Toward a Unified Account of Rationality in Belief, Desire, and Action

Peter Railton

Preliminary draft of November 2010—please do not circulate without permission

Introduction

Consider Robert Stalnaker's well-known, elegant formulation of the functional nature of belief and desire:

> Belief and desire … are correlative dispositional states of a potentially rational agent. To desire that *p* is to be disposed to act in ways that would tend to bring it about that *p* in a world in which one's beliefs, whatever they are, were true. To believe that *p* is to be disposed to act in ways that would tend to satisfy one's desires, whatever they are, in a world in which *p* (together with one's other beliefs) were true. [Stalnaker (1984), 15]

In this picture, belief and desire play co-equal, complementary roles in shaping an agent's spontaneous tendencies to act. These roles are part of what makes belief and desire the states they are, and so even in those individuals who are, as Stalnaker suggests, potentially but not fully rational, their beliefs and desires will work together in this coordinated way to impart a great deal of shape and recognizable purpose to what they think and do.

Under what conditions will agents be not just potentially, but *actually* rational in thought and action? I know of no short answer. But we might begin piecewise. Some modicum of rationality in how they form and revise their beliefs—"whatever they are"—would surely help. And fortunately, we have a lot to say about this—though no consensus exists, a number of systematic, well-developed accounts of rationality in belief are on offer. What about the agent's desires—"whatever they are"?[1] As co-equal participants in shaping dispositions to act, must they, too, show a modicum of rationality? That sounds reasonable enough. After all, in everyday life we often attribute the irrationality of an act—our own or that of another—less to irrational beliefs than to irrationality in the desires that shaped or motivated it. In clinical practice, likewise. Such desires might be arbitrary, groundless, obsessive, phobic, infantile, compulsive, delusive, incorrigible, uncontrollable, disproportionate, self-defeating, distorted by mood or emotion, …—it is a very long list.[2] Calling these forms of *irrationality* in desire seems about as natural as saying that arbitrariness, groundlessness, obsession, phobia, immaturity, compulsion, etc. are forms of irrationality in belief. Moreover, even though those of us who have (what would normally be called) irrational desires typically also have associated irrational beliefs, it does not seem that we can reduce all forms of irrationality in desire to irrationalities in belief. In many cases, it seems the other way 'round—irrationality in desire is driving irrationality in belief. Neither do all apparently irrational desires derive from *false* beliefs. Individuals in the grip of addictive or compulsive desire can have a frighteningly clear view of their situation.

[1] At the outset, I will follow conventional usage among philosophers and treat the term *desire* as a generic term for motivating states of all sorts—from bare urges and appetites to considered preferences and goals. Aversions, too. Later I will identify something more specific for desires to be.

[2] I mean this list to be focused upon the more serious forms of irrationality in desire. I therefore pass over the many relatively harmless spontaneous fancies and spur-of-the-moment urges that leaven our lives, even though these are sometimes spoken of as irrational because they lack any obvious grounding in experience or considered thought.

[2] Here, too, I follow standard usage in speaking of beliefs and desires as propositional attitudes. There are other, perhaps better, models of these states. This question will also arise later on.

Where, then, are the systematic, well-developed accounts of rationality and irrationality in desire that might underwrite such everyday or clinical ways of talking? Instead we find among philosophers a widespread sense that desires and desiring by their nature cannot be evaluated as more or less rational. This claim is typically somewhat qualified—intransitive preferences strike almost everyone as rationally problematic, and as do failures of means/end coherence in what one wants. But these flaws seem traceable to self-defeating relations *among* one's desires, leaving untouched the question which, if any, of the desires involved have greater or lesser rational standing.

Neo-Humeans offer a principled explanation of why this is so. The core operations of reasoning—such as deductive and inductive inference—apply only to bearers of truth conditions, e.g., propositions as classically conceived. Of course, beliefs themselves are no more propositions than are desires—both are propositional *attitudes*.[3] Yet it is said to belong to the nature of the belief attitude to "inherit" the truth conditions of its constituent proposition. If Doreen believes she has a drink of water when she does not, her belief is false—beliefs have a "mind-to-world" direction of fit. If, on the other hand, she *desires* to have a drink of water when she does not, there is no falsehood at all. On the contrary, this is just the sort of condition in which such a desire seems fitting, since desires serve to motivate us to change things to make their propositional objects "come true"—desires have a "world-to-mind" direction of fit.[4] To be sure, we do say that desire is *satisfied* its object is attained, but such satisfaction does not make the desire true any more than its unsatisfied state made it false. Reason and

[3] Here I follow standard usage. We will discuss the question whether belief and desire are always best interpreted as *propositional* attitudes, below.

[4] For seminal discussions, see Humberstone (dddd) and Smith (1994).

reasoning thus lack the requisite purchase on desire. For example, holding beliefs with incompatible objects is said to be contrary to reason because one has, by one's own thoughts, guaranteed *a priori* that one is making a mistake. But there is no such guaranteed error if one holds desires with incompatible objects—this might simply reflect the fact that not all good things can be had together.

Unbound by the constraints of reason and reasoning, desire becomes a wild card in the "orthodox belief-desire model" of intentional action, and this helps to explain the bad reputation this model has acquired in recent years. It is easy to see how desires—"whatever they are"—could be the *causal* reason why someone acts. But how could mere motive force, not subject to rational discipline or assessment, have any claim to the authority necessary if action from desire is to constitute action for a *normative* reason? Warren Quinn famously invited us to consider a man who turns on all the radios he encounters, not because he wants to hear music or the news, nor because he enjoys flipping switches, but simply because he feels the urge and acts from it.[5] Where is the reason in that?

The philosophical case against rationality in desire is compelling, but I think it needs re-examining. I am inspired not only by the hunch that everyday and clinical talk of irrationality in desire seems to be onto something, but also by the broad thought that Nature would never be so feckless as to endow animals, humans included, with refined senses and complex cognitive capacities, but take no care with how or what they were moved to do with this equipment. Of course, it will be argued that, were our sensory and cognitive capacities not somewhat reliable as guides to the world, we humans or

[5] See W. Quinn, ref.

our animal ancestors would long since have disappeared from the face of the earth. (Insert here the mandatory remark about not becoming lunch for a charging tiger.) Hence there is good reason to think that the *representations* of the world produced by our sensory and cognitive systems would have some epistemic and action-guiding authority. But since desires have only a world-to-mind direction of fit, and thus are non-representational in function, there can be no question of their being selected for anything like reliability.

This line of thought, however, overlooks the fact that Nature is not the College Board. Nature rewards correct answers to questions only insofar as they translate into advantageous action, and in the translation from thought to action desire plays a role no less central than belief. Had our ancestors' motivational systems not been capable of operating in tandem with perception and cognition to supply the interest, impetus, and aims necessary for such beings to meet their needs, and moreover to flourish and promote the flourishing of their kin and kith, we would not be here today to celebrate our beliefs and denigrate our desires. For better or worse, we humans have overrun the planet, making ourselves at home in the most varied and difficult environments. For success in such settings, no set of basic instincts or drives could have sufficed—indeed, the key to rapid human expansion and adaptation has been culture, and the extraordinary proliferation of human social and symbolic forms has been marked as much by change and nuance in what people *seek* as by change and nuance in what people think. It would be bizarre if our senses and our capacity to form beliefs were not well-coordinated, the better to serve the need for effective action. Yet there can be no action without motivation, and no *effective* action without well-adapted motivation. So here, too, we should expect good coordination and sensitivity to experience. The question is, can we make philosophical and psychological sense of this idea?

Generic rationality: attunement to reasons

To get started, we will need a working notion of rationality broad enough not to beg the question against rationality in desire. For example, in contrast to the neo-Humean view just discussed, it cannot simply stipulate that only states strictly capable of truth or falsity can be assessed for rationality. That might seem too large a concession to grant up front—I would be making things too easy for myself. But we are very familiar with speaking of the rationality or irrationality of actions and feelings (such as fear or trust), and neither of these are strictly capable of being true or false. Consider even the case of belief. One cannot know whether a belief of mine is rational or irrational simply by knowing whether it is true or false. It could matter, for example, whether it is based upon good evidence or the result of coherent thought. And, once this is known, it is entirely possible to determine that some of my false beliefs are rationally held, while some that are true are not. Moreover, it typically matters whether the *strength* or degree of confidence of my belief is *proportional* to the strength of my evidence. We readily speak of irrational certainty or rational doubt. But degrees of confidence, certainty, or doubt are not themselves assessable as true or false. So seeking an initial idea of rationality that does not stipulate that only states capable of truth or falsity can be rational is hardly special pleading on my part—we need such a notion for many purposes.

Following upon the idea of belief that is proportional to evidence, my suggestion is that, as a first approximation, we think of rationality as a capacity to respond aptly to reasons—in kind and degree. Of course, making any positive use of this idea requires that we have at least some agreement upon what sorts or reasons there are, and what would constitute apt or proportionate responses to them. And that is indeed what I will be assuming. Rationality is a normative notion, and we cannot avoid the need for normative opinions in trying to determine its nature or operation.

But, at least at the outset, the normative opinions should be familiar or non-controversial, and they should afford a representative sample. Some cases will be abstract. For example, we might agree that a relatively stable perceptual appearance as of *p* is, other things equal, a *prima facie* reason of some weight in favor of belief that *p*, or that a regular pattern of past observations that *F* is accompanied by *G* is, other things equal, a *prima facie* reason of some weight in favor of expecting the next *F* will also be *G*. Other cases will be particular. For example, we all might be able to agree that, absent further information, Quinn's radio man has a *bona fide* urge but not a *bona fide* reason for behaving as he does. And we might be able to agree that an acrophobic individual's crippling fear of open spaces is not an apt response to the actual degree of risk he faces. Moreover, we might be able to agree that a novice gambler, who has won the first bet she places on a roulette wheel and decides that this is her lucky night, doubling and redoubling her bet, is exhibiting an irrational degree of confidence in a favorable outcome on the next spin of the wheel, and acting irrationality as a result. However, we might credit her with somewhat greater rationality than a novice gambler who wins his first bet, thinks "Daddy told me to quit while you're ahead—lightning never strikes twice in the same place", concludes that he *won't* win on the next bet, but finds himself excitedly raising his stake regardless.

Rationality in the "apt responsiveness" sense can be thought of as a form of non-accidental *attunement* to reasons. It will often involve the use of *reasoning*, but other forms of responsiveness to reasons will also be important, and reasoning won't always help, as our second gambler shows. Consider the case of beliefs or feelings issuing "non-inferentially" from perceptual experience. While perceptions themselves do not appear to be rational or irrational, perceivers can be, and so can the perceptual beliefs and spontaneous feelings they form—one might come to believe that *p* in proportion to the robustness and reliability of one's perception that *p*, or, by contrast, one might jump to extravagant conclusions like a novice gambler, or form exaggerated fears like an acrophobic. Of course, our generic characterization is too broad. Not every capacity that contributes to apt responsiveness to reasons is part of rationality. Having good eyesight, for example, helps one to respond well to available perceptual evidence, but it is not in itself a defect in rationality to have bad vision. However, it *would* be a defect in rationality to be unaffected by abundant evidence of poor eyesight, and fail to gain any discernment about when or how much to trust one's eyes. I won't attempt here any more precise characterization of which capacities belong to rationality, but instead will attempt to operate in relatively uncontroversial territory.

But not all controversy can be avoided. There are those who are outright skeptical about reasons for action, or about reasons of all kinds, belief included. *They* certainly won't find it uncontroversial to begin with bland assumptions about the existence of reasons or the aptness of our responses to them. So the account to be given here must be understood as conditional in form: *If* there be reasons to believe, act, or feel of the kinds we normally suppose, *then* a unified approach to rationality in belief, action, and feeling is possible if we take rationality as, to a first approximation, a matter of possessing capacities that enable us, non-accidentally, to cotton onto them and respond appropriately.

Rationality in belief: learning from experience

Let's take our first steps toward filling in a unified account of rationality by beginning where there is greatest agreement that rationality is possible, and greatest clarity about what such reasons or apt responsiveness to them might consist in: belief.

How might we be non-accidentally attuned to reasons for belief? A pre-established harmony between mind and world would do the trick, but who is so lucky? Second best: we should be able to learn—and unlearn—from experience. That seems straightforward enough, but understanding how learning from experience is possible turns out to be non-trivial and instructive.

One might think, for example, that the rational way to learn from experience would be to cast aside any initial bias or preconceptions and diligently record exactly what one observes. Asked what to expect next, one would then review these observations and report their import. Surprisingly, as Rudolf Carnap's studies of confirmation functions showed many years ago, this is in fact a recipe for *inability* to learn.[6] He began with a fairly elementary and uncontroversial characterization of what it

[6] For discussion, see Carnap (1955). Page numbers in the text are to this source.

is to learn from experience: "other things equal, a future event is to be regarded as the more probable, the greater the relative frequency of similar events observed so far under similar conditions" (p. ref.). He then considered the following, seemingly natural way of understanding what it would be to begin one's observations without initial bias. *A priori*, each possible sequence of individual events over time—each possible total "state description" of the history of the universe—is to be treated as equally likely.[7] Next, one begins one's observations of the actual world, taking careful note of each experience and striking off one's list of possible state descriptions all those inconsistent with what one has observed thus far.[8] If the sun comes up in the East this morning, for example, one will strike off all descriptions of the universe's history in which that event does not take place on this day. Suppose now that it is New Year's Eve 2020, that one has been making observations of sunrises for many years, that each and every day so far the sun has come up in the East, and that one has scrupulously updated one's list accordingly. If one is now asked what to expect from the sun on the morning of New Year's Day, one need only take the ratio of the number state descriptions remaining on one's list in which the sun *does* come up in the East January 1, 2021 to the total number of state descriptions still on the list. This way of heeding the voice of experience has, however, a curious result. Despite the perfect regularity of the sun's previous observed behavior, one's assessment of the chance that it will rise in the East this New Year's Day

[7] In order to get a workable notion of "possible way the world could be", Carnap considered all possible complete assignments of individuals to predicates, using a finite number of individuals and predicates. If the set of predicates were {*P*} and the set of individuals {*a*, *b*, *c*, *d*}, the possible "state descriptions" would be <*Pa*, *Pb*, *Pc*, *Pd* >, <~*Pa*, *Pb*, *Pc*, *Pd* >, <*Pa*, ~*Pb*, *Pc*, *Pd* >, … , <~*Pa*, ~*Pb*, ~*Pc*, ~*Pd* >.

[8] For simplicity, assume that each observation is perfectly confident. This assumption is not strictly necessary for the argument that follows.

morning is 0.5, "indifference", just as it was before one had made any observations.[9] What went wrong? The initial, unbiased, perfectly flat probability distribution across possible state descriptions in effect made all individual events probabilistically independent of one another, so that no sequence of past events could have any bearing on what to expect in the future.

If we are to learn from experience, Carnap concluded, we must instead begin with a bias toward finding regularities or similarities. Indeed, as Nelson Goodman later emphasized,[10] this bias must favor certain kinds of regularities or similarities rather than others. In Carnap's scheme, we might assign equal *a priori* probability to certain regular *patterns* of outcomes extending uniformly over time—certain "structure descriptions" of the universe.[11] The *a priori* probability of particular events will then depend upon their distribution within these structure descriptions. Once we begin to make observations, striking off individual events and patterns inconsistent with what we have observed, we will find that the ratio of favorable events to total possible events also begins to change, reflecting which patterns remain on the list. By the time we reach January 1, 2020, the expectation of an eastern sunrise can be far from chance.

Carnap's models of confirmation were based upon "logical probability", an *a priori* measure that itself remained static in response to the new experience. Since then, more fully dynamic models of learning have been developed. In Bayesian learning, for example, initial prior probabilities are revised in response to experience via an updating

[9] If there were *n* possible total states of the universe at the outset, then, at the beginning of our observations, there will be *n*/2 in which the sun rises in the East on that day, and in *n*/2 it does not.

[10] Nelson Goodman, *Fact, Fiction, and Forecast* (1955).

[11] In our toy example, the structure descriptions might be, e.g., "The sun always rises in the East", "The sun rises in the East every other day", "The sun never rises in the East", and so on.

rule that takes into account how much the new evidence was expected on its own vs. how much it was expected in light of one's going hypotheses, to yield posterior probabilities that then become the "priors" on the basis of which the next bit of new evidence will be assessed. As evidence grows, biases present in the original set of prior probabilities will tend to exert progressively less influence on one's evolving expectations. Under certain favorable assumptions, as evidence grows the initial bias will tend to "wash out" altogether, and inquirers starting from a wide range of starting points will tend to converge in their expectations, which in turn will tend to approach arbitrarily close to the actual relative frequencies in the world.[12] Viewed sideways, this process can be seen as a paradigmatic example of learning, not just in Carnap's technical sense, but in the broad sense of *attunement* described above: agents whose beliefs are in this way responsive to experience will tend to overcome initial biases and become more in tune with one another and with the world they cohabit.

Since the early work by Carnap and others, it has become a central theme in formal learning theory that experiential learning is driven by a process of *expectation*, *observation*, and *discrepancy reduction*.[13] The learner starts, not empty-handed and empty-headed, but rather equipped with certain assumptions about the world—such as dimensions of similarity (an "implicit quality space"), prior probabilities, or implicit causal models—which generate expectations about what will be observed. These expectations then confront actual observations, and, via some algorithm or other regular process, forward-going expectations are readjusted, if necessary, to reduce discrepancy with observation. This process is then repeated until—if the world is sufficiently regular in

[12] See I.J. Good (ref.) and (ref.).

[13] Reinforcement learning in animals takes a similar form, where expectation includes reward. We will return to reward learning, below.

relevant respects—discrepancies no longer occur, expectation value matches observed value, and "belief' stabilizes. Rationality in belief, on such a picture, is embodied by dynamic process of attunement of this kind.

Bootstrapping and default self-trust

If the foregoing is right, then it cannot be a demand of rationality that one refuse to attribute any epistemic authority to oneself or one's expectations—one's priors, as it were—until one has some positive evidence for doing so. For rationality cannot command us to destroy our best chance to learn.

We can see this most clearly by considering one's own mental faculties. Suppose that one were to insist that, before relying upon one's perception, memory, or inferences, one must have some evidence of their reliability? How could one acquire such evidence, understand it, store it, or assess its bearing, if one were still awaiting permission to use one's faculties? One could not even hope to start off from "self-evident" truth, since nothing is evident to you except through the use of your own faculties—if they possess no initial credibility, then neither does what seems evident to them.

Talk of "self-attribution of epistemic authority" might mislead, since it can sound like a form of normative judgment, but there can be no act of *judgment* until something can be taken as its ground or content. So such self-attribution must be *non-judgmental*—a precondition that makes acts of judgment possible. How should we conceive it, then? It is a *non-derivative*, *default disposition* to rely upon one's faculties: initially, one takes the content of what one perceives, remembers, or thinks at face value. ("You want we should begin with what we *don't* believe?" David Lewis was fond of asking.) Even if this results in conflicts, at least we will have something to work with—including a potential wealth of new experiences and thoughts, taken initially at face value—in trying to resolve those conflicts.

We should, then, obsess less about whether one can give a justificatory reason for everything we believe and instead understand justification and rationality in belief more dynamically, as a matter of how one proceeds from wherever one happens to be at present—justified or unjustified as it might be. Isn't this cutting oneself too much slack? On the contrary, it is what makes trial-and-error learning possible. For the same default reliance that induces expectation and action ("trial") also sets us up to take at face value the result, even when it is contrary to expectation ("error"). Default reliance encourages epistemic risk or experimentation, while also allowing adverse outcomes to hit home (ignoring evidence is *not* default reliance upon experience), and to remain a source of conflict until some doxastic adjustment lessens the discrepancy. As a result, default credences will tend also to be *revisable* and *defeasible*, as one starts to catch on to when, and when not, to believe one's eyes, or to rely upon one's memory or mental arithmetic, or to retain or revise one's initial thoughts or methods.

A dynamic conception of rationality thus enables us to see both how it is possible to escape the epistemic self-liquidation that would result from demanding reasons at every point, and how we might be able to avoid a life sentence of epistemic imprisonment in our initial convictions. "Bootstrapping" is a feature, not a bug. But the point is to pull up, not down.

Our first conclusion, then, is that rationality in belief involves according some measure of default, defeasible epistemic authority to the content and capacities of one's own mind. Put more simply, rational believers are initially disposed, other things equal, to rely directly but defeasibly upon what they see, feel, and think.

A second conclusion is that the need for such default, defeasible self-reliance does not go away once we have gotten inquiry up and running. Even the most advanced, critical reflection at every moment takes for granted at least as much as it puts into question. One cannot both rely upon *none* of one's faculties or beliefs and at the same time engage in the sort of sustained, organized thinking necessary for genuine self-criticism and reflection. This point is well known in the case of *inference*. Reflection often involves reasoning, and reasoners must possess some default, defeasible dispositions to follow certain patterns of thought directly and non-deliberatively, transmitting credence as they go.

Consider for example the transition from a belief that (*p* **v** *q*) and a belief that ~*p* to a belief that *q*. Suppose that no such mental transition could count as *reasoning*, or as made *for a reason*, unless it were mediated by a judgment or other mental state with a content like 'that (*p* **v** *q*) and ~*p* is a reason to believe that *q*'. This proposal would be hard on young children, who seem capable of understanding the content of simple propositions and of mastering simple patterns of reasoning such as "the process of elimination"—"Spot is hiding in the basket or the box. Let's look. He isn't in the box. Where is he?"—well before they can articulate their thinking in terms of reasons for belief.

But the proposal has a deeper flaw. To explain the difference between reasoning and a mere—and therefore purportedly non-rational—transition from one thought to another we have introduced an additional, intermediating step between thinking that (*p* **v** *q*) and ~*p* and thinking that *q*, e.g., thinking that [(*p* **v** *q*) and ~*p*] is a reason to believe that *q*. That seems harmless enough until we ask, "Was the step from *that* intermediating step to believing that *q* made for a reason? Or was it just a non-rational transition from one thought to another?" If we say it was not made for a reason, then we seem to have broken the chain of reasons in trying to fix it. If we say it was made for a reason, then there must be some state or judgment intermediating this transition in thought, perhaps, the thought that {[(*p* **v** *q*) and ~*p*] and [(*p* **v** *q*) and ~*p* is a reason to believe that *q*]} is a reason to believe that *q*. But now we have introduced a second intermediating step. And we cannot escape the question, "Was the step from *it* to our conclusion made for a reason, or just a transition in thought?" Soon, as the Tortoise pointed out to a much-annoyed Achilles, where we had one step, we'll have two, then three, then … . In the name of rational self-governance, we'll have incapacitated ourselves for reasoning.

Is the only alternative positing mindless, reflex-like dispositions to make direct mental transitions? No. The dispositions involved are sensitive to the *contents* of the propositions figuring in reasoning including the logical *rules* or *patterns* tacitly encoded in an understanding of *or*, *and*, and *not*.[14] For example, occurrent belief that (*p* **v** *q*) and ~*p*

[14] What entitles me to speak of *propositional content* or *logical rule* here if the mind is no more than a very complex bundle of dispositions? Here is another kind of skepticism I won't be trying to overcome. So I will assume in what follows that *some* sets of dispositions can indeed make it the case that one's mental states have a certain propositional content or that certain of one's mental transitions are indeed rule-governed. Understanding the nature of the mental states involved in belief, I think, can make this assumption more credible, as can neuro-physiological evidence that, for example, inference draws upon

will potentiate some inferences, such as the inference to *q*, but not others. Compare: Why, when Lucy responded with exasperation and incredulity to the airlines baggage agent's casual report to her that, "Your luggage is still in Denver", did Lucy use 'not ever even', phonetic emphasis, and the third-person plural past tense in exclaiming, "What?! Our bags weren't ever even *loaded* in Denver?"? Rapidly and with no deliberation, a bland present tense singular assertion became, in her mouth, something with a definite logical relation to that original assertion. This was not accomplished by a mindless reflex, but rather reflects her mastery of content of the words involved (e.g., 'luggage' and 'bags' are near synonyms), and the rules for verb agreement, for temporal order and the past tense of 'to load', for negation, and for the conventional expression of emphasis in English, as well as various conversational norms. (Compare what would have happened had the airline agent been French, and Lucy had tried to express her exasperation and concern using her own very imperfect French. She might well have stumbled, and needed to recall explicitly conjugation rules.) The dispositions underlying Lucy's spontaneous remark are complex, and sensitive to a large number of grammatical, rhetorical, and conversational norms or reasons. She had good reason to say what she said, and to say it in just the way she did, and her speech was aptly responsive to—*attuned* to—these reasons *via* the complex, context-, content-, and rule-sensitive dispositions that constitute her linguistic and social competencies. All this is to say, her spontaneous speech was very far from a "brute", mindless response to her situation, even though it was produced by the unmediated operation of dispositions to think and speak rather than by a conscious "recognition of reasons" or judgment.

areas of the brain specialized for learning and deploying rule-like patterns. But I won't have much more to say on this general question in what follows, though I do say more in Railton (2001) and (2006).

Moreover, to say that a disposition to follow a certain pattern of thought operates by default is not to say that it is *indefeasible*. The mind is a connected, structured web. Just as the occurrence of the mental contents (p **v** q) and ~p will tend to potentiate directly the content q in someone who has mastered *or, and,* and *not*, so occurrence of the mental content q will tend to potentiate directly q-related links. If the individual reasoner has no antecedent reservations about q, the inference may go through without a hitch, which seems to be an appropriate response in his doxastic state. But suppose the individual's recent experience has led him actively to doubt whether q. Then the occurrence of q in his inference will potentiate q-related links that tend to *inhibit* the direct translation of belief from (p **v** q) and ~p to q, and may result instead in *hesitation* or *puzzlement*, which in turn might lead the individual's thoughts in the direction of (p **v** q) if recent experience has also made ~p quite firm in his mind. These, too, seem to be appropriate responses. Seeing the mental transitions underlying inference and the transfer of belief as operating by default thus is not incompatible with seeing them as sensitive in ways tied to content and to the individual's wider epistemic situation, including defeating information. Here, too, we can say that such dispositions can attune the individual to available reasons.

Other kinds of conditions with epistemic relevance can also affect the operation of default, defeasible mental transitions. One can learn through embarrassing or regrettable experience that one reasons less well when fatigued, distracted, in a rush, or upset. And one's dispositions to infer can come to be sensitive to such performance-degrading conditions as well. For example, tacit or explicit learning can lead one to become more hesitant about relying upon, or encouraging others to rely upon, one's first responses when one is upset and distracted. The dispositions that subserve inference function to transfer belief from step to step, true, but just how *strongly* they do this is a matter or degree. Conditions insufficient to defeat belief transference might nonetheless appropriately moderate it, or make one disposed to second-guess it.

In short, default reliance—in its shape and scope, conditions, and degree of confidence—can be understood as sensitive in its operation to a wide variety of reasons. And as we grow in experience, maturity, and savvy, it can become downright intelligent.

Of course, even if basic inference proceeds by direct, unmediated steps, it is entirely possible for an individual who has the full panoply of meta-representational capacities and normative vocabulary to render a particular inference explicit, taking note at each step of the reason for making the next one. Similarly, individuals are often in a position to reconstruct their tacit reasoning after the fact—the directness of basic inferential processes does not prevent them from having recoverable content, as we have suggested. Individuals need not be in the dark about what they are thinking and why. Since, as we have argued, basic inferential processes are contentful, there will be some guiding ideas at work, regulating the course of thought. But the crucial point for the present discussion is that the basic steps of a higher-order thought process of explicitating or reconstructing an inference, or the basic regulation of connected thought processes by content and logic, must be direct and dispositional.

A complex array of default, defeasible dispositions is therefore essential equipment for learning and reasoning, and in effective learners and reasoners these dispositions will typically have been hammered into more reliable, context-sensitive forms by experience. That said, however, we should be clear that this does not constitute a *validation* of such self-reliance. A given epistemic starting point might be so benighted that no one who begins by taking it at face value could find her way from error to truth, even if she were ever-so-sensitive to experience and thereby became more practically successful over time. Or an ill-fated epistemic agent might be cursed with undetectably defective faculties, or surrounded by evidence so devilishly misleading that careful observation and thought only leads him further astray. These unfortunate souls might never become in tune with the facts. It seems to me, however, wrong to take this as a mark against the dynamic conception of rationality. Truth is always a gamble, and we must play with the cards we are dealt.[15]

Belief basics

These arguments suggest that belief must originate in, and be sustained by, default credence in one's faculties and in the content of one's experience and thought. Remove this, and our learning and reasoning will come to a halt, as we wait without hope for some sign that it is OK to resume.

What sort of state might belief be, if it is to function as described? As we have seen, it must be capable of arising and doing its job *spontaneously*, without needing some additional judgment—at least, if we understand judgment to be a mental act that by its

[15] Cf. Isaac Levi, *Gambling with Truth* (ref.).

nature needs reasons.[16] That seems to fit with the fact that one can form large numbers of beliefs simply by keeping one's eyes and ears open while walking down the sidewalk at lunch hour, even if one is at the same time fully engaged in a conversation with a friend. One does not need to *do* anything extra or effortful to keep the beliefs coming.

