her discussion by the choice of whether or not to fund the distribution of long-lasting insecticide-treated anti-malarial bed-nets, taking into account the potential indirect effects on population dynamics. Cases of simple cluelessness are characterized by evidential symmetry, whereas cases of complex cluelessness are characterized by conflicting evidence that we do not know how to weigh up. Greaves argues that in cases of simple cluelessness, we can rely on some suitable restriction of the classical Principle of Indifference and so differences in expected values between our actions depend entirely on their short-term, foreseeable consequences.

In discussion, some people have suggested to us that mundane cases of the kind that concern us are handled by Greaves's discussion of simple cluelessness. However, this is a mistake. Mundane cluelessness is not simple cluelessness. Greaves herself notes this: she argues that complex cluelessness also arises in mundane cases, such as "how much to spend on clothes" and "whether or not to give up caffeine" (25). In general, there seems to be no reason to expect that cases of mundane decision-making must involve simple cluelessness, although it may well be true that all cases of simple cluelessness pertain to mundane decisions.

Here's the fourth and final response we'll consider in this section. Some may think the best way to address the problems we've raised is to adopt what Monton (2019) calls 'Nicolausian discounting'. This involves discounting suitably small probabilities down to zero for the purposes of decision-making.

One motivation for Nicolausian discounting is that expected utility maximizers are vulnerable to being enticed by bets offering small probabilities of astronomical value. To use one of Monton's examples - a variation on the St. Petersburg Paradox (Bernoulli 1738) - suppose you are currently expected to die on what we'll call 'the Critical Date', slated 1,000 days from now. You are offered the following gamble. A fair coin is to be flipped until it lands heads after n tosses. We will extend your life beyond the Critical Date by 2^n days. If the coin has not come up heads after 1,000 flips,

we stop and your life is extended by $2^{1,000}$ days beyond the Critical Date. In order to play the game, you must be willing to move up the Critical Date by x days. If you value days of life linearly and maximize expected utility, you should be willing to move the Critical Date forward by $x = 999$ days in order to play this game. But this seems extremely reckless, given that it leaves you with a 7/8 chance of living less than 10 more days.

In response, some say we are permitted or even required to treat possible outcomes whose probabilities are suitably small as if they were impossible for the purposes of decision-making (Buffon 1777, Smith 2014, Monton 2019; compare Buchak 2013: 73-4). This may be thought to allow us to avoid the problems discussed in Section 2, since those problems can also be traced to very small probabilities of very large amounts of value. Mundane decisions, like when to get up, have some probability of significantly affecting the overall shape of the long-term future. The probability is very small, but because there is so much value at stake across the very long run, it seems we ought to try to choose the option that gives the highest probability to the better long-run outcome, insofar as our aim is to maximize expected moral value. If we go in for Nicolausian discounting, we can ignore suitably small probabilities of astronomical value. Doesn't this solve the problem?

We don't think so. We set aside general doubts about the rationality of Nicolausian discounting (Hájek 2014, Isaacs 2016). Our discussion will just focus on whether Nicolausian discounting can solve the problems highlighted in Section 2. We don't think it can.

To see why, suppose Nicolausian discounting is rational: there is a threshold below which probabilities are to be discounted down to zero for purposes of decision making. Where is the threshold? It can't be too high. No rational person treats as impossible any and all states of the world with a less than 1 in 1,000 probability, for example. Presumably, no one knows exactly at what level it ceases to be irrational to treat the very improbable as impossible for purposes of decision-making. Even if we did, we wouldn't have a good handle on whether the probabilities attaching to

possible long-term impacts associated with getting up earlier or later fall above or below it. We can't reliably make such fine discriminations among astronomically small probabilities. Like the precise Bayesian approach, Nicolausian discounting leaves rational decision-making hostage to the whim of decision-makers operating far beyond their discriminatory capacities.

