Rational Desire and Rationality in Desire: An Unapologetic Defense

Peter Railton

Preliminary draft of October 2008—please do not circulate without permission

Introduction

Consider Robert Stalnaker's well-known, elegant formulation of the interlocking, 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; emphasis added]

This characterization gives a picture of a potentially rational agent as, so to speak, at the mercy of his beliefs and desires, *whatever they are*. Whether an individual agent is *actually* rational, one might naturally think, will depend on what fills in for 'whatever they are'—for this will determine the shape of her life and, roughly, who, if anyone, is in charge. Put another way: Autonomy will depend heavily on *how* 'whatever they are' is filled in—what role does the agent herself play? And rationality will depend heavily on *what* is filled in—will the agent's beliefs and desires be as they rationally should be?[1] The thought is: If your beliefs and desires are as they rationally should be, and you as agent have played the right sort of role in acquiring them, then insofar as they are functioning normally—namely, as Stalnaker indicates—to guide how you act, then your actions, too, will be autonomous, and as they rationally should be.

However, a common view among philosophers, traceable at least back to Hume, is that beliefs can be rational, but not desires. For desires, there simply is no "way that they rationally should be". [2] Many modern decision theorists seem to accept a similar view; apart from a few

---

[1] I speak of 'rationally should' rather than simple 'should' because it seems to me at least conceptually possible that the way our lives *morally should be*, or *prudentially should be*, or *aesthetically should be* could differ from how they rationally should be.

[2] I am speaking here of basic or underived desires, rather than means/end desires, which are often ascribed a kind of "relative rationality". Initially, I will use the term 'desire' following established philosophical convention, according to which it is something of a catch-all term for motivations of very diverse kinds. Later I will try to mark some distinctions among motives, and give a clearer idea of what is distinctive of desires as opposed to, say, appetites or drives.

formal conditions, preferences are treated as "exogenously given" in the dynamics of rational choice. One explanation of this difference is that beliefs are representations, with "mind-to-world direction of fit", and so capable of truth or falsity. Desires, by contrast, are not representations. My desires do not present to me an image of how the world *is*, but rather, of how the world *is to be*, insofar as this is in my power. Desire thus have "world-to-mind direction of fit" and so are not candidates for evaluation as true or false, correct or incorrect. Since the canonical operations of reason—drawing inferences and gathering or assessing evidence—take as their objects bearers of truth value, basic desires are excluded from these ways of staking claims of rationality.[3]

It was not always thus. In ancient thought, a good deal of attention was paid to being "rightly motivated", and a number of commentators on Aristotle think 'rational desire' is an appropriate translation for *boulesis*.[4] Aristotle writes in *De Anima*:

> … it is reasonable that these two appear the sources of movement, desire and practical thought. For the object of desire produces movement, and, because of this, thought produces movement … . [T]he intellect does not appear to produce movement without desire (for *boulesis* is a form of desire, and when one is moved in accordance with reasoning, one is moved in accordance with *boulesis* too) … . [*DA* 433a17-25]

And in the *Nicomachean Ethics*:

> Now the origin of action (the efficient, not the final cause) is choice, and the origin of choice is appetition and purposive reasoning. … [A]n action is an end in itself … and the object of appetition. Hence choice is either appetitive intellect or intellectual appetition; and man is a principle of this kind. [*NE* 1139a32-b5]

The idea that "intellectual appetition" is the very "principle" of the practical agent is taken quite seriously:

> [B]rutes have sensation, but no share in action. Pursuit and avoidance in the sphere of appetition correspond exactly to affirmation and negation in the sphere of intellect … so that, since … choice is deliberative appetition, it follows that if the choice is a good one, both the reasoning must be true and the desire right; and the desire must pursue the same things that the reasoning asserts. We are here speaking of intellect and truth in a practical sense … the

[3] A contrasting contemporary view holds that desires are amenable to rational assessment because they are (or are very like) beliefs or judgments. For a variety of reasons—logical and psychological—I think this view untenable. Fortunately, I will argue, there is yet a third view of desires that makes them amenable to rational assessment, even though they are neither beliefs nor judgments.

[4] See, for example, the discussion in Irwin (1999).

> function of practical intellect is to arrive at the truth that corresponds to right appetition. [*NE* 1113a20-28]

And lest one have too intellectualist a view of "right appetition", Aristotle insists:

> Pleasure and pain are also the standards by which—to a greater or lesser extent—we regulate our actions. Since to feel pleasure or pain rightly or wrongly has no little effect upon conduct, it follows that our whole inquiry must be concerned with these sensations. [*NE* 1105a3-5]

About desire and rationality, I'm on Aristotle's side rather than Hume's,[5] although I hope to show that one need not abandon the larger Humean project in order to think this way.