Moreover, these freshly-acquired beliefs typically will, without further effort, effectively do their Stalnakerian job of shaping how one is disposed to act—directly contributing to the guidance of one's walking in light of one's other beliefs and one's destination-seeking, conversation-continuing, and collision-avoiding desires. One will weave through on-coming foot traffic, slow down or speed up to match the pace of one's friend, and steer around the occasional fire hydrant or lamp-post, all the time staying largely on course for the restaurant where one has it in mind to eat. And in addition to the many new beliefs formed, many existing beliefs will be revised with equally little effort. One's sense of the time is more or less continually being updated, as is one's sense of the distance remaining to the restaurant and one's confidence that one understands what one's conversation partner is getting at, or has good answers to the hard questions she is posing.

Belief formation and revision also seems for the most part *nonvoluntary*. Wishful thinking seems to work for some beliefs and some contexts, and we do seem capable of shaping to a considerable degree what we pay attention to or recollect. But even if I

[16] We are here using the term 'spontaneous' in this philosophical sense, as meaning, roughly, "without conscious deliberation and decision". Beliefs that are spontaneous in this sense are also spoken of as "non-derivative" or "immediate". None of this implies, however, that the brain-processes underlying the formation of such beliefs lack all computational complexity.

The term 'judgment' also requires some comment. Some philosophers hold that judgment is not the name for a mental action, but for the *product* of such an action—the "product of pondering", as Ryle put it in *The Concept of Mind* (ref.). But why, apart perhaps from behaviorist motivations, try to legislate against the ordinary usage of the term for the pondering itself? Here and elsewhere in the paper, I am understanding a judgment to be a distinguished kind of mental action, involving essentially an *activity* on the part of the self, akin to affirming, asserting, endorsing, etc.

successfully manage to ignore the content of the cell phone conversation of man sitting just ahead of me on the bus, and perhaps even to *pretend* that he is speaking in an unintelligible tongue, I do not thereby come to *believe* that he is fact silent or indecipherable. Queried as to whether I sat behind a noisy man speaking clear English while on the bus, I will have to confess that I was, whether I liked it or not. We should not, however, overlook the fact that we often have some degree of voluntary control over the processes that lead to belief. Scientific and philosophical belief often exhibit this. One cannot believe a theory at will, but one can, even on the strength of various non-epistemic attractions it holds, decide to accept it as a working hypothesis, and thereby dedicate oneself to finding evidence or argument in support of it, and looking for ways to rebut claims against it. And all of this *will* shape what one comes to believe.

The worries about regress and the conditions for learning we rehearsed above suggest that it would be unwise to say that either deliberation or judgment is necessary if a belief is to count as being held *for a reason* in the normative sense. To respond aptly to some reasons—even the complex reasons of extended inference—some measure of default, defeasible trust is necessary, and this attitude does indeed make possible spontaneous belief. If I have a good measure of default trust my eyes, or memory, or wits, then I will tend to believe what they yield, and this does seem to be the normal state of affairs for most believers most of the time. The overwhelming proportion of belief as a result is spontaneous—not because we never ask for or seek reasons, but because we cannot do either of these effectively without presupposing a good deal that isn't currently in question. Yesterday's conclusion, derived after long and thoughtful reflection, is today's memory. If I had to deliberate equally long and hard before taking that memory at face value, or relying upon it without further ado in today's experience and reasoning, yesterday's work would be very much less useful to me and building up epistemic capital would prove elusive. Similarly for coming to trust an external source of information. If I have acquired implicit trust in you, or the dictionary, or NPR's daily report of the Dow Jones Industrial Index, I will tend to believe the information provided directly, with no need for validation on my part. That is not mere laziness or arationality. I might have very good reason for such trust, and, if nothing seems amiss in the information provided, continually questioning it in the name of epistemic autonomy would handicap me greatly as an inquirer—and, in effect, attribute more authority to my own experience or thought than I think it deserves.

We have thus far been discussing the role and function of spontaneous belief—but how could such a state realized in actual psyches? The relevant psychic state must, as we have been seeing, have a quite special cluster of features. It cannot be strictly voluntary, yet it also cannot simply be a passive or "inert representation". Belief is *action-orienting*, in that it involves expectation and potentiates inference and action. Moreover, belief is *dynamic*, in that it is responsive to experience and argument in its own right, without needing to call upon higher-order deliberation or the deployment of normative concepts—consider, for example, the rapidity with which very young children learn. Belief sets the stage for its own growth and revision, that is, it sets the stage for learning. At the same time, belief is not wholly "automatic", since it is *judgment-sensitive*—we can, by assembling evidence or constructing arguments, holding ourselves to standards, and deliberating our way through to conclusions, change what we believe, or the degree to which we believe it. Belief must be such that reflection and reasoning can, at least when conditions are right, *entrain* belief or *transfer* it from one thought to another. Normative epistemology does seem possible, after all. It is notable, however, that various kinds of reasoning—purely pragmatic reasoning, for example—do not seem capable of inducing belief on their own. Something about belief makes it resistant to this kind of instrumental guidance, and renders it selectively sensitive—brainwashing and drugs apart—to certain kinds of direct influence but not others.

If belief sets us up for learning, what does learning set us up for? Humans surely have been known to seek knowledge for its own sake, but this admirably disinterested interest is hardly enough to explain belief's centrality and governing role in our mental architecture. As William James emphasized, and as Stalnaker's functional characterization makes clear, believing is for doing. Someone who believes that *p*—expressly or tacitly—with some measure of firmness will tend to *rely upon p*, acting as if "in a world in which *p* (together with [her] other beliefs) were true".

This 'as if', and the subjunctive mode, should not convey the idea that belief is a mere pretense or half-hearted acceptance. Belief is different from "as if" acceptance, even when "as if" acceptance accompanied by the normative view that such acceptance is warranted. Someone who believes that *p* in some measure takes *p* to heart—*credence* comes from *credo*, and both derive from a word for the heart. If I genuinely and firmly believe a branch is sturdy—perhaps I have climbed out on it before, or seen others do so—I will confidently shinny out on it to retrieve a child's lodged kite. That confidence is likely to be lacking if I simply have judged that my evidence warrants taking the limb as strong enough. Or consider Joan, who judges that the thing to do is to accept her foreman's explanation of why she cannot yet be moved up from the shop floor—"not enough experience"—even though she feels that she deserved the promotion. After all, she tells herself, the foreman seems like a decent fellow, she has no hard evidence to suggest any bias against her, and the man who was promoted is a good worker with more experience. She doesn't want to imagine things or malign her co-workers. So she decides that, on the basis of her evidence, she is warranted in treating the claim that no discrimination was involved as true. But does Joan therefore take the foreman's story to heart—does she truly *believe* it? We don't yet know. Acceptance does not automatically entrain all of the elements of belief, and Joan's involuntary thoughts, feelings, and implicit behaviors might well not be those of a true believer. If, next time, she is passed over for a man with *less* experience, she will be outraged, but might also feel that part of her expected this all along.

We are constantly going out on limbs on the strength of what we believe, counting on our beliefs in the face of uncertainty, and being surprised, or worse, if things do not go as expected. When someone believes that *p*, it enters intimately into her thoughts—what she notices, infers, anticipates, feels, dreams, and wonders or worries about. She will reason directly from *p*, and, other things equal, go out on a social limb for it as well, being disposed to avow it to others in frank conversation, thereby encouraging them to rely upon it, too.

Of course, one can also believe that *p* without this sort of awareness. As Freudians have long maintained, and cognitive psychologists have recently begun to think, our beliefs extend well beyond what we consciously acknowledge, and perhaps beyond what we could call into consciousness through ordinary introspection. These subterranean credences nonetheless continue to do most of the normal work of conscious belief in shaping how we think, act, and feel. Indeed, given how much the perceptual information available to us outstrips our capacity to attend or notice, we had better hope that it is not all lost on us, that much of this information is being encoded at a level below awareness. This is also the secret of such tricks as subliminal advertising. Nor is tacit belief acquisition confined to the perceptual realm. Many of our most fundamental presuppositions, convictions, and prejudices are largely tacit—the result of unspoken messages picked up over the course of our personal, familial, and social lives. And the influence of such beliefs can be pervasive—shaping what seems to us natural or unnatural, normal or abnormal, comfortable or threatening, thinkable or unthinkable, "us" or "them"—even though, were they made explicit, these beliefs might be vigorously rejected by us. Here, then, is another feature of belief—in origin, persistence, and operation, it can be a more or less completely tacit attitude.

Classically, it is conceived of as a *propositional attitude*—schematically, belief that *p* is not seen as a mental state with the structure **[representation that *p*]** but rather, **attitude [representation that *p*]**. I will often speak of belief this way in what follows, though the account offered below will be latitudinarian on the much-vexed question whether other sorts of representations can figure as the complement of belief or other "propositional attitudes". (For reasons that will soon emerge, it seems to me important to leave room for the possibility that the belief attitude can take a variety of objects.) Now on the standard view, belief is one of the *fundamental* propositional attitudes. It is distinguished from other propositional attitudes—such as supposing, imaging, accepting, pretending, hoping, desiring, and so on—not by its representational content, which can be shared with them, but by the way in which this attitude affects how we will go on to think and act with respect to that content. To use the language discussed above, if one merely pretends or imagines that *p*, for example, one typically will not by default rely upon *p*—one will not simply *expect* the world to conform to it, and or stake one's safety on it.

Talk of a *representation* as the immediate object of a belief should not, however, mislead us. Beliefs that *p* is not *about* the mental representation of *p*, but about what that representation presents, *p*, the *intentional object* of belief. One expects that *p*. This said, however, it is also distinctive of belief that it involves expecting a state of affairs *under a mode of presentation* or *idea* of it. Were that self-same state of affairs to come to consciousness under another guise, one might not even notice that one's expectation had been fulfilled. Lois Lane has a well-developed idea of Clark Kent as a klutzy reporter who seems never to be around when you most need him, and an almost equally well-developed idea of Superman as a hero who carries out thrilling exploits and always comes to the rescue in the nick of time. If we wish to understand her expectations or actions, we will need to attend to the difference between these two guises even though, as we know and she does not, that they pick out but one individual. These ideas *frame* her view of things—they give her practical deliberation a representational content to work with, prime certain inferences and memories rather than others, carry distinctive associations, and orient her in the world. She thinks and acts *through* them. She thus would be very surprised if Clark Kent leapt over a tall building in a single bound, and certainly would not count on him to deflect speeding bullets. But this is not because she would be surprised if an idea in her mind could leap so high or stop a bullet, although these things would indeed surprise her. Rather, she would be surprised because she represents the object of her Clark Kent thoughts Clark-Kentishly—mild-mannered, weak, and hardly bullet-proof. If such an individual were to come between her and hail of bullets, and escape unscathed, she would find this miraculous—or suspicious.

The beginnings of belief

Our question then has then become, if belief is a contentful attitude that sets us up for learning and frames our thinking and doing in these ways, what sort of attitude might it be? It would have to be an attitude capable of eliciting and guiding in a coordinated way a large number of mental phenomena—attention, expectation, association, inference, memory, surprise, assurance, error detection, action-planning, and action-monitoring, among others. Moreover, it would have to be able to orchestrate such effects even when tacit or unconscious, without the benefit of explicit oversight, guidance, deliberation, and choice. It would have to be inherently sensitive to evidence and argument, but resistant to flat-out decision. And, as Hume pointed out, it could not add any "new idea" to the representation that gives its content[17]—for the truth conditions of a belief are precisely the satisfaction conditions of its content, taken at face value.

[17] *Treatise*, ref.

We might well despair of finding any simple attitude to fill this complex and demanding job description. And I will in fact be arguing that no *simple* attitude can fill this bill, though a certain, structured *compound* of two simple attitudes we have already discussed can.

The first of these two attitudes, the "front end" of belief, is an attitude capable of directly inducing default reliance on its object: trust. Trust in the sense that I have been using here, is a species of *affect*, widely viewed as fundamental in human psychology—indeed, in mammalian psychology more generally. Why focus on affect? It is characteristic of affective states that they can arise directly ("spontaneously") from perceptual experience without need for judgments. Moreover, affective states are also distinguished functionally by the central organizing and mobilizing role they play in our psyches. Fear is a paradigm example. Suppose that I am shinnying out on the familiar branch I confidently believed to be sturdy, but suddenly there is a loud crack and the branch drops slightly beneath me. Within a few hundred milliseconds, even before I am able to say what is happening, the sensation of the loud sound and sudden sag will have activated my limbic system and begun to set in motion a complex suite of mental and physical responses. Directly, my sympathetic nervous system will spring into action, releasing stress hormones, accelerating my heartbeat and breathing, stepping up my metabolism, slowing down my digestion, and increasing bloodflow to the brain and muscles. My attention will be heightened and drawn from the kite to my own immediate situation, and relevant memories and motor responses primed. In this way, fear presents my new situation "evaluatively" in the sense of "at risk" and "calling for a response", and orchestrates my reaction accordingly. It *attunes* me to the unexpectedly heightened risk in my environment by *retuning* my manner of relating to it, from benign business as usual to code red. And with good reason.

Trust, unlike fear, is a *default* state, and so not typically state of arousal. Rather than put us on our guard against what we see, trust induces us to *rely upon* it, whether it be a situation or a person. Thus do infants come to trust their caregivers. That which we implicitly or explicitly trust we take at face value, which directly induces an expectation that it will be as it appears to be. As in the case of fear, there is no need for any intermediating thought or judgment in order for trust to affect such phenomena as attention, cognition, memory, motivation, feeling, even physiology, in a coordinated way. When one perceives that an oncoming figure in a dark alley is in fact a trusted friend, one's heart function, metabolism, and vigilance will more or less rapidly return to normal, tensed muscles will relax, and a narrowed gaze will soften. Here is a changed "evaluative response" from fear that calls for avoidance or defensive action to trust that calls for approach and acceptance, as one greets one's friend warmly. The greater the trust, or the more acute the antecedent fear, the more palpable this sense of assurance will be.

In the ordinary case, however, trust is not very palpable. For example, we implicitly trust our own eyes, and follow their lead. If they present me with a clear impression of hallway when the hotel elevator door opens I will "non-inferentially" take this to be the case and act accordingly, immediately stepping forward. The timing, direction, and shape of my behavior will all directly reflect the content of my visual representation of the scene, as I act with assurance *through* this representation to exit the elevator car. I would be surprised indeed were I suddenly to encounter an obstacle to my movement where my eyes see none, or were my foot to discover, not firm floor, but soft oose. From assured approach and undistracted thought I would shift in a trice to avoidance and bewilderment, jumping back, alarmed. By contrast, if I *distrust* a memory, or do not trust my ability to do long division in my head, then, if anything hangs on it, I won't count on this memory or quotient without seeking further confirmation. I won't be immediately and "automatically" oriented toward the world through these distrusted representations, and if they turn out to be wrong, I will not be surprised or puzzled.

As a form of *positive* or *approach* affect, trust is a way of taking to heart. Like belief, it is distinct from simply acting "as if", or from the sort of cognitive acceptance that can be an object of decision. Recall the case of Joan, who, it seems, had somehow come implicitly to distrust the foreman, even if she could identify no corroborating evidence. Perhaps trust induces reliance and taking to heart, without need for an intermediating judgment, because trust itself *is* a judgment, say, of trustworthiness? If this were so, trust would not be properly psychologically basic, and it would appear to depend upon belief rather than being a basic component of it. Trust *does* sometimes result from judgment, yet it seems clear that trust can arise, be sustained, or be lost in the absence of any such judgment, explicit or implicit. A crawling child, who surely lacks the concept of *trustworthiness* (and perhaps also the "faculty of judgment"), can come through experience to trust the gentle old family dog whose presence at first caused him distress.[18] Over time, he has learned to approach the dog without fear, and now

[18] Indeed, animal psychologists attribute trust even to non-human primates and various other mammals, including dogs. This attribution has more recently been buttressed by neurological evidence of similarity in the brain structures, neurotransmitters, and hormones that mediate trust—as well as various other species of affect, such as fear and pleasure—across a wide range of mammals, ourselves included.

calmly cuddles against it while it sleeps. He will be greatly surprised and alarmed if the dog ever suddenly turns on him and snaps.

Just as trusting can exist in the absence of a judgment of trustworthiness, a judgment of trustworthiness can fail to yield trust. Asked whether I deem my employee Chuck trustworthy, I can reflect on the evidence of two years of working with him, with nary a problem, and judge, "Yes, he's entirely trustworthy". I will write this in a letter of recommendation—he has recently informed me that will be applying for another job—and sign it without reservation. Surely, then, having made up my mind on the strength of so much evidence, I trust Chuck? Yet why, when I learn just a day after writing my letter that Chuck has been pilfering cash during the last few weeks, do I find that I am not really surprised? Unbeknownst to me, I had recently been picking up some subliminal or barely perceptible cues in Chuck's behavior and facial expressions—a certain change in his posture or furtiveness in his movements, a split-second facial grimace or instantaneous deflection of his eyes while speaking with him—that betrayed his deceit.[19] Surprised that I was not surprised, I reflect a bit and realize that only last week I had, on the spur of the moment and, it seemed at the time, for no particular reason, given a somewhat delicate task that would normally have been Chuck's responsibility to another employee.[20] Without realizing it, I had recently begun to hedge my bets on Chuck. Despite my judgment, it seems he no longer really had my trust. A neuropsychologist looking into my brain would see that Chuck's presence, and even the thought of Chuck, trigger forms of negative affect associated with distrust, running in parallel with whatever activity is going on in my higher cortices when I produced the

[19] See Ekman.
[20] See Damasio.

letter of recommendation. "This man *thinks* he trusts Chuck, but his brain has been taking in evidence for the last few weeks and knows better," she might say. There is no paradox here, or self-deception on my part. My conscious deliberation simply had access to a narrower and less subtle body of evidence than my feelings, and I was faultlessly unaware of this until I noticed a conscious clue of their presence, namely, my absence of surprise. Had the pilfering and furtiveness gone on for longer, my nonconscious unease might eventually have surfaced as a "hunch" that something was up with Chuck, and moved me to look into inventory or check out the security films, and thereby discover the thieving.[21] This degree of independence of trust and distrust from "top down" regulation by judgment thus is an important asset in learning—it permits "bottom up" seepage of evidence into our doxastic dispositions, and thus permitting us to acquire knowledge of things our conscious judgment had mistakenly ruled out, or not yet noticed.

To make a judgment is a mental action, and is voluntary in the sense of being a matter of *decision*. But trust is neither of these things, and neither is belief. Someone who is inclined to say that I didn't *really* deem Chuck to be trustworthy because my heart wasn't really in it, or that, when it wrote, "without reservation" in my letter for him I didn't really believe it because, at some level, my mind harbored distrust, is in effect conceding the point that belief is not a "pure cognition" transparent to thought, or an "inert representation" free of affect. To issue in fully-fledged belief, a judgment must engage the psyche more than cognitively—it must entrain his confidence as well. As Hume emphasized in discussing the lines of reasoning that he thought lead

[21] Damasio, ref.

inescapably to skepticism, a natural resistance on the part of belief to purely cognitive rulings from on high is an important defense mechanism of our experience-steeped epistemic nature against abstract, speculative reasoning.[22]

We can also see the distinction between trust as a basic attitude, on the one hand, and judgment of trustworthiness, on the other, by examining trust from the psychological side. There it is seen as an elemental psychic kind, one of the essential building blocks of psychological development, sociability, and a sense of self-confidence, identity, and place. Developmental psychologists have argued that the early establishment and maintenance of infant-parent trust are important for the child's cognitive and emotional development across a wide range of domains.[23] Failure to "bond" and establish such trust can result in the child experiencing chronic stress or anxiety, slower learning, difficulty in forming relationships with others, less openness to novelty, and a "cognitive style" in which events tend to be given a negative interpretation.[24]

Mechanisms of default trust that are so automatic for most of us can also be impaired by traumatic experiences later in life, making for a syndrome in which highly stressful, fearful responses to certain kinds of situations cannot be "unlearned" even when those situations no longer pose any danger.[25] Neuropsychologists have made some progress identifying the brain systems that subserve trust and distrust, the hormones and neurotransmitters that tend to increase or decrease them, and the ways

[22] See *Treatise*, ref. "too important to trust to uncertain workings"

[23] Erickson, E. *Childhood and Society*. New York: Norton, 1963. For a recent review of the literature on interpersonal trust, see (Simpson, "Foundations of Interpersonal Trust", in A.W. Kruglanski and E.T. Higgins (eds.), *Social Psychology: Handbook of Basic Principles*, 2nd ed. (New York: Guilford, 2007), pp. 587-607.

[24] Ref

[25] Ref

in which various deficiencies, lesions, or genetic disorders may yield dysfunctions of these systems, producing a susceptibility to mood disorders, social and behavioral difficulties, and lessened ability regulate behavior planfully. The capacity for trust is so basic and central psychically that breakdown of its normal operation can be devastating.

Although trust, as a fundamental psychic kind, is a form of affect, it need not be experienced as a conscious feeling. Indeed, as the default attitude of a healthy psyche toward its own faculties, thoughts, and experiences, trust is an "unmarked" condition, and should be expected to be largely implicit, and to have a relatively low phenomenological profile even when explicit. Instead, we should expect that departures from this default state will be marked by more pronounced and distinctive "feelings" and patterns of arousal which tend to alert the individual, and shift attention, thought, and motivation from their ordinary operation and into a different mode.[26] Thus *distrust*, *doubt*, *fear*, *surprise*, *anxiety*, and *disappointment* are "marked" states, which, when conscious, all have much more distinctive phenomenologies than trust. And even when implicit, they are associated with states of arousal that tend to disrupt business as usual and bring on a more vigilant, guarded, or defensive mode of operation.

Although trust's phenomenology should be expected to be thin, we should not confuse this with a state of *zero affect*. Individuals with little or no positive affect or ability to generate such affect, e.g., those who are chronically depressed or suffer some other deprivation of the neurotransmitters or hormones associated with trust, find trust and self-trust hard to come by, and as a result suffer declines in learning, memory,

[26] See the evidence on emotional tuning.

complex cognition, decision-making, planning, social relationships, and motivation.[27] Default trust, as a floor of confidence in oneself, an openness to others and to new experiences, and a capacity for learning and making stable decisions or judgments, gives to life a definite cast, distinctly unlike the "lived world" of those lacking in basic trust, anxious, or chronically indecisive. Moreover, because trust is a form of positive affect associated with approach and acceptance, systematic attenuation of positive affect tends to result in distinctively different patterns of behavior, marked by social, intellectual, and emotional withdrawal. The usual, easy flow of new information into fresh belief and action functions badly, and thought tends to become repetitive, devoid of positive expectation or hope, and circular.[28]

Modeling belief

Trust, then, a species of positive affect, will be our candidate for the lead role in the attitude of belief, essential to generating the many functional and qualitative differences between belief and other attitudes such as supposing, imagining, pretending, or accepting,   As Hume put it:

> … belief consists not in the nature and order of our ideas, but in the manner of their conception, and in their feeling to the mind.

Hume, however, despaired of further explanation:

> I confess, that 'tis impossible to explain perfectly this feeling or manner of conception. …. [I]n philosophy we can go no farther, than assert, that it is

[27] Diagnostic and Statistical Manual of Mental Disorders

[28] Again, see Damasio.

> something *felt* by the mind, which distinguishes the ideas of judgment from the fictions of the imagination. It gives them more force and influence; makes them appear of greater importance; infixes them in the mind; and renders them the governing principles of all our actions. [*Treatise*, Appendix, 629][29]

Can we, with the help of trust, do better?

The first question to address is whether talk of trust can shed any light on belief, if 'trust' and 'belief' are too nearly synonymous. This I doubt. For example, it may be a conceptual truth that trust is a form of affect or feeling, but the same is hardly true of belief. I would cite as my evidence the many distinguished philosophers who will contest this claim, and reject out of hand any account of the kind being offered here by insisting upon a sharp contrast between cognitive and affective states.[30] They are wrong, I believe, but they are not misusing English.

The first thing to notice is that trust is a more elementary attitude than belief as it is ordinarily understood. Trust can exist in mental systems sufficiently rich to have mental representations, but insufficiently complex or developed for these to represent propositions. And the mechanisms by which a very young infant comes to trust the family dog might not differ much at all from those by which a puppy brought newly into the house might do so. Trust, and acquired trust, seem to be in place in a very wide range of cases in which questions of *reliability* can arise. Thus, while one can of course both trust that *p*, one can also trust one's eyes, memory, intuitions, best friend, footing,

[29] Note that here Hume is using 'judgment' not in the sense of an act of reasoning—which he famously held could not in itself produce belief—but in the sense of a non-voluntary, firm "perception". [*T* 456, 609]

[30] This, for example, is often said by neo-Humeans, though, by a nice irony, Hume himself seems to have considered it one of his chief and most controversial discoveries that, as against the rationalists, belief *is* a form of feeling. (More on this, below.) For a characterization of the neo-Humean conception of belief, see Smith (1994).

aim, or swimming ability, or a sign, map, or recipe. This polyvalence is important for the generation of belief because we must be able to make immediate mental transitions from trusting one's eyes to believing that *p*, where *p* is how one's eyes present things as being; or from trusting a person or source of information to believing that *q*, where *q* is how this person or source says things are; or from trusting a method to believing the results of following that method. Trust by its nature tends through *affective association* and *affective transfer* to transmit confidence immediately from an object of trust to what that object says, shows, or leads to.

This feature of trust is essential to the explanation of how belief can get going without already requiring belief, deliberation, or judgment. If—as infant or adult—one has default trust in one's senses, this will tend directly to induce reliance in thought and action upon the sensory representations they furnish. Similarly default trust in one's memory or reasoning capacity will tend directly to induce reliance upon the operation of these faculties and the representations they furnish, making possible transfer of confidence along a thread of memory or chain of thought without the intermediation of judgment.

Humans do so much not *transcend* the general-purpose mechanisms of trust and affective transfer, as *redeploy* them in a mental economy sufficiently complex to permit the representation of concepts and propositions, and, very importantly, the representation—meta-representation, really—of the operation of trust, affective transfer, and like attitudes and processes in the creation and revision of belief. This redeployment helps account for the remarkable *productivity* of belief—humans are capable of forming and relying upon beliefs about situations they have never experienced, as well as highly general, abstract, modal, or hypothetical representations. This permits human cognition to expand far beyond the realm of native instinct and learning by conditioning and association. Consider what we can think of as "top-down" trust formation. If I implicitly trust you, then your remark that a map is *trustworthy*, along with my default trust in my senses and my understanding of language, can induce in me a readiness to rely upon the map without the need for an intermediating judgment on my part. The common currency of the cognitive economy remains trust, and the baseline condition default trust, for without these, belief will not follow where experience, inference, or memory would take it.

But still, one might wonder, why think that *feeling* is involved in belief? To be sure, we do speak of beliefs as more or less *confident*, *certain*, or *unsure*, and these in ordinary talk are species of feeling. Perhaps, however, we should see this as a metaphorical overlay, not relevant to the nature or function of belief itself. We cannot of course think of beliefs as *mere* representations, for this would not distinguish beliefs from idle thoughts. We have seen that we must talk of an *attitude* of belief, capable of greater or lesser strength. But perhaps it is enough that this attitude be no more than a probability or expectation value, as might be expressed in a betting quotient.

However, there might be important aspects of the ways in which we work or reason with belief that suggest that more than a simple probability or degree of expectation is involved in belief. Consider the difference between two stages in the evolution of a belief about the probability of obtaining *heads* on a single flip of a coin. In the first case, you have just given me a coin in an unfamiliar currency, and urged me to make a bet on the outcome of a single toss. The coin looks fairly ordinary to me, and you allow me to toss it a dozen or so times. It seems to come up *heads* and *tails* about equally and in no particular pattern. Since I see no reason to expect *heads* more than *tails*, my "degree of belief" that the next toss will land *heads* is one-half, and I will elect even odds on the bet. But I will be reluctant to bet much, since I won't be much inclined to rely upon this expectation. Nor will this belief about the chance of *heads* be very firm or resilient in the face of new evidence—if, after the first bet, the next dozen tosses of the coin yield a high frequency of *tails*, my expectation of *heads* on the next toss will be adjusted downward. In the second case, I have been fooling around with this selfsame coin for quite some time, tossing it many times and keeping careful track of the results. The ratio of *heads* to *tails*, I observe, is almost 50:50. When you now ask me the strength of my belief that the coin will land *heads* on the next toss, I will again say one-half and elect even odds. But this time I will be much more willing to rely upon this belief, and so prepared to bet more. Further, I will also be much more reluctant to modify this "degree of belief" downwards even after the next dozen tosses yields a high frequency of *tails.* What has changed? Something about the *strength* of my belief concerning *heads* on the next outcome, to be sure, but not as measured by what probability I attribute to *heads* or what odds I'd take. It seems most natural to say that this changed strength in belief is a matter of how *confident* it is, how much *trust* it involves, and that this accounts both for my greater willingness to rely upon it in taking on risk, and my lesser inclination to modify it significantly in response to a particular bit of new frequentist evidence. We do seem quite naturally to understand these important features of belief, its relation to action, and its responsiveness to new evidence, in terms of attitudes such as confidence, trust, willingness to rely, inclination to revise, and so on.

