Synthese (2010) 176:5–21
DOI 10.1007/s11229-009-9481-4

Human bounds: rationality for our species

Adam Morton

Received: 6 February 2009 / Accepted: 6 February 2009 / Published online: 5 March 2009

Abstract Is there such a thing as bounded rationality? I first try to make sense of the question, and then to suggest which of the disambiguated versions might have answers. We need an account of bounded rationality that takes account of detailed contingent facts about the ways in which human beings fail to perform as we might ideally want to. But we should not think in terms of rules or norms which define good responses to an individual's limitations, but rather in terms of desiderata, situations that limited agents can hope to achieve, and corresponding virtues of achieving them. We should not take formal theories defining optimal behavior in watered-down bounded form, even though they can impose enormous computational or cognitive demands.

Keywords Bounded rationality · Mistakes · Optimality · Cognitive limitations · Intellectual power · Accuracy · Norms of reason

Is there such a thing as bounded rationality? Don't the boundedness and the rationality pull in opposite directions, boundedness lowering standards and rationality raising them? The question is not well-formed, if only because both "bounded" and "rational" are such slippery terms. In this paper I first try to make sense of the question, and then to suggest which of the disambiguated versions might have answers. I shall defend one very general positive theme and two very general negative themes. The positive theme is that we should take seriously the very daunting task of constructing an account of bounded rationality that takes account of detailed contingent facts about the ways in which human beings fail to perform as we might ideally want to. How radical this

A. Morton (✉)
Department of Philosophy, University of Alberta, 2-40 Assiniboia Hall, Edmonton, AB, T6G 2E7, Canada
e-mail: adam.morton@ualberta.ca

project is, and how daunting, is brought out by the two negative themes. The first of these is that we should not think in terms of rules or norms which define good responses to an individual's limitations, but rather in terms of desiderata, situations that limited agents can hope to achieve, and corresponding virtues of achieving them. The second is that the standard formal theories defining optimal behavior - theories of probability, statistics, utility, and truth-in-a-model - should *not* be taken in watered-down bounded form but in full unrestricted form, even though they can impose enormous computational or cognitive demands. An interesting feature of the way I structure the issues is that the second negative conclusion is supported by the first, rather different, one.

1 The bounded landscape

It is important to distinguish two different ambitions for an account of bounded rationality. We may be looking for a theory which describes the best actions and patterns of thinking within the reach of agents with specific kinds of limitations, whether or not they have any knowledge that these are their limitations and whether or not they can understand the patterns of action and thought that would be of benefit to them. Or we may be looking for a theory which describes good reactions to the fact that one's capacities have specific limitations. By good reactions I mean reactions that one could recommend to others, or try to hold oneself to. These are different. The first suggests ways of cutting down the ways of thinking that unlimited agents could use to sub-procedures available to mortals, while the second suggests specific adaptations to the finite condition, and indeed to the condition of our particular species. In fact there is not much point engaging in the latter project unless one is prepared to consider the particular hazards of being a human being. So it will not be a purely apriori project. An example of the first kind of theory would be an account of constrained maximization, cutting full expected utility theory down to size. An example of the second would be the advice that sometimes it makes sense to ignore or even forget facts that are of no apparent use and which are cluttering up precious memory space. A type one theory might explicitly not use more information than needed, but it would not suggest throwing it away, while a type two theory might.

It is also important to have some kind of an explicit idea what kinds of limitations we are talking about. It is easiest to do this from a very abstract point of view, but the classifications we get from considering rational agents in general may not be the most apt for type two recommendations. It may be necessary to take limitations of long term, short term, and working memory separately, for example. And each of these may be a cluster of separate capacities, with different limitations which are best accommodated in different ways. But even thinking in the abstract we can distinguish intellectual or computational power, on the one hand, and intellectual correctness, accuracy or responsibility on the other. It is easy to think of a very powerful mind that makes many mistakes, or is just plain silly. No one who has spent time around a university should find the concept difficult. More subtly, it is obvious that one person can be more intelligent than another but make worse decisions and be less epistemically reliable. The complementary pattern is equally plausible though a bit more puzzling:

one person can be relatively low in intelligence and relatively high in rationality. The sensible plodder.

The power/correctness distinction can be puzzling because we realise that limitations on thinking power will make some good but difficult reasoning unavailable to a limited person. But since we all have less thinking power than we might that only shows that we all often do not use the best reasoning. No doubt there is a quick way to schedule my eight competing obligations for next week, involving the independence of the axiom of choice, if only I could see it. So in a rough vernacular way we do have space in our thinking for reasonableness in the face of intellectual bounds. It is significant that one characteristic of the sensible plodder is an appropriate modesty: he does not take on tasks that are too hard for him. (The concept of intelligence cries out for a critical philosophical analysis. Once it is separated from the concept of rationality, and once we realise that high intelligence is consistent with absence of crucial intellectual virtues, hard questions about the relation between measurable IQ and intelligent behavior as informally conceived. See Gardner (1999) and the essays in Sternberg and Pretz (2005). In this paper I use the concept of intelligence in a naïve way, wishing that I had a more sophisticated tool.)

