knowing what to think about: when epistemology meets the theory of choice

Adam Morton

two moving targets A very traditional epistemology studies changes in belief, given the information and reasoning available to a person, but ignoring her desires and values. A somewhat more realistic epistemology studies changes in belief, given information and reasoning and some relevant desires and values. The question in both cases is: if some of your beliefs change how should you change the rest? Parallel to this, a standard approach to rational decision will study changes in intention given a person's beliefs and desires. Counting intentions as a special kind of desire the question is: if some of your desires change how, given your beliefs, should you change the rest? But in reality we can hold neither beliefs nor desires constant while changing the others. We change both simultaneously. We move from one complex of beliefs and desires to another, hoping that both the transition and the final state have some relation to how they ought to be. The distinction between epistemology and decision theory, the theory of change of belief and the theory of change of desire, is not a very natural or helpful one. Both should gain from being part of a single theory of change of intentional state.

What would such a theory look like? Might some traditional epistemological topics look different from this perspective? In this paper I cannot do more than suggest how complicated and interesting the questions that arise are. I shall focus on two closely related mysteries. First, the factors that determine what combinations of belief- and decision-directed strategies will pay off. We really know very little about what these combinations are, and if we understood them better we would have a better grasp on what makes a reasonable belief and a sensible choice. We might even be able to say something useful to people pursuing epistemic-practical projects. Second, the characteristics of agents that allow them to negotiate these combinations. I do not think we can have an adequate account of reliability, intellectual virtue, or rationality until we have an informative model of how a creature generally like a human being can distribute its limited resources between competing and interlocked cognitive demands.

the rationality of strategies One does not simply open one's eyes and record how the world seems; one does not open one's box of whims and consider what one might do. One follows *strategies* of investigation, which in part determine what evidence get considered, and one follows strategies of choice, which in part determine what options get considered. Some belief-acquiring strategies are more reasonable than others. (Or more promising, less insane - not to get hung up on any particular loaded terms.) Which ones they are clearly depends on one's aims and needs. If your car is stalled on the railroad tracks you should not be pondering the twin primes conjecture. Nor should you be thinking about whether to shift the balance in your retirement portfolio. A special case of this fact is recognized in contemporary epistemology as the distinction between error avoidance and ignorance avoidance, or equivalently informativeness versus accuracy. The idea is that a completely safe – error-free – method of acquiring beliefs is unlikely to answer the questions that most interest us. In order to relieve our ignorance on matters of intellectual or practical concern we will have to take some risk of acquiring some false beliefs along with the true ones. A Cartesian project of guaranteed accurate and informative beliefs is simply not achievable by finite human beings in the likely span of human history. So – as a matter of general intellectual strategy and on a topic-by-topic basis – one has to decide what risk of false beliefs one is willing to take in order to satisfy one's desires for particular kinds of true ones[1].

The error/ignorance balance one strikes will depend on one's desires, for knowing truths of a certain kind and not believing falsehoods. This may be a very practical matter. Suppose you are preparing chemicals to use in a demonstration for a school chemistry class and you want to know how pure the magnesium powder is, because if it is not then the demonstration will fizzle and the students will make fun of you. You will take some care, but you will want a method that gives an answer within half an hour. If it is likely to take longer you can do a different demonstration. But suppose on the other hand that you are analyzing the substratum for a batch of vaccine, and if it is not pure then the vaccine will be dangerous. Then you take much more care; you are willing to follow a procedure that won't give any answers for a week. As is often remarked, you will be willing to accept a conclusion on the basis of weaker evidence in the first case than the second, but it is more important for my purposes now that the method you employ may be entirely different. Epistemic

[1] A good exposition of the crucial distinction between avoiding falsity and avoiding error is made in chapter one of Alvin Goldman *Epistemology and Cognition* (Cambridge, Mass.: Harvard University Press, 1986.) The distinction, though, goes back to Isaac Levi's *Gambling with Truth* (New York: Knopf, 1967)

strategy is sensitive to matters of non-epistemic importance. (There are connections with epistemological contextualism here, which I take up in the ‘rationality’ section below.)

In this case belief-acquisition is tuned by desires. Often the influence is the other way round, and decision procedures are tuned by beliefs. Suppose for example that you are walking down the main street of a strange town looking for a restaurant, and you believe that very soon you will come upon the town’s only four restaurants, in close proximity. (Your guide book may tell you this.) Then your strategy for making your decision may be one of running through all the possibilities taking in much of the relevant information: you visit each one, read the menu and stick your head inside to sense the atmosphere, delaying a decision until you have done this for all four. Or, on the other hand, you may believe that there are many restaurants, of varying quality, scattered randomly for a long way down the street. Then your strategy may be to satisfice, and, say, to sample the first three restaurants with some care and then choose the first one after that which is at least as good as any of the first three. Decision strategy is sensitive to belief.