Rational belief and rational choice have been subject to wonderfully broad and deep investigation for several centuries. But wouldn't it be a feckless nature that would endow us with exquisite capacities for rationality in belief and choice, as well as the power to formulate and follow complex intentions, strategies, policies, and plans, but somehow left to chance the desires that furnish the very aims and objects in the service of which we choose, plan, and act? Indeed, looking at the matter from an evolutionary standpoint, natural selection could not have operated on our capacities for representation and choice except through their expression in behavior and action—where desire has joint say in setting the agenda. Unless the desires of those among our ancestors with more accurate representational capacity or greater decision-making and planning power were by and large harnessed to the pursuit of desires in great measure *fit* to human needs, potentials, and natural and social circumstances, selection would not have favored the evolution of rationality in belief or decision. Natural selection awards points for reproductive success—which involves physical, mental, social, and sexual success—not for formulating correct theories of the world as such, while seeking or doing *whatever*. So it would be astonishing if there were no connection between human possession of characteristics that favor rationality in belief and decision-making and human possession of characteristics that favor forming feasible, need-responsive, self-advancing, life-enhancing, sociable, cooperative, mate-enticing, context-sensitive, and rewarding-to-satisfy desires.

[5] It is not easy to say what Hume's fully considered view on desire was. When he speaks of 'reason' in a narrowly deliberative sense in Book III of the *Treatise*, he clearly classifies desire as non-rational—but there he also argues that belief also is not narrowly deliberative, but rather a "feeling". And elsewhere, where he uses the notion of 'reasonableness' less narrowly, as in his discussions of the sentiments in Book II, Hume seems happy to speak of motivating attitudes as more or less reasonable.

By way of analogy, consider the Rule of Significant Digits we learned in school: a numerical operation, such as a sum or product, can never have greater expected accuracy than its least accurate numerical input. Thus if we have a string of numerical addenda, each accurate to three decimal places, but one further addendum accurate only to the nearest power of ten, then the sum will also have an expected accuracy no better than the nearest power of ten.

Consider, next, a fairly standard belief/desire model of intentional action:

belief + **desire** + **deliberation** → **choice** → **intention** → **action** (1)

Now, turning to the question of rationality in action, introduce as much rationality into the *inputs* of this model as the standard view permits:

rational belief + **??? desire** + **rational deliberation** → … (2)

Surprisingly, proponents of the standard model often go on to speak happily of the *outputs* as:

… **rational choice** → **rational intention** → **rational action** (3)

If output can be no more "accurate" than the least accurate input, one would rather have expected:

… **??? choice** → **??? intention** → **??? action** (4)

Can practical reason beat the limitations imposed by the Rule of Significant Digits even though theoretical reason cannot?[6] If so, it would be nice to have an explanation of how.

An alternative route would be to try to resurrect the notion of rational desire. In giving his account, Aristotle availed himself of metaphysical resources we no longer allow ourselves, notably, expressly normative notions of natural essence and proper function. For good reason, we moderns reject natural teleology, and draw upon only such *telos* as can be found in human aims, aspirations, and conditions for flourishing. But there might be more there to work with than is commonly allowed. In particular, we need to see if we can rehabilitate what Dennis Stampe called "the authority of desire" (Stampe 1987).

The conclusions we reach will have a conditional status. I do not seek to demonstrate to the skeptical mind the possibility of rational desire, or, for that matter, the possibility of rational belief or rational action. Maybe (4) is the best we can do. However, I will present an argument that suggesting that rationality in action might require according "bare" beliefs and desires normative standing. This argument has much in common in form and spirit with Kantian transcendental arguments, though it will be carried out within a low-budget, Humean framework. (Why should Kantians have all the fun?) This Poor Man's transcendental argument, first

[6] Selim Berker has pointed out to me that, for example, perceptions can play an important role in the shaping of rational choice and action, yet we do not think perceptions themselves can be more or less rational. This is so, though we do distinguish between rationality and irrationality *in* perception, e.g., in the formation of perceptual beliefs more or less rationally. On the standard conception of practical reasoning, it is perceptual belief, not perception itself, that enters into deliberation.

broached in section 3, below, will furnish the "negative" side of my defense of rational desire and rationality in desire. But the argument also needs a positive side that presents, at least in a schematic way, a picture of *how* rationality is possible within the domain of desire – how even basic desire can be an object of reasoning and learning, and can contribute to the rationality of action. That is the job of the final sections of this paper.