There are obviously bounds on rationality that do not derive from limited thinking power. We are all subject to wishful thinking, for example, and some resist it better than others. This includes epistemic wishful thinking, intellectual partizanship. But that example shows how subtle this irrationality is. It is not irrational to have a favourite hypothesis, and to look zealously for evidence supporting it; what is irrational is to ignore evidence against it. But it is often not obvious when evidence tells against a hypothesis: how much energy does one have to put into exploring the implications of evidence that seems insignificant but might conceivably threaten a cherished conjecture? And then there are the notorious statistical blind spots—gambler's and base-rate fallacies and getting conditional probabilities the wrong way round—that we all fall into if they are slightly disguised or the terrain is unfamiliar. These can be avoided by thinking slowly and from first principles. But this makes for greater combinatorial complexity and raises the likelihood of mistakes from simple overload. So we need, here and elsewhere, the capacity to choose between on the one hand fallacy-prone short-cuts and on the other hand overload-threatening first principles, as appropriate. Is choosing appropriately between these an attribute of rationality or some other intellectual virtue? I don't think this question has a definite answer. The language of rationality has been too deeply shaped by the image of the powerful systematic thinker who conquers problems by logically bulldozing through complexity, for it to adapt easily to such real life subtleties, central and ubiquitous though they are.

We now have two dimensions of theory. On the first dimension we can be either cutting down perfect rationality or, contrasted with this, describing good adaptations to imperfection, recommendations to imperfect agents. And on the second dimension we can be considering limitations of power, limitations of accuracy, or both. These are not inevitably in competition with one another. They are different theoretical ambitions, though not all of them may be achievable. There is a third dimension that should be mentioned. We may or may not be concerned with agents who are not only limited in thinking power and rationality but also do not follow the principles that such limitations suggest. Suppose that a theory of bounded rationality suggests that a person

should follow a certain procedure, and then as a result of not following it they get into some predicament, such as contradictory beliefs or intransitive preferences. Should the theory say what the person should do then? Should it take into account bounds in our ability to exhibit good bounded behavior? I mentioned sensible plodders, people who though limited in their intellectual capacity manage to stay out of trouble. There are foolish plodders too, people who from not understanding or managing their limits court trouble. A theory may or may not describe sensible procedures for foolish plodders, including incurably foolish plodders. (See Taszka and Zielonka (2002) for examples of agents who are and are not aware of the inaccuracy of their predictions. Thanks to Ray Dacey for discussion on this point.)

A simple consideration that reveals differences between these projects is the possibility of slips (random mistakes, performance errors, "mentos".) One can imagine a coherent and illuminating account of procedures available to a limited agent, respecting limitations of memory and speed, which ignored the possibility that the agent would sometimes just goof: think "3.41951" when it should be "3.14159" or "Avicenna" when it should be "Averroes". It would be a theory of a faultless limited agent. One place the difference would show up would be with procedures that degrade badly, working well as long as they are strictly adhered to but leading to disaster if implemented imperfectly. Such procedures are not attractive from a point of view that describes sensible human reactions, if in fact the slips in question are likely when humans attempt to follow the procedures. Suppose that there is a way of avoiding some class of slips, and the theory describes it. Would it be part of the task of an ideal theory to suggest ways of recovering if out of ignorance or foolishness a person ignores this slip-minimizing strategy? It would on some conceptions of the project of bounded rationality, and would not on others.

Slips are very contingent aspects of human psychology, though their presence affects us deeply. Rational choice theory is traditionally not a very naturalistic discipline. Contemporary behavioral economics attempts to bring a whiff of psychological realism to the issues, but balks at the task of separating the profitable from the unprofitable in the maze of quirky details about our actual thinking habits. Some contemporary epistemology does not shirk from the task. It is a daunting task, and one can sympathise with philosophers who would rather not tackle it, but by the end of this paper I will have urged us to embrace it. Here is an example that brings out the distinctive contingent flavour of human-sized bounded rationality. Work by the psychologist Ap Dijksterhuis suggests that some hard problems are best solved when one is distracted. Dijksterhuis gave subjects problems to which there are unique best solutions (usually about the choice of an apartment meeting a set of criteria). Some were given very little time to think, some were given longer and allowed to concentrate, and some were given longer but distracted by being given a quite different topic to occupy their conscious attention. The people given very little time did worst (found the right answer least often.) For easier topics the people given more undistracted time did better than the people given more but distracted time. But beyond a certain level of difficulty, measured in terms of the number of variables to compare and the number of options to choose between, the people given more but distracted time did better. (See Dijksterhuis and Nordgren 2006. I am grateful to Patricia Churchland for telling me about this work.) Dijksterhuis interprets these results in terms of the greater complexity-handling capacity of uncon-

scious processes and the tendency of conscious attention to interfere with the smooth operation of these processes. This is a very plausible interpretation but it does not matter for my purposes whether it is the right one. The importance of this data for our purposes now is that, in order to get the best results, a person should sometimes concentrate hard and give a lot of conscious meticulous attention, and sometimes allow a suitable level of distraction. Sometimes you should take a long walk, or go to a movie, or sleep on it. But not always. Which problems are of which kind, for you? That is very hard to know, in many cases. Suppose that a person given a hard problem thinks very hard and diligently and fails, when she would have solved it had she spent the time at a party. But, suppose, it is not obvious that this is such a case. Is the person irrational, for using a sub-optimal technique? Can we say that she is not irrational while still condemning as irrational her sister who does go to the party and as a result does not solve an easier problem? I doubt that these questions have semantically naïve answers.