We also note that otherwise plausible justifications for Nicolausian discounting may have no purchase in cases of mundane cluelessness. Monton (2019) argues that Nicolausian discounting is justified because this allows us to avoid being required or permitted to choose options which yield a high probability of our lives going very badly. If you maximize expected utility and move forward the Critical Date by 999 days in order to gamble on living 2^n days past the Critical Date, this makes it very likely that you'll die within a fortnight. This, presumably, is why taking the gamble seems so reckless. Monton writes: "Because one only lives once, one has good reason to avoid choosing an action where one has a high probability of having one's life go badly, regardless of whether or not the action maximizes expected utility." (14) Note, however, that the mundane decisions on which we're focusing aren't cases where you are offered an option that has a very small probability of yielding astronomical value, but will almost certainly end up costing you or other people very dearly. We are dealing with choices like sleeping in or getting up early. Neither choice makes it likely that things will go very badly.

In this section, we have considered and rejected four solutions to the Problem of Decision-Making Demandingness and the Problem of Decision Paralysis: gathering evidence, denying cluelessness, denying EAAL, and Nicoulasian discounting. In the next section, we consider and reject a fifth possible solution.

4. Imprecise Bayesianism

The past several decades have seen an increasing interest in forms of (what we call) *imprecise Bayesianism* (Elga 2010; Joyce 2005, 2010; Rinard 2015; Schoenfield 2012; White 2010). Imprecise Bayesians represent an agent's beliefs using a *credal set* or

representor: i.e., a set of probability functions, rather than a unique subjective probability function. A key motivation for imprecise Bayesianism is the claim that under severe uncertainty, rational agents should not be as opinionated as precise Bayesian models require them to be.

Some authors have suggested that the belief states of clueless agents may be best represented by a credal set (Greaves 2016; Mogensen ms). Given the depth of clueless agents' uncertainty, their belief states should be compatible with a wide variety of probability assignments to potential long-term future outcomes. For example, we should include some probability functions on which getting up early today is more likely than sleeping in to reduce extinction risk, and other probability functions on which the opposite is true.

An immediate concern for this solution is that it will not solve The Problem of Decision-Making Demandingness. On its face, developing imprecise probabilistic models of the long-term effects of our actions and making decisions on the basis of these models is no less time-consuming than precise Bayesian decision-making.

However, our main objection to the imprecise Bayesian solution is that it is extremely difficult to find an acceptable decision theory for clueless agents whose beliefs about the long-term impact of their options are represented by a set of probability functions.

More specifically, we think that all proposed decision rules for imprecise Bayesian agents suffer from one of four problems. The force of these problems increases as the expected values assigned by the agent's representor become more dispersed. For agents whose options' expected utilities are tightly constrained by their representor, these problems may not be devastating. But clueless agents represent their uncertainty by adopting a dispersed range of probabilities over the long-term effects of their actions. Once the full extent of a clueless agent's uncertainty is packed into her representor, it is difficult to say anything specific and plausible about what rationality requires of her.

Some imprecise decision rules suffer from *callous liberalism*: they say that almost any action is permissible. For example, consider the Liberal decision rule on which an action is permissible just in case it maximizes expected utility by the lights of some probability function in the agent's representor. For clueless agents, Liberal makes almost any action permissible. Some functions in your representor take sleeping in to be best; others recommend waking up early. Some recommend going to work; others recommend instead spending the day poisoning pigeons in the park. Intuitively, this is not the right thing to say. Although the choice of whether to wake up early or late might be left up to the whim of the decision maker, the choice of whether to go to work instead of poisoning pigeons should not be.

Other imprecise decision rules suffer from the opposite defect. They generate *widespread rational dilemmas*: in most choice situations, no action is permissible. For example, consider the Conservative decision rule on which an action is permissible just in case it maximizes expected utility by the lights of all probability functions in the agent's representor. For clueless agents, neither sleeping in nor waking up early will be permissible since each action is better than the other on some probability function. Similar conclusions will hold in most decision contexts. However, most ordinary decision problems do not pose rational dilemmas.