1. A quick argument for the impossibility of what I hope to show

Although action theorists used to speak of desires as both causes of, and reasons for, action, the view that desires as such cannot be reasons for action has recently had distinguished advocates.[7] There are certain things, it is said, that we have reason to do regardless of whether they are desired by anyone (e.g., preserve the great works of lost civilizations), while there are various other things that are desired—perhaps even "natively" desired—to which we have no reason to accord any normative weight (e.g., a spiteful desire to see a successful rival brought low, even though—or rather just because!—he has succeeded by his merit).

Here, then, is a quick argument for the thesis that desires as such cannot be reasons for action. Suppose that I desire to go outside for a walk. For simplicity, suppose as well that I recognize this fact and that my psychic state seems to me in no way abnormal at the time. To have this desire is to be in a particular psychological state, perhaps a disposition to notice certain things, be moved to act certain ways, etc. *A priori*, the mere fact that I am in such a state is, from a normative standpoint, just another feature of the world in which I am deliberating. Like any other bare fact, it does not as a conceptual matter have any necessary authority for my deliberation or action. I must, if I am to act on reasons, determine what to make of it, and what, if anything, to do about it. After all, some desires are problems to be overcome, rather than solutions to the question, "What to do?" Instead of satisfying them, one should strive by whatever means to hold them at bay, diminish, or eliminate them.

Thus, the argument runs, although desires do have the causal power to move agents to act, thereby helping to explain their behavior, such *motive force* should not be confused with *normative force*. To explain behavior is not to justify it.[8] Such claims are supported by a fairly clear intuition. After all, chickens also have, and are their behavior is shaped by, internal states

[7] See for example Scanlon (1998), Darwall (2001), and Parfit (2006).
[8] For an extensive indictment of modern moral philosophy for confusing normative with motivational force, see Parfit (2006). I won't be directly addressing Parfit's view below, since he rejects the forms of "normative internalism" that predominate among those who reject desires as reasons. If my positive case for desires as reasons succeeds, however, this would be an objection to Parfit's account.

with motive force. If that sufficed for normative guidance, we would have to concede its presence in chicken life as well as human life.

Perhaps the intuitive comparison with animals will help sharpen the point, and suggest what is required of a solution that captures normative force.[9] As a result of sensory stimulation, animals are thought to form complex brain states that model in detail aspects of their environment and their internal condition. Call these, generically, *representations*. With regard to the visual field, for example, these representational brain states have been extensively mapped in some animals. Moreover, these representations appear to be *coded* in the brain as positive or negative, shaping animal attention, arousal, behavior, and reward accordingly.[10] These representations and their coding are continually updated in the animal brain to yield intelligent goal-directed behavior, such as seeking food, migrating, or joining a fight to protect an ally. But (or so the intuitive comparison goes) this as yet affords no ground for attributing to animals representations with *normative* content. Animals do not appear to take themselves to have *reasons* for acting as they do, or to make value judgments. We should not impute to them animalese versions of "This hunger is getting unbearable, I really *should* step up my efforts to find something food" or "That ape is big, but my ally isn't being *unreasonable* in looking to me to help fight him off". Yet thoughts of these kinds, usually tacit to be sure, are common when humans act for reasons.

Thus, the argument goes, when behavior qualifies as done for a reason in the normative sense, the path between desire and behavior is not simply that of direct, functionally-organized causation. The *agent* must enter into path between desire and behavior in the distinctive way agents can.[11] This could not be the passive contemplation of an unfolding causal spectacle, for then the desire and its sequelae would remain no more than psychic events that happen to occur within sight of the agent. That would not be *acting* at all, for the agent would be epiphenomenal to the story of why the behavior occurs. If the agent is to enter this story in such a way that she acts for reasons, she must play an indispensable role *under some idea* of what she is doing and why—a *normative construal* of her situation. The agent *takes X to be a reason to F* or *treats Y as counting in favor of doing G*, and this in turn guides her activity. This distinguishes actions done for reasons not only from simple reflexes, but also from such situation-appropriate, motivated

[9] Here and throughout I will follow the smug practice of using 'animal' and 'human' in the sense in which they are contrastive. Worse, I will be making a number of assumptions about the nature and limits of animal cognition, affect, and motivation, that are loosely based on some recent research but almost certainly too ungenerous to the animals. For some correctives, see Tomasello and Call (1994), Hare, Call, and Tomasello (2001), Hauser (2000) , Hauser, *et al.* (2003), and Chafee and Ashe (2007).
[10] See Shultz, *et al.* (1997).
[11] We need not imagine that the agent's intervention is in some sense contra-causal, or that the intervention is not "causally realized".

behavior as shifting my weight, unawares, from foot to foot, to improve circulation while chatting on the phone. Or, for that matter, from your dog's climbing in bed with you to keep warm on a cold night.