(One of the few philosophers to have engaged seriously with the fact that the most intuitively rational of us still make simple mistakes is Paul Weirich. In chapter 6 of Weirich (2004) he wrestles with the question of evaluating acts which are well-thought-out on the basis of flawed information or preferences. His analysis is based on the idea of a corrigible mistake. He might say that the mistaken party-avoider just mentioned acts rationally if she could not have corrected her mistake, since she could not have known where the threshold for appropriate distraction lies.)

2 Good procedures versus good advice: Pollock

The first two dimensions of bounded rationality that I described above can be taken as differing in that one describes procedures that if followed would solve our problems at a reasonable price and the other describes procedures that we can profitably try to follow. Both are valuable aims, but they are certainly not the same. There may be—there almost certainly are—cognitive patterns which would give us good results given our limitations, but which we are most likely to bungle if we try following them. For one thing, they may be so complicated that the effort of following them would overload our working memory and information-storage capacities, using up the very resources they were meant to economize. Moreover, how likely mistakes are in following a given procedure will often be a matter of detailed contingent human psychology, very hard to anticipate from an abstract perspective. For example, recent studies indicate that human males, but not females, tend to be more risk-prone in the presence of sexually stimulating material, even when the risks involved have no connection with sex. So a financial decision-making method that involved labelling the options with, say, names and images of film stars might work very badly for ways that are invisible to an abstract conception of bounded and fallible rational agents (See Knutson et al. 2008).

One response, explicit in writings of Paul Weirich and implicit in the practice of John Pollock, is to renounce all connections with formulas that one can deliberately follow to good effect. (See Weirich 2004, Chap. 3; Pollock 2006, Chap. 1) On this approach, reminiscent of a certain rule-utilitarian line, the aim of theory is to identify good routines for agents, whether or not trying to follow them would have good results.

The insight here is just that the two aims are different. There need not be a single aim of theory. Different projects can target optimal processes and optimally followable advice. The important thing is to realise that these are different.

The issues are illustrated by the account developed in chapters 9 and 10 of John Pollock's recent *Thinking about acting* (2006). Pollock grapples with the tension between two facts. On the one hand the consequences of doing A and B are likely to be very different from doing A alone, or B alone, so that there is not much point asking whether one should do a finely specified atomic act. This suggests that the objects of evaluation ought to be larger scale plans. But on the other hand the larger a plan becomes the harder it is to evaluate or even to specify in a form that allows a detailed evaluation. So "move my little finger" is too small, but "adopt plan of life P" is too big. Traditionally, decision theory focuses on acts between these in size: bringing red wine or white, taking this job or that, making this investment or that. But it is not obvious what this natural target is, and attending only to options that are intuitively within it leaves important questions about the other scales unanswered: how to generate candidates for detailed evaluation, how to form larger-scale plans. Pollock's reaction is to describe a complex procedure for evaluating acts in a roughly expected-utility way in the context of larger-scale plans. The procedure is meant to take the agent's limitations into account, and relies on computations that while intricate do not require unbounded processing power. The results are only reasonable if the larger-scale plans are good, though, so these too have to be evaluated. Expectational analysis would be too cumbersome here, so Pollock adopts a good-enough satisficing-like technique to rule out crazy plans of life.

For present purposes there are two interesting points. The first is how the difficulty of optimizing within cognitive bounds creates a situation in which complexity is ultimately unmanageable. The other point is the way in which what is explicitly meant as a description of a set of cognitive procedures, rather than an intellectual strategy, implicitly describes such a strategy. For if Pollock's procedures are good ones, that would suggest that a good general strategy is to think as precisely as one can about one's actions, one by one, taking into account as much as possible of the limitations of one's information and one's depth of thinking, directing the process in a general motivational direction that one does not subject to such intense scrutiny. Contrast this with the opposite: thinking as hard as one can about one's general aims in life and then choosing actions if they fit in with it. (This has sometimes been suggested by high-minded philosophers: one should think hardest about values.) There are obviously many positions in-between. If we take the extremes to be meticulousness act-by-act and only cursory examination of Life (not Pollock's position), and meticulousness about Life and only cursory examination act-by-act, then it is intuitively obvious that the best policy is somewhere in the middle. But it is not at all obvious where in the middle it is, or how we might specify it. And it is not at all obvious that there is a unique stable point in the middle. (I have developed these points further in a review of Pollock: Morton to appear.)

There is another way of putting the second of these points. Take a procedure that evaluates an action by associating a quantity with it, such as its expected utility. Suppose that this procedure is cognitively expensive. Cut down the procedure to something more manageable, but suppose that it too defines a quantity, for example

"approximate expected utility". Then we have the resources to define a cognitive goal, that of maximizing approximate expected utility. And we might think that this goal is a replacement for the goat of maximizing expected utility, a goal that is appropriate for limited agents. That is exactly what we should *not* do, in my opinion. We should not aim at maximizing scaled down utility, or evaluating theories in terms of scaled down force of evidence, or evaluating consistency relative to a scaled down universe of models. When we think in terms of utility or force of evidence we should do our best to find out what they in their un-scaled real form actually are. We should also understand that our estimates of them are often imperfect, and we should act accordingly.