In fact most of the time there are influences both ways. The restaurant example illustrates this. As you walk down the street looking into restaurants you are picking up evidence of the distribution of restaurant quality in the town, this affects your intentions for the evening meal. For example you may decide that it is just hopeless to expect a decent Indian meal in a Midwestern American town, but that Thai is a more promising possibility. So you don’t cross the road to look into the occasional Indian restaurant, but make efforts to check out the Thai restaurants. Your initial beliefs and desires lead to strategies for revising beliefs and desires – finding out what kind of a town it is, food-wise, and choosing somewhere to eat – which themselves lead, jointly, both to further belief-revision strategies and further decision-strategies – cross this road to check out this place, choose the first Thai place as good as the last two we’ve seen. This is the way it generally is, though when we reflect on our thinking we often focus only on one side of the picture[2].

In these examples the modulations of the thinking leading to choices have gone deeper than those leading to beliefs. There is a reason for this. Rational strategy is driven by resource-allocation more than by anything else. Intellectual resources (time and working memory, centrally) are scarce in comparison to the complexity of most intellectual problems, so we must distribute what we have efficiently

[2] Why is decision theory full of restaurant examples? I suspect it is because they summon basic foraging problems: this way to get this food, or that way for that?

towards the best attainable outcomes. The study of constrained rationality, as Herbert Simon, called it, is well under way in decision theory, but is hardly visible in epistemology[3]. But in fact the same issues arise, and the same responses are attractive. (And run into similar obstacles.) When making a decision one can economize by sifting more hastily through a preliminary list of options to get a "short list" to give more intensive consideration. Or one can satisfice, that is, pick a threshold of acceptability and then choose any option (or the first that arrives) that is above the threshold. The epistemic analogs of both of these are often reasonable ways of proceeding. If you are reasoning by inference to the best explanation of a phenomenon, you do not give equal attention to the pros and cons of all candidates. You quickly focus on a few potentially powerful explanations and compare them carefully. (You want to understand the pattern of the tides. You consider for a moment, but only a moment, the possibilities that schools of fish move in an almost-daily pattern that causes the ocean to shift, or that the patterns are random so as to generate an illusion of high and low tides, and move quickly onto hypotheses involving the moon and the sun, gravitation and momentum. If none of these work out you may go back to considering the fish.) Or for analogs to satisficing suppose that you have a series of hypotheses of decreasing initial implausibility. You do not know whether any of them will explain all the phenomena in question, and it may take forever to consider them all. So you begin with the most plausible, and reject it if it does not explain "enough" of the data, moving on to the next. Eventually, with luck, you come to a hypothesis that leaves few enough mysteries and anomalies that you are satisfied with it, and you accept it. The balance of factors could also be the other way round. You could have a large body of hypotheses, all of which explain the data adequately. You run through them as they occur to you, and accept one when its initial plausibility is high enough. Even if it explains everything you are not going to believe something that seems crazy to you, but there comes a point when high explanatory force overcomes implausibility, even though if you wait you or someone else may come up with a simple intuitive explanation that is just as powerful[4].

[3] I am not going to cite the now enormous and varied literature on bounded rationality. For an accessible exposition with unexpected philosophical connections see Michael Slote *Beyond optimizing: a study of rational choice* (Cambridge, Mass.: Harvard University Press, 1989) and for advanced work in decision theory see Rubinstein, Ariel *Modelling bounded rationality* (Cambridge, Mass.: MIT Press, 1988.) Chapter four of Richard Foley's *Working without a net* (New York: Oxford University Press, 1992) is also very stimulating in this connection.

[4] This is not to say that the situations with respect to belief and choice are fully symmetrical. Some differences come from the fact that a choice once made and acted on is normally irrevocable, while beliefs can be revised. (If the hypothesis that

The influence of standing desires, and desire-changing strategies, on epistemic strategy in examples such as these, that is to say in nearly all real cases, is clearly not confined to setting the balance between risk of falsity and possibility of informative truth. As the examples above show, the influence is much more pervasive. What we learn on any occasion is shaped in all respects by the action-choosing strategies we have adopted, which are themselves shaped by belief-acquiring strategies. Everything happens at once. This presents a problem of normative regress, though. As long as we consider beliefs to be fixed when considering desire-development, and vice versa, we can suppose that individual agents can, when the situation calls for reflection, consult some fixed principles which tell them what to choose or what to accept. But if the methods, the procedures, are themselves among the variables then the task for the reflective agent is very different. Should she reflect about which reflective criteria govern her choosing and her learning? If so what principles govern *those* reflections[5]? It seems clear that an intelligent human agent would very rarely gain by going down this path. (And with hindsight we can see that the problem was there all along in the form of the question: what tells a person when it is an appropriate time to reflect on – apply explicit norms to – her thinking?)

Feedback routes There is a dogma that when you change your beliefs some of your desires may change, because you now see the consequences of achieving them differently, but that when you change your desires your beliefs should be unaffected. The thoughts in the previous section do not challenge this doctrine because they show two-way connections between *strategies* for changing beliefs and desires, not those changes themselves. Yet there is a feedback from questions of method to first-order questions of what one should believe and what one should choose. I shall describe two feedback routes.

First, a variation on a Humean theme: the potential atypicality of all samples. You are testing a coin to see if it is fair, by tossing it ten

is both plausible and powerful comes along, you switch to it.) But this does not affect the point I am making here. A subtle difference between the choice and belief situations with respect to thresholds comes from the fact that as part of choosing an act one can choose a decision method. But as part of coming to believe one does not deliberately choose anything. Rather one *believes* in advance that e.g. a hypothesis with high explanatory power and more than "enough" initial plausibility is likely to be true. This is an issue that calls for a lot more attention.