On this picture, a desire is not in itself a reason for action, though what one desires typically plays a role in how one construes one's situation. Theorists vary in how they understanding *taking to be a reason*, but among the suggestions on offer: the agent *endorses* the desire, *assents* to it, *throws her weight* behind it, or *identifies* with it. Moreover, it is typically the *object* of desire, the *end* in sight, not the psychic state itself, that the agent endorses or identifies with. I usually need a good reason for getting up before dawn, and my reason for doing so yesterday was needing to catch that day's only flight home, or (for short) needing to be home, *not* needing to satisfy a desire to catch the flight or be home.[12] Of course, there are exceptions. If I check into a treatment program for alcoholics to help rid me of an overpowering desire to drink, or take a new route home while bicycling just because I feel like it, then my action may be aimed primarily at eliminating or satisfying a desire as such. But these are unusual cases.

Three comments are in order. (1) We need not saddle this view with the idea that "taking *X* as a reason to *F*"—embodied in an act of endorsement, assent, identification, etc.—must be a full-fledged judgment concerning the reason-giving status of *X*. It can of course take that form when the agent pauses to reflect or deliberate explicitly, but ordinarily an agent's construal of her situation is formed and guides what she does without pause for reflection or extensive internal commentary.

(2) Although "taking *X* as a reason to *F*" need not be a full-scale judgment, it is nonetheless a distinctively normative way of seeing things. It requires at least an implicit grasp of the concept of a *reason to act*, where this is contrasted both with reasons *to believe*, *to hope*, or *to imagine*, say, and with acting *for no reason*. This implicit grasp must afford the agent an in-principle possibility of reflection and criticism, and thus be must more than a sense of how strongly one happens to be motivated. Moreover, it requires at least an implicit grasp of the concept of a *good* reason to *F*. Such a tacit understanding is manifest, for example, whenever an individual queried about why she has acted a certain way offers a rationale that she expects to *defend* or *justify*, and not merely explain, what she is up to.[13] So far as we know, animals lack

[12] See Pettit and Smith (1990) on the "backgrounding" desire in deliberation.

[13] The growing literature on "dual process" models of the mind suggests that we are often mistaken about the reasons for, and causes of, what we do. *Post facto* rationalization is thought to be pervasive, even though, from the agent's perspective, this often presents itself as self-understanding. See Bargh and Chartrand (1999), Haidt (2001), and Hassin *et al.* (2005). If correct, these models pose no small problem for how to think realistically about action and practical rationality. For some preliminary remarks, see Railton (2006 and forthcoming).

this implicit normative understanding of what they are doing, and do not feel our powerful human need to offer ourselves and others rationales—or rationalizations.[14]

(3) For an agent genuinely to act for reason *X*, her normative construal involving *X* must actually play a *regulative* role with respect to her thought and action. This need not involve explicit formation of an intention to act or an express exercise of will, but it does require the existence of a prior construal involving *X* in the agent's mind, and also the right sort of active involvement of this construal in the agent's shaping of her action.[15] Without this, how are we to distinguish doing *F* for reason *X* from simply doing *F* (which conforms to reason *X*) but not for this reason? The agent's shaping of her action in light of her normative understanding of the situation is, on this account, something *real* even when it is more or less implicit. It will, for example, be experienced as effortful when the agent faces obstacles and temptations.[16] Note, however, that figuring in this shaping role is independent of whether *X* is a good or bad reason to *F*. Rather it is what acting on good and bad reasons have in common. Acting for a seeming reason is not seeming to act for a reason.

To become the reason for an action, then, a "bare" desire requires the active collusion of an agent, who makes the desire into part of the explanation of what she does through taking it or (much more commonly) its object to be a reason for acting. My practical autonomy—my capacity to be more than a creature pushed and pulled across the landscape by desire in proportion to its strength or urgency—is seen as depending upon this gap between desires and reasons. *Pace* Hume, reason enslaved to passion would simply not be practical reason at all. Or, to take a certain liberty with a well-known religious motto, "Desires propose; you dispose."[17]

This strikes me as a very persuasive argument. Yet I think it cannot be right.

2. Reasons for belief

To show why, it is easier to turn first to reasons for belief, and consider the fate of a parallel line of argument.

---

[14] "What's the most important thing in life?"—*Rationalization.* "Rationalization? Are you crazy? More important than sex?"—*Sure. Anyone can go a day without sex, but just* try *to go a day without a rationalization.*

[15] For a fuller idea of what such normative guidance might involve, see Railton (2006).

[16] Baumeister, *et al.* (1998) gives experimental evidence of this effortfulness, which among other things suggests that repeated, unsuccessful encounters with obstacles can "deplete" the ego, and thereby lower an agent's capacity for self-control.