3 Three fallacies

Here are three wrong assumptions. If any of them were right, bounded rationality would be in some respect more like ideal rationality than it is, and the difference between scaled down ideal rationality and human-centred bounded rationality would be less great.

> *the approximation fallacy* "Humans are very imperfect, and we shouldn't expect to get perfect results. But we can try. We can determine what an ideal agent would do and then approximate that to the best of our abilities. Then the results will approximate as closely to ideal results as is possible for a mere human."

The mistake here is in the last sentence. It is not true that if one approximates to an ideal agent one will do as well as one can. The reason is seen most easily with an analogy to the moral case. Some people ask "what would Jesus do?" (Or Mohammed, Ghandi, Mandela, …) And then they do or recommend it. But this is foolish. Jesus could safely put himself in the way of temptation that it would be fatal for you to expose yourself to. Mandela could resolve conflicts that you should stay out of. In general, we often meet situations in which the best action is to do A and then B, the second best is not to do A, and the third best is to do A and not follow with B. But B may be an act that you can expect yourself not to perform. The dynamic choice literature is built on such examples. (See for example Jackson and Pargetter 1986; Rabinowicz 1995) In such situations the ideal agent would do A and B, an approximator to the ideal would do A and not B, and an agent well-adjusted to her limitations would not do A. Situations of this general form occur in non-moral situations, too. We do not wisely undertake projects whose later stages would require more intelligence, speed, memory, or imagination than we have. Instead we settle for second best, which often means starting off in a very different way.

The approximation fallacy is closely related to the reasons why we should focus on full utility, evidence, and so on, rather than scaled-down surrogates. Consider a situation in which a person is faced with a very large number of options, some of which are very hard to evaluate. Represent this as a tree with 1000 nodes immediately above the initial node, referred to as nodes 1 to 1,000, with branches above them of increasing length and complexity. Suppose that all trees growing from the n'th of the nodes immediately above the origin are of height n and that nodes of height m have 2^m successor nodes, of equal probability. Payoffs are at terminal nodes, so that a branch

cannot be evaluated until one has explored it to its end. The payoffs are all values from 1 to $2^{1000} - (= \sum 2^s, 1 \leq s \leq 1000)$, distributed randomly among the terminal nodes. The complexity of the whole tree is enormous, well beyond human capacity for exhaustive search, but well within the limits of actual situations, if they are taken with all their potentially relevant details. The mitigating feature is that some easily identified branches are much simpler. An ideal agent would explore all branches and select the one with the highest payoff. A bounded agent might use any of a great variety of approximations to full exploration. She might begin with node number one, "the leftmost", and explore all trees branching from it, proceeding rightwards to explore successively taller trees until she reaches nodes too large to process, and then choosing the option with the greatest payoff among those she has considered. Or she might set a bound in advance on the numbers of initial nodes to consider and a bound on the number of successor nodes to any node to consider, always proceeding from left to right. The point of proceeding from left to right is in both cases to ensure that at least some terminal nodes are reached, so that the best of these could be chosen. Or she might proceed in indefinitely other ways, all approximating to a full search of the tree. (See Pearl 1984 for a taxonomy of tree-searching procedures, and a discussion of their comparative merits, which usually depend on details of the situation.)

Now compare any of these to a non-approximating procedure. We choose ten level one nodes at random and explore each of them by following just one successor node at each stage, chosen at random, so that we have ten randomly chosen complete branches. Comparing these three procedures—the complete search, the approximation to it, and the guesswork method—it is obvious that the complete search will always do better than either of the others. (Obvious too that it is usually out of the question as a practical procedure.) And it is clear on a little thought that the guesswork method will usually do better than the approximation, as long as the cognitive bounds are relatively small relative to the number of options. For since the majority of terminal nodes will be on branches that the approximate procedure will never reach, unless the bounds are so high that the agent is almost ideal, a random selection of terminal nodes will usually contain terminal nodes with higher values than are considered by either of the two approximate procedures. In fact, I expect that one can make plausible definition of "bound", "approximate procedure", and "random procedure" such that most random procedures, for most not too large bounds, most often do better than most approximations.

The paradigm dynamic choice situation, in which trying naively for the best outcome will lead to a worse result than renouncing it, is thus reproduced at the level of decision-methods. Naively approximating to an ideal procedure can lead to a worse result than thinking in entirely different terms. Or, to link to a theme of this paper, trying to maximize scaled down utility (the number that comes out of an approximating algorithm) can result in less real utility than following a procedure that one has reason to believe very often gives a good amount of real utility.

There are many ways in which one could have reason to believe that a procedure gave good results. One would itself incorporate a kind of approximation, to an imperfect model. If some model agent shares many of your limitations but makes better choice of means and ends, then you will do well to imitate her ways of thinking. The choice of a model and the choice of which aspects to imitate can both be subtle questions.

Suppose that you can do any act that the model can, but cannot make decisions as well as the model can. Then you should imitate the model in the sense that if you know the model will do a certain act in a certain situation you should at least seriously consider doing it yourself. But you should not imitate the model's larger procedures, such as those described as "collect data of kind D, list options according to criteria C, and then pick an action", as you may emerge having picked different actions than the model would have. You should imitate the results of the model's decision-making but not that decision-making itself. Knowing what those results would be is going to be difficult, though, if you cannot find out by duplicating the decision-thinking yourself. A theory giving the acts of the model as a function of the circumstances would not be subject to this objection. But there is something very weird about having access to a theory that gives you the effect of thinking you cannot do. One would expect that in general when a pattern of thought is too hard then a theory describing it is too hard. (In the case of ideal rationality, when for example a deduction is too complex for ordinary humans, the proof that it follows from a set of axioms for logic will be at least as complex.)