Similar language figures prominently in another important way in which we work or reason with our beliefs: the process of reflective equilibrium. Here is Nelson Goodman's initial characterization of the process as applied to rules of inference:

> *A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend.*[31]

More generally, reflective equilibrium is seen as a process in which we seek a balance between the generalizations or principles we find most compelling, and the particular cases in which we are most confident, such that we are, as Norman Daniels puts it, "un-inclined to revise any further because together they have the highest degree of acceptability or credibility for us."[32] 'Credibility' here is clearly not simply a matter of finding a collection of beliefs that assign coherent or high expectation values to outcomes—it is a matter of finding an overall balance in which we have greatest confidence or trust. Reflective equilibrium, when it comes to credibility or confidence, is affective equilibrium. Moreover, it is important for such reflective processes that the dimension of strength of belief corresponding to confidence *inhere* in particular beliefs themselves, rather than being entirely a matter of what we believe *about* those beliefs—we must have some initial weights with which to start. After all, if deliberative weights were not part of beliefs as such, then we would also have to ask what we believe about our beliefs about our beliefs, and so on, without ever coming to beliefs that themselves carried the weight of confidence or conviction.

[31] N. Goodman, *Fact, Fiction, and Forecast*, 3rd ed. (Cambridge: Harvard University Press), p. 64.
[32] N. Daniels, "Reflective Equilibrium", *Stanford Encyclopedia of Philosophy*.

One might of course doubt the appropriateness of bringing talk of "feeling" into belief on logical grounds. After all, belief, but not feelings, is capable of truth or falsity, and with this is associated the possibility of evaluating belief as more or less rationally justified. But here it is important to note that truth or falsity attaches in the first instance to the *objects* of the belief attitude, and that there are important features of rational assessment of belief that depend not simply upon content but upon *strength*. We cannot tell simply from the truth or falsity of the object of a belief whether that belief is held rationally. To know this we must know whether there is evidence for given belief, and, if so, whether the strength of the belief—whether true or false—is proportional to our evidence. And there is nothing amiss in speaking of feelings such as confidence or fear as more or less based upon, or proportional to, or *warranted by*, evidence. Had I, at the beginning of our coin-tossing example, been entirely *certain* that the odds of *heads* is one-half, then even though the object of my certainty indeed turned out true, as later evidence would show, my doxastic attitude toward it would not have been warranted or rational.

Understanding the function of belief in thought and action thus requires understanding how beliefs can possess *weight* as well as content. Moreover, this weight must be of a kind to mobilize directly the relevant faculties enabling belief to play its Stalnakerian role. Human mental architecture appears to have been able to draw upon an inherited system of experience-shaped trusting affect to give belief the just the sort of heft, efficacy, and responsiveness it needs. Moreover, such affect can, in humans at least, be consciously experienced. And this makes it possible for us to guide our deliberation by the strength as well as content of our beliefs. Thanks to a shared currency of trust—which can exist not only in propositional contents but in our senses or faculties, or in our sources or methods—thoughts acquired through the various sensory modalities, or retrieved from memory, or received from others, or resulting from reflective equilibrium, can acquire the status of beliefs, and possess varying degrees of strength. Such weighted thoughts in turn will tend to exert a direct force in shaping subsequent expectation, inference, deliberation, and action. This extension of the animal system, which presumably co-evolved with language, enabled trust to confer its affective potency or "ertness"—its capacity to induce reliance and expectation, to influence attention, etc.—on mental representations of spatially and temporally remote states of affairs, or mere possibilities, or abstracta. Humans thereby acquired unprecedented power to counteract the power of the actual, proximate, and perceptually salient features of their environment by making inferences, imagining alternatives, developing systematic ways of recording the past, forming and testing generalizations and explanatory theories, and laying out conditional plans and strategies.

This enlarged not only our ability to think—why would evolution care what we merely think?—but our ability to *do*: to adapt to new circumstances, to make provision for the future and act strategically, to augment our personal, local experience by pooling information with others, to develop new technologies, and to experiment with new forms of individual and social action that enabled us to go out on limbs no one had ever climbed before, and reach fruits—and poisons—that otherwise would have been inaccessible.

A model

As a first approximation, then, belief has the following functional character:

> **Belief that *p*** (first version)**:**
>
> A degree of trust in a representation, *p*, gives rise to, and regulates, a degree of expectation that things are or will be as *p* portrays them.[33]

This of course is a simplification. As we have already noted, trust in a representation typically gives rise to, and regulates, much more than expectation—attention, association, inference, memory, feeling, and so on. But expectation seems to be of the essence in belief, since it is this forward-looking projection that one relies upon most directly when translating thought into action. In our simplified model, then, we will let expectation stand for all the rest.

If this first model exposed the full "essential" architecture of belief, however, belief would be a dangerous state to install as "the governing principle of all of our actions". For this architecture is all "feedforward" and no "feedback", all default reliance, with no defeasibility. It would keep me calmly crawling out on the limb I initially trusted to be safe, however disquietingly it bent or groaned under my weight, doing nothing to revise this confident belief even after the branch began to fail. Trust and belief, we know, are not like that. Rather, they tend to respond directly and "automatically" to on-going experience by wavering or failing as untoward and unexpected events arise, or by remaining solid or growing in strength as events bear our

[33] Note that 'trust in a representation' is to be understood as explained above. Thus *trust in* a representation that *p* is *trust that p*. Expectation is here understood as it is by psychologists, an elementary state found in humans and animals alike. Why is the expectation included in the model of belief, if the attitude of trust is sufficient to produce it? The answer is that trust will sometimes fail to produce corresponding expectation, and, when it does, the attitude seems to be defective as a belief. We will discuss this below. Note, too, that trust *mediated* by a representation is nonetheless trust in the *intentional object* of the representation.

expectations out. This is, as we noted, a characteristic feature of affective states such as trust or fear, and not something we ordinarily must *do to* these states, or to our beliefs—the way we must balance our checkbooks, floss our teeth, or fold the laundry. Let us therefore make this explicitly part of our model of the essential architecture of belief (see also fig. 1, below):

> **Belief that *p*** (final version)**:**
>
> A degree of trust in a representation, *p*, gives rise to, and regulates, a degree of expectation that things are or will be as *p* portrays them; and this degree of trust is modulated by whether, and to what extent or in what direction, these expectations are met or violated.

"Aiming at truth"

Without modulation by feedback, belief would be like an archer who aims and shoots, but then shows no interest in where the shaft lands, so that even if it falls wide of the target, she pays no mind and holds the bow in just the same way the next time. After a while, we'd begin to question whether she's really trying to hit the target after all. Belief, we are told, by its nature "aims at truth". If so, it seems, it had better by its nature pay attention to whether the arrow of expectation hits its target.

Sometimes such "aiming at truth" is spoken of as a fact about *believers*—about the norms by which they regulate their attitudes. Just as the archer had to hold her aiming accountable to hits or misses in order to count as aiming at the target, an agent's propositional attitude will not count as a belief unless he holds it to responsible to evidence of truth or falsity. An archer who pays no attention to where her arrows land is credited at most with *pretending* to aim at the target, and an agent who pays no attention to whether an attitude is borne out by evidence is credited at most with supposition, imagining, or pretense.

This seems plausible, but it cannot be the full story. For it does not explain how the belief attitude *by its nature* aims at truth—instead it delegates this task to the believer. As a result, it makes belief into a complicated and demanding enterprise. To believe one would need not only representations, but meta-representations, containing such normative concepts as *evidence*, *relevance*, and *support*. This would appear to preclude belief for the very young or conceptually naïve. And each new belief would impose a fresh burden upon the believer, requiring her to allocate scarce higher-order mental resources to the task of keeping it accountable to on-going experience, much as parents have to allocate scarce parenting and familial resources to each additional child. In the case of parents, we know that there is as a consequence a rather low upper bound on the number of children they can manage to raise well at the same time. Something similar would have to be true of believers and beliefs.[34]

It seems to me that if the idea of that belief aims at truth is to be vindicated, the aiming must be *inherent* in the belief attitude itself. Only then would evidence-responsiveness seem genuinely *constitutive* of belief, and only then would the acquisition of large numbers of beliefs be manageable, or an asset. Aiming at truth must be part of

[34] Compare the criticism often made of subjective Bayesianism that its model of continuous updating of credences by conditionalization is highly unrealistic psychologically.

the "design specification" of belief itself, whether belief occurs in epistemic sticklers or slackers, in children or adults, consciously or unconsciously. David Velleman has written:

> What distinguishes belief from other states that take their propositional objects as true is that, unlike assumption or fantasy, belief tends to track what *is* true, when its regulatory mechanisms are functioning as designed. … Belief thus aims at truth in the same sense that the circulation aims to supply body tissues with nutrients and oxygen. Not just any movement of fluids within the body counts as the circulation, but only those movements which are under the control of mechanisms designed to direct them at supplying the tissues.[35]

Our revised model of belief exhibits just this sort of internal regulatory mechanism, which helps us to see why possessing the capacity to form beliefs might have been selected for. The mechanism is "autonomic" in the sense that it makes no demands upon higher-order cognition. But it is not "blind", in that it is sensitive to the testimony of one's experience regarding whether, or to what extent, one's expectations have been met. In this way, it is just like the experientially-sensitive regulation of circulation by our autonomic nervous system. Indeed, the two regulative systems are closely linked. Surprises directly "upregulate" our pulse even as they "downregulate" our confidence in the associated beliefs.

An engineer looking at the functional architecture of our model would see a *variable control circuit*, "designed" to attune feed-forward regulation of expectation to an independent environmental variable by means of feed-back to reduce the discrepancy

[35] Velleman (2000), p. 17.

between the predicted value and the observed value. A formal learning theorist would see an automatic learning mechanism of roughly Bayesian design.[36] And a neurologist would see a functional design characteristic of, for example, the layers of neurons responsible for dynamic "orientation tuning" through perceptual inputs in the visual cortex and for "retuning" our affective register perceptually to the relative amount of risk present in our environment.[37] This sort of dynamic, distributed regulative architecture is exactly what we need to find in beliefs themselves if we are to give a psychologically realistic account of how belief might, by its nature, aim at truth in the "design sense" suggested by Velleman.[38]

We receive through experience a vast amount of information, and continuously form and store a large number of new perceptual beliefs and associations, as well as updating existing beliefs. This occurs without special cognitive effort and without conscious notice, leaving us to keep our mind on things requiring more complex thought. Through these processes there is built up in our minds an elaborate distributed representation of the world and people around us. Thanks to the inherent dynamic in belief, this representation can undergo constant revision as new perceptual information is received and previously potentiated neural networks that embody the representation are selectively activated and altered. Just as one's computer works continually, even while one is pausing to think what to write, to update its file indices

[36] The similarity of the functional model of belief to Bayes' Rule will not have escaped many readers: our initial degree of trust that *H* is akin to the prior, p(*H*), the strength of expectation of the new evidence *E* supported by this prior is akin to the likelihood p(*E*|*H*), and the surprise value of the occurrence of *E* is akin to the denominator p(*E*). However, as we'll see below, there is a duality in familiar ways of talking about "degree" or "strength" of belief that suggests that simple Bayesianism needs supplementing.

[37] For a recent brief discussion of perceptual tuning, see Ringach, et al. (1997). For a discussion of "retuning" the "affective keyboard", see Berridge (ref.).

[38] This also supplies the Bayesian with a response to the criticism of unrealism—the perceptual system, for example, is remarkably Bayesian in its design and learning performance. For discussion of Bayesian models of human language learning and cognition, see *Trends in Cog Sci* Ref.

and effect linkages among newly obtained and already stored information, so that, for example, 'search' commands can work rapidly and efficiently, so does one's brain continually update its data-bases, with no need for higher-order thought or agential intervention—and even while one sleeps.[39] As a result one's many-dimensional representation of the world is largely up-to-date and ready to be recalled and relied upon it in thought and action without prior notice. New beliefs therefore are resources for conscious thought, not new burdens. Thanks to the long-term potentiation of neurons in learning, when trust makes us receptive, a given bit of visual input can lead almost instantly to messages being sent down tens of thousands of neural pathways into dozens of brain sub-systems. And thanks to the functional architecture of the brain, selective readjustments of attention, priming of associations and memories, and updating of underlying representations will ensue of its own accord, linking this new bit of experiential information with a vast network of information already acquired. The self-tending character of these processes makes handling such a volume of information possible, and frees conscious and higher-order thought to concern itself with what is less predictable, proximate, or local.

This sort of "bottom-up" learning from experience is well-studied in the case of animals. The fact that such learning is rapid, affectively-mediated, and not "mindful" should not lead us to think—as some have recently done—that this default system must crude, or "quick and dirty", or limited to a few statistically-inaccurate "heuristics". This distributed system is the core of animal learning, and it has been honed toward greater accuracy and efficiency through millions of generations. Not only can it take in, retain,

[39] Sleep and memory consolidation.

and make use of much more information than our more focused higher-order thought, but it can yield patterns of expectation and behavior that correspond quite closely to formal learning models and quantitative decision theory.[40] Clever experiments monitoring neuronal activity in the animal brain show how patterns of experience can lead to the formation of detailed internal maps of surroundings.[41] Feedforward-feedback control of on-going movements (e.g., reaching and grasping) in the cerebellum compares actual information from perception and proprioception to expected sensory states, and issues new instructions to motor areas in the cortex to adjust movement to correct errors (e.g., overshooting the sought-after object, or insufficiently opening the forepaw).[42] Recent work with macaques has yielded evidence of neuronal encoding of expected value and risk as acquired through learning, and the magnitudes of neuronal activity can be plotted closely against standard expected value and risk curves. Distinctive neuronal "error signals" are sent when expected values are not obtained or risk increases.[43]

Through such means, in more naturalistic settings, animal expectations can become so precisely attuned to the location, reliability, energy value, and energy cost of the food sources in their environment that their behavior very closely approximates the predictions of optimal foraging theory. [44] This is akin to solving a problem in multivariate linear programming of considerable computational complexity. It would be churlish to deny such highly refined forms of attunement to the environment and to the prospects of behaviors the status of "aiming at truth" in the *design* sense, just because

[40] Dugatkin, Bayesian.

[41]

[42] Jeuptner et al., 1996; Pockett 2006.

[43] There are several competing interpretations of these results. See Schultz (1999) and Berridge (ref).

[44] Dugatkin dd

the animals never think of it that way. After all, accuracy and reliability of representation—robust attunement—is a large part of what these systems were selected for.

Human evolution put this "aiming-at-truth" architecture it in the service of building and revising internal representations that take advantage of the vastly larger and more flexible expressive resources of language. Language permitted not only representation, but meta-representation, and making it possible to hold views about our relationship to the world as well as the world itself. Such views then became subject to learning through the feedback inherent in belief. Principles to govern such learning could also be formulated, and our successes or failures in relying upon those principles could bring us to have greater or lesser confidence in such norms. This meant that the trust needed to entrain credence could attach to self-conscious, norm-guided epistemic deliberation and judgment, which in turn could serve as an important counterbalance to bottom-up processes. We could expressly hold ourselves and one another to standards of consistency and warrant, and develop ways of tying hypotheses concerning the most far-flung, minute, and immense features of our universe, or the inner workings of our own psyches, to available proximate evidence.

Such top-down mechanisms, we argued earlier in the case of reflective equilibrium, also redeploy the common-currency role of affect and the basic system of trust and expectation. After all, we have no more direct access to the balance of reasons than through our credences themselves, the complex affective upshot of our long, shared history of trial-and-error engagement with the world. Merely mentally representing a rule, and even reasoning with it—as we do when we are hypothetically supposing a rule for the sake of seeing where it might lead—, are not enough to bring belief along with them. For although such reasoning is *rule-following*, we have, as yet, none of the confidence in the rule or in our application of it that is necessary to bring belief along with it. The genius of the "scientific method" has been to find ways of enforcing the basic experiential discipline of feedback inherent in belief with ever-increasing rigor and precision upon even our most ambitious thoughts and principles.

Affective primacy and rationality in affect

Our model of belief asks us to put a species of affect, trust, as primary in cognition. Is this psychologically realistic? Do we have evidence that the affective system comes into the cognitive process as early and systematically as this would require? The cognitive psychologist Jonathan Haidt writes:

> … social psychologists have increasing embraced a version of the "affective primacy" principle … [in light of] evidence that the human mind is composed of an ancient, automatic, and very fast affective system and a phylogenetically newer, slower, and motivationally weaker cognitive system. … [The] basic point was that brains are always and automatically evaluating everything they perceive, and that higher-level thinking is preceded, permeated, and influenced by affective reactions … . [Haidt 2007]

This is not an idiosyncratic view, but rather is an integral part of the large body of research, accumulated over recent decades, supporting "dual process" models of the mind (see fig. 4, below). It makes perfect evolutionary sense that the brain would not be designed simply to register experience passively. To be successful an organism must be able, at least in a preliminary way, to sort out continuously, and in response to changing needs and circumstances, what calls for attention, what resources and responses are to be readied, what new threats or opportunities may be on the horizon, what is worth remembering, and so on—all while carrying out the tasks at hand. These are evaluative, not merely recording, tasks, so a system is needed that evaluates incoming information immediately and effortlessly—"automatically"—without placing demands on higher-order cognition. Such evaluation must, if it is to have sufficient reliability, flexibility, and range, be responsive to evidence and capable of learning. It must also be capable of immediately orchestrating an array of mental and physical actions and reactions.

Core affective centers of the brain, with their direct and extensive inputs from perception, bodily condition, and memory, are capable of integrating information from a wide variety of physical and mental sources.[45] And with their direct and extensive projections to cognitive, somatic, and motivational areas of the brain, these affective areas are also capable of regulating a wide variety of mental, physical, and physiological responses.

But affect is more than a "nerve center" for the translation of information into action. In beings with consciousness, such as ourselves, it *presents* the world and our possibilities to act in various distinctive ways—frightening, reassuring, pleasant, attractive, disgusting, angering, interesting. We readily speak of such affective attitudes as fear, trust, anger, etc. as more or less warranted, and in this sense they can be

[45] Insula research. B. Craig, *Nat Rev Neuroscience* (2009) no. 1.

thought of as attuning us more or less reliably to reasons, and thus as part of our "emotional intelligence" and, ultimately, our rationality.

If affect is a form of evaluation, then what does *trust* evaluate? Fear, we say, evaluates *riskiness* of all kinds. As Aristotle pointed out, the virtue of courage is not fearlessness, for that is reckless, but fearing the right things to the right degree and being properly guided by this, rather than undone by it. Trust and distrust, we might say, evaluate reliability and unreliability of all kinds. The person with practical intelligence, in Aristotle's sense, would trust the right things to the right degree, distrust the right things to the right degree, and be properly guided as a result. For this to happen, trust and distrust, like fear, must be both good learners (to become properly attuned through experience) and effective leaders (to properly attune the individual's responses). In consequence, practically intelligent individuals will develop not just a physical map of their surroundings, but an *evaluative* map, with objects and options, people and places, coded for caution and reliance. It would be a map that presents to the agent prospects and possibilities. Not just the world as it exists or has been, but also of what to expect of it going forward in time, and what is risky or reliable. Just the sort of map an *agent* most needs. Think, for example, of what you'd most want to know going into a meeting crucial for your future with people unknown to you.

Function and dysfunction

A good model of belief should help us to understand key features of belief when everything is working normally, but also to understand some of its characteristic *dysfunctions* and *disorders* when this is not the case. And these distinctive forms of abnormality should tend to be viewed in practice as *irrationalities* of belief.

We have already discussed some basic features of belief: how belief differs from animal expectation in virtue of the involvement of representations with semantic content; how, thanks a redeployment of the "experience-tuned" animal architecture of trust and expectation, belief can have an inherent learning dynamic, so that, even though it is non-voluntary, it can be intelligent and capable of learning; how belief thus can by its nature "aim at truth"; how belief, as involving affect, can come in degrees and directly affect our dispositions to think and act, i.e., why belief is "taken to heart" in thought and action, and implicitly relied upon; how inference can bring credence along with it; and why the phenomenology of belief of belief is so thin, even though disbelief has a more distinctive "feel". Before we turn to ways in which the normal functioning of belief becomes disordered, it might be useful to identify one further key feature of belief that will help us to understand these dysfunctions.

Belief as represented in our model is *compound* in character, involving two coupled attitudes, trust and expectation, each capable of differences in degree. If this is a good model of belief, then there should be, in effect, two dimensions of "strength" in any one belief. In the normal case, it can be difficult to tease these two dimensions apart—a strong belief that *p* involves a high expectation—but as our earlier discussion of coin-tossing suggests, we can indeed locate such a distinction *within* a given belief attitude. Interestingly, neurological evidence now suggests that the brain encodes probability, utility, and risk separately—much as is done in so-called financial decision theory.[46] Since riskiness is a good reason to limit one's stake and perhaps also seek more information or counterbalancing risk, it is clear why both financial decision theory and the human brain give this dimension separate attention.

There are many more features of belief to be considered, but not now and not here. For we still need to ask, Can our model afford us any insight into the ways in which belief goes awry?

Our model makes belief out to be a compound state of a distinctive kind, incorporating a regulative or control structure. One sort of control is feed-forward—degree of trust tends to regulate degree of expectation. The other is feedback—discrepancies between expectation and outcome tend to modulate degree of trust.[47] Like most psychic processes, such regulation is a matter of dispositions that do not always yield precisely the same result—instead, they centered around an average or modal response, a "norm of response", corresponding to normal functioning. For simplicity, let's set aside cases in which the belief in question concerns a probabilistic process, such as coin-flipping. Then it seems appropriate to say that, other things equal, one's degree of expectation that *p* *should* reflect one's degree of confidence that *p*, and that, other things equal, one's degree of confidence that *p* *should* reflect one's past experience of discrepancies between expectation and outcome with respect to *p*. Put another way, the normal, "healthy" functioning of these regulative processes is part of

[46] Probability is at a maximal value when an outcome is most certain, while risk is at a maximum value when an outcome is most uncertain or shows the greatest variance. A course of action might have high expected value (= probability **x** utility) while at the same time having high risk. See P.N. Tobler, et al., "Reward Value Coding Distinct from Risk Attitude-Related Uncertainty Coding in Human Reward Systems", *J Neurophysiol* (2007) 97: 1621-1632 and S. Quartz, "Reason, Emotion, and Decision-Making: Risk and Reward Computation with Feeling", *Trends in Cog Sci* (2009) 13: 209-215.

[47] We have seen that in the case of chance outcomes, this simple description of feed-forward control does not do full justice to the complexity of belief. There is a corresponding simplification in our description of feed-back control—the brain systems involved in forward expectation can also show direct feedback modulation in response to predictive error.

what enables us to be *rational in belief*. It should follow, then, that when these regulative processes break down or do not function normally, there will be something anomalous, non-rational, or irrational about the beliefs in question. And there should be at least two kinds of cases: failures of forward regulation of expectation, and failures of feedback regulation of trust. Do we observe anything like this?

Let's consider forward regulation of expectation first. When might expectation escape the control of trust? Consider a phobia. Bob is now in his 40's. Until his mid-20's had no special fear of bridges. As a child, he even enjoyed playing on them—leaning over the railing to drop things into the water below, watching with fascination as they fell or floated downward. Starting in his 24th year, about the time his first child was born, Bob began to feel very fearful when crossing bridges, especially when they are high and across water. No particular traumatic experience was the cause—such phobias do sometimes spontaneously occur at about this point in life.

Bob is not, however, delusional. He's never had an actual problem on a bridge and he knows all the statistics—about which he harbors no secret theories of manipulation or cover-up. He is quite convinced that, mile for mile, crossing bridges is statistically as safe driving on solid ground, about which he has no special fear. Ask him whether to be wary of bridges and he will insist that you should not. Now, if this high level of trust in the idea that bridges are safe were functioning normally in Bob, then his expectations about crossing bridges would effortlessly be regulated by it. Indeed, at nearly every moment of his life when he is solid ground, that is precisely what happens. He does not expect to see stories in the newspaper about bridge collapses, and he's never worried about friends or loved ones crossing bridges. Bob plans family trips to take the shortest and safest routes, avoiding bridges only when all else is about equal, and then simply because he hates the feeling he has when crossing bridges and worries that it might impair his driving. Despite all this sane regulation of expectation, however, when Bob himself is crossing a bridge of any size, his expectations start to run away with him. He begins to sweat and his pulse rises sharply. He becomes hyper-vigilant and eager to be across. Time, however, seems to slow down inexorably. A sense that disaster is about to happen seeps into his mind despite his best efforts to focus exclusively on the task of driving.

If belief were simply a matter of expectation, then Bob would appear to be experiencing a sudden change in belief—or at least the sudden emergence of a salient and powerful new belief that the bridge *is* about to collapse, competing with his long-term conviction that he is at no special risk on bridges. And this change or conflict in belief would, just as suddenly, go away once he's back on solid ground. But that doesn't sound right. Bob will say, "Look, I have this irrational thing about crossing bridges—it's something I can't control and can't seem to get rid of. So when I'm actually on a bridge, I can't help but feel it's about to collapse. But I'm not so idiotic as to *believe*—even for a moment, even when these terrible thoughts and feelings are very strong—that the bridge really *is* about to collapse. If I believed *that*, I'd warn the passengers and start trying to turn around. Instead, I just keep driving, and try to keep my eye on the road and my feelings and fantasies to myself."

Bob, then, does not experience a change of mind or form a new belief. We can see this in the fact that he doesn't *rely upon* these intrusive fearful thoughts and feelings in what he does or says. Yet although Bob manages to control himself in these ways, he also experiences a loss of doxastic control over his own expectations and associations, an epistemic *incontinence* or *akrasia* that is *experienced as* a kind of irrationality.[48]

What makes Bob's self-control possible is not recklessness—he *would* be reckless to ignore an actual belief that the bridge is about to collapse—but his *distrust* of his own feeling of fear and the expectations to which it gives rise. Because he knows that the fear and dreadful expectations he experiences are entirely predictable whenever crossing a bridge, whether or not there is any risk, there is no surprise in feeling the fear or surprise when no collapse occurs. These experiences have no news value, and this "automatically" discounts the characteristic effects of such fear or expectation in shaping his thoughts and actions. In effect, the default trust we normally have in what we feel and in our own capacities has come through experience to be defeated in the case of Bob's phobia, so the fearful thoughts are not really accepted at face value and taken to heart. The fear and associated expectations have come to feel "alien" to him, even it they clearly come from within, and retain a dramatic capacity to affect him psychically and physiologically.[49] We need not posit a rational "inner Bob" who identifies with one set of facts or expectations rather than another in order to explain why this would be so. For Bob remains essentially of one mind—it is *his* unchanging beliefs about bridge safety that are failing to regulate expectation fully, even though he has managed to get enough psychic distance from his phobia to separate the

[48] Whether a given instance of epistemic incontinence or akrasia genuinely is irrational, or rather is a form of "bottom-up" learning, it will nonetheless always feel to the agent like a non-rational process, one "out of his hands".

[49] Young children with phobias, by contrast, apparently often fail to recognize that their fear is excessive or to understand its origin within them, and so it does not seem "alien" in this way. As a result, they tend to be more distressed and panicked the object of fear, but do not express distress at *having* this tendency to uncontrolled fear. DSM 446.

bulk of his thought and action from it. His doxastic control feels *defective* and *weak*, but it is not altogether gone.

Bob has *learned* through experience not to take his fear of bridges at face value, but still he cannot seem to *unlearn* the fear itself—in contrast to the way one normally unlearns fear (of driving, of the neighbor's dog) through repeated experiences without harm or incident. This, then, is a second kind of regulative failure in phobias—the normal control *feedback* exerts over affect is of limited efficacy. Despite the fact that his basic beliefs never change, Bob will freely admit that he still can't bring himself to *fully* trust bridges to be safe. This, too, feels to him like a control failure. Not because one normally can trust or believe at will, but because he has done all he could to exercise the chief forms of rational control over what he thinks and feels about bridges—the fullest possible consideration of evidence and argument—and still there is this recalcitrant, residual fear and distrust. Bob thus naturally feels that his epistemic autonomy is compromised by his phobia. It won't "listen to reason".

Intractability and loss of epistemic autonomy of this kind can occur in other cases as well. Someone who has undergone a traumatic experience may suffer recurrent attacks of anxiety and fear that correspond to no continuing risk, and that resist all attempts at unlearning through experience and reasoning. In a somewhat similar way, a person who has suffered a terrible personal loss may find—even years later, and despite decisive evidence from her own eyes—that she still somehow keeps expecting the person to return, or to be alive. Such thoughts may cause her real distress and anguish, and she might well wish to put this loss behind her and get on with her life, yet she cannot free herself of them through any ordinary process of rational control. We might well hear her lament that she is haunted by an "irrational conviction", that won't "listen to reason".