[5] In this connection see the discussion of the "AEA" pattern in Adam Morton 'Saving epistemology from the epistemologists: recent work in the theory of knowledge', *British Journal for the Philosophy of Science*, **51,** December 2000, pp. 685-704.

times, recording (H_{eads}-T_{ails}), tossing it ten times again, and then after you have done this twenty times calculating the total. If it is less than 15 you intend to announce that the coin is fair, and if it is more than 30 or less than –30 you intend to announce that the coin is biased. You have got through nineteen of the twenty sets, and the total balance is 11. In none of the previous sets was |(H-T)| more than three. This is pretty strong evidence that the coin is fair, and if your original plan had been to do nineteen sets you would have announced just that. But the twentieth set is about to begin and you know that the coin might just come down heads four or more times out of the ten. It might do so even if your inclination to think it fair is right. So you do not conclude that the coin is fair. Your attitude is "wait and see" until the last set is over.

This example does not turn on the effect of practical on epistemic deliberation. But it does show one way that the consequences of the support that evidence gives to a hypothesis can depend on the strategy in the course of which the evidence was obtained. Very often you should not perform the temporary closing of the file on a topic that we describe as belief (or as concluding or reporting with conviction) until you have completed your investigations[6]. (Bayesians may complain that your degree of belief in the hypothesis should be independent of the investigatory strategy. I'm not sure even of this, since it assumes that your prior probabilities are independent of your strategy, which I'm not convinced of. But, be that as it may, the Bayesian world-view just doesn't have a place for belief or acceptance.) And since the choice of strategy itself usually depends on some larger practical (practical plus epistemic) context, the point at which it is appropriate to form a belief on a topic is very often a result of the practical context, among other things. To return to the example, you might have set twenty sets as your target because you were intending to place a large bet on the coin and needed a certain minimum assurance of its propensities for the risk to be reasonable.

The second feedback route has a paradoxical air. There are many topics on which you have no opinion, though you may have some ideas about the general character of the evidence. In the course of working out how to satisfy a desire you may acquire a reason to investigate. And then it is very likely that you will acquire an opinion. Sometimes, indeed, you can tell which way the opinion is likely to lie. Suppose for example that you are an agnostic by reason of a complete

[6] One reason for saying "very often" rather than "always" is situations in medical experimentation where a partial tally of the evidence suggests that it would be wrong to continue with the experiment as planned.

lack of interest in religious questions. You undertake to give some lectures in the philosophy of religion, though, to help out a colleague, and as a result plan to read and think about arguments for and against the existence of God. Your general impression before really going into the matter was that the arguments against were stronger (the dubious intelligibility of the concept, the problem of evil, and so on.) So you now think that very likely in a month's time you will be an atheist. Does that give you reason to revise your beliefs in the direction of atheism now, before considering the arguments? Intuitively it does not, although it is slightly puzzling why. In other cases the existence of evidence you have not yet seen can itself count as evidence. (A person in whom you have great trust and who is in a position to know assures you that an envelope, which will be unsealed tomorrow, contains overwhelming evidence that X is the murderer. You should feel pretty sure now that X is the murderer.) If you do have reason now to incline to atheism then the decision to help your colleague will have made it rational to change your religious beliefs. Even if it does not, the decision will have given expectations about what your future beliefs will be together with expectations that these beliefs will be better founded than your present ones, which comes to something very similar[7].

These two feedback routes may be connected, in that our reluctance in the atheism case to think that "you" should change your opinions now may be connected with the fact that you have not yet gone through arguments and evidence in the way you intend to. It is only at some stages that we count our inclinations as beliefs. This cannot be the whole story, though. I suspect that there is a body of principles waiting to be articulated here, governing the effect that engagement in strategies for developing one's beliefs and desires should have on those beliefs and desires.

virtues of intelligent activity In matters of both belief and decision people can be responsible, careful, sober, and prudent. They can also be adventurous, stubborn, and brave. These are all characteristics that can lead to good results in some circumstances. Characteristics in the first list are often thought to be valuable in all circumstances, while those in the second are valuable when they appear at the right moment. For this reason it helps to distinguish between character traits and virtues, though they often have the same names. A

[7] The issues here are connected with the issues about the reflection or principal principle discussed by Timothy Williamson in chapter 10 of *Knowledge and its Limits* (Oxford: Oxford University Press 2000.)

character trait is a disposition to some way of thinking, acting, or feeling. A virtue is a disposition to exhibit the thought, action, or emotion *when it is appropriate*. So someone may have character marked by intellectual courage, frequently defending unpopular positions, and often going out on limbs even when he knows it may expose her to ridicule or disagreement. But this is not an intellectual virtue unless the positions she defends are not transparently lunatic ones, and her defenses and conjectures often result in interesting truths and profitable decisions. I would argue, in fact, that we must make this distinction even for characteristics from the first list. You can be too responsible, if it makes you boring; or too careful if it makes you miss worthwhile opportunities. A virtue must embody two kinds of tacit knowledge: of when it makes sense to exhibit the trait, and to what extent.