The natural use of other imperfect humans as models is traditional: one learns various virtues by being in the presence of suitable role models and letting their habits rub off on one. One does not have to understand how their thinking works: one simply has to come to do it. "Simply" from the point of the view of the learner; the actual process must be subtle and complicated. It is natural to conjecture that it involves observing the model's use of a taxonomy of situations and responses and building up a similar pattern-recognition skill in oneself. This must be a very topic specific business: one learns a particular virtue with respect to a particular topic by contact with a particular exemplar. And it must be a very contingent psychological matter which virtues can be learned in this way.

Though approximation to the ideal is not generally the best strategy, the best non-ideal strategy is usually hard to find, sometimes so hard that it is not worth searching for. That is because recognizing a situation as being of the crucial type, one in which to approximate imperfectly to the best is worse than to take a completely different route, often requires that one know that an anticipated problem will be too hard for one. But in general it is very hard to know how hard a problem is (see Morton 2004). And as a result the reasonable bounded agent will often not know whether or not it is best to approximate. Sometimes it will be clear that finding the best non-approximating sequence is too hard a task. And in those cases the best course may be to throw away sophistication and do one's best to act as if one was ideal. Life is complicated.

> *the composition fallacy* "if a task breaks into two parts, and I do the first part as well as I can, given my limitations, and I do the second equally efficiently, then their combination will be done as well as I can do it."

The fallacy is obvious, if we take "as well as I can, given my limitations" in terms of staying within a fixed resource budget. For clearly the best way of doing part 1 may use most of the resource, as may the best way of doing part 2, but the combination of these is impossible within the budget. Combining efficient processes to make efficient processes is not a trivial business. The situation is complicated by the fact that determining how expensive a procedure is can itself be a difficult matter. So sometimes a

person cannot determine whether their whole proposed sequence is manageable, only that the parts are, relative to one way of breaking it into parts. That leaves the person ignorant of the total bill.

The point can be illustrated with Pollock's proposed solution to the small-scale/large-scale problem illustrate the point (not that Pollock is committing the fallacy.) Each stage of increasingly large-scale planning may be manageable by some sort of limited utility maximization, until a point where the scale is too big, and a very rough satisficing is needed to settle large aims in life. That may not be the best way to perform the large-scale thinking, but it is the best that remaining resources allow. It might have been better to inject some satisficing into a smaller-stage element, even at the cost of worse results, in order to free up resources for the larger scale. It is hard to tell.

A simpler example comes from the winnowing of decision-options. In most situations there are countless possible acts that might be chosen, far too many to give all of them more than a moment's consideration. Some seem crazy, but it is always possible that on close examination one of these will turn out to be a brilliant solution. Some seem brilliant, but on closer examination some of these will be crazy. The intelligent reaction to this has to be some sort of sequential process, in which a very quick pass through the whole set leads eventually to a detailed consideration of a small subset. There are many ways of doing it, and selecting the right one in a particular situation is a hard and delicate problem. (So hard that it is best not to think about it: but that is to anticipate the discussion of the metaresource fallacy below.) Some ways, satisficing-like, may not require one to give any attention at all to some members of the set. Consider a two stage procedure, involving an initial filtering which feeds options to a detailed action-chooser. Suppose now that there is a way of performing the initial filtering that works as well as could be expected: most of the time it will leave for later consideration the option that a really detailed consideration would reveal to be best. Suppose moreover that it does this in a way that is feasible given the agent's resources and time constraints. Now suppose there is a way of doing the second stage, the choice of an action to perform, that is also good and also feasible given the constraints. Can we be sure that the combination of the first stage and the second stage is a good way of choosing given the constraints? Not at all. In fact it may not meet the constraints. This is clearest for the time constraints: each may be within the overall constraint but take so long that it does not leave adequate room for the other. It also applies to memory limitations and constraints on the total effort committed to the decision. For example the initial winnowing may along the way produce information about features of actions, for example their worst case outcomes, that would be useful to the final choice of action. So the combination of the best ways of first finding a subset which have no obviously lead to disaster, and then finding the member of this subset with the best detailed expected consequences, may be less good that the combination of a more cumbersome first stage that passes on more information to the second stage.

(In more detail: the first stage may involve searching a tree to width w and depth d, as a result of which branches are selected which constitute a tree of width $w' < w$ which at the second stage is searched to depth $d' > d$. A full description of the first stage would specify whether the information passed on to the second stage consists just of a list of initial nodes or of some information about the results of the bounded

search. Which is better will depend on the details of the tree, and the situation in which the decision is being made.)

> *the metaresource fallacy* "if a procedure is the best, then we can know it is best, and so the best decision will be to use it."