Supervaluationists split the difference between liberal and conservative rules by retaining their most plausible verdicts and leaving the rest as silence (Rinard 2015). For example, we might say that an option is determinately permissible just in case it maximizes expected utility on all probability functions in the agent's representor, determinately impermissible just in case it maximizes expected utility according to none of the probability functions in the agent's representor, and has indeterminate rational status otherwise. Supervaluationism is a form of *quietism*. It avoids callous liberalism and does not posit widespread rational dilemmas. But supervaluationism achieves this only by refusing to say anything at all about what rationality requires in most circumstances. Getting out of bed early has indeterminate rational status, as does getting out of bed late. So too for going to work or poisoning pigeons in the park.

A final class of decision rules return specific but *implausible verdicts*. For example, the Hurwicz criterion scores options by a weighted sum of their minimum and maximum expected utilities across the agent's representor, where the relative weights are left up to the tastes of the agent. Options are permissible just in case they maximize this weighted sum. The problem for the Hurwicz criterion is that it is entirely insensitive to anything but the best- and worst-case expected utilities. For clueless agents pondering the long-term future, the best- and worst-case expected utilities will be quite extravagant. In most cases, the worst-case expected utility involves putting significant probability mass on futures filled with astronomical suffering, while the best-case expected utility estimate involves putting significant probability mass on outcomes that involve flourishing on a galactic scale. The Hurwicz criterion focuses only on these extremes and ignores everything in between. And that is implausible.

Summing up, imprecise Bayesianism is not meant as a solution to the Problem of Decision-Making Demandingness and cannot solve The Problem of Decision Paralysis either. Imprecise Bayesian decision rules for clueless agents are either callously liberal, generate widespread rational dilemmas, lapse into quietism, or return implausible verdicts. But if imprecise Bayesianism cannot help us, what will?

5. Constraints on acceptable solutions

To get a handle on how rational decision-makers should make decisions with the potential to affect the long-term future, it will help to examine some cases. We think that in each case it is relatively clear how rational decision-makers ought to proceed. By examining these three cases, we are able to point the way to a solution to both problems.

Suppose first that you are given a billion dollars which you can donate to any cause. In this case, the first quick solution was not entirely wrong: you should gather substantial evidence to lessen your cluelessness. You should compare a wide variety of

options. Some of these options will be *longtermist interventions* aimed to address existential risks, promote flourishing and alleviate suffering on a long time-scale. You should evaluate these options using a detailed model of their long-term consequences, and possibly also by considering their impacts on the short- and medium-term future. You should also consider a variety of *benchmark interventions* aimed to address short- and medium-scale challenges such as global health and poverty. These interventions should be evaluated by detailed models of their short- and medium-term consequences, although it is unclear whether much would be gained by modelling the impacts of benchmark interventions beyond a timescale of several hundred years.

Second, suppose that you are given five hundred dollars which you may donate to any cause you like. Here, we think that you should pass the buck if possible, giving control of your decision to a larger organization with more decision-making resources. If this is not possible, you should treat this case like a hurried version of the first. You should gather some evidence to alleviate cluelessness, but less than in the first case. You should compare a range of options including some longtermist and benchmark interventions. You should evaluate the longtermist interventions based on a sparse model of their potential long-term consequences, and evaluate the benchmark interventions based on a sparse model of their short- and medium-term consequences.

Finally, suppose you are deciding whether to get out of bed in the morning or sleep in. Here, we think that you should gather no evidence. You should choose quickly, based on a sparse model of the short-term consequences of your choice.

Let us suppose that these intuitive verdicts are correct. Although high-stakes longtermist interventions ought to be evaluated by detailed consideration of their long-term effects, as the stakes decrease and options become less explicitly directed at affecting the long-term future, long-term effects become increasingly less important to rational decision-making. How can this be reconciled with EAAL? If the vast majority of ex ante value is determined by an option's long-term effects, how can it

ever be rational to reduce or eliminate our reliance on long-term considerations in decision-making? We take up this question in the next section.