[17] I am told that Harry Frankfurt has written, "A desire is a *problem*, not a reason."

Suppose that I lift my eyes from my laptop. I have been wholly absorbed in making nit-picking changes to an overdue text, and it appears to me as if night has fallen. Suppose also for simplicity that I am aware of having this experience, and further that there is nothing "off" about it or about how I currently feel. To be clear: I do not *assess* this experience to be coherent, rather, it simply is the case that my perceptual experience is coherent (the windows all are dark, the clock says 8:30, etc.); neither do I *judge* my condition to be normal, rather, I simply am at the moment receiving no perceptual or proprioceptual cue of anomaly (no dream-like appearance, sense of wooziness, etc.).

Can this "bare appearance" or "mere experiential state", *in itself*, be a reason for me to believe that night has come? If, as the causal upshot of this normal, coherent experience, I immediately ("without thinking"[18]) come to believe that night has fallen, can we say that this is an instance of rational belief formation, or that I have come to believe that night has fallen for a normative reason?

Well, consider the following quick argument. To have a conscious perceptual or proprioceptual experience is simply to be in a particular psychological or physiological state, perhaps the possession of certain *qualia* along with certain related dispositions to expect, infer, etc. Now the mere fact that I am in such a state is, from the standpoint of theoretical reason, just another feature of the world within and about which I am deliberating. Like any other bare fact, of itself it has no *a priori* normative authority.

Sensory experiences might directly and reliably cause me to form mental representations of the world around me, but this presumably is just what happens in the perceptual experience of cognitively sophisticated animals. They, too, "immediately" form mental representations and expectations as the causal upshot of sensation and proprioception. For example, many mammals are nocturnal and especially attentive to whether night has fallen. Imagine a raccoon waking from its slumber an hour after sunset and peeking up from its nest in an abandoned chimney. It experiences a distinctive internal sensory state that is cued to reduced ambient light and that triggers a "night-specific representation" in her brain. This representation, in turn, is linked by instinct or association with an array of night-specific expectations and dispositions. To keep things simple, imagine that the visual process cuing the night-specific representation is highly reliable.

What does this raccoon lack that an epistemic agent in a similar situation has? A raccoon, however clever, lacks conceptual equipment necessary to form propositional attitudes, such as belief, or to have even a tacit self-understanding of herself as responding to relevant

[18] "Non-inferentially" in the terminology of mid-20th-century philosophy of perception.

evidence or reasons for belief. So it would be rank anthropomorphism to say that she forms her mental representations for reasons, or that she takes her night-ish experience to be a reason to think that night has fallen. To be sure, the reliable process by which she has come to have a night-specific representation in the brain is much more than a "mere causal process" or reflex. It is part of a learning system that generates reliable information about the state of the world that, in the presence of certain other cues, such as hunger, triggers apt shifts in behavior, such as beginning to move about and forage. In short, it belongs to a form of intelligence that attunes the raccoon and its behaviors to its needs and the world.

Intelligent as the raccoon may be, the argument goes, it is not rational and does not think or behave as it does for reasons. What is missing? What must human belief-formation be like, beyond intelligent, that can qualify it to earn it the name of rationality? According to the view under consideration, the believer himself must enter into path between perceptual appearance and belief in the distinctive way an *epistemic agent* can. This cannot be a matter of passively noticing a perceptual appearance and its causal sequelae, for that would not be *forming a belief*—it would be watching a belief form, the way in which one might watch a realization slowly dawn upon a child. On this view, what makes a belief *my* belief, and differentiates my *beliefs* from more or less forceful or vivid stray bits of thought in my mind, is a matter of the involvement of my epistemic agency in forming or retaining the belief. It is up to me what I make of a given experience, and whether I take it as evidence or a reason to believe.

Hume thought of belief as a vivid, firm impression, but there is a difference between having a vivid, firm impression that night has fallen, on the one hand, and *assenting* to this impression or *accepting* its objective purport, on the other. It is a question not of the face value of this experience, but of whether I *take* the experience at face value. One does not always do so, e.g., when one has doubts about the reliability of one's vision without glasses. To take an experience *e* as a reason to believe that *p* is to have some idea of one's epistemic situation, a normative construal of *e* as relevant to, and counting in favor of, *p*. Without such license from the epistemic agent, we cannot count an experientially-based belief that *p* as held for a normative reason.