The mistake here is in ignoring the costs of discovering that it is the best procedure in the given circumstances. This is illustrated by the "distraction" phenomenon described above. Some sufficiently complex decisions are best made when one has prevented one's conscious mind from interfering too much. How complex? You are unlikely to know; in fact you are unlikely even to have the conceptual means to describe a problem in suitable terms. So a rule of thumb that very often results in a person's avoiding too-explicit reflection on too-complex problems may well also miss-assign problems of marginal complexity. Then a person may be operating in the best way she can but as a result using a method that is not best for the situation at hand.

There is a complex relation between the metaresource fallacy and the approximation fallacy. It is hard to avoid both simultaneously. A wise person does not approximate to the behavior of an ideal one when this would involve facing problems she can expect not to solve. But telling which problems these are, is itself hard, sometimes so hard that a sensible person will not try, and will instead use the ideal behaviour as a rule of thumb guide. So in some classes of problems she will use a simple rule of approximating the ideal, knowing that this will sometimes have bad results, but also knowing that the effort of discovering which cases these are is likely to be greater than the benefit of identifying them. For example in the example above of the 1000-1 –branched decision tree a suitable guessing strategy might be best, given a suitable probability distribution of payoffs over terminal nodes. But discovering what the distribution is, whether it has this consequence, and what guessing strategies it has the consequence for, may be more trouble than it is worth.

4 Against norms

Profitable patterns of reasoning that ignore the agent's cognitive limits are generally much simpler than those that do not. The reasons why the three fallacies are fallacies should make this clear. In some ways infinity is simpler than the finite condition. This is a theme that emerged in the philosophy of mathematics in the last century, when intuitionists and other logicians were trying to reformulate mathematical proof so that it takes account of the gap between being able to deduce a contradiction from the denial of an existence assumption and being able to produce an example satisfying the assumption. The resulting systems of proof are much more complex than classical mathematics. Similarly, a set of rules of inference or of decision-making procedures that takes into account the agent's cognitive limits is likely to be much more complex. Relevance logic, meant to capture the patterns of deductive inference that are intuitively natural to our minds, is more complex than classical logic, which can take short cuts through model-theory. Prospect theory, embodying just some of the quirks of actual human decision-making, is more complex than classical decision theory.

This is a general tendency rather than a universal law. (though I take it be a very general tendency.) But it does suggest that a set of principles that say what reasoning agents should do in various circumstances given that their resources and their abilities are limited in specific ways, will be extremely complicated. Complicated enough that the metaresource trap will apply, and agents having those limitations will be unable to apply the principles. So they will not be much use as a tool for figuring out what to do, or as a guide to critical reflection on whether one's thinking has gone astray.

The complexity of any full theory of bounded rationality should have a big impact on any attempt to describe good ways limited agents can think. It drives a wedge between the first and second dimension of theories of rationality that I described above. For it means that a full account of the ways of thinking that it would aid a person to use is too unwieldy to be the content of recommendations. The ways in which hard-to-predict details of human psychology can affect our performance has a similar impact. Should a reasonable person burden her mind not only with a complex theory of bounded rationality, incorporating among other things ways of assessing the difficulty of tasks and ways of anticipating the combined cost of individual components of many-part processes, but also with a full account of the many quirks of human thinking that may affect her performance? I think that the accumulated difficulty of designing efficient procedures for limited agents, which I have been adumbrating so far in this paper, suggest that there is little hope of describing ways of thinking that limited agents can profitably embrace, in terms of any normative principles. (In fact I doubt that there are non-trivial norms of thinking of any kind, but that is another matter.) But there are intellectual virtues, and I shall spend the rest of the paper explaining the difference.

Norms are patterns of behavior—in this case patterns of reasoning—that we hold ourselves to. That is, we expect one another to try to exhibit them, we monitor our own behavior to ensure that it does, and we criticise others when their behavior does not. A norm can be widely broken; what is essential is that there is something between an expectation that it will be often adhered to, and an awareness when it is not. Both of these require that we be able to represent the pattern to ourselves. But this is what is unlikely for profitable patterns that take agents' limitations into account. It is unlikely because of the considerations that make the three fallacies fallacious. Consider yet again the distraction phenomenon, which illustrates many points. If there were a norm of using one's brain as efficiently as possible, then we would have to try to solve problems with the right mix of conscious and unconscious attention. Now for a given individual there may an optimal pattern of attention, of having the right mix for each particular problem. And there may be some general relationship between the attention one should give and the difficulty of a problem, expressed in some objective terms that are unfortunately not available to present day humanity. But such patterns and relationships cannot be articulated by us, not in any usable way, just because our cognitive powers are limited. The optimum patterns for agents of some degree of cognitive boundedness could only be part of the content of expectations and awareness of agents whose cognitive powers exceed that bound.

This is not to deny that there are principles that can be used for good advice to bounded agents. Modus Ponens and the sure-thing principle can generate "If you believe P and believe *If P then Q* and have no reasons to disbelieve Q then consider believing Q," and "If you would prefer A to B if P, and you would prefer A to B

if Q, then consider preferring A to B if P or Q." But this is not advice to bounded agents *as* bounded agents: it does not tell them how to take account of their limits. To put the point differently, the same theory that produces this advice also generates advice to heed inference patterns so long that they would impede the progress of any human inquiry and utility calculations so complex that they would leave any human decision-maker too bewildered to act. The obstacles I am discussing lie in the way of principles that would explicitly block these confusing and counterproductive pieces of advice.