Nearly all the intellectual virtues that we have everyday names for are virtues of intelligent activity generally, and not specifically of belief formation, decision, or some other category of thought. Even epistemologically oriented virtues such as respect for evidence are applicable generally: someone with no respect for evidence will make disastrous decisions. And decision-oriented virtues such as prudence have epistemic relevance: in planning and carrying out a belief-acquisition strategy one has to look forward as carefully as in any other activity. I think there are two closely related reasons for this, the ubiquity of strategy and the centrality of limitation management.

It should be clear by now that strategy is everywhere. Whenever we think we do so as part of a plan, even if sometimes a simple one, in which getting clearer about some things and making some decisions takes one along the way to getting clearer about some targeted facts and making some targeted decisions. But since that is so, the capacities to make and carry out suitable plans are everywhere, and are everywhere essential. So the epistemic virtues, in particular, are pointless unless they coincide with or cooperate with virtues of epistemic strategy. And these virtues are just virtues of strategy in general. Epistemic care, for example, requires that one not overlook slight possibilities of evidence against one's intended conclusion. But this is a particular case of not overlooking slight possibilities in general, particularly slight possibilities of unwelcome outcomes, and this is not epistemic care but prudential care. The pattern is the same.

This is not to say that someone who has the virtue of epistemic care, for example, will have the virtue of prudential care. The discrepancy may be crude, in that someone might rarely be disposed to be careful in prudential matters though they often are in epistemic matters. (Though given the inextricability of the two, a failing in one will lead to problems in the other.) Or it might be subtler, in that someone's dispositions to care in respect to evidence gathering might be more or less appropriate than her dispositions to care in respect to scouting out dire possibilities. But, then, for that matter, someone who often exhibits epistemic care may exhibit it with regard to scientific matters, say, and not religious ones, or may exhibit it as a virtue in one of these and not the others. Virtues are like this: their instantiation as virtues in any person are very scattered. There is a deep problem here, a discrepancy between a simple interpretation of what they might seem to involve and what they could in fact possibly be, and it is not resolved by distinguishing between epistemic and prudential virtues[8].

There is a reason why most intellectual virtues are virtues of making and carrying out plans. It is that we are so very finite; our working memories are so small in comparison with the complexity of the intellectual projects we can set ourselves. Many stages of most processes involve searches: one has to consider a fact, assess it, then consider further facts suggested by ones assessment, assess them, and so on through a ramifying tree of possibilities. For obvious combinatorial reasons (a binary tree has 2^n branches to depth n) no real agent can search both thoroughly and deeply. But different stages of an enquiry may require one to

- search the consequences of an act or a proposition for advantages or plausibilities in general
- search the consequences of an act or a proposition for disadvantages or implausibilities in general
- search the consequences of an act or a proposition for advantages or plausibilities of a particular kind
- search the consequences of an act or a proposition for disadvantages or implausibilities of a particular kind
- search the consequences of an act or a proposition to a great depth, looking for advantages/disadvantages/plausibilities/implausibilities generally/ of a particular kind
- search the consequences of an act or a proposition very thoroughly, looking for advantages/disadvantages/ plausibilities/implausibilities generally/of a particular kind

This is just too hard to do by brute force. We have to content ourselves with doing some aspects of it, with many shortcuts most of

[8] For the variability of behavior that any realistic concept of a virtue will have to take into account see Gilbert Harman 'Moral philosophy meets social psychology: virtue ethics and the fundamental attribution error', *Proceedings of the Aristotelian Society* **99**, 1999, pp 315-331, and John Doris *Lack of character: personality and moral behavior* (New York: Cambridge University Press, 2002).

which are inconsistent with doing other aspects of it. We have to learn which short cuts pay off, for us, and when[9].

Intellectual care, for example, has at its heart searching comprehensively, to a shallow depth if need be, carrying out subsidiary searches when necessary to check the relevance of facts as they emerge. Intellectual daring, on the other hand, has at its heart deeper and usually less comprehensive searches, trusting that details will not be missed that invalidate the whole procedure. Both are necessary, and they can very rarely be combined. Each, then, needs to be accompanied with the capacity to employ it when it is needed and not when it is obstructive.

Most intellectual virtues have essential connections to capacities to do search in some particular manner, and capacities to know when that kind of search is a good idea. These are the capacities that we make names for, because they are the ones we need names for. They are hard-to-acquire and vary from individual to individual, and they cannot be summoned on a particular occasion unless the ground has been prepared by practice and self-modulation. So we need to name them and become friends with them. And since their important characteristics apply to searches in many kinds of thinking, they are multi-purpose virtues of intelligent activity.

rationality: justification and knowledge One “virtue” has particular historical importance, and has played a large role in the development of western culture. That is rationality, and I have scare-quoted its claim to virtue not because rationality does not have many of the characteristics of an all-purpose intellectual virtue, but because part of the point of thinking in terms of intellectual virtues is to avoid begging questions about the relations between the qualities that make for intellectual success. They may not have much in common; they may often act contrary to one another; it may not be possible for one person to cultivate all of them. In standard epistemology the idea of rationality is reflected in the concept of a justified belief. At first this seems simple: a belief is justified if it is acquired in the right way, with no bad reasoning involved. On second thought complications arise. A justified belief may be acquired in an irrational way, as long as the reasons the person continues to hold it, or would defend it with, are good ones. This clearly opens up a large amount of vagueness, to add to the vagueness of good reasons or reasoning. Further thought raises further complications. We count the beliefs a person acquires by