6. From substantive to procedural rationality

Questions about rational choice can be posed at two levels (Simon 1976). At the level of *substantive rationality*, we ask normative questions about the first-order options facing an agent, which in this case are base-level actions like sleeping in or donating money to a specific charity. Questions about substantive rationality concern *what to do*. At the level of *procedural rationality*, we raise normative questions about the process of decision-making. For example, we ask how agents ought to make up their minds about whether to get out of bed or about which charity to donate to. Questions about procedural rationality concern *how to decide what to do*.

Substantive and procedural rationality are distinguished by the objects they consider, as opposed to the questions raised about those objects. At each level we can ask the *evaluative* question of what the best option or decision procedure would be. We can ask the *deontic* question of what option or decision procedure agents ought to take. We can ask *culpatory* questions, such as which options or decision procedures agents can be blamed for taking. And we can ask *aretaic* questions, such as which options or decision procedures a virtuous agent would use.

We take Sections 3-4 to suggest that deontic questions about substantive rationality are often intractable under conditions of cluelessness. But deontic questions about procedural rationality may be more amenable to study. The lesson of Section 5 is that we often have a reasonably good handle on the procedures that rational agents should use to make decisions of the kind that interest us. Moreover, we think that both our motivating problems are ultimately best understood as posed at the procedural level. The concern raised by the Problem of Decision-Making Demandingness is that agents apparently ought to use decision-making procedures which are highly demanding. The concern raised by the Problem of Decision Paralysis

is that there are often no clear procedures by which agents ought to make up their minds.

Some readers are likely to doubt that we can make progress by thinking in terms of procedural rationality, as opposed to substantive rationality. They will deny that there is really a separate subject matter here. It can seem natural to think either (i) that a (token) decision-making process is procedurally rational if and only if the (token) decision it yields is substantively rational or (ii) that a (token) decision-making process is procedurally rational if and only if the (token) decision to deliberate by implementing that (token) decision-making process would be substantively rational. If (i) or (ii) is correct and conditions of cluelessness induce ignorance about the criterion of substantive rationality, then clearly no progress can be made by thinking about criteria for procedural rationality instead (compare Lenman 2001: 360-1).

However, we think there is good reason to reject both (i) and (ii). More generally, we think questions about procedural rationality are not simply questions about substantive rationality under a certain guise.

Against (i), we note that a decision can be substantively rational but procedurally irrational. For example, suppose Frank has available a range of investment opportunities. In light of his evidence, buying stocks in Acme Corp maximizes Frank's expected utility. Frank invests in Acme Corp. However, his decision is not made by carefully weighing the evidence of which he is aware. Instead, he is guided by the mysterious voices in his head. Frank chose the rational option, but he nonetheless chose irrationally. His decision is substantively rational, but procedurally irrational.

Against (ii), we note that there are possible cases in which it would be rational to decide to rely on an irrational decision process. Parfit (1984: 12-13) describes a well-known case of this kind: *Schelling's Answer to Armed Robbery*. Imagine that a man breaks into your house and orders you to open the safe in which you hoard your gold, threatening that he will otherwise kill your children. As Parfit notes, if it were within your power, then it would be rational for you to render yourself incapable of

responding rationally to the man's threats and to everything else going on in this situation. In your state of madness, it would be impossible for the man to coerce you into opening the safe. Furthermore, he could count on you not to record the number of the car in which he drives away, and thus has no incentive to kill you in order to cover his tracks.

Both (i) and (ii) should therefore be rejected. We conjecture that other attempts to offer a reductive explanation of procedural rationality in terms of substantive rationality also fail. Conversely, ignorance of the criterion of substantive rationality need not mean we are ignorant about the requirements of procedural rationality. There is a whole new territory here, waiting to be explored. We think it is worth exploring.