The essential point is that two things come apart in a bounded landscape. One is prescription: things it is useful or important for people to insist on to one another and to themselves. The other is performance: effective patterns of thinking and planning, given the facts about a particular individual. The more these come apart, the more problematic rules and norms become. So how else can we serve the function that norms of rationality serve? My full view is that we need a sophisticated and structured account of intellectual virtues. If the word "virtue" is to be significant here it must be backed up with a theory of intellectual and other virtues. That is a delicate job, and I hope to make some progress on it in a forthcoming book. Where we have virtues, though, we have desiderata. Courage is a virtue in part because it very often it is good that people stand up to threats and resist their fear. Caution is a virtue too: very often it is good that people pay attention to threats and resist their courageous impulses. To list both kinds of desiderata is not to say when it is that threats are to be given in to and when resisted; it is just to say that there are situations of both kinds. Intellectual virtues are associated with desiderata too. Imaginativeness is linked to the value of original hypotheses and plans of action. Care is linked to the value of finding flaws in hypotheses and plans. To give the barest outline of how an account of intellectual desiderata can give useful information about bounded rationality, here are three basic facts about intellectual desiderata.

4.1 Opposites coexist

It is good to have original ideas, and also good to have carefully worked out safe ideas. An unlimited agent could try for both, permitting wild fantasies but combining them with meticulous attention to their flaws. We don't have that luxury, and have to distribute our resources between the two. The distribution is rarely optimal, in that much of the time a little bit more originality or a little bit more care would have been beneficial. That does not negate the value of having the concepts of originality and care, of praising people for them and berating oneself for one's shortcomings with respect to them. And it does not negate the necessity of having both cognitive processes that provide novel ideas in some circumstances, especially when one is at a theoretical or practical impasse, and processes that check for flaws in one's reasoning and instigate detailed examination of the pros and cons of ideas. (One way of managing this in human life is to distribute the virtues between people—some are careful and act as a brake on their more imaginative friends—but that is not the topic here.)

There are complementary virtues associated with the tug between full maximization and various alternatives to it. In many situations a person can either follow a

full maximization procedure, taking the risk of errors from data-overload, or a rough approximation to maximization, or something of a quite different form. The $2^{1000} - 1$-branched decision-situation described above illustrates this. Depending on the details of the situation one or the other may pay off for a given person. So there are virtues of comprehensiveness and detail, associated with full maximization, virtues of self-preservation, associated with approximation, and virtues of self-trust, associated with randomising procedures. And, standing behind these, there are virtues of being able to select appropriately among processes, so that the person uses the one most likely to bring good results. Usually the virtue of selecting an appropriate process, like most virtues, does not make use of much conscious thought. But sometimes a person deliberately opts for a particular approach to the problem. Then the person must decide whether to think in terms of direct calculation of quantities such as expected utility and degree of evidential support, or in terms of short-cut approximations to them, or in terms that ignore them but which are trusted to result, in the long run, in profitable decisions and well-supported beliefs. There are varied virtues here too, of comprehensiveness, efficiency, and intuition.

4.2 Implementation is unspecified

To say that an outcome is desirable is not to say how to achieve it. So if for example we impress on someone the importance of greater care in checking for flaws in her suggestions, we leave it up to her to find out what things to check and when a particular level of care is called for. We may point to particular other people as models of doing these things well, but they may do them in different ways, as may she when she takes the message on board. The differences between people achieving similar cognitive results may not be describable in folk psychological terms, or in the traditional concepts for talking about rationality. So an account of bounded rationality along these lines need not be hostile to accounts of the mechanics of cognition in very un-intuitive or physiological terms.

The variability of implementation is in fact an advantage. The gap between prescription and performance arises from the cognitive burdens of describing in effective detail the processes that balance economy and results. But if we focus just on the desiderata we do not incur that cost. In effect, the burden is passed from explicit verbal theorizing to learning processes specific to the individual person. So the process of learning as well as its results can be expected to be implemented in indescribably many ways.

4.3 Some desiderata are more important

It is nice if your problem-solving does not leave you with a headache, but it is essential that you get adequate solutions in time, even at the price of a headache. Both low probability of error, given the situation and the available data, and high explanatory value are desirable, but in some contexts one is vital and the other incidental. So, in the inevitable cases where desiderata conflict, there is an essential skill of focussing on the more important one (not pondering number theory when you smell smoke).

The skill amounts to attending to the over-riding but vague desideratum of solving problems essential to survival and basic well-being.

An interesting question is the importance of maintaining consistency among one's beliefs and between one's beliefs and desires. Logical consistency is a basic requirement on a traditional conception of rationality. Once it is lost there is little point trying for anything else. But that is because consistency is a necessary condition for the other desiderata of a traditional conception, such as truth. But considerations of bounded rationality introduce a wider set of desiderata, for not all of which is consistency a necessary condition. So we can ask where it ranks in importance among this wider set. Several recent writers, most notably Harman (1986) and Foley (1993), have argued that consistency is not of over-riding importance in a well-ordered set of beliefs. While I am sympathetic to these views, my aim now is not to assess their force but to make a remark on how best to understand them. It is tempting from a traditional epistemological point of view to express such claims by saying that one should not strive for consistent beliefs, or that it is permissible to allow inconsistencies among one's beliefs. But of course those claims are much too strong. Often it is crazy to continue to hold both of two beliefs when one realises that they contradict one another. And often when one discovers a contradiction it is a top priority matter to resolve it. The problem comes from the language of norms: ought, must, may. One way of putting the point of Harman and others in a sane way is in terms of relative priority: sometimes when one finds an inconsistency eliminating it is a lower priority than achieving some other desideratum; often the fact that inconsistencies can result from a belief-forming process is not a strong objection to using it. Putting the claims in terms of desiderata instead of rules allows them the measured formulation that the topic requires.