Failure of feedback to do its normal regulatory job is also implicated in another familiar form of irrationality in belief or failure to "listen to reason", dogmatism or fanaticism. In such cases, of course, the diagnosis of dysfunction or irrationality in belief is likely to come from *outside* the agent. For the true dogmatist, positive evidence is everywhere, while facts that any independent observer would say are contrary to the dogmatist's views inspire in him only defensive reinterpretation or contemptuous dismissal. (Near the end of his long life, a famous 20th-century philosopher of science was asked by a reporter, in the course of an increasingly frustrating interview, "But hasn't *anyone* ever given you a good argument against your view?" Only to receive the emphatic reply, "No. *Never.*")

In dogmatism, an important downside of the affective character of belief becomes manifest. The dogmatist's *idée fixe* typically serves multiple important psychic functions, including maintenance of a favorable self-image. The dogmatist typically has "invested" a great deal of positive affect in his *idée fixe,* perhaps even become identified with it. Positive evidence will be immediately liked, welcomed, and taken in. Whatever it might lack in evidential force would be more than made up by the positive psychic attraction and "compellingness" it will have for the dogmatist. Negative evidence, by contrast, will be immediately disliked, feared, or distrusted, making it difficult or impossible for him to take it to heart.[50] Moreover, anything that seems to discredit

[50] In a series of experiments, Caccioppo showed that individuals asked to identify words during rapid exposure could respond more quickly to positively-valenced words (like 'baby' or 'soft') if the response took the form of pulling a lever *toward* oneself, and more quickly to negatively-valenced words (like 'spider' or 'sharp') if the response took the form of pushing the lever *away*. Caccioppo, ref.

supposedly contrary evidence—even if the discrediting information is far-fetched by ordinary standards—will tend to receive a very positive, welcoming reception, and therefore seem especially credible. Affective defense mechanisms like these can operate unconsciously, on the very front line of cognition, and thus require no higher-order internal deception cleverly "masked" from consciousness. Small wonder, then, that dogmatic belief is so effortlessly self-perpetuating once it has become established, even in relatively simple, guileless psyches. As Winston Churchill put it, the fanatic is someone who "can't change his mind, and won't change the subject."[51] Unfortunately, the dogmatist may have no sense that he is being unresponsive to good evidence or arguments, and will instead be constantly amazed that others cannot see what is to him so obvious. A fanatical Republican will ask, "How can the Democrats possibly be taken seriously? How can they themselves fail to see the complete hypocrisy of their position?" And of course, dogmatic Democrats will experience the same sense of wonderment when thinking about Republicans.

Faith, and especially religious faith, self-consciously represents itself as able to confront contrary evidence without wavering. The proof of deep and sincere faith in God, or in one's children, is the steadfastness of conviction no matter how badly things are faring or how sorely one is tested. Sensible people—not ideologues—can approach this psychic limit when the object of their faith, God or child, is deeply loved. Sufficiently robust and pervasive positive affect swamps the negative, rejecting feelings necessary for doubt or mistrust to creep in. Those who love, and have built their lives

[51] Ref. Churchill.

around faith in God, or a person, or a cause, will see this as a good, not bad, thing. But I doubt they will defend it as pure epistemic rationality.

Now, according to our account, all belief involves *some* element of faith, namely, a measure of self-trust that is not grounded in independent reasons. Moreover, every belief involves some positive affect—indeed, the root of the word 'belief' is the same as 'beloved'.[52] This would suggest that the mechanisms just described, by which we become attached to our beliefs and tacitly welcome positive evidence, while resisting the entry of negative evidence, should be present to some degree in all belief. Owing to affective transfer or association, one form of positive affect tends to entrain others. Where we have trust, positive, confirmatory, supportive evidence will in general be better liked, more easily accepted, more credible, and more readily taken to heart than negative evidence. Do we in fact observe this sort of generalized positive evidence bias in believers? A substantial body of research in cognitive psychology has found this to be the case.[53] Since this selective receptivity occurs on the front line of perception, it usually is implicit, fast, and unnoticed. As a result, it can be quite invisible to the believer, who may feel himself to be wholly fair-minded, and quite refractory to higher-order thought even among those who "know better". Statisticians and cognitive psychologists are said to be prone in their everyday reasoning to the same bias toward positive evidence as the uninitiated.

The phenomena of affective transfer and association also have the effect of making us especially vulnerable to social stereotyping, and to manipulation by propaganda and advertising. Trusting and liking are close bedfellows, as are distrust and

[52] The shared Indo-European root is *be-* + *leubh-*.
[53] See Nisbett and Ross.

dislike. If certain groups are subject to pervasive social stigmatization, even fair-minded individuals—indeed, even members of these groups—will tend to pick up some negative affective associations with membership in these groups. These give implicit distrust an emotional head-start over implicit trust, with the result that individual members of such groups will find it more difficult to establish their credibility or reliability, and evidence placing them in a negative light will tend to have readier psyche entry and greater initial credibility than evidence of something positive.[54] Socially valorized groups, by contrast, will have an emotional head-start in winning the trust of others, and in avoiding being "branded" by negative evidence. Stereotypes will thus tend to be especially recalcitrant to evidence. For example, numerous studies have shown that c.v.'s, grant proposals, and article submissions will be seen as less credible and evaluated more severely, even by panels of academics and scientists firmly committed to equal treatment, if the names attached to the documents are feminine or suggest membership in a stigmatized minority group.[55]

Propagandists and advertisers have long taken advantage of this Achilles heel of our affect-based belief system. Triggering positive affect of any kind—through words, images, music, events, implicit appeals to status or evocations of the past, etc.—even if unconnected with any evidential value, will nonetheless tend to "bleed over" into trust and acceptance of the message.[56] Triggering negative affect, by contrast, will tend to mobilize distrust and rejection of whatever is portrayed. Since most of these effects occur prior to the level of conscious awareness, they are scarcely visible and especially

[54] On implicit attitudes and their effects on cognition, see ref. On stereotyping, its recalcitrance, and its effects on cognition and cognitive performace even in the minds of members stigmatized groups, see ref.

[55] Ref on cv studies. Similarly for credit reports.

[56] See Schwartz, Gilbert on how systems believe.

difficult to counter through learning and higher-order reasoning. We tend to be "taken in" by propaganda and advertising, again and again, precisely because advertising and propaganda are designed precisely to be attuned to our "affective keyboard", and thus to win ready acceptance of their message, which colors even subsequent, more self-aware and critical evaluation.

We also tend to be "taken in" by flatterers, it seems, whose welcome, favorable remarks tend to be experienced as more credible than the unwelcome, unfavorable remarks of critics—even when we know the flatterer has an ulterior motive.[57] And similar mechanisms affect self-evaluation. We seem to have no difficulty learning that most people are mostly average, and yet continue rating ourselves "above average"—in intelligence, driving skill, personality, and the like. These internal evaluations will *seem* or *feel* right, to have a credibility that seems recalcitrant to normal feedback from the actual outcomes of our lives.[58] Evidence of our competence and loveableness will always have a head-start over evidence to the contrary.

Unless one has suffered a serious blow to one's self-esteem. By operating directly at the level of affect, such a blow can have an immediate effect on one's beliefs even in the absence of relevant new evidence. Someone who has just been told he scored much worse than she expected on a vocabulary test suffers a slight loss in self-trust, and momentarily, and will be less confident in his opinions on a range of unrelated matters.[59] In major mood disorders, such non-evidential change in belief is often pervasive and severe. Those suffering a major depression have radically diminished capacity to feel positive affect of any kind, including trust and self-trust. They are

[57] Flattery

[58] Above average. We'll return below to positive self-bias.

[59] Confidence on unrelated matters

exceptions to rule that we tend to overestimate how favorably others view us,[60] but this gain in accuracy is accompanied by systematic cognitive losses. Suffering a deficit their ability to generate positive affect, depressives experience difficulty in learning new facts or recalling recent events. They have difficulty in concentration and in regulating their own thought processes, lose "attunement" to their social environment, and become absorbed in repetitive, circular, self-undermining thoughts. Someone who has been a world authority in a given domain may—without receiving any external evidence that her previous views were wrong—become incapable of saying or thinking anything definite on the subject. External signs that would normally confirm her authority, such as awards or honors, lose their capacity to produce any positive conviction—given the pervasive negativity of her inner affective world, such signs are instead likely to be distrusted and resisted, and seen as "mistakes". Any evidence of failure, by contrast, is "mood congruent", and so is immediately taken as to heart.

A complementary mood disorder, mania, has the opposite non-evidential effect on belief. Manic individuals experience episodes of extreme excess of positive affect, greatly inflating self-trust. They become very confident of their opinions even on matters where they have no real experience or expertise. A previously prudent individual will bet everything on a risky venture about which he knows next to nothing, fully expecting success. His sensory perception will be heightened and sped up, and his thoughts will race ahead. Episodic experiences and thoughts that happen to cross his mind will feel entirely compelling, so that his beliefs tend to accelerate off in all directions, unmodulated by negative feedback. For the positive affect that buoys self-

[60] Ref. In many other respects, however, chronic depressives have excessively negative, "hopeless" expectations.

assurance and certainty in one's convictions is present in such abundance, and the negative affect necessary for doubt or distrust in such short supply, that normal human resistance to the "entry" of negative evidence has been strengthened many times.

A purely cognitivist view of belief does not have a systematic way of explaining why mood disorders have these characteristic dramatic, non-evidential effects on belief and cognition generally. To be sure, if one feels less or more self-confident then one might take this to be evidence that one's beliefs are less or more credible, respectively. But this higher-order inference is rather tenuous and could hardly have pervasive, systematic effects on one's entire doxastic corpus, style of thought, perceptual sensitivity, and ability to learn or form memories. On the present model of belief, by contrast, such systematic cognitive effects of affective disorders is readily intelligible—they follow from the nature of belief itself.

These various types of irrationality in belief—phobias, dogmatism, bias toward positive evidence, evidence-resistant stereotypes and self-bias, mood disorders—are cases in which the regulative dynamics inherent in belief are not functioning normally, or tend to show some systematic asymmetries or unbalance owing to "external" affective factors. Recent years have seen a great deal of work on the ways in which affect and emotion contribute to cognition and rationality in thought and action.[61] I have been arguing that we should see belief itself in this light. The intimate involvement of affect in belief, however, must also be acknowledged to carry certain distinct and important liabilities. Evolution must work with the materials it has, and cannot be expected to optimize with respect to a single design criterion such as "aiming at truth". Many

[61] For examples drawn from a range of areas of thought and action, see ref.

cognitive strategies that presumably worked quite well for our hominid ancestors in acquiring reliable information and promoting group cohesion—giving greater trust to members of one's own group, paying greater to heed to what is said or done by individuals who appear more successful, healthy, and friendly, emulating the thought and behavior of those who are socially dominant—also place us at considerable risk of prejudice and manipulation. Not just by others, but by ourselves.

Fortunately, the affective and representational mechanisms of belief also make possible reflection and systematic inquiry into causes and effects. We have learned a great deal in recent decades about the ways in which belief works, and the ways in which it can go awry. It remains for us to incorporate this greater understanding into our conceptions of rationality and irrationality in belief, and into our ideas of what it is to achieve autonomous epistemic agency.

What is distinctive about desire?

It is high time we got to desire. Was our long detour worth it? Can we put what has been said about belief, and rationality in belief, to good use in understanding what desire and rationality in desire might amount to? I have already tipped my hand as to the key claims I will be defending, namely, that something deserving the name of *learning* is possible in desire, and so, too, something deserving the name *rationality*.

To make progress in understanding what learning or rationality in desire might amount to, we will need first to inquire into the nature of desire—just as we inquired into the nature of belief. In philosophical discussions, unfortunately, a motley crew of types of motivation—urges, drives, appetites, inclinations, cravings, wants, goals, wishes, etc.—are often simply lumped together under the term 'desire'. Initially, we will need at least one important distinction among them, a distinction with a philosophical pedigree going back at least as far as Aristotle, namely, the distinction between *appetite* (*epithumia*), and what we will be calling *desire* (*boulesis*).

For a start, desire, like belief, has an intentional object, while simple drives and appetites do not. Hunger and thirst are forms of motivation we share with animals, for example. They can arise in a response to a physiological condition, such as a deprivation or deficit state, and they motivate relevant kinds of directed attention and activity. Such states, at least in animals capable of feeling, sets the organism up for restless discomfort until some food or water is found, and pleasure or satisfaction when it is. A hungry mouse, for example, achieves this sort of *goal-directedness* with respect to food without needing the concept *food* or a self-representation as *food-seeking*. It can follow its excellent nose and memory for food locations to reach the nearest food source, and will arrive there prepared to eat, enjoy the food, and be rewarded for finding it.

Contrast my desire to serve an endive salad to friends visiting from Brussels this coming Friday evening. This, too, can lead to goal-directed behavior, but the goal—*that I serve an endive salad to my friends*—is now *represented* as an intentional object in the content of the desire. This goal can be absent, and never before have existed, yet the desire enables it to shape my thought and action. Thanks to this representation, I now can have what Aristotle called "deliberative appetition". Nor need I feel any hunger, or even anticipate soon feeling any hunger, for this to be so. And, unlike the mouse, I can begin my pursuit of this goal deliberatively, by reasoning out from a representation of it.

If I am to serve a proper endive salad on Friday night, I realize, I must find the right sort of cheese. Now I can draw upon memory or consult a book, and use this information deliberatively to arrive at beforehand. I can then reason out from what I know about Roquefort and cheese acquisition that I will need to find a shop that carries French cheeses. Arriving at this thought, I now want to find my local business directory, and use it to identify such a shop. Thus can desire translate motivation from an end to a considered means, once that means is seen in light of the end. Identifying a promising cheese shop, I do not follow my nose to it, or rely upon a conditioned association between its locale and cheese. I might never have been to this shop or neighborhood, but I can rely upon another representation—a map. And so on. Earthy as desire may be, it differs from mere appetite in virtue of the fact that its motive force is mediated by a representation, which provides the desirer with a potential starting point for reasoning, and permits motivation to be articulated into a plan of action. Moreover, the representations that mediate desire also set their satisfaction conditions. If the cheese shop I visit has only domestic blue cheese—which I might be unable to distinguish by taste or smell from Roquefort—I will not be wholly satisfied. That is not what I had in mind for my guests. A mouse with an appetite for cheese, by contrast, cares only for taste and smell—the absence of a certificate of authenticity will in no way impair its satisfaction. Desire thus involves a special kind of pursuit—"idea-guided" pursuit.

But not because we pursue the idea. Rather, we pursue what the idea represents by acting *through* this idea to bring about its realization. The idea *frames* our world-directed efforts. As Marx put it:

> The operations carried out by a spider resemble those of a weaver, and many a human architect is put to shame by the bee in the construction of its wax cells. However, the poorest architect is categorically distinguished from the best of bees by the fact that before he builds a cell in wax, he has built it in his head.[162]

This is the famous "world-to-mind" direction of fit of desire. In desire the goal-directed animal mechanism of appetitive striving is not replaced, but rather *recruited* on behalf of ideas—just as belief redeployed the animal mechanism of expectation.

It is not enough, we saw, to *believe* that *p* that one have a suitably rich mental representation that *p*—for this same representation is found in mere supposition and idle thought. To generate a real expectation owing to one's belief that *p*, one had to have some measure of *trust* or *confidence* that *p*. Similarly, it is not enough, to *desire* that *p* that one represent the proposition *p*—for this same representation can be found in belief, supposition, and idle thought. To generate a real motivation to bring it about that *p*, something warmer is needed.

What was it about the idea of an endive salad, for example, that moved me to take the trouble to pursue Roquefort—and authentic Roquefort at that? Not hunger. Rather, it was thinking about having my Bruxellois friends over, and hitting upon the idea of putting together an authentic endive salad to surprise and please them. Once I hit upon this idea, I *liked* it at once. And once I liked this idea, I *wanted* to carry it out. Did I know that I could? Not yet—but it seemed possible, so I was ready to give it a try.

As in the case of belief, a species of positive affect, in this case *liking*, can bring an otherwise inert representation to life in psyche, giving it force in shaping how I would go

[62] K. Marx, *Capital*, v. 1.

on to think and act. Affect by its nature is positioned to mobilize and guide all the requisites for pursuit—attention, expectation, attraction, association, inference, memory, and motivation. *Liking an idea* sufficiently thus can induce all of this coordinated psychic activity. Of course, in the typical case of desire, it is not the idea itself, a mental state, that is the object of liking—though in wishful fantasies and day-dreaming we find something like this. Rather, it is the intentional object of the idea—the *object as represented* or *under a mode of presentation* that is liked.

I am here using 'like' in what seems to me a natural and also somewhat generic way, as standing in for a wide range of species of *positive affective interest*. 'Like' will sometimes sound too strong. If I am *mildly interested* in hearing a talk, or *somewhat curious* about the speaker, then I will (in the present sense) to some degree *like* the idea attending it, and to that rather limited extent will be moved to go, and be to that extent frustrated and disappointed if I arrive and find the lecture room already filled to overflowing, so that there's no chance of my hearing the talk. 'Like' will other times sound too weak. If I *love* the idea of seeing an old friend while in Baltimore, then I will (in the present sense) very much like the idea of seeing him, and thus want very much to do so. If unable to follow through, I will feel correspondingly frustrated and unhappy and eager to find an alternative. 'Like' as we will be using it is a generic form of *positive affective interest*, and we must be careful not to be misled by the normal association between liking and pleasantness or pleasure. I do not need to think it would be *pleasant* to see my old friend—perhaps for various reasons it is bound to be quite painful, for she recently has suffered an irremediable loss—in order to be strongly attracted to the idea of seeing her again, and definitely want not to leave Baltimore without doing so. A kind of reward is in play, of course, namely, the reward associated with managing to do something one cares about or values. But the explanation of this valuing or reward need not be founded upon an expectation or experience of pleasure. *The Godfather* and *Psycho* can be someone's favorite films, or *MacBeth* and *A Streetcar Named Desire* his favorite plays, all of which he experiences as rewarding to watch, without any of them being *pleasant* to watch. One can intrinsically like the idea of sacrificing oneself for God or country, or making this world safe for generations yet unborn, and so find giving one's all in such a cause rewarding, with no thought of personal pleasure.

Typically, if we like something directly—the taste of sugar, the warmth and comfort of home on a cold day, the physical excitement of dancing, the novelty of travel, doing fine woodwork—we also like the idea of it, and often with corresponding strength. Mammalian behavior in general is heavily governed by such first-order liking—native (a taste for the sweet and distaste for the bitter) and acquired (a dog's excitement about riding in the car). In humans, primates, and perhaps other mammals, liking can also be a *conscious* experience, and can be *generalized* from a case of first-order, object-directed liking to a liking for related *categories* or *kinds* of things. As a result, the mere thought that something in such a category or kind is available can produce motivation and pursuit ("And you say there's a beach?"). And, in humans at least, it is possible to persuade ourselves, or be persuaded, to like the idea of something through words and images alone.

This gives human desire a creativity and freedom from the bounds of animal appetite that is surely one of our most distinctive features. Tell someone other than an analytic philosopher that you're studying desire, and the conversation will soon turn to sexual desire. But human sexual desire is very different from a mere “sex drive”. Erotic thoughts, occurring entirely out of context, can produce in humans a powerful cascade of physiological changes that other species must be in heat or rut to experience. And erotic thoughts need not take their cue from primary and secondary sexual characteristics, or pheromones, or hormones—they can take a dazzling variety of forms, including some of the most symbolic and sublimated ideas ever conceived. Nowhere is the importance of the mode of presentation of the representation that is at the center of desire more evident than in the erotic. The very same physical activity presented so enticingly by an erotic image, or thought, or letter, or song, or fantasy, could be presented with greater accuracy and detail through the language and anatomical illustrations of a textbook on reproductive biology, or by a series of flatly explicit photos, and not be in the least erotic, attractive, or “liked” in prospect. Moreover, the erotic does not simply dispose humans to various stereotyped displays of courtship or fixed action patterns of mating. It can inspire people to everything from Baroque religious architecture to jazz improvisation to Congressional politics to purchases at the mall. Such is the power and rapture, the boundless, seductive imaginativeness, of desire.

Modeling desire

Let’s move from X rating to PG-13 in order to look in a bit more detail at the *dynamic* operation of human desire in a typical case.

Sunburned and exhausted at the end of day on the water, I am thirsty (appetite, focused attention), but the idea of drinking the tepid water that remains in my water-bottle inspires me not at all. Instead, the thought of a cold beer starts to insinuate itself into my mind, thanks to various fond associations primed by my thirst and my awareness of it (spontaneous mental imagery and suggestion, affective association, memory priming and activation, Madison Avenue). Now *there's* an idea I like (positive affective interest). I can almost see the beer in my mind—golden, frosty, alluring (affective mode of presentation, incentive, more Madison Avenue). This charming idea energizes and sharpens my previously diffuse and languid mind (focused attention, cognitive salience, motivated cognition). Liking the idea, I find it pleasant to dwell on this thought (pleasures of anticipation and elaboration), but that alone does not satisfy me (felt lack). I really *want* a cold beer (incentive salience,[63] directed motivation), and begin to feel *impatient* to set about obtaining it (discontent with the *status quo*, restlessness, incipient frustration). My mind has already started reaching ahead, thinking through the question of the quickest way to obtain a beer once ashore (more motivated cognition, means-end reasoning, partial planning). Another idea comes to mind (spontaneous instrumental association), and suddenly you find me showing renewed purpose and energy in stowing the gear and cleaning the deck. This might seem surprising, given that a moment ago I was exhausted, and that clearing away is a job I actively dislike and usually postpone to the last minute. But now I am eagerly drawn to doing it, because I see this in the favorable light of shortening the time that separates me from my beer, a light that owes its glow to the radiant idea of the cold beer that lies at

[63] See Wyvell and Berridge (2000, 2001).

journey's end (affective mode of presentation, affective transfer from end to perceived means). And, once the clearing away is complete, I feel a certain extra satisfaction (instrumental gratification) to see the task completed just before we dock (progress monitoring). Looking up, however, I see that the shoreside bar is closed (more progress-monitoring, active frustration). Fortunately, I recall that there is another tavern nearby, a few blocks back from the water (memory priming, affective association, more motivated cognition and partial planning). I am delighted to find it and see that it is open (yet more progress-monitoring and instrumental gratification), and to discover that it has a favorite beer on tap (spontaneous pleasure, strengthened incentive and motivation). … You know the rest (pleasures of achievement and consumption, relief of drive reduction). All this, and more, is what happens when we wake to desire.

As in the case of simple appetitive pursuit, there is here a pattern of directed attention and motivated activity in pursuit of realizing a state that is associated with some sort of reward. But in the case of conscious human desire, as in the example above, there is something more: a favorable mental representation of what I am doing, an idea of it that I like, which plays a central role in eliciting and guiding this activity and gives it intentionality and self-intelligibility. I will have a ready answer to the questions, "What are you doing? And why?"

And I will be able to do with associative thought and reasoning what an animal might accomplish only through reward-based learning. For example, no sooner did it occur to me that clearing the deck in advance of docking would enable me to get ashore more quickly, than this inherently-disliked task became something I took an active interest in, genuinely wanted to do, and found it satisfying to accomplish. This behavior did not need to be conditioned into me by past positive association—perhaps I previously had always lazily waited until we were docked, where cleaning up is easier. I needed only the well-liked idea of a beer, a belief that cleaning up before docking would speed things along, and a capacity to see this means in the positive light of the end. The result is a remarkable gain in behavioral flexibility over animal appetite, and a capacity to adapt much more rapidly to new circumstances.

Here, then, is our first model of desire:

Desire to *R* (first version)**:**

> A *degree of positive affective interest* (liking) *toward* a representation *R* of an act or outcome functions to elicit and regulate a *degree of positive motivation* (wanting) toward bringing about or maintaining the act or state of affairs that *R* portrays.

A few questions and comments are in order. First, we ordinarily think of desire in two ways, as occurrent and as latent or dispositional. Moritz can be said to have a desire to climb the corporate ladder even when he's asleep, or totally caught up in the thrill of a rock concert. This is latent desire. But latent need not be "mothballed" or sequestered. For latent desire disposes the agent to find certain things salient, and to respond in distinctive ways to certain novel cues furnished by his environment. For example, Moritz, while enthusiastically jumping up and down along with pounding beat of his favorite indie band, might suddenly notice his boss in the crowd, jumping up and down just as enthusiastically. Even in the midst of his musical transport, this galvanizes Moritz's attention and starts his thinking going. He might ordinarily find her company tiresome, but it takes only a second or two of mental association for the idea to occur to him that he might be able to score some points with her by showing that they share a taste in music—if only he could catch her eye in all this noise and confusion. Almost instantaneously, that idea appeals to him and he's primed to swing into action. This is occurrent desire. He suddenly *wants* something urgently that before was the furthest thing from his mind, and he will be actively frustrated if he cannot get it.

Transitions from latent to occurrent desire typically involve at least two species of affect—for trust must be present as well as liking. For the agent must have some degree of confidence in his faculties, and in his sense of the situation as ripe for certain possibilities, for the well-liked idea to generate active motivation (wanting) as opposed to mere favoring (preferring) or imagining (wishing). Rather than complicate our model unnecessarily, we will simply assume that the term 'functions' in our model can itself be read dispositionally ("normally functions in suitable conditions to elicit and regulate, and to dispose to notice when conditions are suitable, and to prefer hypothetically when in suitable situations") as well as occurrently ("actively functions in the present circumstance to elicit and regulate").

It is also an important feature of desire that it can yield motivated pursuit even without a fully-formed plan or express intention. Moritz, for example, wanted to catch his boss' eye the minute the thought occurred to him. The strong initial appeal of an idea can draw one forward without ever taking an explicit decision or forming a complete plan. One simply begins to act, trusting, in effect, that one will deal with whatever happens next. This is typical of desire, and helps explain why we often find ourselves in the middle of actions—making a certain spontaneous remark, for example—without ever having decided to perform them, and with no clear idea of how things will end. Such desire-based action is nonetheless intentional and intelligible to us, however, thanks to the initial, perhaps fleeting, favorable representation under which we began it. We might not have intended to insult, or upset, with our remark, but we did intentionally speak the words we uttered, to our subsequent regret. Moritz will soon feel such regret, when he realizes in an instant from the expression on her face that, having caught her eye while she was indecorously jumping and shouting, this is the last thing that she would have welcomed.

Second, it might be wondered whether *liking* and *wanting* are really distinct attitudes, as the model presupposes. Are they not very nearly synonymous? And are not both nearly synonymous with 'desire'? We certainly do sometimes use these three terms almost interchangeably, speaking indifferently of Moritz liking the idea of advancing his career, or wanting to, or desiring to. It seems to me, however, that when we ask the terms 'liking' and 'wanting' to bear a little weight, we can see that liking and wanting are importantly distinct, and that each contributes something distinctive to desire without being equivalent to it.

To *want* is to feel a lack—as the etymology of the term suggests[64]—and be moved to fill it. Hunger and thirst are paradigms, and as they show, wanting is not an inherently pleasant or attractive state. A once-very-popular song speaks of a condition of "no more hunger and thirst", and the audience knows at once that this is a good thing. But would "no more liking" or "no more desire" be understood this way? Good food policies achieve "freedom from want" or "overcome want", but only bad kitchens produce "freedom from liking" or "overcome desire". Those who have known intense,

[64] From **eu-** (to lack) via the Old Norse *vanta* (lack).

unrelenting wanting, such as extreme hunger or thirst, or incessant craving for a cigarette, drink, or drug one cannot have, or desperation to get a breath of air while someone holds your head under water, will attest that this is a painful, wretched condition one is directly moved to escape. That, indeed, is wanting's function—it is a restless state, often linked in animals and humans with deprivation of a core need, that makes one unsatisfied with one's current condition, focuses and motivates activity to change that condition, and rewards one for success at doing so (gratification) while "punishing" one for failure (frustration). The usefulness of such an uncomfortable state in goading us to meet our needs is plain enough.

Liking, by contrast, is an affective rather than motivational state, though it typically does, as the model suggests, induce and regulate motivation. Unlike wanting, active liking or positive affective interest is in itself often pleasant or enjoyable to experience. Window shoppers, recreational perusers of catalogues, and those of us who cannot resist reading gorgeous magazine ads for luxury items for which we have no need, all enjoy the experience of thinking about something attractive and likeable, mentally exploring what one likes about it, elaborating upon it, and coming to know and savor distinctions in liking. (Such activity is not as pointless as it might look—it is important to discover what one likes in life and to develop some discernment in the evaluation of goods. Imaginative simulation is a chief, albeit often imperfect, way of doing this. The window displays, catalogues, and ads are props in this imaginative play.) Active liking, then, is not a state that motivates one to escape or remove it, but an *enticing* state, which *lures* you further into it. The characteristic, immediate mental and behavioral response to liking is *approach*—an open, welcoming, accepting state.[65] As we noticed earlier, liking is thus a close cousin of trust, while disliking is characteristically expressed in *avoidance* and *rejection*, much like distrust. Liking sustains the pleasures and rewards of *contemplation*, which need not depend upon the existence of any felt lack or active pursuit. We can find it intrinsically rewarding to daydream about something we like the idea of—making a stirring speech to the Constitutional Congress, or flying to glory in the R.A.F. in the Battle of Britain—even though there is no prospect of ever obtaining it. Liking also underwrites the pleasures or positive excitement of *anticipation* and the enjoyment or rewards of *consumption* and *attainment*, insofar as these are distinct from the gratification associated with successful pursuit. One can be gratified to attain a fancy position one has coveted for years, yet find it harrowing and unrewarding; a highly sought-after object, once the gratification of sheer acquisition is past, may bring nothing but consumer's remorse. We do, of course, directly *like* the gratification afforded by successful pursuit, and so also like the idea of it and look forward to it. But we can also like, and therefore enjoy, an unexpected gift of something it had never occurred to us to seek.