[9] For more on this theme see Adam Morton “Epistemic virtues, metavirtues, and computational complexity”, *Nous,* **38**, 3, Sept 04.

honest inference from misleading evidence as justified. But suppose the person could easily have realized that the evidence was misleading, if only she had followed up a line of investigation she was too lazy to pursue. Or suppose a person follows a line of reasoning that is actually incorrect but which is generally taken in her circle to be alright, and whose mistakes are too subtle for her to notice. Are her beliefs going to be justified or not? (This can easily happen with statistical reasoning, which can produce problems that will baffle anyone, however sophisticated and intelligent.)

I think that no contemporary epistemologist takes the idea of a justified belief to be a very clear or useful one, unless it is hedged about with prescriptive definitions and made into a technical term of a well-developed theory[10]. At this point the concerns of this paper open up some new positions addressing traditional issues. I shall state these possibilities, without defending them. The important thing is that there are more possible reactions to the issues than we had thought.

What concepts of rational belief may we deploy without making dangerous assumptions about the psychology of belief and action or fudging the complexities of human thinking? The two below seem to be intelligible, if not always precise.

First there is the possession of evidence or similar grounds for a belief. A person may have considerations in favor of a claim, that she can produce to persuade others. People are rarely very well placed to assess exactly how strong the evidence they possess is, but the strength of the support that it gives to a claim is a relatively objective matter. The conditional probability of the claim given the evidence conjoined with relevant background beliefs is the most precise measure of support. It's precision and objectivity is admittedly qualified by the indefiniteness of what is to count as relevant background belief but, still, in most cases we know how to begin assessing whether given evidence supports a claim strongly, weakly, or not at all.

Second there is the reasonableness of change of belief. A person has a set of beliefs at one time and another set at a later time. The transition from the one to the other may or may not be in accord with the way a well-constructed human agent would operate in the given situation. There are many aspects of the situation that might be taken into account. The ones that interest me are the belief- and desire– forming strategy that the person is pursuing and the person's whole prior complex of beliefs and desires. If the changes in belief –

[10] As for example in Alvin Goldman's *Epistemology and cognition* (Cambridge, Mass.: Harvard University Press, 1986.)

deletions as well as additions – result from a strategy that is a reasonable one for that person to follow, given everything she believes and wants and her particular capacities, then the change of belief is reasonable. I am not going to give any analysis of the central attribute here: the acceptability of a belief-and-desire-evolving strategy. It obviously has vague edges, but my claim is that it is relatively objective and unparadoxical, and grounded in human psychology. Given this very substantial assumption, there is a piece of normative human psychology that needs a name: I'll call it directed big state rationality. (Directed because it concerns changes that result from a strategy; big state because it concerns both beliefs and desires.)

There are other ways of trying to understand change of belief. In particular, one could study the reasonableness of changes of belief alone, not in relation to desire. And one could study 'unmotivated' changes of belief, considering for example cases in which a person passes from one complex of beliefs to another as a result of the random play of neurons. These are rivals, and if they give robust and useful ways of understanding belief change then the line I am developing in this paper is less interesting. To sharpen the rivalry I shall state a thesis, an aggressive claim which if true makes the study of strategies for simultaneously changing beliefs and desires central to epistemology.

The big state rationality thesis: the possession of evidence for one's beliefs and their acquisition by directed big state rationality are the only sustainable concepts of rational belief.

According to the whole state transition thesis there is only one concept of rational belief, and only one concept of rational change of belief. Only one kind, that is, that stands up to analysis and handles complex examples without ad hoc requirements, coheres realistically with actual human psychology, and sheds light on the forms of thought that are worth cultivating. Moreover the two are very different. At the heart of evidence there is an abstract relation between propositions, taken as semantical objects, evaluated ultimately in terms of the possible worlds in which the claim-proposition is true which are excluded by the evidence proposition. At the heart of big state rationality are details of very contingent human psychology: what patterns of thought in the pursuit of what projects our brains can follow effectively. The claim is that there is nothing in between[11].

[11] Gilbert Harman's *Change in View* (Cambridge, Mass.: MIT Press 1986) can be read as claiming something similar. Harman does not stick his neck so far out though.

When we keep these two parts of rationality apart, and don't glance between them, we find that many puzzle cases can be described quite simply. For there is no problem with an irrational acquisition giving a perfectly rational (well-evidenced) result. Consider for example cases in which a person adopts a perfectly idiotic belief forming strategy and comes up with a well-supported belief. The kitchen is on fire and the flames remind someone of the distribution of primes, so that he muses on a conjecture in number theory instead of thinking how to get his children out of an upstairs bedroom. At the end of the story the children are dead, the person is heartbroken, and there is a good proof of a new result. We should have no compunction about calling this person irrational, and describing the thought process that led to the number-theoretical belief as very defective, while allowing that the person's great losses are mitigated by the gain of one small rational belief.