5 Conclusion: prospects for theory

At the beginning of this paper I said that I would try to make sense of the question of whether there is such a thing as bounded rationality, and then to suggest which of the disambiguated versions might have answers. The disambiguation happened early in the paper, and the evaluation of prospects later. Making a grid out of the distinctions early on—it is actually a three dimensional grid—and inserting the upshot of the following discussion, we get the following.

	Limitations of power	Limitations of accuracy
Reduced ideal rationality	SP possible // NSP unlikely	SP unlikely // NSP doubly unlikely
Recommendations for imperfect agents	SP general virtues // NSP specific virtues	SP general virtues // NSP specific virtues

There are eight possibilities here. Within the 4 × 4 grid I have marked off two possibilities in each cell. First we have theprospects for a theory that applies only to

Sensible People who conform to the principles of sensible limitation-management, whatever they are. Then, after the //, we have the prospects for a theory that applies also to Non Sensible People, who get into various kinds of typically human trouble that conformity to the theory would prevent.

Reduced ideal + limitations of power + SP: that is the theory that several worthy thinkers, such as Pollock and Weirich, are working on. Perhaps it can be constructed. It won't yield helpful recommendations, though.

Reduced ideal + limitations of power + nonSP: inasmuch as such a theory describes an "ideal non-ideal" agent, it does not apply to misbehaving non-ideal agents, even though that includes most of us most of the time.

Reduced ideal + limitations of accuracy, with or without SP: I'm very skeptical about the prospects for a systematic theory here, because of the relevance of quirks of human psychology which in fact vary from person to person.

Recommendations for imperfect agents (with either kind of limitation and SP or NSP): We can begin such a theory just by listing the desiderata associated with various kinds of situations. The recommendations to agents take the form "work your way towards (away from) a situation of this kind". It is up to the agent to find a way of doing this. It may take a lot of training, and the mechanisms that eventually subserve the task may vary from one person to another. Essentially the same applies to foolish people who have got themselves into pickles, except that some special damage-limitation advice may apply. ("If you are in the middle of a chain of reasoning so long that you've forgotten the premises and the intended conclusion, forget about it and concentrate on remembering what you were thinking about.")

There are only two promising areas in this table, reduced ideal rationality applied to limited agents who do not make mistakes, and virtue theory. The first is a theory of the patterns of thinking that bounded agents can profitably instantiate, and the second is a theory of the descriptions of thinking that bounded, specifically human, agents can profitably aim to satisfy. These are not competing aims. There are numerous connections between them. One subtle connection, which has recurred throughout this paper, concerns the status of abstract quantities such as expected utility and evidential support (defined in terms of conditional probability, say). One option for the first project is to define patterns of thinking that will approximately maximize these quantities. That is not an option for the second project. No one should aim at approximately maximal utility. One should aim at maximal utility *tout court*, though the means one can adopt in aiming at it are extremely varied. A person having intellectual virtues that allow her to react gracefully to the fact that she has limited intelligence, memory, control, and tends to many kinds of mistakes, will adopt means that are appropriate to her situation, sometimes explicitly calculating, sometimes approximating, and sometimes guessing.

References

Dijksterhuis, A., & Nordgren, L. F. (2006). A theory of unconscious thought. *Perspectives on Psychological Science, 1*, 95–109.

Foley, R. (1993). *Working without a net*. Oxford: Oxford University Press.

Gardner, H. (1999). *Intelligence reframed*. New York: Basic Books.

Harman, G. (1986). *Change in view*. Cambridge: MIT Press.

Jackson, F., & Pargetter, R. (1986). Oughts, options, and actualism. *Philosophical Review, 95*, 233–255.
Knutson, B., Wimmer, E., Kuhnen, C. M., Winkielman, P. (2008). Nucleus accumbens activation mediates the influence of reward cues on financial risk taking. *Neuroreport, 19*(5), 509–513.
Morton, A. Review of Pollock. *Thinking about acting*. Mind, (to appear).
Pearl, J. (1984). *Heuristics: Intelligent search strategies for computer problem solving*. Reading, MA: Addison-Wesley.
Pollock, John (2006). *Thinking about acting*. Oxford: Oxford University Press.
Rabinowicz, W. (1995). To have one's cake and eat it too: Sequential choice and expected utility violations'. *Journal of Philosophy, 92*, 586–620.
Sternberg, R. J., Pretz, J. (Eds.) (2005). *Cognition and intelligence: Identifying the mechanisms of the mind*. New York: Cambridge University Press.
Taszka, T., & Zielonka, P. (2002). Expert judgments: Financial analysts versus weather forecasters. *The Journal of Psychology and Financial Markets, 3*(3), 152–160.
Weirich, P. (2004). *Realistic decision theory*. Oxford: Oxford University Press.