If active liking is a positive state, and active wanting frequently negative, what is active desiring, if, as I claim, it is made up from both? The poets and song-writers will tell us in an instant: it is *bittersweet*. Or perhaps, in light of our model, *sweet-bitter* since the entry into desire is attraction and allure, while the state once entered involves wanting, which can, especially if intense or chronic, be painful and yield strong feelings of incompletion and frustration. Yet if true desire is present, the urge is not simply to be

[65] 'Like' and 'alike' share the same root, **[līk-**, body or same. This suggests the thought: that which we like is to be joined with us in body, is not foreign or to be resisted.

rid of the discomfort of wanting, for desire organizes our thought, expectations, and actions around a positive idea, an idea promising something attractive, which would be lost if wanting were simply stilled without completion. And yet the fulfillment of desire is also the loss of something—the enticement or *fire* is gone, and that which was imagined or aspired to, which burned so brightly in the mind and was so longed for, is now, merely, real. This sweetbitter tension is inherent in desire, making it a more complex and fascinating condition, and a much better subject of drama or poetry, than mere liking or bare wanting. The arc of desire can be the stuff of epic narratives. Indeed, the very etymology of the word 'desire' contains both elements within it—it comes from *de-* + *sidere*; somewhat freely rendered: "toward, but separated from, a star".

Liking and wanting thus seem to me quite distinct, and it also seems to me that neither on its own captures the whole of the *concept* of desire. Neither does either state suffices on its own for the *property* of desiring, or for its distinctive *phenomenology*. But poetry, metaphors, and etymology aside, is there any reason to take seriously the idea that these two attitudes, liking and wanting, are basic building blocks in a scientific account of the underlying psychology of motivation? Neuropsychological research on addiction in animals and humans yields evidence that motivation involves at least two distinct brain systems and substrates, which have come to be characterized in the literature, interestingly, as "liking", a positive affective state or "gloss" associated with the hedonic value or intrinsic reward of a stimulus, and "wanting", a motivational state that drives pursuit and gives "incentive salience" to perceptual cues.[66] Typically, these two systems are coupled in the way suggested by our model: animals and humans tend to want what they like, and want more what they like better. In the normal case, the prospect of something liked—whether natively liked, or liked as the result of learning—elicits some corresponding motivation to seek or select it when it is in the offing. But liking and wanting can also come apart in various ways. Sugar, the taste of which is natively liked, will be taken up in the mouth of a newborn child or mouse in a way that requires no antecedent wanting—no hunger or drive. Indeed, the infant or mouse can be thoroughly *sated* and yet still welcome and enjoy the experience of a drop of sugar on the tongue. We can also find wanting without liking. Addicts who have become so "tolerant" of a drug through heavy use that it has lost any hedonic value or "liking", may nonetheless continue to crave the drug overpoweringly—the drug's take-over of their motivational system requires no positive contribution from affect in order to function. A mouse, for example, can be gotten to sip a liquid so bitter that it is intensely disliked, and would normally be refused to the point of gagging, if it contains a trace amount of an addictive drug to which the mouse's motivational system has become sensitized. This thrashing struggle to swallow the liquid will consume the mouse even if the drug itself has no apparent impact on its hedonic system.[67]

Third, notice the structural similarity between desire as thus modeled and belief as modeled earlier. Both are compound attitudes connecting two fundamental psychological attitudes and associated systems in a functional architecture. A species of positive affective attitude takes a representation as its immediate complement, and

[66] This pleasant discovery was for me a happy accident—very much liked even though not sought. Berridge (2004).

[67] Berridge, ref.

functions to elicit and regulate an action-oriented, projective attitude that takes a state-of-affairs as its satisfaction condition. Mind is, in this way, oriented in the world *via* its proxies: representations as proxies for the things represented, and degrees of affect as proxy for the reasons at stake—reasons to believe or reasons to do. Thanks to the pivotal role of representation, belief and desire alike permit inference and imagination to enter, and shape, the basic processes involved in behavior—expectation and pursuit. Action can occur intentionally, "under an idea", and how much we trust or like various imagined or hypothetical situations will profoundly shape which ideas we will act under. Aristotle also noted the parallelism between belief and desire. "Pursuit and avoidance in the sphere of appetition correspond exactly to affirmation and negation in the sphere of intellect", he wrote, and, "the desire must pursue the same things that the reasoning asserts." (*NE* ref.)

The parallelism of belief and desire and the room they together make for ideas in the governance of action also has a noteworthy downside, since our ideas can run away with us. We can envision, and gain trust and affection for, states of affairs that could never exist, or that would in fact be awful to bring about. Led on by positive but misguided affect, we will journey into absurdity or make a hell on earth. At the same time, flat-footed "realism" in belief and desire is not always a good thing—to confine one's imaginative and affective horizon to the tried and true can impede personal growth, romance, new discoveries, technological innovation, and the overcoming of entrenched injustices. As Joseph Conrad wrote in enumerating the few resources (Lord) Jim took with him as he disappeared into the backlands:

> He, on his side, had that faculty of beholding at a hint the face of his desire and the shape of his dream, without which the earth would know no lover and no adventurer. [Conrad dddd, pp]

In Jim we see the "feedforward" control of pursuit by desire at full throttle, and this brings us to a fourth point about our initial model of desire. It would be a disaster for the organism to have *only* feedforward control in desire—nothing to moderate or discourage the pursuit of an attractive but futile or overly costly end, or to draw us away from courses of action that are not living up to expectations and toward those that are proving more rewarding than anticipated, nothing, even, to turn off pursuit once a particular goal has been achieved. The lability and suggestibility of human desire, and its independence from mere instinct, permit desire to promote innovation, but unless such innovation is subject to some pressure from positive or negative feedback, we could have little or no confidence that desires will help attune us to reasons for action. Without some suitably evidence-sensitive process to selectively reshape existing desires or encourage new ones, with no discipline from experience, we would soon be at the mercy of our own imaginations or the imaginations of others.

Affect, as we argued earlier, by its nature functions to orient the organism to its environment in ways important for its well-being or reproductive fitness. It thus tends not only to shape attention, expectation, cognition, and motivation in a coordinated way in response to changing circumstances or needs, but also to be subject to feedback from continuing experience, enabling it to adjust and readjust rapidly in response to actual outcomes, opportunities, and perils. Unqualified trust in a car's reliability is immediately undermined when the blasted thing won't start on a cold morning and one is running late for an appointment. Fear quickly dissipates when a figure looming ahead in a dark alley is recognized to be a friend. And attraction to the idea of eating in a charming-looking hole-in-the-wall restaurant instantly dims when one spots cockroaches scuttling over the counter or detects an off smell coming from the pile of "Fresh Caught" fish displayed prominently near the entrance. Similarly, attraction to the idea of a holiday at a beach resort, which was enough to get one to pay hard-earned money for a package deal, will start to fade the minute one walks out onto the hot sand in swimsuit and flip-flops and discovers that one can no longer stand being out in the glaring, broiling sun. These changes in feeling tend to come of their own accord, unreflectively. As Robert Zajonc, a pioneer of affective primacy put it, "preference needs no inference".[68]

Sometimes, perhaps even often in some contexts, it makes sense to plumb one's feelings self-critically, and to deliberate explicitly about what values are at stake. But such higher-order thought and judgment take time and mental resources, and are by no means guaranteed to increase one's reliability or effectiveness. Like belief, desire projects the individual into the future, guiding action and putting one out on a limb on the strength of an idea of how things will go. But such ideas may not pan out. Just as one's trust in a tree branch continuously and "automatically" varies as it bends deeper under one's weight, so does one's attraction to the idea of crawling out on this branch to retrieve one's kite. Such non-deliberative "changes of mind" are exceptionally rapid—a matter of a few hundred milliseconds[69]—and immediately affect the relative

[68] Zajonc.

[69] According to one recent study, the brain generates a positive or negative evaluative response to losses and gains in a gambling situation within 300 milliseconds of the arrival of the sensory stimulus. This is well before any information from that stimulus reaches "higher" cognitive regions of the mind. Gehring and Willoughby (Science 2002)

allocation of one's mental and physical energy. With no need to pause for reflection, one's thought and effort shift from focusing primarily on implementing the once well-liked idea of crawling out to retrieve the kite, to devoting increasing attention to a different and now better-liked idea of extracting oneself from this situation as quickly as possible.

Desire thus parallels and complements belief in setting one up not only for action, but for something like trial-and-error learning through action. Recall our earlier discussion of optimal foraging in animals. Achieving optimality in foraging involves not just an accurate map of what food is available in one's environment, but of its relative *value*—its nutritive value but also its "hedonic value", that is, how well it is liked. Optimality also requires allocating attention and motivation to implement this "evaluative map" efficiently. That is, it is as much a task for motivation as it is for cognition. Indeed, in the wider, evolutionary setting, since an animal's likes and dislikes will tend to shape it food-seeking and food-consuming behavior, we should expect its tendencies to like and dislike—native or acquired through normal learning mechanisms—to reflect its caloric and nutritive requirements. Over time, how well-liked something is will tend to be a reliable indicator of food value, while how much other things are disliked will be a reliable indicator of their lack of food value or toxicity. Likes and dislikes, moreover, cannot be static, but must dynamically change with the condition of the animal (hunger or satiety, sickness or health, fatigue or energy) and with differences in the availability or cost of different food resources in its environment. Studies of animal learning show that the brain signals arising in response to the receipt of various foods take the form of a value function, varying in strength with respect to the amount of food, its gustatory qualities, and whether it meets the needs of the moment. It appears that what the animal "likes" (the food it selects in a free choice or "prefers") is continually being adjusted in response to its state of hunger or specific nutritive needs. Such signals are combined with other signals, which keep track of overall bodily condition and on-going changes in relative risk and effort, to shape action tendencies, selectively motivating certain behaviors and inhibiting others to yield a global pattern of optimal foraging behavior.[70] This is a system "designed" to enable the animal to respond effectively to what we might think of its changing food-related "reasons for action", much as its system of expectation-formation, attention, and learning through feedback was designed to enable it to adjust its internal representations of the environment to respond effectively to what we might think of as its continually changing "reasons for belief".

Moreover, successful foraging for food is but one of many domains of essential requisites of successful animal life and reproduction. The more complex the animal's life and needs—the greater the importance of social relations, the greater the skills and flexibility in devising ways of obtaining food or warding off predators, the heavier the demands of successful mating and parenting, and so on—the more evaluative information an animal must be able to integrate efficiently and translate into appropriate behavior. No such system could be perfect, or optimal in all dimensions. But when one considers the vast number of problems an animal as complex as a mammal or primate must solve in order to live well and fruitfully, it seems clear that this "value foraging" system—the behavioral manifestation of which is crucially mediated by relative degrees

[70] See Dugatkin (ref.)

of positive and negative affective interest, likes and dislikes—must have remarkable representational capacity, learning power, and behavior-guiding sophistication. It is clearly capable of attuning them to a wide range of animal "reasons for action".

Humans possess this same system, made significantly more flexible—and therefore also, more unreliable—owing to our greater representational capacity, speed and generality of learning, and sophistication in guiding behavior through strategies and plans. This makes it easy to see why human desire has a reputation for caprice rather than a sober reputation as a reliable learning mechanism—at least, in those societies where most people live in conditions of surplus rather than scarcity. But it is typical of any well-designed and normally well-functioning system that our attention is called to it primarily when it misbehaves or lands us in trouble. In their transition from helpless infancy to self-directed adulthood most humans manage in most environments, short of absolute scarcity, to allocate their motivation in ways that largely tend to meet their basic needs and enable them to secure a place in the natural and social world from which they derive substantial subjective well-being and intrinsic reward. In the well-developed societies in the West, the median adult report of "happiness" or "satisfaction with life" typically hovers around 7 on a scale of 10.[71]

Our desire system, I claim, is an important part of our "emotional intelligence", and of our practical competence, fluency as agents.[72] Ordinary language does not blanche at speaking of desires as proportionate or disproportionate to our evidence (the grandiose desires of Lord Jim come to mind), or as more of less warranted ("sadder but wiser"). Moritz, for example, showed his inexperience in forming and

[71] Comparative SWB measures.

[72] See Railton Pract Compete, Fluent Agency

immediately acting upon a desire to make it known to his boss that he was at the same concert as she. Most of us could have told him that bosses seldom like to be seen by their employees jumping up and down in a frenzy at a rock concert. As soon as he saw the look on her face when their eyes met, all enthusiasm for his plan drained from him. He is "sadder but wiser", and won't find a similar idea attractive for a long time.

Spelling out the feedback regulation inherent in desire (see also fig. 2, below):

> **Desire to *R*** (final version)**:**
>
> A *degree of positive affect* (liking) toward a representation of an available act or outcome *R* functions to elicit and regulate a *degree of positive expectation* (affective forecast) *and positive motivation* (wanting) toward maintaining or bringing about the act or state of affairs that *R* portrays; and this *degree of positive affect* (liking) is subsequently modulated by whether actual experience of performing, realizing, or moving toward *R* is better than, worse than, or in conformity with, the *degree of affective expectation* of it.

Like its predecessor, this model is a simplification, since, as an affective attitude, desire functions to elicit and regulate not only expectation and motivation, but also attention, cognition, memory, anticipation, planning, emotion, and so on.[73] But so long as we keep this in mind, perhaps the model does perhaps well enough to capture the essential core of desire. In any event, I will proceed on that assumption.

Desire and deliberation

[73] Moreover, the regulative structure of desire is more intricate than we can portray it here. Thanks, for example, to the multi-dimensional sensitivity of mammalian motivation systems, feedback can affect wanting directly as well as indirectly.

To understand desire, think of advertising. It is the goal of advertisers to induce in you desire for their product. Often, this is something you have never sampled, with which you have no history of positive experience that would leave you already with such a desire. The advertiser cannot *make* you like her product—but she can make *think* you do if she can make you like the idea of possessing it. The advertiser thus uses words, images, music, etc. to present the product in a positive light—just the sort of light liking the product would spontaneously produce within you. If you can be gotten to "take up" or internalize this positive representation, this will be a convincing *simulacrum* of the "front end" of desire, and will act upon you in much the same way as indigenous liking of an idea—it will lead you to *want* the realize this idea, to have some degree of focused interest and motivation toward it, increasing your likelihood of actually pursuing it. Through the normal operation of affect and affective association, the very seductiveness and attraction of the advertiser's images induces in us the attention, interest, and openness that give their positive representations easy entry into our minds. Advertising thus is the theater of desire—a special world in which beer and whiskey glow from within, fabulous-looking people drape themselves with, or around, fabulous-looking things and fix their gaze upon us, restaurants are full of laughing, attractive young people, and automobiles and wristwatches exude wealth and power. Like theater, it is designed to draw the viewer into the drama and provide him with a vicarious experience of something she is not living first-hand. By studying our actual likes and dislikes, fascinations and fetishes, an advertiser learns what props, people, and plots are needed for her theater of desire to work its wiles upon us. In desire we are moved to act under a favorable or attractive idea. Advertising's job is to plant such an idea in our minds.

The "invitation" to desire, the positive affective that lies at the core of desire, can be as varied and complicated as human feeling itself. We might like the idea of doing something because we're simply curious, or because it promises pleasure or happiness, or because of what or whom we admire, love, respect, revere, value, appreciate, envy, find titillating, or appreciate aesthetically. Thus we can desire, and appetitively pursue, intentional objects that promise something whose intrinsic or extrinsic reward is far from pleasure, and perhaps very obscure to us. Of course, desire can misrepresent as well as represent, so that it does not depend upon whether we *actually* like its intentional object. Dislike, too, has multiple sources and forms, corresponding to a wealth of negative representations— aversion (to hospitals), distaste (for a bland salad with duck), disgust (at drinking tepid water), fear (of infection), resentment (of arbitrary authority), hostility (toward foreigners), hatred (of public humiliation), etc. All of this would seem to make *deliberation* impossibly difficulty. How to compare or weigh all these diverse forms of "affective evaluation"? Here is where the "common currency" of positive and negative affect come into their own. Admiring, loving, envying, despising, fearing, etc. are all forms of positive and negative affective interest, ways and degrees of (among other things) liking and disliking.

Over the course of our lives, we build up a vast repertoire of ideas and scenarios, associated in our minds not only with certain characteristic causes and consequences, but also with positive or negative evaluative information, or, often, some of each. What was an event, or person, or relationship, or activity like?—Rewarding?

Painful? Exciting? Demanding? Hard to bear? Amusing? Boring? Meaningful? Discouraging? Energizing? Forgettable? Embarrassing? Frightening? Admirable? Beautiful? Guilt-inducing? Or some combination thereof? We come to make these evaluations through first-hand experience, but also through a great raft of more or less trusted sources—family, friends, colleagues, stray conversations, movies, stories, reviews, guidebooks, websites, proverbs, and pieties. This array of affective information is synthesized into positive or negative "feelings", with the result that certain possible actions or outcomes will "seem" more interesting, attractive, promising, exciting, welcome, frightening, or repellant, than others.[74] We will like certain ideas more than others. And these "seemings" will directly affect what we want and how we are disposed to act—including how we are disposed to deliberate reflectively.

For example. Often we find our desires unclear or conflicting. We're unsure what we want—as individuals, or together. What is deliberation like in such cases? We will need to sound out ourselves, and each other. What seems to matter in its own right? and how much? We will deliberate not so much to find a means to a known end, but to begin to give some shape to an end—to *specify* our aim, at least enough to take the next step or two.

Imagine that you and I are sitting at home, both a bit at loose ends. Neither of us has had a particularly good day, and we can't stand the thought of doing nothing but frittering away the evening doing e-mail and paying the bills. So we both hit on the idea of going out. "How about going out for a drink?" I suggest. But this idea falls flat. A

[74] It is not a foolish demand for unity, the work of overly-theoretical philosophers, that drives this need for integration and synthesis. It is a vital task that requires continual, evolving solutions if the organism is to respond to the enormous number of factors that must be taken into account in order to function well. Evidence concerning the insula.

moment's reflection on what it would be like to sit at bar trying to come up with interesting conversation in our current moods makes us realize that this is *not* what we want to do. "How about a movie?" you suggest. That works. "Good idea. But let's go to a theater—the thought of staying home to watch another DVD on the couch leaves me cold." You, too. So that's readily agreed. But which film? This is never easy. We consult the listings. "I'd like to see something that isn't a Serious Film," I say, "I'm really not in the mood for probing the depths of the human condition." You look a bit puzzled, "But didn't you say last time we went to a frivolous film that hated the idea of going to all the trouble of seeing a movie for something so mindless and predictable?" " 'Not Serious' doesn't mean frivolous," I hedge. "What I had in mind was something less *heavy* than that Bergman film I dragged us to last month. I was thinking of something worth watching and talking about, but more engrossing than a bunch of people wandering in the fog, looking miserable." We're stymied until an ad in the paper catches my eye. "Look—here's that new Israeli film Ori recommended. From what he said, it isn't heavy—and it's bound to be interesting just to watch. But we'll have to hurry, it's about to start." "Well, what's it about?" you ask, willing to give the idea a hearing. "It says here, 'A subtle and heart-felt work, but not without moments of black comedy. Slow paced—almost nothing happens—but excellent ensemble acting makes it clear that all the action is underneath the surface.' " This somewhat quirky description appeals to you, conjuring up vague memories of other films you've liked. So does the thought that we'd be able to discuss the film with Ori and Joan next time we see them. Our thinking aligned, we now have our choice. "Let's just grab our coats and go," you urge. And we do.

Here we have deliberated our way to a practical conclusion—the beginning of action—from the starting point of a somewhat inchoate discontent, a desire for change in search of a definite object. We worked our way to this object—worked out what we wanted to do with the evening—not by trying to fathom (how?) the strength of the various motivational forces within us and combine them into a vector sum, but by tossing out ideas to see if there was one that held sufficient attraction for both of us, that we both *liked* well enough. Affect and ideas were the currency of our exchanges, and our eventual agreement, not drive. But why did we not need to worry about drive? Why were we done with our reasoning when we identified an idea with sufficiently strong attraction, without needing to work out a vector sum of motivational force to see if I pointed in the same direction? (The way we might have to search our pockets to see if we have enough cash to buy the tickets.) Because desire was working normally in both of us, and thus we could rely without a thought on the inherent tendency of positive affective interest to elicit and direct motivation. Affect entrains motivation, and as our thinking converged upon on an idea with strong shared appeal, requisite motivation simply followed suit. "Straightway, we acted," as Aristotle would say.

We were able to identify a shareable aim even though the solution we hit upon was a novel one—a film we'd never seen—in part because we could draw upon the compositional nature of complex ideas. Various aspects of the ideas proposed corresponded to relevantly similar past experience, the affective coding of which, when explicitly or implicitly called to mind, colors our responses to the ideas we entertained. Neuropsychologists speculate that we manage this by a rapid, non-conscious ("off line") empathetic mental simulation of the option under consideration, which apparently plays itself out on the "test bed" of our own affective system and memories.[75] What emerges from such "sounding out" of how we feel, if we're lucky, is a shared, positively-valenced representation of the action or outcome—an idea we both like (for the same or different reasons). We can deliberate with desire, in short, by performing our own little theater of desire—imagining scenarios, and "feeling them out. If we finally settle upon one choice, we will have arrived at somthing of the form **affective attitude [representation]** that is manifested in part in a *projected* a degree of liking or disliking for there various scenarios, reflecting our beliefs and the strength and valence of our affective interest. This process of "preference formation"—with its associated expectations or "affective forecasts"[76]—in turn manifests itself in behavior in weaker or stronger dispositions to seek out, select, invest effort in, or avoid certain options. In the case of our discussion of the evening, the task of empathetic simulation of our own future experiences was not difficult, because we could exploit rather close similarities with other films we have liked, or have enjoyed discussing. But the same mechanism can work even for outcomes much further from our actual experience, as when a teenager runs through various scenarios of her adult future life and "makes up her mind", at least for now, about "what she'd like to do with her life". It is possible for to consider, and even become devoutly attached to, goals whose only reality, for the moment, lies in our minds.

Desire, like belief and language itself, is thus endlessly *productive* and *innovative*—and we ourselves show our origins from foragers by being characteristically *exploratory* and *curious* as we reconnoiter the actual and potential evaluative landscape before us,

[75] Empathetic simulation and choice, ref.

[76] Gilbert, affective forecasting.

leveraging old experience as we seek out the new. We do, I think, take ourselves to be more or less well-positioned or well-equipped by past experience to engage in prospective assessment of any given way of acting, and so can think of the affective forecasting implicit in desire as more or less warranted by, and proportionally responsive to, our evidence. This, combined with facts about the situations in which we find ourselves, and the needs or exigences we face, supports notions of when it makes sense on our part, or does not, to take our affective representations at face value, and to rely directly upon them in action. The forms of affect in question are, of course, not limited to those of positive or negative interest mediated by ideas. To the mix we add *direct* likings for people, places, and things. And in order to deliberate at all, we must possess some ground-level trust in our faculties, and in our sense of the content of our thoughts and experiences. We must not only have pro or con feelings keyed to representations of possible actions or outcomes, but also some measure of trust in those representations and feelings. Fortunately, trust is our default attitude. But as in the case of belief more narrowly construed, default trust is defeasible, and under various conditions one can lose or gain confidence in our ability, particular or general, to represent possibilities accurately, or to have reliable affective reactions to them.

Even when we are considering important choices, such as where to live, what career to pursue, or whom to marry, we typically must rely in a fundamental way upon something like this sort of affectively-guided deliberation and judgment. This does not mean that our deliberation about such matters is solipsistic, concerned only with our own feelings. On the contrary, as we saw in the case of belief, the affect we feel toward a representation typically is not the ground of our deliberation and judgment, nor does it appear that way to the agent, who sees the world *through* her affect. Rather, positive affective interest *presents* its propositional objects in a certain light, "singing their praise" and "lighting them up". And it is to this illuminated content that we primarily attend in deliberating. Desire is thus an "internal" advertising, an "intimating" of something we'd like to *have* or *do*, and not just think about. My glowing mental image of a cold beer advertises the beer itself as cool and refreshing, not the idea of the beer. This internally-supplied "glow" is simply a proxy representational device for the consummatory delight that awaits me, my way of "seeing" and thereby deliberating with it, so as to guide current, non-delighted pursuit in the world. Shaped by native likings and past experience, made more or less salient or acute by internal feelings that are responsive to my personal condition (hunger, anxiety, excitement), and cued by features of our current situation, these evaluative proxies vary in strength, vividness, and urgency, coloring and recoloring an internal map of the favorableness or unfavorableness of the possibilities that seem open to me at the moment. Distant prospective goods or harms can, if they possess sufficient affective interest, enter into the scales of my deliberation right alongside the blandishments or threats of immediate experience, and exert a real influence on choice. Such appearances of future gain or loss can of course *mislead*, but the important point to notice in the present context is that, real or imagined, they nonetheless can *lead* in the here-and-now. Eva's powerful attraction to the idea of becoming a doctor draws her forward through a long and intensely difficult period of schooling and training. And her aversion to the thought of Joe as a husband, or the father of her children, supplies the resistance she needs to hold at bay his constant pressure to deepen their relationship. If her ideas of the future, and her affective reactions to them, have been developed and refined through relevant experience—growing up in a father like Joe, summer internships in hospitals, years of experience developing increasingly mature relationships with others, previous romantic involvements, etc.—, then, in following them, she may well exhibit practical intelligence and rationality, a quite sensitive attunement to reasons that might be much too complex to put into a simple decision matrix.[77] Those unable to simulate empathetically what it would be like to live with various outcomes—who lack the ability to form reliable representations future possibilities, or to respond to them affectively with any sensitivity—appear to be severely handicapped in both practical intelligence and rationality.[78] Caught in the incentives of the moment, they may make a long-term decision on short-term grounds, against which nothing within them rebels.

Normative guidance and the problem of regress

This talk of deliberating via "affective evaluation" is all very well, one might object, but don't *norms* and *values* play a central role in deliberation about action in their own right? Yes, and no. They do play such a role, but thanks, in part, to the operation of affect. Let us consider the case of values first, then norms.

[77] I once asked an expert on the psychology of human decision-making what, to his mind, was the best way to make a decision like which job to take or where to live? He replied that making lists of pros and cons might help one's thinking at the beginning, but that it's impossible to think of all the things that matter, or how much they matter to you. Lists may make us pay attention to too few factors. And often, it seems, we end up fudging such lists to yield the decision that "feels right" anyhow. We have to be able to live with our decision, to embrace it, to "feel right" about it and have some enthusiasm for it. The key thing, he said, is to use your feelings *well*. "Learn as much about the alternatives as you can, then simulate them in your mind as concretely, vividly, and fully as you can … then read your somatic markers." We are being advised, in effect, to treat our feelings as our best indicators of reasons for action, and as what will best enable us to respond to the reasons we have.

[78] Damasio, Bechara

"People experience the world as infused with many different values," Elizabeth Anderson began her pioneering study of value and valuation. While philosophers have alleged that talk of experiencing value is mysterious, Anderson has a straightforward reply:

> We can dispel this mystery by recalling what ordinary experiences of value are like. We experience things not simply as good or bad, but as good or bad in particular respects that elicit distinct responses in us. There is nothing mysterious about finding a dessert delectable, a joke hilarious, a soccer match exhilarating, a revolution liberating. We can also find someone's compliments cloying, a task burdensome, a speech boring. To experience something as good is to be favorably aroused by it—to be inspired, attracted, interested, pleased, awed. To experience it as bad is to be unfavorably aroused by it—to be shocked, offended, disgusted, irritated, bored, pained. Evaluative experiences are experiences of things as arousing particular positive or negative responses in us.[79]

Our discussion up to this point has largely been concerned with this sort of evaluative, affective experience. Just as our eyes, ears, nose, and so on are our ground-floor window onto the world of objects, our affective experience is our ground-floor window onto the world of values. And we have been maintaining that desire is one dimension of that affective experience, and perhaps the most important one in the guidance of action, just as vision is one dimension, and the most important dimension, of the physical experience that guides action. Through the frame of intrinsic desire we see much in the

[79] Elizabeth Anderson, *Value in Ethics and Economics* (Harvard 1993), pp. 1-2.

world to like for its own sake—pleasure and happiness, to be sure, but also such intrinsic goods as accomplishment, loving relationships, and autonomy. Small wonder philosophers have often held that desires are reasons for action in much the same way as perceptions—both *intimate* a reality of central concern to us.

Sometimes, however, we not only experience something evaluatively, but *value* it. As Anderson observes, "To value something is to have a complex of positive attitudes toward it", which include not only ground-floor positive affective attitudes, but also higher-order *regulative* dispositions shaping one's "perception, emotion, deliberation, desire, and conduct" toward the valued object.[80] One does not simply love a friend, the way a very young child might quite simply love a parent, but also *values* her and the friendship. One therefore hates the idea of abandoning this friend when she is in difficulty, and would be disgusted and angry with oneself if one began to yield to temptation to do so. One also admires the idea of a friendship strong enough to resist such temptation, and will feel in some measure ashamed of oneself if one fails to live up to that ideal, or proud of oneself if one succeeds.