7. Heuristics for decision-making

Many who study procedural rationality think that agents should often decide by using heuristics (Gigerenzer et al. 1999, Gigerenzer and Gaissmaier 2011) Heuristics are distinguished from Bayesian decision procedures in at least two ways: they process only a subset of the available information, and they process that information according to simple decision rules. For example, suppose you are asked to predict the winner of a tennis match. A Bayesian agent incorporates all of her relevant knowledge about each player and the match conditions to predict the winner. But suppose that you only recognize one of the players. Then you will do quite well by applying the *recognition heuristic* and predicting that the recognized player is the winner (Serwe and Frings 2006; Scheibehenne and Bröder 2007). The recognition heuristic uses only a single item of information, namely whether a player is recognized, and processes that information according to a simple decision rule.

Both distinguishing features of heuristic decision-making are reflected in the examples from Section 5. Decisionmakers should sometimes partially or fully ignore information bearing on the long-term impacts of their actions. And decision-makers

should consider fewer options, using sparser models of their relevant effects as the stakes decrease. This suggests that standard justifications for heuristic decision-making will shed light on the justification of decision procedures for longtermists.

There are three standard justifications for heuristic decision-making. The first invokes *cognitive abilities*: agents are not always capable of using more complicated Bayesian methods. In this paper, we will mostly be concerned with two further arguments. The second invokes *accuracy-effort tradeoffs*: processing a larger amount of information more completely often increases decision quality at the expense of cognitive and physical effort (Johnson and Payne 1985). Heuristics typically strike the best balance between decision quality and decision costs. The third argument invokes *less-is-more effects*: sometimes processing more information more fully predictably decreases decision quality (Gigerenzer and Brighton 2009, Wheeler forthcoming).

Less-is-more effects can be quite surprising. It seems almost a truism that more information is always better. How can incorporating more information make our decisions worse? In the remainder of this section, we characterize and explain the conditions under which less-is-more effects should be expected. Then in Section 8, we use less-is-more effects and accuracy-effort tradeoffs to characterize seven factors bearing on rational longtermist decision-making and apply these factors to solve both of our motivating problems.

Suppose you are trying to predict the longevity of cars, measured in number of miles driven during the lifetime of the car. You aim to predict longevity by linear regression, that is by building the best linear model relating longevity to observable features of a car. Your model should include highly relevant features such as make, model, and design characteristics. But what would happen if you included additional tangentially relevant variables such as the weight of the driver?

It might seem that your model would become more accurate. More information is always better, right? Wrong. The method of adding tangentially relevant inputs to improve regression fit is often parodied as *kitchen sink regression*. Kitchen sink

regression is not a good idea. Incorporating tangentially relevant inputs into regression models tends to decrease the predictive accuracy of these models.

How can this be so? The quantity of interest is the expected predictive error of a statistical model. This measures the expected inaccuracy of the model's predictions at a novel data point drawn randomly from the population. In our example, we want to assess the expected inaccuracy of the model's predictions about the lifespan of a new car. The relevant statistical fact is illustrated by the *bias-variance decomposition* of expected predictive error:

$$\text{Expected predictive error} = \text{irreducible error} + \text{bias}^2 + \text{variance}.$$

Here we see that expected predictive error is driven by three factors. The *irreducible error* is a function of noisiness in the data and cannot be controlled. The *bias* of a model measures the tendency of the model to systematically return answers that are either too high, or too low. The *variance* of the model measures the tendency of the model to return different predictions when refitted to a novel data set.

The bias-variance decomposition gives rise to a *bias-variance dilemma* (Geman et al 1992). Increasing the complexity of a model tends to increase model variance and decrease model bias. More complex models are less biased because they are better able to capture relevant statistical regularities. But more complex models tend to have higher variance because they are in a better position to overfit themselves to spurious statistical regularities in the observed data.

We can increase model complexity by changing the functional form of the model, as in moving from a linear to a polynomial model. But we can also increase model complexity by allowing the model to use all - as opposed to only some - of the available input variables. This was the problem with kitchen sink regression. Adding additional variables such as driver weight increases model variance, and this increase is not sufficiently compensated by decreased bias. That is because information about driver weight is not sufficiently informative, given other available information, to

significantly lower the model bias. As a result, kitchen sink regression tends to increase the expected predictive error of a model by increasing its variance.