As before, three comments are in order. (1) "Taking *E* to be a reason to believe that *p*" need not be an explicit judgment, though it can be. (2) Although no full-dress judgment need be involved, the agent must have at least a tacit grasp both of the concept of a reason to believe, or evidence, and of the distinction between good and bad reasons to believe. This conceptual competence will be manifest whenever he is queried or challenged: the agent knows how to produce a justificatory rationale or bit of evidence in favor of what he believes. (3) In order for a

belief that *p* genuinely to be arrived at for reason *X*, the individual's normative construal of his epistemic situation involving *X* must have played an active, regulative role in the agent coming to believe that *p* — a role that is common to believing for good and bad reasons. Of course, it is not uncommon to paint a rosier picture of one's epistemic *bona fides* than the facts warrant, construing oneself as holding a belief for justificatory reasons that do not correspond to the causal reasons that actually explain why one holds it. Raccoons, I am sure, never kid themselves in this way about their reasons for thinking what they do. The fact that we are capable of such feats, and so powerfully motivated to perform them, attests to our grasp of the normative character of belief.

Let's explore a bit further the difference between the explanatory and the justificatory perspective in belief formation. Change our example a bit. I have been obsessively working away at my laptop in a windowless office, look up, and wonder whether night has fallen. I enter the next room, where there is window, look out, find the sky already dark, and immediately think, "It's nighttime". An observer seeking to explain my change in belief (from being unable to say whether it's nighttime to having a definite view), would point out that PR has just now moved to a location where he *is having an experience in which it looks as if night has fallen*. But from my perspective as an epistemic agent, I credit my change in belief to the *content* of this experience rather than the fact of it, i.e., I come to think night has fallen on the basis of what my senses present to me: *it looks as if night has fallen*. Beyond this, any reference to the fact of the experience drops out.

The same holds for beliefs. Suppose that my next thought is, "Rats. Now my jet-lag will last longer" – I have just flown to a new time zone, and, reading the science section of the *Times* on the plane, I have just learned that one should get exposure to daylight at one's destination in order to reset one's biological clock to local time. My travel companion has also spent the day indoors, and is right now is looking out different window in a different room, and noticing for the first time that it is night. Unlike me, she does not immediately think this will affect her jet-lag – she got the front section of the *Times* on the plane. Suppose an observer now is asked to explain why I immediately formed a jet-lag thought upon looking up from my computer. He would say that *PR has just come to believe that getting out of doors helps reset one's biological clock*. But from my standpoint, I do not reason from the *fact* of my belief, nor do I see it *recentness* as relevant to my conclusion about jet lag. Rather, I reason solely from the belief's *content*: *getting out of doors helps resent one's biological clock*. Likewise for the content of my belief that it is nighttime, which also figured in my little inference. After all, I was not so foolish as to think that the *Times* was trying to tell me, "Be sure to get outdoors while you still *believe* that it is daytime".

Although one does not – and should not – always take one's experience or existing beliefs at face value, when one does, the "face value" one is treating as a reason is not the fact of having a certain experience or belief – experiences and beliefs do not typically have *this* on their face – but rather what the experience portrays or belief represents. My epistemic autonomy arises from the fact that it is up to me what I make of the content of a given experience or belief—will I trust it or question it? In this sense, however much I depend upon my sensations and existing beliefs, I still am not a slave to their onslaught, dragged about by whatever image or thought happens to enter my mind with the greatest force at a given moment.[19] I might be a captive audience for my experiences and thoughts, but I still am able to hold them at arm's length, exercise tacit or explicit judgment, and confer or withhold assent. If reason should not be the slave of our passions, neither should it be the slave of our impressions or thoughts.

According to this argument, then, having a "bare" experience or finding oneself with a "bare" belief is never *in itself* a reason to believe. "Experience and thought propose; you dispose."[20]

Once again, we have a very plausible argument to a seemingly compelling conclusion. However, this argument cannot be right. And now I think I can say why.

3. The problem

An initial objection to the argument just considered would be that if an agent always had to perform a mental act of accepting or rejecting her perceptual experience or her existing beliefs, in order to form a new belief for a reason, then believing for reasons would be impossibly cumbersome. No sooner would I have begun to examine one experience or thought critically than a fresh experience would assail me or thought would appear – indeed, the very thoughts I am now thinking and experiencing I am now having. Should *those* be taken at face value? To be sure, we have already agreed that assenting to an experience's purport or a thought's content, or taking it as a reason to believe, need not involve full-dress reflection and judgment. So we need not imagine this as a very time-consuming process – it could be quite quick.