If one has a true friend, then she *merits* or *warrants* these first-order and higher-order affective attitudes and responses—they are a *fitting* response to her and to the friendship. When it comes to value, our affective responses can get things right, or wrong. And when we have the affective attitudes that befit something of value, and these attitudes are functioning in their normal way to regulate what we pay attention to, how we perceive it, what weight we give it in our deliberation, and what force it has in motivating our actions, this *is* what it is to respond aptly to value in practical

[80] Anderson (1993), p. 3. Anderson gives a special role to *standards* in her account of these higher-order attitudes. We will turn to the question of standards next.

deliberation. Excruciating pain, we say, is intrinsically bad. To intrinsically dislike such pain, and be moved to prevent it in oneself and others, befits its nature. It amounts to intrinsically desiring to prevent something intrinsically undesirable. And if empathy causes one to experience through internal simulation the pain of another, so that one intrinsically dislikes the idea of the other undergoing such pain and thus wants to prevent it, this, too, is a fitting response. And if, in addition, one stably desires that one's desire to prevent pain be *regulative* in one's actions, so that one's concern with the pain of others not depend entirely upon the uncertain action of empathy, and one feels admiration for those who minister to the pain of individuals with whom it is difficult to empathize, and so on, then one might be said not only to desire the prevention of pain, but to *value* this. And this amounts to intrinsically valuing that which is intrinsically valuable, another fitting response. First-order and higher-order desire and aversion, along with other affective attitudes, thus are part of our basic equipment in being aptly responsive to value in deliberation and action. To be motivated to prevent pain, in oneself or others, *without feeling*, or *without any positive interest* in doing so, or *without any dislike of pain in itself*, would be a less, not more, fitting response.

Of course, this does not mean that value *concepts* are identical with affective concepts, or that we can give the full content of deliberating in terms of value entirely in the language of first-order and higher-order affect. The conception connection is rather *normative*—values *merit* certain affective responses. But it does mean that when we are deliberating evaluatively we typically are able to do so by deliberating affectively—as we saw in the case of our decision about what to do with our evening.

Anderson also emphasizes that neither evaluative experience nor valuing are *judgments*.[81] One can judge that something is valuable without experiencing it that way, or responding to it in one's sentiments as valuing requires. As we saw in the case of trust and belief, affective attitudes are involuntary in the sense that they are not subject to direct decision or legislation. And, as in the case of trust and belief, this resistance to "top-down" decision might be a very important evolved protection mechanism—to counterweight the potentially arbitrary operation of reason with feelings based upon a lifetime of experience. This also means, however, that there is a distinctive role for value judgments in deliberation. Our value judgments can prove essential in forcing us to keep certain values in mind, or take their measure in deliberation, or hold ourselves to them in action, even when we don't feel like doing so. This, too, is an important countervailing force, which helps us to complete and render more coherent and stable our evaluative map of the world. Value judgments do this, however, with the help of *other* affective attitudes. We must, at least implicitly, *trust* our judgment and its issuances, and have a positive attitude toward giving them weight in thought and deed. We must want to be consistent and find inconsistency uncomfortable. If, through bitter experience, we have come to distrust our judgment or judgments, or to feel disgust with ourselves for taking them or questions of consistency seriously, our value judgments could hardly shape practical reasoning in an appropriate way. Just as, in the case of belief, seeking reflective equilibrium is typically a matter of seeking an affective equilibrium with respect to confidence, so, in the case of value, is seeking reflective equilibrium typically is also seeking an affective equilibrium with respect to care or

[81] Anderson (1993), p. 3.

concern. Combining these two perspectives, we can see how this would make sense if the point of reflective equilibrium is not merely making up our minds theoretically, as a way of shaping what we expect, but also practically, as a way of shaping how we will go on to act. For if a reflective equilibrium judgment represents the best we can do in "feeling our way toward" what we most firmly trust and care about, then it will have the requisite action-guiding force; otherwise, it would remain a purely hypothetical exercise. A psychic structure in which deliberative judgment engages one's first-order and higher-order affective system permits such judgment to entrain motivation directly—without need for a further judgment or act of will that could launch a fatal regress.

Wouldn't a Kantian, however, protest that in true practical reasoning, the practical force of moral principles must be product of a judgment of rational necessity, not a feeling of any kind? Not if the Kantian is Kant himself. We noticed in the case of belief that one can be aptly guided by a norm of reasoning thanks to, rather than in spite of, an affective attitude toward it, namely, trust. That which I trust I implicitly *defer to*, and it will tend to regulate how I go on to think and act without need of any intermediating attitude or judgment, thus avoiding regress. But a key further point is that such deference can itself be *intelligent* rather than brute or blind. For trust by its nature is sensitive to discrepancies between expectation and reality, and thus is informed by learning and unlearning. Moreover, nothing about the trusting attitude as such would destabilize the force of a norm. To learn that we have followed a putatively epistemic norm simply out of political ideology would tend to undermine the authority of that norm, other things equal. But to learn that we have followed such a norm because we have come through experience to trust it sounds to my ear like an epistemic *commendation*, not condemnation, of it. (Should we instead follow norms we do not trust?) Thus, we need not face a choice: intelligent and self-intelligible or transparent normative guidance (with the threat of regress) vs. brute, blind affective guidance.

What, then, of norms in the practical sphere? It is, of course, important that we trust them, and trust ourselves in applying them. But trust alone is insufficient to account for their full role in shaping action. As Kant observed, we must distinguish a "merely theoretical cognition" of a rule from a practical appreciation of it (*MM* 6:ref). I can, for example, theoretically cognize a rule of proper etiquette, and see how it applies "categorically" to a choice I am about to make, without the rule thereby acquiring any action-guiding force within me. But what, Kant asks, could be the appropriate incentive to make the moral law one's effective maxim in action? It would have to move us to act on behalf of the moral law *in itself*, with no promise of any gain. And it would have to be fully appreciative of the dignity of the moral law and the value beyond price of a good will. Kant's answer is clear, this "moral feeling" must be respect ("*reverentia*") and "non-pathological" love, i.e., love that is not self-centered or based upon pleasure. If one reveres and loves nature, one will be moved to protect natural habitats for their own sake, without further purpose, with no reference to the future course of one's own life or experience, and even in the face of opposing interests and inclinations. And if one loves and respects a person as an end-in-himself, then coming to see that a certain act would destroy him, betray him, or degrade him, or treat him as a mere means, would directly elicit motivation to reject or oppose this action—giving this cognition *practical* as well as theoretical force. It would be wrong in these cases to say that, because affect is involved, the "real" aim of the action is self-centered, say, the satisfaction of preference or avoidance of a discomfort. For it is a core feature of reverential respect and non-pathological love that they direct the agent's concern *outward*, toward that which is loved or respected, seen as ends-in-themselves. Neither are either of these affective attitudes inappropriate for the regulation of *moral* conduct. For a good will and the moral law *merit* the feeling of respect in themselves: "*Respect* is a *tribute* we cannot refuse to pay to merit, whether we want to or not; we may indeed withhold it outwardly but we still cannot help feeling it inwardly" (*CPrR* 5:77). Kant, moreover, is clear that is only because respect and love are species of affect, not judgments, that practical regress can be avoided:

> Respect (*reverentia*) [for the moral law] is, again, something subjective, a feeling of a special kind, not a judgment about an object that it would be a duty to bring about or promote. For, such a duty, regarded as a duty, could be represented to us only through the *respect* we have for it. A duty to have respect would thus amount to being put under obligations to duties. [*MM* 6:402-403]

The power of judgment alone, unaided by affect, could not turn us into moral agents—we must also possess "moral feeling": "If a person entirely lacked any receptivity to this sensation, he would be morally speaking dead …" (*MM* 6:399).

Kant's virtuous person is therefore not someone who acts "legalistically", or purely for the sake of respect for the law as a kind of demand of obedience (*observantia*). To have a good will one must find "an allurement, which invites" (*MM* 6:219) in the moral law—it must be an *ideal* one loves and reveres, like a "jewel that shines by its own light" (*G* 4:394). This feeling is like nothing so much, he suggests, as the experience of the *sublime* in nature (*CJ* ref.). Kant's virtuous person thus shares at least this much more with Aristotle's—she does the right thing, for the right reason, in the right way, from the right desire. Love and respect for the moral law guide us *categorically* in way that matches the categorical character of duty. Affect itself has the power to mobilize the requisites for action, and unconditional affect to do so unconditionally—needing no further incentive. We see this most clearly, perhaps, in the case of unconditional parental love.

I have focused here on moral norms, and on Kant's account of how affect permits the appropriate sort of guidance by moral norms, since one might at first have suspected that no affect-based account of our susceptibility to normative regulation could possibly accommodate morality, especially as conceived by Kant. A fuller appreciation of Kant's account would bring in other forms of affect that Kant takes to work alongside respect in supporting our moral practices, such as love of one's neighbor, respect for one's self, and non-pathological self-love. In the real world of normative guidance, this is surely right. What we experience as the categorical demands of morality is bound up not only with a respect for principles, but also with self-respect and a concern for others as ends in themselves, as well as various other first-order or higher-order affective attitudes—for example, being averse to the idea that one might be unable to withstand one's own scrutiny, or being attracted to an idea of living with others on terms of equality or voluntariness, and so on. To account for other forms of normative guidance, such as guidance by social conventions or conversational norms, would involve yet other affective attitudes.[82] But let us leave such questions for now, and draw a few simple conclusions.

First, effective normative guidance properly so-called is possible thanks attitudes of the form **affective attitude [norm]**, so long as the affect is of the right sort—e.g., trust, respect, love, admiration, so on.[83] Second, although such attitudes permit normative guidance without regress by disposing us to make direct, unmediated transitions in thought, nothing about this requires that the dispositions themselves be brute, opaque, or unintelligent. Trust, respect, and love no less than fear directly shape what we attend to, notice, infer, associate, remember, and are moved to do, without need for any intermediating judgment or supplementation by a further aim or motive. This brings us to a third point. For these affective attitudes are not blind or mute. They present their intentional objects in a distinctive "positive light", conscious or unconscious, with the effect that being guided by this representational content can be *intelligible* to us. Moreover, each of these attitudes is *intelligent* in the sense of being capable learning in response to experience. We can learn to trust, respect, or love (or to distrust, disrespect, or hate) just as much as we can learn, or unlearn, fear. Such learning is not confined to one's individual experience. For example, surveying the many failed utopian communities of the 19th and 20th century may remove much of the luster from highly rigid, ultra-pure ideals of human conduct. And greater equality in gender roles over recent decades has made possible accomplishments by women that command new forms of respect in men, despite themselves—as the "tribute we cannot refuse to

[82] See Railton (ref.) Normative guidance.

[83] An interesting way of seeing this is to consider cases of the sort discussed by Antonio Damasio and others, where the connections between affective and cognitive regions of the brain have been severed, with the result that an individual who is very capable of moral reasoning nonetheless has persistent problems regulating his behavior even in matters of simple prudence. See Damasio (ref.)

pay to merit, whether we want to or not". Evaluative affective responses can, no less than fear, be subject to evolution through experience, becoming more or less *warranted* or *appropriate* as a result. And this brings us to a third point. The affective attitudes that lie behind normative regulation can be as experienced, thoughtful, and reasonable as we ourselves are. It might be true at some basic neurological level that attitudes such as trust and love are built up from brute, opaque dispositions—I don't suppose that individual neurons see things in a positive light—but there still is all the difference in the world between a parent who holds back an impulse to interfere in her adult child's life out of love and respect for him as a person, and the parent who does so from an opaque, brute disposition, unintelligible to herself, who sees nothing positive in doing so, and can only answer, "That's just what I do" in response to the question, "Why?" Being normatively regulated by the right sort of affect is not a "second best"—on the contrary, it is precisely the way in which that which is precious can be given its full due.

Acting *from*, and learning *in*, desire

Desire as modeled here, I have been claiming, exhibits a compound structure involving both feedforward and feedback control, making possible a learning dynamic inherent in our motivational system. It is not hard to see how desire might "learn" in the case of instrumental desires, since desire sets us up to become frustrated and impatient with ideas for how to realize our ends that prove unsuccessful in practice, to be motivated to look for better ideas, and to be rewarded if we hit upon ones that bear more fruit. When we act from desire, then, an expectation-based "monitoring" of progress and motivated readjustment of behavior takes place in us continually, without need for self-conscious deliberation or higher-order cognitive effort. Consider someone who is walking to class, dodging on-coming students, carrying on a conversation, trying to reassure a worried colleague, balancing a heavy load of books and papers, opening doors, looking around for a clock, and adjusting the angle of an umbrella to meet the on-coming wind. It would hardly be extraordinary if someone were to succeed at all these tasks over the course of a five-minute stretch on a given morning. But it would be quite extraordinary if one did it by deliberating about all these things, authorizing each and every constituent action—from the arc and angle of the next swing of a leg to the choice of the content and tone of each sentence before one utters it—while at the same time monitoring how one's previous acts are playing themselves out. Conscious, deliberate control can take charge of only so many things at once without inducing mutual interference, and it makes no sense to squander it on goal-directed activities that belief, desire, memory, and habit can intelligently and smoothly regulate on their own. When they do so, our behavior is no less responsive to reasons. After all, the transition from being a novice to being skilled at an activity involves just such a transfer of control of the bulk of what one is doing. We do not wish to suggest that, in becoming skilled, we have lost contact with the reasons for which our training process was initiated.

But what of learning in the case of non-instrumental or intrinsic desires? A convincing example must be somewhat complex.

I'm languidly reading the back sections of the Sunday paper after a prolonged brunch. My eye comes to rest on a review of a new Singaporean restaurant, by a food editor whose recommendations I know many take seriously. She presents a vivid, intriguing description of a dish I have never tried, made with fish I have never heard of and a sauce based upon a fruit I have never tasted. The fruit is durian, and the dish, she writes, is especially remarkable because it concentrates the essence of the mysterious "fifth flavor", *umami*. I know nothing about this taste, but curiosity makes me excited by the idea of trying it.

By the normal working of desire, these favorable representations induce directly in me a new motive—I now *want* to visit this restaurant and try this dish, and my mind has already begun to crank away at forming a plan to do so. When I call for a reservation, however, I am frustrated by a constant busy signal, and about to give up when I finally get through. I am delighted to hear that I can have a reservation in two weeks' time, but the minimum party is two. So now I want yet another thing—to find a friend to go with me. Momentarily stymied, I'm delighted to see that Ellen, a friend who also enjoys trying new things, is still at the brunch. She seems intrigued, and will check her calendar when she gets home, and let me know. And I'm pleased and relieved when she calls that afternoon to say she can come.

Several remarkable things have happened, at least, if you compare me with a moderately intelligent gerbil. Despite being stuffed with food and the furthest thing from hungry, I suddenly began actively wanting to obtain food. But not food anywhere nearby in space or time. And not generic food—a very particular kind of food, yet not a kind I have ever tasted before, or particularly expect to like. A mere *idea* of a food. Moreover, I began my pursuit of this food not by following a sensory trail, but by thinking and planning. My attraction to the idea of trying this food transferred effortlessly to the task of obtaining it. Though I don't like calling restaurants cold, I now did so readily. All this sounds rather cerebral, but I was also, in fairly quick succession: eager to call the restaurant, frustrated when the phone was busy, pleased when I finally got through, satisfied to be able to make a reservation, anxious about finding a dinner partner, delighted Ellen was there to be asked, and relieved when she said yes. All this affect, with its attendant effects guiding how I go on think and act, has arisen from an idea that caught my fancy.

The night of the reservation arrives, and I rendezvous with Ellen at the bus stop. To my surprise, her eleven-year-old son Max is with her. Max doesn't at all like going out to dinner at fancy restaurants or tasting strange things, but when the arrangement for a sitter fell through at the last minute, he knew that his mother, who obviously wanted very much to go, could do so only if he tagged along. A worry crosses my mind—now we're a party of three. But Ellen sees my puzzlement, and suggests that there is unlikely to any problem about making space for a child at the table. With this sensible thought, my worry goes away. At least, *that* worry. But what about Max? Last time they all had dinner together, a little over a year ago, he was immediately bored by the adult conversation and very unhappy about having to sit still, winning the concession of playing his pocket video game throughout the remainder of the meal. Here's a worry I share with Ellen—she, too, has no clear idea of what to expect and her mind is racing ahead exploring various possible ways of coping. When she realizes that neither of them remembered to bring a video game, she flushes with consternation as her pleasant image of the evening seems to be at risk of vanishing altogether. She badly wants to think of a solution. Seeing another child at the bus stop playing a game on his cell phone, she suddenly brightens—"That's it". Now she has a fall-back, and her mind relaxes. Here, too, we see the working of desire.

The bus arrives, a bit late, and cross-town traffic is slow. Max senses Ellen's anxiety, and keeps us entertained with a very long story about his gym teacher. We arrive at the restaurant a quarter-hour late and with one additional person, but miraculously (or so it feels) the headwaiter has saved a table for us despite the number of people in vestibule waiting for an opening. "But of course!" he smoothly intones, gracefully shepherding us to our spot. Although this turns out to be a tiny table jammed into a back corner by the busy kitchen door, we are pleased and most grateful, and tip him handsomely. Headwaiters understand about the arc of desire.

Thus far we have gone through planning, discomfort, anxiety, and spent not a little time and effort on what?—An intriguing idea? Not exactly. For while we have taken our guidance from an idea, that idea only *mediates* or *regulates* our pursuit—its true object is the meal we now believe are about to have. If we had wanted only the attractive idea, we could have spared ourselves all this trouble.

And now the long-awaited fish entrée arrives, beautiful and pungent. The smell is, however, odd to say the least. "Oh boy!" Max exclaims, "Smells just like my old socks," getting a nervous laugh. We have to concede his point, though Ellen adds, as if to reassure herself, "Well, it's no worse than some cheeses I really like." At last, my first taste—surprising, incredibly intense, and wonderfully complex!—though with a long "finish" that's hard to pin down. Another taste and the strong, savory flavor—one technical description is "sweaty green"—has become dominant. A third taste and, my face reddening, I realize that I will actually have trouble eating this. Not wanting to lose face—and, after all, this way *my* idea in the first place—I manage to swallow down whole enough chunks of the fish to finish a respectable amount. By the end, I feel as if I have had enough durian for a lifetime. The boundary of my gustatory curiosity will not be extended further in *this* direction, I think. "I'm glad I tried it," I say, deflated and apologetic, "but I think I won't make a habit of it. Sorry if I dragged you here …".

"Not at all," Ellen interjects. She has a much more sensitive nose and palate than I do, and no pretensions about the catholicity of her taste, but nonetheless is having a quite different reaction. "I really had no idea what to expect," she continues, "but this is something else. What a smell! But the taste … . Not pleasant—this is obviously an acquired taste. I can't exactly say I *enjoy* it. But I could learn to—it really is fascinating, with lots of possibilities. I think I can begin to see why people make fuss over it." In any event, she seems to be enjoying herself—focused mostly on the conversation, not the food. Max, who no doubt has the most sensitive nose and palate of the three, had wisely ordered steak, which he finished with gusto, ready for dessert. He, too, seems to be enjoying himself and, buoyed with the success of his story about the gym teacher and his joke about the fish, has become surprisingly voluble. The conversation ranges widely, and never flags. So the question of finding a suitable distraction never comes up, and the cellphone remains in Ellen's purse. At one point, upon urging, Max accepted a forkful of fish from his mother. Hamming it up by holding his nose, raising the fork, and conspicuously touching the food with no more than the very tip of his tongue, he says "Yum. It *tastes* just like my old socks, too!" Another, less nervous laugh. Very acute, this Max, I think. When we are about to go our separate ways after the dinner, I suggest to Ellen, "Let's do this again soon, shall we?", adding, before I knew what I was doing, “And you’ll join us, Max, won’t you?” Ellen looks a bit taken aback, and I’m suddenly abashed at having barged ahead like this without consulting her. But Max doesn’t miss a beat. He agrees with an alacrity that seems to surprise even him. “Deal,” he says, extending his hand to shake mine. Smiles all around as we part company.

Thanks to desire, we all feel somewhat gratified by the end of the evening despite our very different experiences. I had my chance to taste umami straight on, and managed to eat enough to keep intact some of my gustatory self-esteem—that is part of what I wanted, anyway. Ellen experienced a new and genuinely interesting taste, was able to keep Max actively involved in the conversation and the spirit of the evening, and had a calm dinner out—more than she hoped for. And Max emerged pleased that he made his mother happy, not bored, and proud of his grown-up behavior and sense of humor—the first two were as much as he hoped for, but the third is a very pleasant surprise. Indeed, we all were a bit surprised at how well the threesome worked out—no doubt that was why I blurted out he invitation to Max. This was something we hadn’t expected, or actively sought, but nonetheless found intrinsically rewarding. As a result of these experiences, each of us will have somewhat changed desires going forward.

Now that I have a more accurate idea of what it is like to taste umami, this idea no longer inspires the same vague positive image in me. Indeed, it inspires an active aversion. I had hoped to like the taste, and, implicitly, wanted to be able to say, if the subject of durian’s notorious taste and smell ever came up, “You think it’s disgusting? I think it’s marvelous.” These self-flattering thoughts were dashed, and on balance I feel rather crestfallen and annoyed with myself in the aftermath. Yet that somehow doesn’t prevent me from recalling the evening warmly. For Ellen, although tasting durian no longer has the appeal of utter novelty, it has acquired a more accurate and focused intrinsic gustatory or aesthetic interest. Unlike me, then, she therefore actively desires to try it again—though perhaps not right away. As for Max, he experienced what it is like, at his current, year-older age, to have an evening meal at a fancy restaurant with grown-ups. He found out that he could actually enjoy such an evening. Max has matured in many ways—physical, psychological, and social—since that previous dinner disaster, and his cognitive and affective repertoire have grown to include new conceptual and motivational possibilities. Without his actually articulating it, a new conception of what it is to be a grown-up kid began to take form in his mind during that dinner—not the idea of being taller, or pleasing adults, or being allowed to do more grown-up things, but the idea of being a more grown-up *person*. Someone who can interact confidently with adults, on terms of greater equality—holding his own in a real conversation, showing a bit of cool and sophistication, and winning some respect for his wit. When, after the meal, I unexpectedly invited him out again, this suddenly "felt" just right to him, and he had said yes without a thought. He even finds himself looking forward to the next time out, and has begun to survey his wardrobe in an altogether new way. The appeal to Max of this as-yet inarticulate new idea of himself is intrinsic—he wants to *be* this new way, not just to please his mother or make himself happy. Without realizing it, he has begun to *respect* himself in a new way, too. Recollecting the evening, he feels a certain pride that before was unknown to him. That next evening, when the menu comes, he'll find he wants to try something he's never eaten before—

something that suggests a larger, more sophisticated world that he is beginning to long for.

These changes in desire all strike me as uncontroversial cases of *learning*—as *warranted* responses to our recent experience. I've been put in touch with certain reasons for seeking the company of Ellen *and* Max. Ellen has been put in touch with certain reasons for exploring the intrinsic aesthetic rewards of the taste of durian and, more importantly, for changing her feelings toward Max and hopes for his future. And Max has been put in touch with certain reasons for wanting to grow up, and experiencing at a new level the intrinsic rewards of conversation, conviviality, and food.

What can we say, then, about the direction of fit of desire? In one sense, desire did its "world-to-mind" job for all three of us, motivating us to change the world to bring it into accord with an idea that antecedently attracted us. But if our changes in desire are to count as learning, then they should also embody something like a gain in mind-to-world direction of fit—brought on by tracking more accurately features of the landscape of intrinsic reward and desirability that had been unknown, or misrepresented, by us. Ellen and I both take an non-instrumental interest in exploring the gustatory realm—we are not, for example, motivated merely by the desire for pleasure—and now we have a better appreciation of the aesthetics of basic taste and clearer sense of what we immediately like or dislike. But what of Max's evolution? He was stimulated by the experience to begin to form a new conception of himself. In virtue of having experienced some of its inherent qualities and intrinsic rewards, he finds himself intrinsically drawn to this conception. As a result, Max now has a number of unprecedented desires—desires that do not serve a pre-existing end of his, and that have begun to attune him to some of the most important values in human life. He hasn't put things this way to himself, of course, and he certainly hasn't *judged* that being a socially-engaged, mature, open-minded individual is "intrinsically desirable" or a good in itself. But he has begun to daydream about such a "grown up" self, to see things from this perspective, and to feel better about himself as a result. New values, as yet only dimly perceived, are starting to appear on his horizon, and, through desire, he is awakening to them.

I have also undergone changes in intrinsic desire as a result of that evening, though it takes me more time than Max to begin to see the light. I came away from that dinner pleased with the evening, but very dissatisfied with myself. This I implicitly attribute to the failure of my palate to live up to expectations. That, or so it seemed to me, is why I felt such anxious embarrassment. But a vague, restless sense of self-dissatisfaction persisted long after that meal, and somehow felt to me to be larger in scope. It was accompanied, moreover, by an equally vague but growing feeling of admiration for Ellen. When the much-awaited next dinner together arrived, it proved to be an even more enjoyable evening than the first one. This time Ellen chose the restaurant—a "slow food", locavore kind of place—and I felt much more relaxed. Somehow, I found that I really relished the simply-prepared food. Max liked his quail, and seemed to have saved up some good stories to entertain us. Over the course of the evening, however, he discovered that he likes discussing a "big idea"—religion—even better than monologuing. We made plans for yet another evening, with Max picking the restaurant. Finally, the next morning, I woke up feeling as if I, too, had made a discovery. "That's it," I suddenly thought, "*Ellen*'s the one with the right attitude toward food—not me, Mr. Aesthete! She actually *tasted* the food that night at the Singaporean restaurant—but after the first few bites, I paid almost no attention. She didn't see it as some kind of personal test for her, but was happy to explore it, taking it for whatever it was. Last night, I was simply *imitating* her—and that's why I was tasting the food instead of judging myself. Once my ego was out of the way, the flavor could come through." Mulling it over further, I began to realize that I was imitating Ellen in another way as well. This time I actually tossed the conversational ball back and forth with Max—talking *with* him, not *at* him. That's how the discussion of religion got started. "No wonder I had such a good time," I thought. "Kids can be quite something," I reflected, as if discovering a heretofore unknown continent. The appeal of these new thoughts grew stronger as I ponder them and they really sink in.

Well before my moment of revelation, I had begun to change how I acted. Following Ellen's model on the second evening came very readily—I had already become attracted to the idea of being the way she was, admiring and respecting it, wanting to be that way myself without realizing it. I recall to my embarrassment that I had previously thought Ellen's attitude toward food to be insufficiently serious and engaged. Now it seems to me that the shoe was on the other foot—my previous gustatory *machismo* now feels faintly ridiculous and has lost all luster.

Ellen, meanwhile also began to see Max in a new light on that first evening, though it would not be for some time that she realized this. Max had always been the apple of her eye, of course, and she'd always been proud of his precocity. But seeing him across the table—joking, teasing me, telling stories, sharing in the conversation—gave Ellen an implicit glimpse of something new: the idea of Max as someone out in the wider world, his own man. This idea, though still embryonic, awakened in her mixed feelings—love, pride, fear, longing, and loss. And it had begun to transform her desires even before the end of the evening. When I spontaneously invited Max to join us at the next dinner, she felt an initial jolt of annoyance and alarm—"What does he think he's doing, asking Max without talking to me first?" But when Max said yes so eagerly, she felt something quite different—not anxiety about how things would work, but pleased and proud. At that instant, it seems, Max's enthusiastic affirmation was exactly what she wanted to hear. In the years to come, this powerful new desire to see Max entering and embracing the wider world will become increasingly central in shaping Ellen's life, an end-in-itself for which she would sacrifice much, though always, it seemed, with a certain wistfulness.

Our intrinsic desires have *matured*, and not simply changed, as we metabolized the experiences of those evenings out. We are each of us now sensitive, and aptly responsive to, some genuine goods that we had never before sought, or even known to seek. This came about as we acted upon our desires, and experienced a complex mixture of satisfaction and dissatisfaction, surprise and reward. We found new things we liked, and shed old ideas that had lost their appeal. Initially, most of this took place implicitly, not through judgment, but through a complex interplay of affect, experience, and imagination. When eventually we could put a name to what we had learned, we were giving explicit expression to what we already tacitly knew, and what had already begun to shape what we wanted, and what we did.

Max, Ellen, and I emerged from our experiences with attitudes that better fit the world of value in which we lived. Not by accident, but through the nuanced schooling of experience, feeling, and unconscious thought. Our desires and beliefs changed for "the right kind of reason"—a dynamic sensitivity to the course of life that is inherent in the nature of desire and belief themselves.

Finally, our intrinsic desires have changed in another way as well. A bond of mutual affection is growing up among us. Is that, too, fitting? Are we, perhaps, discovering here yet another intrinsic good, a form of companionship none of us had previously experienced or known to anticipate? Time will tell.

"Desire aims at the good"

If something like this conception of how experience can make possible genuine learning in desire could be sustained, it would explain the persistence over the history of philosophy of the view that desires formed under conditions of greater information, wider experience, and richer and more vivid imagination have greater practical authority, and are better approximations of one's good, or of what one has reason to do.[84] It is perhaps easy to understand the appeal of such theories when it comes to instrumental desires—correcting the facts upon which they depend would seem like a clear gain in normative authority. After all, what remains to be said for an instrumental desire that isn't instrumental? But what of intrinsic desires, which are, in the end, the most important? "Ideal observer", "idealized agent", and "ideal attitude" theories would make more sense if intrinsic desires did not simply *change* as a result of exposure to greater information, experience, and imagination, but *learned*.

[84] See Sobel for refs.