More generally, the bias-variance dilemma explains less-is-more effects. In some environments, simple rules such as the recognition heuristic reliably outperform more complicated models such as linear regression, and sometimes they even outperform Bayesian methods (Gigerenzer and Brighton 2009). This is not because simple rules are less effortful, but rather because they effectively control model variance without unduly biasing the model. The claim is not that simple rules are always better than complex rules, or that any particular simple rule performs well across environments. The project is rather to study problem environments and statistical prediction and decision-making rules in order to understand when and why a given rule is appropriate (Gigerenzer and Todd 2012). In many cases such as recognition-based decision-making, we can give detailed mathematical (Davis-Stober et al 2010, Hogarth and Karelaia 2005, Katsikopoulos 2010) and empirical (Goldstein and Gigerenzer 2002, Pachur et al 2011) characterizations of these conditions.

Summing up, decisionmakers can sometimes make better predictions and decisions by employing simple decision rules due to the bias-variance dilemma. Simpler rules prevent overfitting by keeping variance manageable, thereby reducing predictive error. In the next section, we put this insight together with the accuracy-effort tradeoff in order to shed light on procedurally rational longtermist decision-making and solve our motivating problems.

8. Application to longtermist decision-making

In Section 4, we gave three examples of rational procedures for longtermist decision-making. In each example, we claimed, it is relatively clear which procedures longtermist decision-makers ought to use. But why should longtermists use these procedures? More generally, which factors bear on the procedural rationality of longtermist decision-making?

We are now in a position to see that two types of factors bear on the procedural rationality of longtermist decision-making. The first set of factors relate primarily to accuracy-effort tradeoffs. We can often increase the quality of our decisions by taking full account of their long-term effects, but after a while the increase in decision quality is outweighed by decision costs. These factors include the *decision costs* of employing a particular decision procedure, measured in physical costs such as money and energy as well as opportunity costs from lost time. They also include the *predictability* of long-term effects. Often we cannot make substantially better predictions or decisions because we are not in a good position to predict the long-term effects of, for example, getting out of bed early. Also relevant are the *stakes* of decision-making: how important is it to make a high-quality decision? While it is quite important to make the best possible use of a billion-dollar charitable endowment, it is less important to ensure that five-hundred dollars are used as well as possible. These factors together favor processes that de-emphasize long-term effects in order to reduce the costs of decision-making.

Another set of factors relate primarily to the *bias-variance dilemma*. These are reasons, independent of decision costs, to de-emphasize long-term effects in decision-making. These factors include the *number of variables* in a predictive model. To accurately predict the long-term future, decision-makers must draw on a much larger range of information than in predicting the short- and medium-term future. Model complexity increases as new input variables are incorporated. Another factor is the *number of auxiliary assumptions* that must be made. For example, we might constrain the functional form of the model by making substantive empirical assumptions about the long-term effects of physical parameters. Perhaps we estimate the value of extinction risks by estimating future population size, and these estimates might incorporate a specific functional relationship between the speed of future space travel, the density of nearby planets, and the future human population. These auxiliary assumptions introduce additional degrees of freedom to the model insofar as they are not made on

the basis of substantial data and are not fully constrained by the arguments available to even the most careful decision-makers.

Another factor raised by this example is *estimation bias*. Often the inputs to our model, such as the speed of future space flight, are not observed but estimated, and these estimations are a source of substantial bias. Finally, we should be concerned with model *testability*. Bias and variance can be measured and controlled by standard tools such as cross-validation (Arlot and Celisse 2010). But these tools require actual observations of the quantities to be predicted. We have never observed the long-term effects of rising early, nor have we observed anything similar which might be used to test our models. Similar problems affect auxiliary assumptions and estimation bias in estimating model parameters. These factors together suggest that even if the long-term effects of options typically dominate their ex-ante values, longtermist decision-makers may not always substantially improve the quality of their decisions by taking long-term effects into consideration.