There are also examples in the opposite direction, in which an irrational belief is acquired by a rational process. Sometimes it is reasonable to acquire a belief although the evidence for it is fairly feeble. Another burning house case. A person realizes that the house is on fire and that his children are upstairs. There are two ways to get to them and out with them. One is up the stairs, but the staircase is already beginning to smolder. The other is to dash out the front door, up the metal fire escape, then in and out through the bedroom window. He decides that the latter gives the best chance, though only if he moves immediately, given the greater distance. Consider his belief, as he begins to run, that the outside route gives a greater chance of getting there and returning with the children. He doesn't have much evidence for it: the stairs were only smoldering rather than flaming, and he hasn't really considered the difficulty of getting two sleepy children through the window onto the fire escape. But to delay while considering the evidence would be to make the situation worse: the right thing is to take the most plausible-seeming possibility and act on it decisively. So his thinking is as it should be, though the belief it leads to is not strongly based.

There is a connection here with an old debate between evidentialists and pragmatists. I'm taking evidentialists to be people who describe claims that are made on less than adequate evidence as irrational or pretend belief. (The view is usually accompanied with firm views about what is adequate evidence.) Pragmatists think that non-evidential considerations can sometimes justify a belief[12]. The

[12] The classic source is the debate between William James 'The will to believe' in *Essays in Pragmatism* (New York: Haffner, 1948), and William Clifford 'The Ethics of Belief' in his *Lectures and Essays, vol 2.* (London: MacMillan, 1901). For a some

example above has a pragmatist flavor. The person's belief-acquisition is as it should be although it does not result in his having adequate evidence. An example in the "feedback routes" section earlier in the paper has an evidentialist flavor. Given a strategy for confirming a hypothesis with a series of experiments, one should not form a belief about the hypothesis until all the evidence is in. But in fact the big state rationality thesis undermines the terms of the debate. In the fire-escape case the person's thinking is fine, but the belief is not justified. And that's perfectly ok. In the series of experiments case belief on less than total evidence is not appropriate, but the evidence is strong before that point. And that's perfectly ok too[13].

The big state rationality thesis concerns the range of considerations that are relevant to what is traditionally thought of as the justification of a belief (though it suggests that the terminology of justified belief is rather confusing: better to talk of rational belief change and of evidence.) The same emphasis on the agent's total cognitive state can be applied to issues about knowledge. Go back to the example early in the paper, in the rationality of strategies section, contrasting the amount of evidence you would collect before accepting that a school demonstration or that a batch of vaccine was ready to go. Suppose that in the vaccine case you had settled for the amount of evidence that would have been adequate in the classroom case. Then you would not be judged to have known the composition of the chemical, in spite of the truth of your belief and evidence which would have been strong enough in a different context. Your procedure would not have eliminated some possibilities that in that practical context were relevant.

Though this example does suggest that big state factors discriminate knowledge from non-knowledge, it is of limited impact because they operate through their effect on what it is reasonable for a person to believe. We get more interesting suggestions by playing with some of the cases from earlier in this section. Suppose for example that in the fire escape case the person does not have adequate evidence to exclude the possibility that the window into the

recent twists on the theme see Jonathan Adler *Belief's Own Ethics* (Cambridge, Mass.: MIT Press, 2002.)

[13] All these cases raise questions about what is to count as belief. My own inclination is not to take the concept of belief too seriously – as I argue in chapter 3 of *The Importance of being understood: folk psychology as* ethics (London: Routledge 2002) – and to say that beliefs are whatever we act on and tell one another. That suggestion is not essential here: for example the fire escape case can be taken to show just that the result of a bit of good thinking can be an informational state about which the evidence is inadequate.

bedroom may be jammed, or that he cannot get the children through it without injuring them, but that these possibilities are in fact false, and the person's belief that the fire escape route will get him in and the children out is true. Not only true but formed in the best way, given human limitations and the situation the person was in[14]. Then, according to my intuitions, the person knows that fact about the fire escape route. Since it would have been irrational to waste time excluding the possibilities that the person didn't waste time excluding, the person's failure to exclude them does not demote his belief from knowledge[15]. On the other hand, in the fire and number theory case the person does, intuitively, know the theorem in question, even though the thought process was all things considered insane. But that does not really tell against the relevance of the whole state to the ascription of knowledge, because the important aspect of the whole state – the person's desire to avoid the fire danger to his children – is not relevant to the exclusion of any possibility in which the number-theoretical conclusion is false.

There are examples like the fire and number theory case, in which the whole state is relevant, though. First a far-out case. A soldier ought to be on guard duty on the ramparts, and in fact he wants to do his duty and protect the city but his alcoholic tendencies have got the better of him and he is in the tavern far from his post. He wonders where his wife is and figures that she would have returned from the neighbors and will be at home putting the children to bed. The possibility that she has been abducted by aliens does not occur to him, so he considers no evidence for or against it. Now, as it happens, just above his post on the ramparts a flying saucer has just circled, on its way to its mother ship with a cargo of human specimens, not including the soldier's wife. Had he been at his post, as he should have been, he would have seen it, and thought about alien abductions, and then when he wondered about the whereabouts of his wife he would have suspended judgement.