But would quickness help? One must always be careful about multiplying mental operations and acts, and not only because of limits on time and attention. Let us adopt the most relaxed attitude possible toward what it is for an individual to assent to an experience or an

[19] Compare Christine Korsgaard's remark that, for Hume, "beliefs are sentiments which are caused in us by perceptions and habits. Reason doesn't really enter into it." (1997, p. 223).
[20] Jerry Hankfurt might say, "To have an experience is to have a problem, not a reason."

existing belief as a reason to believe, consistent with this assent being a something rather than a nothing. Still, this "something" must, if the argument is right, comprise some sort of *doing*, a bit of agency organized under a governing idea or normative construal. Not simply an appropriate receptivity in the absence of countervailing experience or thought. Not even an intelligent or learned response to relevant information. There latter are, after all, within the repertoire of raccoons.

Moreover, this doing must be of the right sort to have some authority for the agent. For example, the epistemic agent might act by instantly assenting to, or rejecting, his current experience or thoughts *just like that*. But even if such an act of assent or rejection were possible – I cannot believe at will, or remove belief by will – it would certainly be no help in understanding what it is to believe for an *epistemic* reason. For how could there be any epistemic authority in such *fiat*? It is hard to see how it could ever answer a reflective demand for reason to believe. Appeal to self-evidence could not help here, since that would simply be accepting one's own epistemic authority on the strength of one's own epistemic authority, nothing more.

This doing, then, would seemingly have to be done for a reason, and a reason of the right sort. But now it is clear that we have reasoned ourselves into a regress. For where is the agent to look for such reasons? He cannot turn to his other beliefs or to some feature of his actual state of mind or prior experience, say, the confidence he has acquired in his senses through favorable experience. After all, these are mere features of his psychology. For them to count as reasons to accept or reject the agent would have to take them, too, as reasons. Otherwise, his assent or rejection would be done for no epistemic reason, and could not be a manifestation of his epistemic rationality or autonomy. What we have, then, is not a recipe for how to become an autonomous epistemic agent guided by reasons, but a recipe for a Zeno-like process in which each act of "taking to be a reason to believe" cannot be completed without another, prior act of just the same kind. Such a process, however tacit and lightning quick, will never—can never—arrive at belief.

4. How to be rational in belief: default, defeasible reasons

Another approach is needed if it is ever to be possible to believe for reasons.[21] Of course, I cannot prove that we ever *do* believe for reasons. Perhaps we should regard the argument just given as a *reductio* of the very idea of believing for a reason – it will always lead us into a

[21] A similar problem arises for carrying out rational inferences. For discussion, see Railton (2004).

regress, or into its opposite, belief for *no* reason.[22] I find this bleak conclusion premature. But rather than argue directly against it, I will aim only for a conditional claim concerning the possibility of believing for reasons, which is surely ambitious enough: There would be a workable notion of believing for reasons *if* it took the form sketched below.[23] This "demi-transcendental argument" will, I believe, lead us to a picture of rationality in belief that meets the strictest Humean standards for purity of ingredients.

The fatal step in the previous attempt was interpreting belief for a reason as requiring a distinguished mental act on the part of the epistemic agent – taking *e* to be a reason to believe that *p*. To avoid this step, or others like it, we must find a conception of epistemic rationality according to which *having an experience* as of *p*, or *having confidence* that *q*, *r*, and *s*, which together entail *p*, puts one in the position of being able to believe that *p* for a reason.

Right away, this suggestion appears to have a fatal flaw of its own, since it leads to *bootstrapping*. Suppose that I happen to be confident that *p*, for no particular reason. Of course, *p* entails *p*. Thanks to the principle just enunciated, I would now have a *reason* to believe that *p* after all! This looks like getting something for nothing. The full answer to this worry cannot yet be given, though it will come soon. For now, let me simply suggest: given the choice between a theory of reasons for belief that lets us pull ourselves *up* by our own bootstraps, and one that requires us to pull ourselves *down*, I'll take up.

First, we need to address another immediate worry: If "bare" experience can support belief for a reason without any intervening exercise of epistemic agency, does a raccoon in whom nightfall produces a night-specific representation count as believing for a reason?[24] If the most we can secure for rationality in belief or epistemic autonomy admits raccoons and hound dogs, perhaps we had better drop the self-aggrandizing talk and concede the *reductio* after all.

Now although this initially sounds cryptic, what I will claim is that the place to look for an answer to the question of what separates those of us who form perceptual beliefs for reasons

---

[22] One way to read Hume in Book III is as having seen and accepted this *reductio* (cf. his remarks about "running up inference upon inference" and upon belief as a "feeling" rather than a judgment; cf. as well his comment in the *Abstract* that the "association of ideas" is his most important discovery, and the sole possible solution to the problem of understanding how chains of reasoning are possible). If this reading were right, then we could not complain that Hume has simply *replaced* normative epistemology with a theory of the natural conditions for stability in belief. Rather, this would simply the best fallback position in the face of a *reductio* of our familiar normative notions—a "skeptical solution" to the problem of how to go on inquiring and believing. One way of reading Wittgenstein's claim (1953, § nn) that certain mental processes, such as following an argument, must be "blind" is to take him as claiming that they are properly seen as done for no reason.