The examples in Section 5 suggest that these two sets of factors combine to determine the procedural rationality of longtermist decision-making. We are not so pessimistic as to assume that less is always more. We think that longtermist decision-makers with substantial resources at their disposal can probably do better by constructing detailed models of the long-term effects of longtermist interventions. However, we think that as the stakes of decision-making decrease, accuracy-effort tradeoffs become more pressing. Even if the best long-term models reliably outperform simpler short- and medium-term models, rational decision-makers should often use simpler models and there is no reason to suspect that these models will be improved by taking quick and dirty shortcuts to account for long-term effects. We also suspect that even in high-stakes contexts, it may be inappropriate to model the long-term consequences of benchmark interventions whose long-term effects may be less predictable and less strong.

These remarks put us in a good position to address our motivating problems. There is no procedural Problem of Decision-Making Demandingness. Decision costs

are straightforwardly relevant to the procedures that rational decision-makers ought to use. If rational decision-makers are occasionally required to use demanding decision rules, that is because we sometimes make important decisions in which steep decision costs are compensated by increased decision quality.

We have also made substantial headway on the procedural Problem of Decision Paralysis. As the discussion in Section 4 illustrates, it is often quite clear which procedures rational longtermist decision-makers ought to use. In Sections 5-7, we made analytical progress on the problem of Decision Paralysis by listing two types of factors bearing on procedural rationality for longtermists. While more work is needed to clarify the precise demands on procedurally rational longtermist decision-making, we hope to have motivated our claim that it is much clearer how longtermist decision-makers ought to make up their minds than how they ought to act.

The procedural solutions to our motivating problems convey a pair of normative lessons. First, demandingness and paralysis are not decisive objections to EAAL. The force of these objections is mostly procedural, but at the procedural level both objections can be answered. Second, the turn from substantive to procedural rationality is a neglected and generalizable normative maneuver. Normative theorists have begun turning their attention from substantive to procedural rationality (Friedman forthcoming). We think that a careful distinction between procedural and substantive questions is a useful device for resolving difficult normative problems.

References

Adams, Fred C., ``Long-term astrophysical processes,'' in Bostrom and Cirkovic (eds.), *Global catastrophic risks* (Oxford: Oxford, 2008)

Arlot, Sylvain, and Alain Celisse, ``A survey of cross-validation procedures for model selection,'' *Statistics Surveys* 4.1 (2010): 40-79.

Beckstead, Nick, *On the overwhelming importance of shaping the far future* (2013), PhD Thesis, Department of Philosophy, Rutgers University.

Bernoulli, Daniel, ``Specimen theoriae novae de mensura sortis,'' *Comentarii Academiae Scientarum Imperialis Petropolitanae* 5 (1738): 175-92.

Bostrom, Nick, ``Astronomical waste: the opportunity cost of delayed technological development,'' *Utilitas* 15 (2003): 308-14.

Bostrom, Nick, ``Existential risk prevention as a global priority,'' *Global Policy* 4 (2013): 15-31.

Broome, John, *Climate matters: ethics in a warming world* (NY: Norton, 2012).

Buchak, Lara, *Risk and rationality* (Oxford: Oxford, 2013).

Buffon, Georges-Louis, ``Essai d'Arithmétique morale,'' *Supplement a l'Histoire Naturelle* 4 (1777): 46-123.

Davis-Stober, Clintin, Jason Dana, and David Budescu, ``Why recognition is rational: optimality results on single-variable decision rules,'' *Judgment and Decision Making* 5.4 (2010): 216-29.

Elga, Adam, ``Subjective probabilities should be sharp,'' *Philosophers' Imprint* 10.5 (2010): 1-11.

Friedman, Jane, ``Teleological epistemology,'' forthcoming in *Philosophical Studies*.

Geman, Stuart, Ellie Bienenstock and René Doursat, ``Neural networks and the bias/variance dilemma,'' *Neural Computation* 4.1 (1992): 1-58.

Gigerenzer Gerd and Henry Brighton, ``Homo heuristicus: why biased minds make better inferences,'' *Topics in Cognitive Science* 1.1 (2009): 107-43.