A more normal case. A young biologist is running an experiment on whether a new antibiotic inhibits the growth of a bacillus in a culture. A colleague is running a very similar experiment with a

[14] This is a conclusion that could also be derived from a virtue-epistemological approach to the definition of knowledge, for example Linda Zagzebski's in "What is knowledge?" in John Greco and Ernest Sosa, eds. *The Blackwell Guide to Epistemology* (Oxford: Blackwell Publishers, 1999). I don't know if many virtue epistemologists would find this a welcome consequence.

[15] This is a sort of converse of David Lewis's assertion that "when error would be particularly disastrous, few possibilities may be properly ignored": 'Elusive knowledge', *Australasian Journal of Philosophy* 74, 1996, pp. 549-67. See also chapter 2 of John Hawthorne, *Knowledge and Lotteries* (New York: Oxford University Press, 2004.)

different antibiotic, bacillus, and culture, and his samples are in the same lab as our biologist. One Saturday morning she has come in to the lab to check on her experiment, and has also promised her colleague to check on his. She intends to do both, but instead gets distracted and while looking at her experiment muses about what it would be like to win a Nobel prize at the age of 26. So she doesn't even glance at her colleague's samples, but does inspect her own and realizes that the bacillus is being killed off. She checks various obvious possibilities and comes to the conclusion that it is the antibiotic that is killing the bacillus. She does not even consider the extremely rare and unlikely possibility that the bacillus is being attacked by a particular virus V that normally does not infect that bacillus. But, though she does not consider it, if that virus had been present the symptoms in the affected bacilli would be exactly as she observed. Now as it happens though V is not present in her samples it is present in the simultaneous decimation of the bacilli in those of her colleague, which are much more susceptible to V. Had she done as she intended and looked at her colleagues samples, she would have immediately wondered whether V was present, and then would have been led to check whether it was present in her own samples. But in fact she did nothing to rule out this possibility, because of an irrational distraction from the course she had wanted to follow.

Does the soldier know that his wife is at home? Does the biologist know that the antibiotic is attacking the bacilli? My intuitions suggest that they do not. I expect these examples to be controversial, but what they suggest to me is that when considering knowledge, too, considerations about how a person ought to be thinking turn on the person's big state, their whole complex of beliefs and desires[16].

knowing when to reflect The big state rationality thesis is a conjecture, which if true connects the enlarged theory of belief and decision that I am looking forward to with traditional epistemology in a particularly simple way. If it is right then epistemology is a simpler and less puzzling subject in the new context. But it is definitely conjectural. To end the paper I shall discuss another aspect of rationality that has played a large role in the history of epistemology:

[16] Both my examples have a moral flavor, which I have tried to keep out of the discussion. But it would be good to explore examples in which whether a person knows something is affected by the possibilities excluded by intellectual strategies they morally ought to be adopting or avoiding. The suggestion here is in accordance with a suggestion that John Hawthorne attributes to Jonathan Schaffer. See footnote 53 of p 188 of *Knowledge and Lotteries* (Oxford: Oxford University Press, 2004.)

the ability to reflect critically on one's own processes of belief formation.

Few contemporary epistemologists will defend the idea that a rational human agent should exercise constant control of her belief-forming processes, always conscious of them as they occur and always checking them against explicit norms of rationality. But skepticism about the ideal of conscious rational control should not make one deny that there is a special importance to the virtue of knowing when one should reflect, and run-over, check, and reconsider one's thinking. In fact it is particularly significant when belief and choice interact. It is Reflection can be a naïve business of pausing and asking "does that seem right?" or it can consist in applying explicit norms of rationality. In either case, it offers as many possibilities of obstruction as of help. Reflection burdens working memory, introduces new sources of error, and generally slows things down. Done at the wrong moment, it can hinder or even sabotage reasoning that would otherwise succeed. So we have to know when to do it. And usually we cannot or should not try to know this by thinking out in a principled way "this is/is not a moment to pause and take stock", for two reasons. Explicit thinking of this kind is very expensive in cognitive resources, which are not likely to be available at the very moments that reflection might be called for. (It is when everyone has been called out to fight the fire that it might be most relevant for someone to tell us that it is a false alarm, but at that of all times we cannot spare someone to find out.) And we very rarely have enough knowledge of our own thinking to give us the cues that such principles would engage with. (Even if we could spare someone to check for false alarms, he would be guessing half the time.) If the big state rationality thesis is right then the barriers to self-knowledge here are even more formidable. It would be asking much to much to demand that one know what strategy one is following, the relevant characteristics of the totality of everything one believes and wants *and* the limits of one's own particular capacities. So in general one will not know if one's belief is formed by an acceptable process; we're better off considering simply how adequate the evidence is.

These two reasons connect, in that even when we could learn enough about our on-going thinking to apply such meta-principles doing so would draw on the very resources whose scarcity makes reflection often a bad idea[17]. This is so even thought there are exceptional cases, where it is easy to tell that reflection is called for. For example it doesn't cost much to follow the principle "when you find

[17] In this connection see chapter two of Hilary Kornblith *Knowledge and its Place in Nature* (Oxford: Oxford University Press, 2002)

yourself driven to a transparently implausible conclusion, stop and see if you've done something stupid". But the conclusion has to be transparently implausible. Usually when a claim is very improbable given one's prior beliefs, or even contradictory, it would take a lot of thinking to make this explicit. So even this mild principle carries the danger that it might lure someone into excessive fussing at the boundaries of what is obvious.