Does desire therefore "aim at the good", as the ancients suggested? If 'the good' in this formula is taken as the concept under which the desirer represents the object of desire—i.e., as the substantive idea under which he pursues it—then the claim seems to me, simply, false.[85] It is quite possible to desire something without thinking of it as substantively good. Even leaving aside all the admittedly nasty things we desire from time to time, there simply are too many different ways in which ideas can inspire or appeal to us. To listen to some venerable philosophers, even the appeal of doing one's duty does not involve any concept of the act or result being good—*fiat justicia ruat caelum!*

If, on the other hand, we understand 'aiming at the good' here in the functional sense discussed in connection with belief, then we might hypothesize that human desire is evolutionarily "designed" with the core function of helping individuals aim at and achieve things that are of some value to her, her kin, or the group upon which her depends. To "design" an organism in which desire has so fundamental and pervasive a role in shaping behavior without building in some feedback mechanisms to discipline it to the organism's needs would be a recipe for extinction, not for a species that has proliferated across the entire surface of the earth. In the design sense, then, we might say that desire tracks benefits, while aversion tracks harms. Optimal foraging and mate selection, which require the fine-tuning of appetitive pursuit, serve as examples of how the affective and motivational systems of our ancestors "aimed at" benefits in the design sense. In humans, this "aiming" system came within the scope of much greater representational capacities, enabling humans to track and pursue—imperfectly of

[85] See, e.g., Velleman (dddd)

course—the requisites for survival, physical and psychic health, social cooperation, reproduction, and even prosperity across a vast range of circumstances. But the same expansive representational capacities permitted the formation of desires that had nothing to do with our health, well-being, or individual or collective survival. Many of these desires are the bane of our existence, and the power of feedback to reign them in seems rather slow and feeble in contrast to the powers of imagination and ambition that generate them. To these non-beneficial desires we must also add tendencies to desire that were beneficial in the late Pleistocene era, but often are harmful now—e.g., excessive fondness for sweetness and fat once such foods have become cheap, or an aversion to social "difference" or out-groups in the context of large-scale societies, or a taste for revenge now that weapons of mass destruction. There are many areas of life where we might despair of thinking that the learning inherent in desire will sort things out for us in some beneficial way. And the bad reputation of desire presumably owes much to these facts. The desires we *notice* are largely the ones that get us into trouble, and that we have trouble getting out of. Lost in this, however, are the legions of desires formed and acted on without much notice every hour of the day, thanks to which the average level of satisfaction with life or happiness reported by individuals in prosperous societies hovers around 7 on a scale of 10.[86] We cannot count on evolution to guarantee that our desires will "naturally" tend toward any identifiable good, but we can count on evolution to have secured within a desire a robust tendency to track a large number of everyday goods, much as belief—for all of its wild excesses—has a robust tendency to track a large number of everyday truths. It would be hopeless Pollyannaism

[86] Happiness ref.

to insist that desire necessarily aims at the good, or even at goods, but there is be more truth than falsehood in the notion that desire does very good work at putting us in touch with a large range of reasons to act.

There is, however, another, more phenomenological sense in which desire, according to our model, can be said to "aim at the good". Not because each desire involves a conception of some substantive good that the object of desire would serve, but because desire always presents its object in *some* positive light, as *liked*, *attractive*, *alluring*, *admirable*, etc. And when we act on desire we act *through* this "mode of presentation" or positive frame. Often this frame is specific enough to highlight certain determinate "desirability characteristics", but often, too, it is just a positive Gestalt—a net positive affective interest the details of which are not readily accessible to us.[87] We noted earlier that psychologists use the term *affordance* to describe the ways in which aspects of one's environment are seen as holding out certain prospects or perils for action. We can think of the affective representations built up through experience as encoding such affordances and making them available through perception and deliberation. Philosophers from antiquity onward have spoken of desires as "perceptions" of an apparent good. More recently, Dennis Stampe has defended this way of talking, and T.M. Scanlon has suggested that desires should be understood them as quasi-perceptions of an apparent reason for acting.[88] Our model of desire helps us to see why such views seem to get something right. For, according to the view advanced here, desire is not a blank drive, but an intelligible, idea-mediated striving—we "see

[87] Affective system and "holistic" processing.

[88] Aristotle, Stampe, Scanlon.

something" positive in what we desire, even if only implicitly, and even if we cannot quite say what.

Rationality in desire

Desire, as modeled here, thus is like belief in having a "self-tending" or "self-correcting" quasi-Bayesian architecture. This permits us to learn a great deal more from life than conscious, self-aware judgment could manage. As in the case of belief, there is no guarantee that this dynamic process will yield an ideal, or even nearly ideal affective representation of the world. There are too many perils and pitfalls, and there is much too much *noise* in any given person's experience. What is important, however, is the dynamic character that desire shares with belief, which makes it possible for desire to attune us to various sorts of reasons for action, so that, as the scope, diversity, and depth of our experience and knowledge grows, we have some tendency to become more sensitive to what matters and more accurate in guiding our action in ways that matter. That seems to me a form of attunement or practical rationality—*rationality in desiring*. We can enhance this rationality by taking care what we come to believe, and by leading our lives in ways that resemble, to a greater or lesser extent, Millian "experiments in living". With luck, we can thereby achieve some measure of intelligent "knowingness" in desire—some depth and breadth of understanding of the sort of world we inhabit and its possibilities, and the sort of beings we are and what makes our lives go better or worse.

Because it can enlist the services of here-and-now concrete appetite on behalf of grand, abstract, and temporally remote ideals, desire can empower one to do the work, overcome the obstacles, and resist the temptations that stand between one and the self one admires. This is the world-to-mind direction of fit of desire. But, as we have seen, desire also sets one up for the "bottom-up" process of learning or unlearning desires through the experience of acting upon them. This is the mind-to-world direction of fit *within* desire. One can even be learning *affectively* from experience while resolutely following a course in the opposite direction. In such cases, it may only be *after* one has achieved one's dream that one learns that one really wants something else, or something more, from life.

Desire may know more about us than we ourselves do, pay closer attention to a greater range of evidence, and be less easily taken in by a persuasive self-narration. In a recent series of experiments, subjects able to self-consciously deliberate did well in matching their preferences when making a choice involving trade-offs between three or fewer variables. But when the number of variables grew larger and the complexity of the trade-offs grew exponentially, they did better if their attempts at conscious deliberation were interrupted by distractions, so that they were forced to make rapid, "intuitive" decisions.[89] The intuitive "feelings" that our affective system can provide for or against certain options manifest the mind-to-world direction of fit within desire, and permit desire to be a fundamental part of our practical intelligence and knowledge, just as Aristotle suggested.

[89] Dijksterhuis et al (ref)

Functions and dysfunctions of desire

With the help of examples and a bit of tendentious rhetoric, I have tried to show how the account of desire presented here can capture a number of desire's familiar features while also imparting to desire functional capacities that might earn it a place among our rational capacities. We have focused on the key role of representation in desire, the allure of desire, its connection to pleasures of anticipation, efficacy, and consumption, its involuntary yet intelligible character, its ability to transfer the attractive and motivational force of an end upon the means to it, its role in explaining how normative guidance is possible without regress or blindness, and above all, why more experienced, more widely experienced, and better informed desire tends to accorded greater normative authority when we are deliberating about how to act—arising, I have claimed, for desire's commendable capacity for *learning* from new information in ways that promise to attune us to reasons for acting. Before turning to dysfunctions and potential irrationalities, let us briefly consider few less familiar features of desire.

First, in our discussion of belief, it was argued that default trust, as a species of positive, accepting affect, makes trial-and-error learning possible in a way that default distrust or indifference would not. A being whose default was indifference would form no expectations that experience might confirm or disconfirm; and a being whose default was to distrust its own faculties, and the content of its experience and thoughts, would be incapable of acquiring evidence from experience, positive or negative, or following any line of reasoning. Is there a similar situation in the case of desire? Is *liking* a default attitude?

Default liking would be, among other things, a spontaneous tendency to have a positive attitude and expectation toward new experience—giving rise to a default motive to explore and sample, that is, a motive for trial-and-error. Parents often complain that their tiny infant will grab at almost anything that comes within reach, and, more often than not, attempt to put it in his mouth—his most sensitive sensory organ. The baby is acting as if it had a positive expectation of each untried thing—as if, even before trying it, he had some reason to think he liked it. Of course, no such thought goes through a baby's mind—but then no such thought is needed in order for positive affective interest to regulate its behavior in this systematic way. An initially undiscriminating pattern of approach and "taking in" or sampling has many risks, to be sure. But it does promote learning from trial and success as well as trial and error. Suppose instead that the baby had, in effect, no positive expectation of untried things. Then even if the world contained many things the baby would like, or badly needs, there would no positive interest in reaching out and trying, and as a result no *default* motivation to explore and sample. Basic appetites might *drive* the child to eat or drink, but once these deprivation states were sated the infant would return to a normal default of zero positive interest in, or motivation to explore, what the world may hold. And if this infant's default instead were dislike, then he would, in effect, expect harm rather than benefit from any new experience. The arm and hand would push away, or shrink from, new objects, not reach out and grasp them. Once hunger and thirst were sated, the mouth would resist the entry of whatever was encountered or proffered, and the eyes would look away, not eagerly observe. This is not a strategy for learning what to want in life.

The notion that *liking*—basic positive affective interest—might be the default condition of our motivational system may sound strange, but not, perhaps, when put against all that we've learned in recent years about the background conditions necessary for successful learning, or communication, or cooperation. In a very wide range of cases, logic and simulation have favored a default attitude of *initially unsecured positive acceptance* or *cooperation*, which is withdrawn only if one does not receive anything positive or cooperative in return, but which again becomes the default more or less immediately afterwards. Such an initially accepting strategy makes learning, communication, and cooperation possible, sustainable, and resistant to "invasion" by more exploitative individuals. Social psychologists have found in the general population a "person positivity bias", such that the more an object resembles a person, the better it is initially liked.[90] Experimental economists have found default cooperation to be widespread in the populations they have studied, and ethnologists have found that default cooperation is typical of a large portion of the small-scale societies they have studied.[91] And there is the beginning of literature in cognitive neuroscience suggesting that the "default system" that is involved in the control and monitoring of ordinary behavior overlaps extensively with systems associated with positive affect.[92] This is congruent with the widely-documented tendencies of adults—whom one might expect to have become jaded—to declare themselves relatively satisfied with their lives, to rebound from terrible experiences, and to have a yet more positive estimation of the future relative to the present or past.[93] Kant once remarked that this positive bias

[90] Person positivity bias

[91] Varies with importance of social interaction. Bowles et al (ref).

[92] Berridge, neural basis of positive affect.

[93] Brickman, others on future. Kahneman et al on general swb.

toward the future—in effect, a positive expectation of continuing life—is the one departure from strict adherence to reason that he would not change in the human frame, aware as he was of how necessary it is to sustain human hope and enterprise.[94]

Second, just as there are two naïve notions of strength of belief, there are at least two naïve ways of speaking about desiring. Recall our recovering hospital patient. It makes sense to say of her that she doesn't *want* to eat. After all, she feels no hunger and finds the thought, sight, and smell of food awful. She does not eat spontaneously when food is offered—eating would have to be forced. Yet at the same time, doesn't she *want* to eat after she learns it is a precondition for discharge from the hospital? And wouldn't she be frustrated and annoyed if she couldn't? Eating still must be forced, but now it will be forced by her. Once we distinguish a bare appetite for food, which she lacks, from a desire to eat, which she has, there is no paradox here. A desire to eat is a matter of being attracted to the *idea* of eating food, which does not require that one be attracted to the food itself, or even the idea of it. Our patient is strongly attracted to the idea of escaping the hospital, and this has cast a favorable light upon eating, regardless of appetite. Because her desire to be home is so strong, and her desire system is working as usual, this favorable light is sufficient to lead her to want very much to eat, overcoming even a disinclination so strong that she almost chokes when she tries to swallow a mouthful of the gluey mashed potatoes or squishy cooked carrots.

Third, because desire is mediated by a representation, competing desires can be outweighed without being cancelled, also resulting in a state of "wanting and not wanting". Ellen could truthfully whisper to me at the beginning of our evening, "Max

[94] Ref. Kant on future

really didn't want to come, but he's being a good boy." Max, at the beginning, hates the idea of fancy dinners at restaurants in the company of no one but adults, and so, *under this description* or *in view of these aspects* of the event, does not want to come and would not spontaneously do so. Yet Max also loves his mother, and so loves the idea of helping her. Under *that* description and in view of *this* aspect of the event, he very much wants to make her dinner possible by coming along. As a "good boy", Max's love for Ellen so much exceeds his hatred for fancy dinners that it is this idea that triumphs and that he feels best about. Max is also a "good boy" in the sense that a younger or more impulsive child might, despite strong love for his mother, not manage when the time comes to resist his powerful aversion to going, made much more salient by leaving the house and heading for the bus—he would dig in his heels at the last minute, or storm and protest if dragged along, spoiling the evening and perhaps feeling very bad about this later. Max, by contrast manages to swallow such impulses much as our hospital patient managed to swallow her food—under the guidance of a powerfully attractive idea. Yet his strong aversion will persist—after all, the considerations favoring going merely outweigh, but do not cancel the considerations against it—and continue to shape his behavior in noticeable ways: in the energy with which he will walk with Ellen to the bus stop, for example, or in the eagerness with which he will view my arrival on the scene and the alacrity with which he will greet me. Despite the persistent conflict, Max will feel "in control" of himself and "freer" than an earlier self who would tantrum and then regret it. Instead, Max has managed to do what he most liked the idea of himself doing, what he "feels best" about. No judgment of *duty* or *rational requirement* was necessary for this eminently rational—and "self-regulating"—bit of behavior, just a strong, clear desire to help his mother, and to be the sort of boy who, thanks to love, can do that even when he "doesn't want to".

This example of successful reasons-sensitive self-regulation through desire can, however, serve to introduce us to cases in which desire fails in its characteristic functions and yields what we typically think of as irrationalities in desire. Consider an "unwilling addict" who is habituated to a drug to the extent that it no longer produces in him any noticeable pleasure, only occasional relief from the gnawing torment of withdrawal. He hates what has become of his life—the oppressive, degrading demand to find money for the next fix has cost him his job, and led him to abandon his children, rob from his friends, and, eventually, sell himself as a prostitute. He lives in disgrace, having made a wreckage of everything he loves and driven away anyone who ever loved him. There is no euphoric pleasure of a drug-induced high to compensate for any of this, just a temporary release from the bondage to merciless drive, which returns as soon as the drug wears off. He dreams of a life free of the drug, longs to return to his children, and admires and envies fellow addicts who have become clean and free again. Yet it cannot be denied that he *wants* the next fix, or that, in seeking the fix, he is acting on his strongest motive. Or consider an obsessive-compulsive who hates the way her obsession with avoiding germs has cut her off from friends, her colleagues, and even her family. She feels trapped in a prison of her own making. True, when she acts on this obsession, she is acting on her strongest motive, what she "most wants" to do. But when, for example, she feels this powerful urge coming between her and her own child—when, for example, she finds herself desperately wanting to escape the touch of her child's hand, which she cannot help but visualize as teeming with horrible germs—

she feels anger and frustration even as she withdraws her hand. Earlier, before her compulsion took over her life, she had a positive passion for cleanliness and found it rewarding to achieve. Now she feels like a drudge obeying an alien task-master—and sees only a hated, bleak, sterile life stretching out ahead of her. She marvels that her life was ever easy or "normal", and wishes nothing more than that this easy relation to the world and the people around her could return.

Cases like these have been used by philosophers to argue that desires in themselves cannot be reasons for action—for are not all these individuals doing what they *most want* to do? And so would have most reason for doing if the desire-based theory were right?[95] Yet we do not see them, nor even do they seem to see themselves, as having good reason to continue to act as they do. On the contrary, they would more likely describe themselves as acting irrationally.

What is striking to me about these cases, however, is that they are examples where the normal regulative and feedback mechanisms characteristic of desire have broken down. In the first place, there is a "dysregulation" or "uncoupling" of affect and motivation that should be experienced as a loss of *control* over one's own actions. Actions that the agent sees with little or no positive affection, but rather with the sort of powerful disaffection that normally would result in strong aversion, remain stubbornly, overpoweringly *wanted*—a motivational drive out of proportion with, and increasingly unresponsive to, any "positive light" in which the individual sees his or her actions. Owing to induced drug tolerance, the drug addict will tend to seek larger and larger doses, taking a greater and greater toll on his life, even as the positive effect of

[95] Similar cases have been used against theories of freedom of action expressed in terms of "acting from desire". The addict and the compulsive are "acting from desire", it is said, but clearly aren't free. We will return to this point, below.

the drug pales dramatically. In such circumstances, normal feedback would *reduce* one's tendency to seek the drug—the way that, if the pleasure and excitement of a novel restaurant or activity dramatically pales, one will tend to seek out other places or other things to do. This "modulation" cannot occur, however, because the drug has in effect hijacked a portion of the addict's motivational system and given it a life of its own. Instead of feeling that he is getting his heart's desire, or living the life he'd *like* to live, the addict will likely feel that his life is out of his control, run not by an unrelenting "monkey on my back". For rather different reasons, an obsessive-compulsive individual has dispositions to act that she cannot control, and cannot alter by learning or reasoning. She may have come to distrust the urges and images that drive her to wash her hands until they are raw, or withdraw from physical contact, and to detest a life of surrendering to them, and yet this mass of negative affect seems powerless to regulate the strength of her drive to perform and re-perform the "rituals" of her compulsion.[96] Neither does her compulsion respond to the most basic sort of feedback from experience—its strength is unreduced by an experience of compulsion, emptiness, and drudgery that would normally sap positive motivation, neither can its force be redirected toward other activities she finds more pleasant. She and the addict may be acting in accord with the demands of their strongest wants, but neither, so far as I can see, are living as they most *desire*. The normal *coherence* of behavior arising from the regulative structure of desire—of acting under a positive idea of what one is doing—has been thoroughly disrupted, and the normal *intelligence* of behavior arising from the feedback structure of desire—of seeking what one has learned to like and avoiding what

[96] Contrast obsessive-compulsive personality disorder. Usual DSM caveats.

one has learned to hate—has broken down. Small wonder they experience their addictive and compulsive actions as out of their control, unfree, irrational.

With this thought in mind, let us return to Warren Quinn's well-known example, mentioned at the outset, of the man who finds himself with an inexplicable urge to turn on radios whenever he sees one, but who can say nothing, even to himself, to explain the point of doing so.[97] Recall that he sees nothing positive in turning radios on, doesn't look forward to doing so, and feels no pleasure or sense of accomplishment when he succeeds in doing so. He isn't even interested in hearing what's on the air. Moreover, competing considerations, such as whether it would be rude or inappropriate to turn on some else's radio, or whether he has other goals with which this behavior would interfere, seem to have no effect on this urge or its manifestation. Still, he plods ahead, methodically switching radios on at every opportunity. What keeps him at it, then? Simply that, at the moment he comes across a radio not in use, this is what he finds himself most *wanting* to do.

From the description, Radio Man resembles an obsessive-compulsive in that he has a predictable, repetitious, irresistible urge to enact an invariant behavioral routine in response to certain circumstances. Unlike many obsessive-compulsives, however, he acts with no favorable gloss at all on what he is doing—he is not avoiding infection, keeping things neat, eliminating fire hazards, or perfecting a history of everything. His action may have more in common with a motor disorder, in which an individual may feel an almost irresistible urge to perform some fixed action sequence—to repeat what others have just said, or to continually retrace her steps while walking—regardless of its

[97] Quinn (1993).

appropriateness to her context or purposes, receiving as gratification only a feeling of release of tension, a "drive reduction", once the behavior has been performed.[98] Perhaps this is how it is for Radio Man, who enacts a contextually-invariant action sequence under the force of a featureless "want".

Is this anything like *desire* in action? Where is the favorable idea through which he acts, that would regulate his wanting, give intelligibility to his behavior, and provide a representation of what he is doing that could enter into deliberation? Where are the pleasures of anticipation or consumption? Instead we find a blank, intractable urge—all push and no pull, providing neither foothold for reasoning nor the possibility of learning through feedback. Of course, we all have some habitual actions we perform without much by way of positive conception—twisting paper clips, popping bubble-pack, doodling, and so on. In general, there is nothing dysfunctional about such habits. If one finds it even a little bit diverting, or engaging, or satisfying, or interesting, or better than doing nothing, then doodling through a boring meeting is a perfectly sensible thing to do. But what if someone found *nothing* in twisting paper clips, had no idea why he did it, would do it even when this involved reaching over onto his boss' desk while she was talking to pick up a paper clip and start bending it? Such unmodulated behavior, insensitive to context, devoid of interest, and not answerable to one's other concerns, strikes me as closer to a motor disorder than to normal, desire-based behavior. It would be unsurprising if Radio Man saw this as a peculiar and isolated *glitch* in his motivational system—"I just have this irrational thing about turning radios on …"—rather than an expression of his own desires.

[98] DSM

One's affective attitudes—what one likes, loves, hates, dreads, hopes for, is interested in, cares about, detests, respects, or aspires to—have a much closer relation to one's self than one's mere urges or bare wants. In cases in which one's motivation fails to be regulated by one's affective attitude because an almost-overpowering motivational urge pre-empts such regulation—as in reluctant addiction, compulsion, or motor disorders—the agent may well feel as if she has diminished control over her acts, that she is in some measure "unfree", "in the grip of" of this drive, unable to "do as I like" or "be as I would like to be". This feeling of compromised control can persist even though, in surrendering to the impulse and thereby obtaining release from the oppressive tension of such an urge, the agent may be acting on her strongest want at the moment. For acting in this way may be very far from what she in her heart most desires, or fervently wishes for.

Run-of-the mill "weakness of will" likewise often pits intrusive inclination or situationally-induced wanting against heartfelt desires and fervent wishes, though less chronically and dramatically. The urge, when hungry and presented with salty, fatty food, to eat to excess; or, when full and presented with a creamy, chocolaty, sweet dessert, to eat much more than one intended; or, when work has become difficult or daunting and an easy chance to shirk or postpone it is presented, to slack off or allow others to do it; or, when facing the future and presented with a quick immediate gain, to take it at the expense of a greater gain later; or, when aroused and presented with a chance, to engage in an inadvisable sexual liaison—such "temptations" yield local, focused, hard-to-resist motivation to act in ways contrary to one's stable, longer-term desires and more deeply-rooted affective attitudes and attachments. Such psychic conflict is ubiquitous, and reflects the situational opportunism we inherit from eons of evolution under scarcity, when rich loads of calories, fat, and salt were hard to come by, energy typically needed to be conserved, more distant future gains were often highly uncertain, and surreptitious ways of doing less than one's share, or reproducing outside one's primary relationship, effective ways of leaving more copies of one's genes. The menu of our most popular "weaknesses of will" is thus far from arbitrary from an evolutionary standpoint. But many of its entries are quite arbitrary once we find ourselves in an affluent, technologically-advanced society where food and physical energy are not limiting conditions, and many social structures are in place that make short-term sacrifice for the long-term a good bet. What had been sources of opportunistic adaptive advantage have instead become leading causes of ill health, disability, and economic precariousness.

What resources do we have to counter such opportunism now that it is often so counter-productive? We have already discussed the important phenomenon of self-regulation through norms, and the contribution of affective regulation to making normative guidance possible. The mere possession of principles, without the backing of powerful affect, would leave us with little to resist the pressure of opportunistic urges. On the other hand, those with a passion for living up to their principles have shown themselves able to mobilize spectacular motivational force in the face of threat, privation, agony, and even death. Resisting everyday "weakness of will" requires no such heroism, and often, so far as I can tell, involves no normative judgment or appeal to principles. It is enough to feel such weakness to have stable, long-term desires—to be fit and healthy, to complete one's work or advance one's career, or to help, not hurt, those one cares most about—that comes into competition with the insistent motivational force of more immediate inclination, ease, and gratification. We face such conflicts many times a day, typically without much fanfare: we resist the urge to consult e-mail or browse the web in order to complete the day's work, we walk up a few flights of stairs in the parking structure rather than take the elevator, we forgo purchasing any of the candies spread colorfully before us at the store check out when we stop to buy milk on the way home, we put down our own work in order to read a bed-time story to our child, and we turn off the late-night TV show in order to get a decent night's sleep. Those most effective in resisting such everyday temptations, and managing to forge ahead despite distractions, I suspect, are those in whom longer-term desires, affective investments in work, and attachments to family are *strong*, not weak—real *desires* rather than mere wishes or abstract principles. That is no doubt part of why Aristotle placed such emphasis on the cultivation in the virtuous individual of the right desires, and in the right strength, so that she is led by *boulesis*, rather than mere *epithumia*. Indeed, if we follow Kant in locating the "will" in the faculty of desire (*MM*: ref., *CJ* ref.), then we can say that strength of will actually *is* strength of desire. For example, in the case of moral conduct, it is "moral feeling"—a "predisposition … on the side of feeling" (*MM* 6: ref) to love humanity as an end, and respect the moral law and oneself—that renders Kantian agents capable of acting directly in the face of the powerful incentives of inclination and self-interest. Rather than see "weakness of will" as paradigmatically a contest between Judgment and Desire, then, we should see that, in its most fundamental form, it is a contest *among* the things that move us, in which that which we most strongly and enduringly care about is at risk from what, at the moment of acting, we most strongly and immediately want, and find gratification in.

Historical models aside, "weakness of will" as it has come to be discussed today is obviously a complex phenomenon, potentially involving as many dimensions and layers as human thought and feeling themselves. My goal here is not to reduce all "weakness of will" to the level of first-order conflict between desire and inclination. Rather, it is show how much of "weakness of will" can be found at that level, and thereby to confirm a prediction of our model of desire. For, according to that model, the regulation of much or most of our everyday action takes place *through* the normal functioning of the affective regulation present in desire. If that normal regulation is disrupted by intrusive, unregulated motivation, this should be experienced as a *failure* of the self-control and control over action present in normal agency, even without invoking any higher-order thought or principles. In countering such "weakness", strong desire can be the key to "strength of will", not the enemy of it.[99]

If desire is composed to two attitudes, with two associated notions of strength of desire, then shouldn't there also be two notions of weakness of desire? We've just considered a range of cases in which desire proves weak in the sense that the affective regulation of motivation characteristic of desire fails in the face of inclination. But desire can also be too weak because affect itself fails, so that the normal processes of

[99] Failures of "full rational self-control" need not be failures in *net* rationality. Sometimes "weakness of will" is all that stands between us and committing an error, injury, or atrocity in the name of some "higher principle" or demand of social propriety. When our perceptually-driven, local inclinations save us from our "higher selves", we may in fact be doing a *better* job of acting in tune with the real reasons at stake in a situation. A parent's unconscious empathy for a confused and frightened child might, despite his principles, stay the hand and spare the rod. So the mechanism by which such inclinations can displace "top down" regulation of motivation by higher purposes is an important asset of practical intelligence and rationality. But such displacement will still *feel* more like personal defeat or failure than rational control. And that is the phenomenon we are here concerned to explain.

feedforward and feedback control break down. In severe depression, for example, individuals suffer a chronic deficit in positive affect. One result is that they often are unable to generate sufficient motivation to carry on with their normal lives. A depressive individual may find that she has lost all interest in her work, or in a cause to which she had until recently been deeply devoted, even though she has learned nothing in the meanwhile that would count as *evidence* against this work's or cause's continuing value or importance. She may start failing to take care of herself or meet her own basic needs, be unable to sustain basic family responsibilities or relations of friendship, and become hopelessly irresolute in the face of choices, losing the "will" to decide and act. Effective motivation and guidance of action drains away as affect drains away. Desire and learning alike become attenuated, as depressives lose interest in life and often suffer a marked decline in cognitive ability and experience great difficulty remembering novel information. Residual negative affect instead drives aversion to taking resolution action, and "motivated cognition" in the form of repetitious negative self-evaluation and thought.

Corresponding to this sort of excessive, undiscriminating weakness of motivation is the excessive, undiscriminating strength of motivation found when positive affect so floods an individual's mind that she becomes manic. She may become enamored of, and convinced by, whatever idea pops into her head—however little sense it might make to others, or would have made to her former self. She might agree to marry a person she has barely met, or walk into a sporting goods store and spend her entire bank balance on skiing equipment, even though she has never liked anything remotely like skiing, and is under doctor's orders to be careful with her knees. Without any fresh evidence, without any real learning, her desires have undergone dramatic changes, and her behavior becomes grandiose and erratic. Rendered largely insensitive to negative feedback by overflowing positive affect, the manic individual will persist, sure of herself and her "intuition", in evidently ruinous schemes.

If, as in our model of desire, motivation and action are typically regulated by affect, then when affect goes haywire, so will motivation and action. The individual may lose the contact with her actual situation, capabilities, and prospects that is afforded by the modulating effect of experience upon affect. Instead, she will behave in ways we readily identify as unwarranted and irrational. Since affect is our principal currency of evaluation, severe disruption of the affect system, whether in a positive or negative direction, will tend to induce severe and systematic distortion in one's evaluations, and ability to estimate or appreciate value. When such distortions arise, reasoning and judgment may have little power to restore normalcy, since these mental operations themselves depend for their operation and force upon requisite, well-modulated levels of affect, and, in particular, trust. The depressive loses trust in his judgment and reasoning ability at just the time when judgment and reasoning are most needed to stop the downward spiral of self-defeating, self-absorbed rumination, and the manic becomes so undiscriminating in what he trusts that his judgment and reasoning lose precisely the discipline they would need to moderate, as opposed to accelerate, his flights of fancy. Our capacity for practical rationality thus depends heavily upon a well-functioning, well-modulated affective system—without it, thought itself can become the enemy of reason.