Gigerenzer, Gerd and Wolfgang Gaissmaier, ``Heuristic decision making,'' *Annual Review of Psychology* 62.1 (2011): 451-82.

Gigerenzer, Gerd, Peter Todd, and the ABC Research Group, *Simple heuristics that make us smart* (Oxford: Oxford, 1999).

Goldstein, Daniel and Gerd Gigerenzer, ``Models of ecological rationality: the recognition heuristic,'' *Psychological Review* 109.1 (2002): 75-90.'

Greaves, Hilary, ``Cluelessness,'' *Proceedings of the Aristotelian Society* 116.3 (2016): 311-39.

Greaves, Hilary and William Macaskill, ``The case for long-termism,'' ms.

Hájek, Alan, ``Unexpected expectations,'' *Mind* 123.490 (2014): 533-67.

Hogarth, Robin and Natalia Karelaia, ``Ignoring information in binary choice with continuous variables: when is less `more'?,'' *Journal of Mathematical Psychology* 49.2 (2005): 115-24.

Hublin, Jean-Jacques et al, ``New fossils from the Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens,'' *Nature* 546.7657 (2017): 289-92.

Isaacs, Yoaav, ``Probabilities cannot be rationally neglected,'' *Mind* 125.499 (2016): 759-62.

Johnson, Eric and John Payne, ``Effort and accuracy in choice,'' *Management Science* 31.4 (1985): 395-414.

Joyce, James, ``A defense of imprecise credences in inference and decision making,'' *Philosophical Perspectives* 24.1 (2010): 281-323.

Joyce, James, ``How probabilities reflect evidence,'' *Philosophical Perspectives* 19.1 (2005): 153-78.

Katsikopoulos, Konstantinos, ``The less-is-more effect: predictions and tests,'' *Judgment and Decision Making* 5.4 (2010): 244-57.

Knight, Frank, *Risk, uncertainty, and profit* (Boston: Houghton-Mifflin, 1921).

Lenman, James, ``Consequentialism and cluelessness,'' *Philosophy and Public Affairs* 29.4 (2000): 342-70.

Mogensen, Andreas, ``Maximal cluelessness,'' ms.

Monton, Bradley, ``How to avoid maximizing expected utility,'' *Philosophers' Imprint* 19.18 (2019): 1-25.

Pachur, Thorsten, Peter Todd, Gerd Gigerenzer, Lael Schooler, and Daniel Goldstein, ``The recognition heuristic: a review of theory and tests,'' *Frontiers in Psychology* 2.1 (2011): 1-14.

Parfit, Derek, *Reasons and persons* (Oxford: Oxford, 1984).

Rawls, John, *A theory of justice* (Cambridge: Harvard, 1999).

Rinard, Susanna, ``A decision theory for imprecise probabilities,'' *Philosophers' Imprint* 15.7 (2015): 1-16.

Scheibehenne, Benjamin and Arndt Bröder, ``Predicting Wimbledon 2005 tennis results by mere player name recognition,'' *International Journal of Forecasting* 23.3 (2007): 415-26.

Schoenfield, Miriam, ``Chilling out on epistemic rationality: a defense of imprecise credences (and other imprecise doxastic attitudes),'' *Philosophical Studies* 158.2 (2012): 197-219.

Serwe, Sascha and Christian Frings, ``Who will win Wimbledon? The recognition heuristic in predicting sports events,'' *Behavioral Decision Making* 19.4 (2006): 321-32.

Smith, Nicholas, ``Is evaluative compositionality a requirement of rationality?,'' *Mind* 123.490 (2014): 457-502.

Wheeler, Gregory, ``Less is more for Bayesians, too,'' forthcoming in Viale (ed.), *Routledge Handbook of Bounded Rationality*.

White, Roger, ``Evidential symmetry and mushy credence,'' in Szabo Gendler and Hawthorne (eds.), *Oxford Studies in Epistemology*, vol. 3 (Oxford: Oxford, 2010).