So how do we know when to reflect? Very often we don't, and do it too often, too little, or at the wrong times. But there is a virtue of appropriate reflection, which some people exhibit on some topics. (We don't have a good name for it: perhaps "rationality" will do, taken as a more subtle thing than simply the capacity to reflect: that capacity plus the sense when to use it.) The cognitive psychology of intellectual virtues is largely unexplored, but however the details work out, it seems to me that it must consist in sensitivities to large libraries of typical cognitive situations built up over a period of time and then recognized as similar to situations as they occur. So one has to be able to build up the library, store it in an accessible way, and recognize relevant similarities with actual events. (The pattern is not unique to intellectual virtues. Chess players build up large mental databases of "combinations", which they have to recognize in actual token combinations of pieces. Courageous, honest, or generous people will have gone through the Aristotelian process of observing and cataloging the admirable actions of their elders and betters, while slowly learning what it takes to imitate them.[18]) The important point is that the working of the virtue will usually be inaccessible to the agent and not tune-able by her on the particular occasion. She cannot simulate it by thinking "what would a well-constructed agent do in this situation.[19]"

At this point we should ask: what makes it an appropriate time to reflect? One should reflect when it helps to, of course. But helps what? Helps one's reasoning to conform to norms of rationality? Helps one achieve epistemic and practical ends? When it is a matter of reflecting on the thinking that concerns us here, thinking that combines belief and choice, the aim of conforming to rational norms is really not an option. For we don't really have any culturally inherited or apriori accessible norms for this general case. We have norms for

[18] For more on this again see my "Epistemic virtues, metavirtues, and computational complexity", *Nous,* **38**, 3, Sept 04.

[19] The virtue of appropriate reflection is a higher-order virtue, consisting in the direction of first-order thinking. Its relation to simpler intellectual virtues is analogous to the relation of the virtue – also vital, also nameless – of being able to feel regret at the right moment and in the right amount to simpler moral virtues such as courage and honesty.

assessing the force of evidence and a few rough and ready norms of good epistemic procedure; we have rough norms of means-ends rationality and more precise rules for calculating expected utility. But we do not have anything explicit to guide us in choosing how to allocate our intellectual resources between competing parts of a epistemic/practical project, or how to choose a procedure for choice of belief or action that will fit best with the rest of an epistemic or practical project. So if there are moments that are appropriate and inappropriate to stop and consider one's thinking, they are surely determined by more externalist considerations, in terms of what procedures are in fact likely to produce what kinds of results[20].

It is interesting that we almost never reflect on how our belief-acquisition and our decision-making fit together, even though their fitting is crucial to almost all of our activities. The virtue of knowing when to reflect has an aspect of benign illusion about it: it directs us at only part of the cognitive situation at any time. It says "rethink the logic here", or "go slowly about the choice of options here", or "there must be further consequences"; but rarely more than one of these, and almost never asks us to reflect on our belief-formation and our choice at the same time. That is just as well, as we wouldn't know how to go about such a complicated reflection, but it leaves us with the impression that we are aware of far more of our thinking than we are. In fact, the choice of intellectual strategy, the way that we nudge ourselves into one or another procedure for coming up with choices and beliefs, guided at suitable points by explicit reflection on tiny parts of our thinking, is by and large a mystery to us, both introspectively and in terms of normative lore[21].

The choice of strategy, then, probably the most crucial element in our intellectual life, is not something that we can evaluate in accordance with any standard norms. But some strategies are clearly successful and some disastrous. And, discriminating more finely, some are clearly more successful than others. There are many indeterminate cases, too. A person values truth above everything and who adopts an epistemic policy which gives beliefs that allow her to achieve many of her less abstract desires and live happily, but at the cost of a large number of false beliefs: is the policy Right because it

[20] But given the length of time it has taken to get at all clear about evidence - from Humean gropings about induction to confirmation theory à la Hempel and Carnap to the Bayesian orthodoxy that I take to be our best current account – one could doubt whether many people have ever made much conscious appeal to correct and explicitly formulated norms of evidential force.

[21] Another aspect of the benign illusion: it tends to direct us to the tractable question "is this reasoning valid?" rather than the more important and much less tractable question "are these claims consistent?"

gets her a good measure of what she wants, or Wrong because it does not achieve the good that she would herself have judged it by? I am not sure there are answers to such questions. But this fine-grained indeterminacy should not obscure the fact that there are systematic factors which make some strategies successful and others not.

These factors are extremely complicated. They would involve far too much searching for individuals to think them through instance by instance. So they are externalist factors; they are not to be applied in reflective regulation of one's own thinking. But they are externalist factors with a twist: they determine the internalistic criteria that we do apply reflectively. For when it is appropriate to reflect and we reflect successfully we apply standards of reasoning that are appropriate to part of our reasoning given the unreflected-on nature of the rest of it. The standards are only appropriate because they fit into a larger pattern, which itself is valuable simply because it works.