Intrinsically irrational desires?

On the present account, desires are seen as *intimating* reasons for action, in much the same way that perceptions *intimate* an external world of entities and features. Each state presents an “appearance" with a certain content, which represents something other than itself as being a particular way—a red apple is situated some five feet away, say, or, that selfsame apple is appealing to pluck and eat. The result is a “perceptual affordance”, which ordinarily gives rise to some motivation to act according—to pluck and eat (or save, if one is not now hungry) the apple—and some expectation that the result will be well-liked. We tend to act by default on these affordances, without awaiting some further validation of what we are about to do. What affordances we find in a situation, what we perceive as real, eligible, or well-liked, is in part the product of past experience—learning to distinguish apples from other sorts of hanging fruits, and learning that one likes apples of this kind when they are ripe, that one doesn’t get stomach-aches from eating them, that they help reduce hunger, and so on. In most of us most of the time, such perceptual affordances are the result of the normal functioning of our belief and desire “systems”, a reliable set of mechanisms for attuning us to genuine features of the world and genuine reasons for action. Being disposed to take these affordances at face value, to rely upon them by default in thought and action, is a core case of acting for reasons. Of course, as we have earlier stressed, neither the perception nor the desire *itself* is the *ground* of our reason for acting desire—we act *through* perception and desire toward the objects and features they place before us and frame as real or attractive—yet being disposed to act by default on perception and desire is a precondition for responding aptly to the reasons they intimate without regress.

To this point, I have focused on the idea of rationality *in* belief and desire, conceiving this idea dynamically. I have tried to show that processes built into the architecture of belief and desire could have the effect of attuning us to reasons to believe and desire, and thus fostering our practical rationality. A natural challenge to any such account would be to show that these processes can generate arbitrarily large numbers of beliefs and desires we would deem irrational in themselves, or irrational to rely upon in action. Of course, all actual processes are in some measure fallible, and in an open-ended variety of ways. So the question is whether there is anything in the processes subserving rationality in belief and desire, according to the present account, that would tend to operate *at all* against paradigmatically irrational beliefs or desires. Let us focus on the case of desire, since that is bound to be the more controversial one.

Consider the following two examples of paradigmatic irrationality in desire: someone who does not want to avoid horrible pain, and someone who is indifferent to what happens to him on future Tuesdays.[100] Suppose that someone began with these as basic desires. Would the processes discussed here have any tendency to work against the reinforcement or retention of these desires? Horrible pain is an experience with a very definite inherent quality. Part of this quality is that one cannot help but want this experience to stop—it is intrinsically aversive and disliked, as anyone who has ever been caught in the searing grip of its iron talons will attest. (Of course, there can be good reasons for undergoing horrible pain, but our concern is simply with the pain, taken in

[100] The examples are, of course, Parfit's.

itself.) Suppose, for example, that someone who has never experienced severe pain declares himself indifferent to it. Still, possessing this attitude would have little or no tendency to prevent horrible pain, should it arrive, from being intensely, intrinsically aversive and disliked. His neutral tacit expectation or "affective forecast" of pain would be starkly at odds with the felt quality of his actual pain experience, and this would normally cause an immediate, downward adjustment of the original affective attitude of indifference. He would begin to dislike the idea of being in horrible pain, and this would induce a wanting to avoid it, other things equal. Something similar would be true of anyone who professed indifference to experiences of all kinds on Tuesdays. If possessed of a normally operating desire system, she would enter Tuesdays with a neutral affective forecast, but soon would be surprised to find that what happens to her on Tuesday is felt just as much as any other day. Once again, having a neutral attitude would not guarantee that whatever happened on Tuesday would in fact be indifferent to her. Instead, she would find that positive and negative things affect her just as much Tuesdays as other days. Through normal feedback, then, her neutral affective attitude toward future Tuesdays would tend to be undermined, so that she would dislike the idea of horrible pain on Tuesdays and Wednesdays (Thursdays, Fridays, etc.) alike. Only a dysfunction of her normal desire would prevent this inherent tendency to readjustment, or attunement, of prospective affect and expectation.[101] Someone in whom such failure of learning were chronic, would indeed be showing a recognizable kind of irrationality in desire and, likely, action as well. Like the woman who laments, "Why am I always

[101] Of course, if there *were* someone whose Tuesday experiences for some reason were always felt to be neutral, then it would not seem at all irrational for such an individual to prefer that normally wretched experiences take place on Tuesdays, and normally rewarding ones on Wednesday through Monday.

attracted to the wrong kind of man—the ones who lead to nothing but heart-break and pain? Why can't I ever seem to learn?"

Belief and desire in action

In normal belief and desire, I have been arguing, learning goes on constantly at the personal and subpersonal level, a process that depends upon the general health and responsiveness of our affective system as a whole as well as the diversity of experiences we seek out or encounter. We may begin life with a fairly narrow band of native likings and drives, as well as certain native tendencies to trust and expect. However, we soon move beyond these—not only by acquiring new, instrumental attitudes, but also by acquiring new, intrinsic aims. When the affective learning system that underlies such learning undergoes atrophy or hypertrophy, as in a mood disorder like depression or mania, belief and desire show alike undergo a loss in discrimination and reliability. Likewise, vagaries or limitations of experience can leave even the most sensible person with bizarre beliefs and regrettable desires. But in normal conditions, when all is working well, belief and desire function together as a system to keep us reasonably oriented and active in the world, exerting shared "feed forward" control over action by regulating what we want and what we expect, and undergoing shared "feedback" as experience reshapes what we like and trust.

Belief and desire are "made for" each other, and made for action, as Stalnaker's characterization made clear. Working together, their normal dynamic operation supplies not only energy and direction, but also a *conception* of what we are doing that casts it or its aim in some positive light. They thus give our conduct an *intentional object*, and make it into intentional, intelligible action, even in the absence of a pre-formed plan or intention. That helps us fill a gap in our account of intentional action, since we need to be able to see how the bulk of our moment-to-moment activity could be intentional—e.g., could be such that we have a ready, non-inferential answer to the questions, "What are you doing?" and "Why?"—even though we seldom actually do form intentions. When we act on belief and desire we act *through* an idea of what we are doing, an idea that we in some measure trust and that has some positive affective interest for us, giving rise not only to motivation to act, but also some expectations about what is going to result. Moreover, the control structure of belief and desire result in an inherent tendency to *monitor* whether expectations are being met as we act, and to adjust what we think and do as a result. Thus belief and desire do not simply *launch* action, they *guide* it as it unfolds and give it a genuinely teleological character—unlike the archer who sends an arrow toward a target but ignores where it lands, and never readjusts her aim. Thanks to the strategic position of affect in our psychic economy, belief and desire are far from inert representations, but rather function to mobilize and guide the full range of elements present in effective, intelligent human action: attention, perception, inference, association, memory, motivation, comparison of outcome with expectation, and consequent modulation of feeling and focus. The operation of this coordination of faculties and activities can be, but need not be, conscious. Implicit beliefs and implicit desires can give rise to implicit *yet intentional* action. Depth psychologists have long argued that implicit agency must occur in order to explain a wide range of phenomena, a form of explanation that has been resisted in many quarters out of suspicion that it requires us to posit a "second self", a separate, implicit, inner agent. But we can now see that no such posit is needed in order for unconscious intentional action to occur, and to play a role—perhaps pervasive and fundamental—in shaping our behavior.

We act under some ideas and through certain feelings that are fairly reliably accessible to us, and others that are not. These ideas and feelings, conscious and unconscious, give shape and meaning to what we do, and make it the case that there is a story to our lives. But for beings as complicated and conflicted as we are, and psyches as layered and full of opponent processing, this story is unlikely to be a single narrative strand. Even so, it will be full of the special sort of meaningful striving that is the arc of desire. Life over time is sequence of strivings on many fronts, in which layers of overlapping arcs, like the overlapping fibers of a rope or cable, produce a connected whole even though none of the parts is the same throughout, and none is the true core. Our practical rationality, insofar as we manage it, is almost certainly due as much to those strands we do not actively call to mind, and that are very largely self-monitoring, as it is to those that happen to come into conscious focus in our mental foreground.

The fact that belief and desire, as understood here, can unite to produce intelligent, intentional agency *directly* is fundamental to our solution to the problem of regress in agency that threatens the very possibility of rational action. For this permits agency to be a *self-organizing* phenomenon, a product of non-agential states that, like belief and desire, can coalesce around a trusted, well-liked idea to crystallize coordinated activity aimed at translating this idea into action (see fig. 3, below). The alternative to this *emergent* idea of basic intentional agency is to posit an inner agent, a homuncular "will", already organized as an agent, who does the intending and organizing for us. And whose capacity to be aptly responsive to reasons we do not seek to explain, but rather take on faith—a philosophical dead end.

If, instead, we can understand how basic agency can emerge from non-agential states such as beliefs and desires, then we can also begin to understand how higher-order, self-conscious forms of agency become possible once beliefs and desires can take as their immediate objects representations of a self-reflexive, meta-representational character. Beliefs about what we believe, and desires about what we desire, permit us to engage in expressly self-conscious deliberation, choice, intention-formation, and planning, and to impart to our actions a new and very powerful kind of organizing force and meaning, essential to our full flowering as rational agents. However, even once it has come onto the scene, self-consciously deliberate, intended action does not supplant, but rather at every step depends upon, basic, non-deliberated intentional action. It must be possible to deliberate intentionally without forming a prior deliberate intention to do so, and to respond to reasons in deliberation and choice without presupposing a reasoned mediation of each step or a "choice of reasons" for each decision. Instead, thanks to the affective representations upon which they depend, our basic, self-organizing intentional processes are themselves intelligent and capable of learning, of attuning us to reasons through thought and experience, trial and error—we need not avoid regress only by becoming stupid and incorrigible.

It is important to emphasize, however, that this model of "self-organized" action need not, and perhaps does not typically, make intentional action *opaque* to the agent. Desire furnishes a positively-valenced representation of what is sought, and belief furnishes confidence that this sought-after condition is available through action. Thus there is a recoverable "idea" under which the agent acts, which shapes the nature of the action and gives it some point, while desire and belief also monitor whether things are going as expected. This is, we might say, desire and belief regulating action "in the right way" to constitute it as intentional. And the agent can thus be in a position to answer the Anscombean question, "What are you doing?" as well as its natural successor, "And why?"

Aristotle made the requisite distinctions between merely appetitive behavior and action "under an idea"—"brutes have sensation, but no share in action" (*NE* 1113a20). But he also realized that ideas alone are not enough:

> Thought by itself, however, moves nothing; what moves us is thought aiming at some goal and concerned with action. … Now desire is for the goal. Hence choice is either understanding combined with desire, or desire combined with thought; and what originates movement in this way is a human being. [*NE* 1139b1-6]

Or, as Kant wrote, with equal sweep:

> The **faculty of desire** is the faculty to be, by means of one's representations, the cause of the objects of those representations. The faculty of a being to act in accordance with its representations is called *life*. [*MM* 211]

6. Conclusion

A plausible approach to rationality in belief holds that we cannot tell the rationality of a belief solely from its content. True beliefs that are held strongly on the basis of poor information or fallacious arguments can be irrational; and false beliefs held strongly in response to significant evidence or owing to good arguments can be rational. Even beliefs in logical truths can be held illogically—not owing to a grasp of the proposition's self-evidence, say, but from reliance on a fallacious line of thought or deference to a putative authority one in fact has strong evidence to suspect.

It is better to speak of rationality *in* belief—in the process of forming, revising, or discarding beliefs—for that does seem crucial in understanding when beliefs are rational, whatever their content. What makes for rationality in belief, then, is a *dynamic* matter, less a question of where one starts than of how one goes on to respond to new experiences, arguments, or ideas. Default self-trust—in one's faculties, and in the content of one's experience and thought—is essential for getting rational responsiveness going. Defeasibility through feedback—unexpected experiences, inconsistent thoughts, inexplicable incoherence—is essential for keeping rational responsiveness going. Fortunately, though hardly by accident, human believers seem to be set up this way, thanks in large measure to the very architecture of belief itself. As a result, we can escape reason-seeking regress, and go about living and learning. The most rational among us do these latter tasks best, sensitive to all that has gone before, but still maintaining open and active minds, eager to try out new ideas, and attentive to how well this goes.

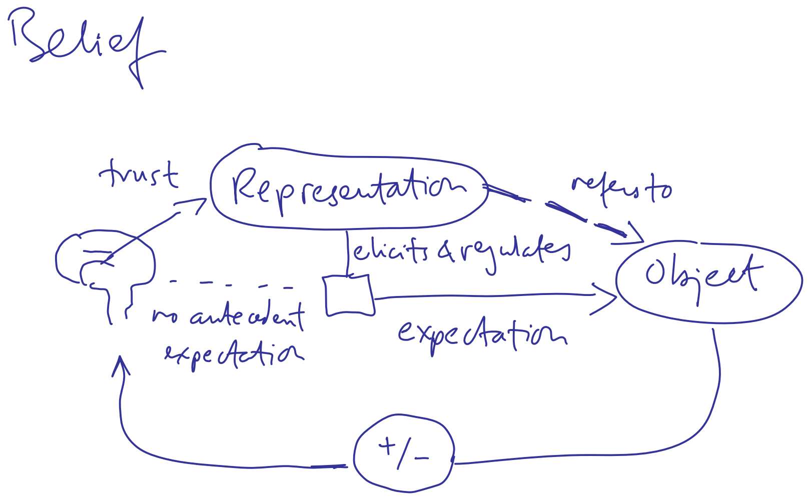
A similar picture, I have argued, can be given of rationality *in* desire. The rationality of a desire cannot be judged solely from its object. A desire for a genuine good that is exaggerated or obsessive, or that is the result of delusion, or that is so inflexibly lodged in the psyche as to owe nothing to experience and be immune to reconsideration, can be irrational; and a desire for something not at all good that has arisen from first-hand, positive experience, or from reflection involving false but justified beliefs, can be rational. Desirers, like believers, display their rationality dynamically—by their sensitivity to past experience, and by continuing to experiment through living, seeking to expand their knowledge of themselves and their world, trying out interesting and attractive ideas, and being attentive to how things turn out. To start this process on its way, desirers, too, must show a degree of default self-trust. They must be disposed to accord some measure of underived, default authority to whatever they find themselves initially liking or attracted to, without awaiting rational certification. But, as in belief, if desire is to be responsive to reasons this default must be defeasible when expectations are not met. For this to be possible, desirers must to some degree trust the content of their experience and their thinking and memory. Fortunately, and again not by accident, the architecture of human desire directly facilitates this dynamic of default, defeasible motivation. In this way, desirers, too, can live and learn.

Of course, we are all believers *and* desirers, and, happily, all enjoy some measure of rationality in both. The combined, interactive default, defeasible dynamic that results makes it possible for us to enjoy some measure of rationality in action as well—some degree of practical attunement to reasons to act. And yet, in the philosophical story of rational action, desire has long stood on the outside looking in. Perhaps it is time to change the story. Perhaps, instead, desire has all along been an integral part of our practical intelligence—contributing vitally, with but few thanks, to our ability to discover, and respond aptly to, the reasons of our lives.[102]

[102] This paper has been very long in the making, and has benefited from the comments and criticisms of so many colleagues and audiences that I despair of according proper recognition to everyone. Apologies in advance to all. Special thanks are due to the faculty and students in the Departments of Philosophy at Stanford (whose invitation back in 1999 got me started on this topic, and whose reception of my initial efforts encouraged me to continue), as well as the Australian National University, Bergen, Georgetown, Iowa, Maryland, MIT, Michigan, Minnesota, NYU, North Carolina State, Ohio State, Oslo, Ecole Polytechnique, Purdue, Rice, Rome-La Sapienza, Rotterdam, Texas-Austin, Texas Tech, Union College, and the Institute Jean Nicod (Paris) and Center for the Study of Mind in Nature (Oslo) among others. I am also especially grateful to the following individuals: Elizabeth Anderson, Nomy Arpaly, Joshua Berke, Kent Berridge, Paul Boghossian, Linda Brakel, Michael Bratman, Aaron Bronfman, John Broome, Sarah Buss, Monique Canto-Sperber, Ruth Chang, David Copp, Jonathan Dancy, Stephen Darwall, Wayne Davis, Jean-Pierre Dupuy, Gabriel Fagnière, Allan Gibbard, Don Herzog, Thomas Hill, Douglas Husak, Nadeem Hussain, Michael Jacovides, Frances Kamm, Christine Korsgaard, Richard Kraut, Douglas MacLean, Jeff McMahan, Tito Magri, William Miller, Richard Nisbett, Howard Nye, Sven Nyholm, Marina Oshana, Derek Parfit, Philip Pettit, Wlodek Rabinowicz, Frederick Rauscher, Joseph Raz, Henry Richardson, Geoffrey Sayre-McCord, Victoria McGeer, Jacob Ross, Thomas Scanlon, Mark Schroeder, Timothy Schroeder, Norbert Schwartz, Michael Smith, Chandra Sripada, Jason Stanley, Sigrun Svavarsdottir, Larry Temkin, David Velleman, Gary Watson, and Susan Wolf, among others. I'd also like to thank the the American Council of Learned Societies, the Solomon Guggenheim Foundation, and the National Endowment of the Humanities for support.

Figures

Figure 1:



14

Figure 2:

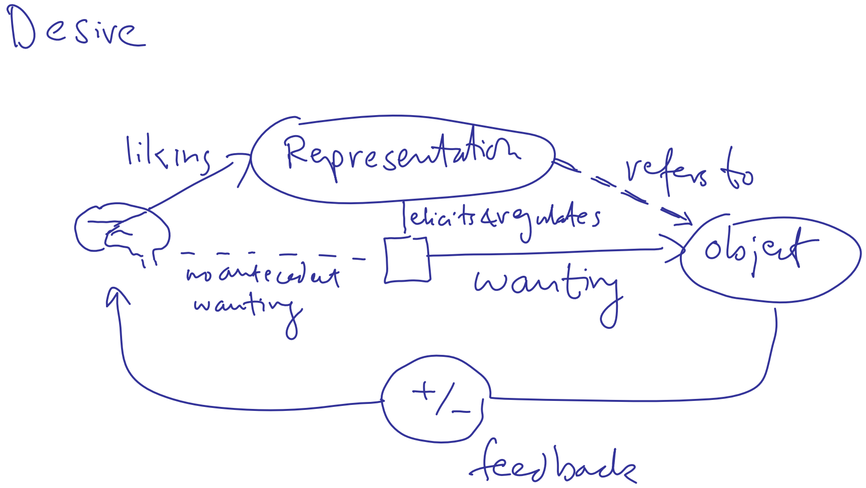
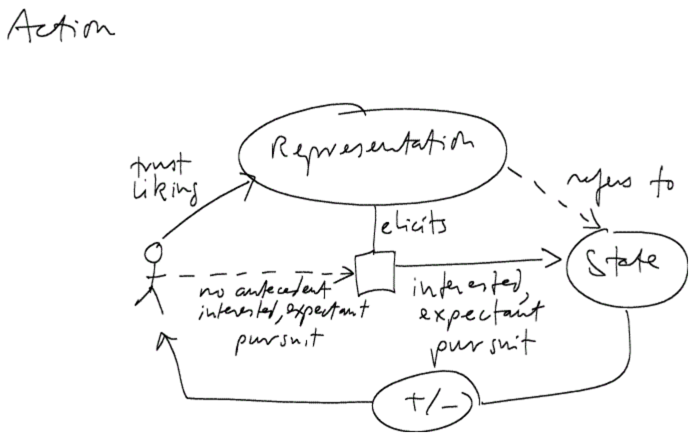

Desire
liking
Representation
refers to
elicits & regulates
object
no antecedent wanting
wanting
+/−
feedback

Figure 3:


Action
trust
liking
Representation
refers to
elicits
State
no antecedent
interested, expectant
pursuit
interested,
expectant
pursuit
+/–

62

Figure 4:

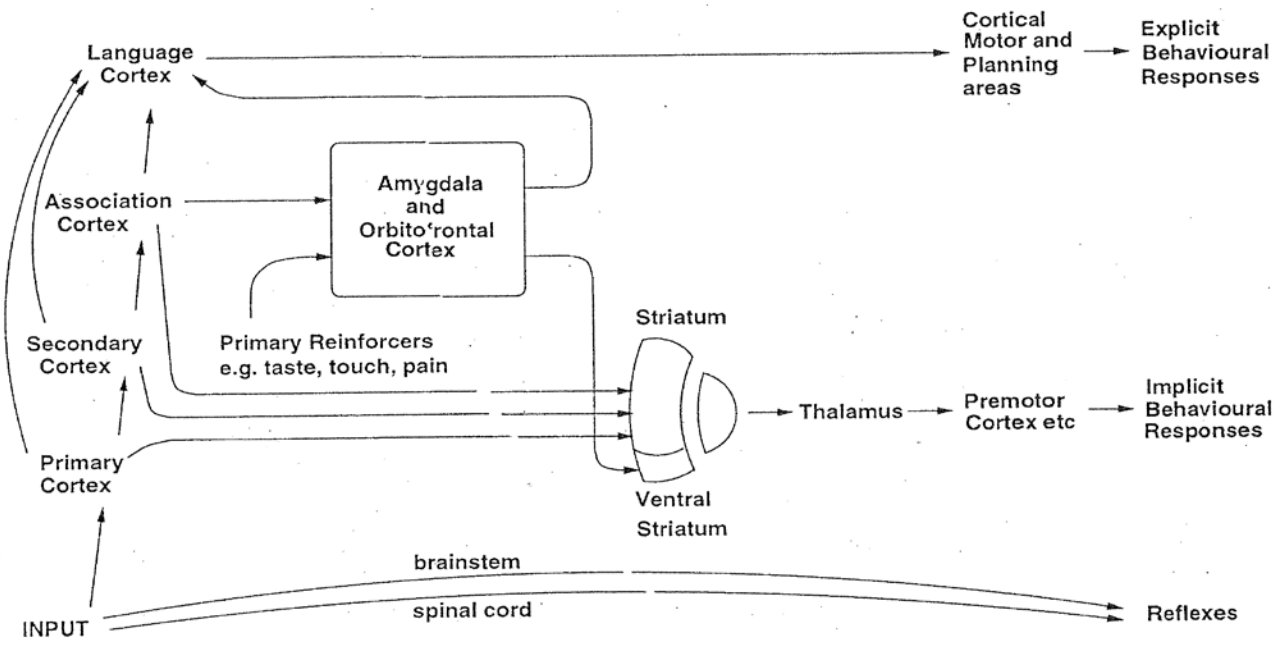

Language Cortex
Cortical Motor and Planning areas
Explicit Behavioural Responses
Association Cortex
Amygdala and Orbitofrontal Cortex
Primary Reinforcers e.g. taste, touch, pain
Secondary Cortex
Striatum
Thalamus
Premotor Cortex etc
Implicit Behavioural Responses
Primary Cortex
Ventral Striatum
brainstem
spinal cord
INPUT
Reflexes

Fig. 9.4 Dual routes to the initiation of action in response to rewarding and punishing stimuli. The inputs from different sensory systems to brain structures such as the orbitofrontal cortex and amygdala allow these brain structures to evaluate the reward- or punishment-related value of incoming stimuli, or of remembered stimuli. The different sensory inputs enable evaluations within the orbitofrontal cortex and amygdala based mainly on the primary (unlearned) reinforcement value for taste, touch, and olfactory stimuli, and on the secondary (learned) reinforcement value for visual and auditory stimuli. In the case of vision, the 'association cortex' which outputs representations of objects to the amygdala and orbitofrontal cortex is the inferior temporal visual cortex. One route for the outputs from these evaluative brain structures is via projections directly to structures such as the basal ganglia (including the striatum and ventral striatum) to enable implicit, direct behavioural responses based on the reward- or punishment-related evaluation of the stimuli to be made. The second route is via the language systems of the brain, which allow explicit (verbalizable) decisions involving multistep syntactic planning to be implemented.

References

Anscombe, G.E.M. *Intention* (Oxford: Blackwell, 1963).

Bargh, J.A. and Chartrand, T. (1999), "The Unbearable Automaticity of Being", *American Psychologist* **54**: 462-479.

Baumeister, R.F. *et al.* (2003). "Intellectual Performance and Ego Depletion". *Journal of Personality and Social Psychology* **85**: 33-46.

Bechara, A., *et al.* (1994). "Deciding Advantageously Before Knowing the Advantageous Strategy". *Science* **275**: 1293-95.

Chafee, M.V. and Ashe, C. (2007), "Intelligence in Action". *Nature Neuroscience* **10**: 142-143, reporting Shima, K., *et al.* (in press) *Nature*

Crick, F. and Koch, C. (1995). "Towards a Neurobiological Theory of Consciousness". *Nature* **375**: 121-123.

Darwall, S. (2001). "Because *I want it*". *Social Philosophy and Policy* **18**: 129-153.

Gibbard, A. (2005). "Truth and Correct Belief". *Philosophical Issues* (*Nous Supplement*) **15**: 338-350.

Gilbert, D.T. and Wilson, T.D. (2000) "Miswanting: Some Problems in Future Affective Forecasting". *Feeling and Thinking: The Role of Affect in Social Cognition*. (Cambridge: Cambridge University Press).

Haidt, J. (2001). "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment". *Psychological Review* **108**: 814-834.

Haidt, J. (2007). "The New Synthesis in Moral Psychology". *Science* **316**: 998-1002.

Hare, B., Call, J., and Tomasello, M. (2001). "Do Chimpanzees Know What Conspecifics Know?" *Animal Behavior*

Hassin, R.R., *et al.* (2005) (eds.). *The New Unconscious* (Oxford: Oxford University Press).

Hauser, M.D. (2000). *Wild Minds* (New York: Henry Holt).

Hauser, M.D., *et al.* (2003). "Give Unto Others: Genetically Unrelated Cotton-Top Tamarin Monkeys Preferentially Give Food to Those Who Altruistically Give Food Back". *Proceedings of the Royal Society, London* **B 270**: 2363-2370.

Humberstone, direction of fit

Irwin, T.H. (1999) (trans). *Aristotle: The Nicomachean Ethics* (Indianapolis: Hackett).

Kahneman, D. and Tverskty, A. (2000). *Choices, Values, and Frames* (Cambridge: Cambridge University Press).

Korsgaard, C. (1997). "The Normativity of Instrumental Reason". Cullity, G. and Gaut, B. *Ethics and Practical Reason* (Oxford: Oxford University Press).

Levi, I. (date). *Gambling with Truth* (ref.)

Lormand, E. (1990). "Framing the Frame Problem". *Synthese* **82**: 353-374.

Parfit, D. (2006). "Normativity". *Oxford Studies in Metaethics* **1**: 325-380.

Pettit, P. and Smith, M. (1990). "Backgrounding Desire". *Philosophical Review* **99**: 565-592.

Quinn, W. (1993). "Putting Rationality in Its Place". Hursthouse, R. (ed.), *Virtues and Reasons* (Oxford: Clarendon).

Railton, P. (2001). "*A Priori* Rules: Wittgenstein on the Normativity of Logic", in Paul Boghossian and Christopher Peacocke (eds.), *New Essays on the A Priori* (Oxford: Oxford University Press).

Railton, P. (2004). "How to Engage Reason: The Problem of Regress". Wallace, R.J. *et al.* (eds.), *Reason and Value*. (Oxford: Oxford University Press).

Railton, P. (2006). "Normative Guidance". *Oxford Studies in Metaethics* **1**: 3-34.

Railton, P. (forthcoming). "Practical Competence and Fluent Agency". Sobel, D. and Wall, S. (eds.), *Practical Reason* (Cambridge: Cambridge University Press).

Scanlon, T.M. (1998). *What We Owe Each Other* (Cambridge: Harvard University Press).

Scheuler, G.F. (1995). *Desire*: *Its Role in Practical Reason and the Explanation of Action*. (Cambridge: MIT Press).

Schroeder, T. (2004). *Three Faces of Desire*. (Oxford: Oxford University Press).

Shultz, W., *et al.* (1997). "A Neural Substrate of Prediction and Reward". *Science* **275**: 1593-1599.

Stalnaker, Robert (1984). *Inquiry* (Cambridge: MIT Press).

Stampe, Dennis (1987). "The Authority of Desire". *Philosophical Review* **87**: 335-381.

Tomasello, M. and Call, J. (1994). *Primate Cognition* (New York: Oxford University Press).

Wallace, R. Jay (2006). *Normativity and the Will* (Oxford: Clarendon).

Wittgenstein, L. (1953). *Philosophical Investigations*. Anscombe, G.E.M. (trans.) (London: Blackwell).

Wyvell, C.L. and Berridge, K.C. (2000). "Intra-Accumbens Amphetamine Increases the Conditioned Incentive Salience of Sucrose Reward: Enhancement of Reward "Wanting" without Enhanced "Liking" or Response Reinforcement". *Journal of Neuroscience* **20**: 8122-8130.

Wyvell, C.L. and Berridge, K.C. (2001). "Incentive Sensitization by Previous Amphetamine Exposure: Increased Cue-Triggered "Wanting" for Sucrose Reward. *Journal of Neuroscience* **21**: 7831-7840.

Zajonc, R. (1980). "Preference Needs No Inference". *American Psychologist* **35**: 151-