[23] And *only if*? Some defenders of dynamic conceptions of rationality in belief might make this claim, but for present purposes I set it aside.

[24] Hume appears to have accepted this conclusion. See *Treatise*, ref.

from intelligent animals does not lie in the directness of the causal path between sensory input and perceptual representation, or between perceptual representation and more or less confident expectation. Rather, the difference will lie in the richness of the content of the representations that can be formed and of the capabilities of the mental architectures in which these representations appear and are used.

We will need to build up to this point somewhat gradually. Let us start with basic learning from experience. It is well known that in order to learn about one's environment (whether one be human, animal, or teachable computer) it is not enough to have ample sensory input and plenty of memory registers to fill. The learner must also bring to the encounter with sensory input some expectations—such as expected dimensions of similarity (the "implicit quality space"). Otherwise, experience will simply accumulate in its infinite diversity, and all experiences will be equally relevant or irrelevant to one another. No lessons will be extracted. Carnap (19xx) gave an elegant demonstration of this point within the theory of logical probability. He asked us to consider a confirmation function that began (sensibly, it would seem) with no "prior bias", i.e., that assigned the same non-zero probability to every possible state of the world (what he called "state descriptions"). This function would, even given indefinitely large amounts of information about past states of the world, still assign the same probability to probability to every logically possible way of extending this history into the future. In a fundamental sense, it could not learn from experience. By contrast, if the confirmation function were initially "biased" toward certain kinds of regularities or similarities ("structure descriptions"), then information about past states input into this function would quickly begin to produce differential expectations toward the future. Learning was now possible.

Subsequent work on formal models of learning, e.g., Bayesian probabilism, improved this picture by removing the static, *a priori* features of logical probability and providing an update function for belief that begins with prior credences (a "bias"), but that can revise these priors in the face of experience. One result is that the particular bias of one's starting point can come to exert less and less biasing effect on expectations as experience grows. Under ideally favorable conditions, the effect of most starting points will tend to wash out altogether as information increases. Inquirers who begin from quite different starting points, but who all follow the update rule in the face of new experience, will tend to converge in their expectations. Moreover, their expectations will tend to approximate the actual relative frequencies in the environment.[25]

A dynamic conception of rationality emerges. Believers display their rationality not by demanding reasons for belief *up front*, refusing to accord any credence experience until they see a

[25] See Good (dddd). For discussion, see also ref.

reason for doing so, but rather by the way they *revise* their original credences in response to new evidence or argument. Of course, how they react to new evidence or argument will be shaped by what they now believe – Should they start out from what they *don't* believe? But if they are systematic and honest in following the update rule, they will not be prisoners of their starting points. Thus they reason, and respond to new evidence, as if they accorded their current credences *prima facie* epistemic authority – innocent until proven guilty. I say 'as if', because this process presupposes no *act* of according *prima facie* authority. It is enough if they *think with what they have*. To be sure, among the thoughts epistemic agents have or tend to develop are credences concerning how they have come to believe what they believe, how reliable this process might be, etc. Nothing bars reflection or reflective endorsement. The point rather is that even in reflective thought, one is also thinking with what one has. If everything were up for grabs, requiring certification or endorsement before proceeding, we'd have chaos, not reflection. This may sound like getting something from nothing, but the something one gets is precisely the sort of thing that renders one vulnerable to experience (expectations might not be met, and update requires revision) and argument (credences might yield incoherence).

Bootstrapping thus is a feature, not a bug. It equips agents for learning and unlearning, and for reflection and refutation, thus making guidance by epistemic reasons possible. Without bootstrapping, epistemic agents would also face the "frame problem".[26] Suppose that I find myself believing that *p*. Is this any reason to go on believing that *p*? After all, I do not wish to be dragged about by whatever mental states I happen to have! So I demand a reason for continuing to believe that *p*. For any given *p*, I might be able to find such a reason more or less quickly: I learned it from experience, I acquired it from a reliable source, etc. But in giving these reasons I am relying upon other beliefs of mine. And just as *p* could have been *any* of my beliefs, it could also be *all* of them. Now I have no where to turn to answer my question. Once again, what seemed a perfectly sensible demand in order to believe for reasons has become the undoing of belief altogether. And once again, the implication seems obvious: we must think with what we have – to bootstrap our way up, not down.

What we take from such arguments is the general idea that effective epistemic agency requires *default, revisable trust*, a form of *underived epistemic authority*. Learning starts out from a set of initial, unlearned default expectations, which give shape and relevance to subsequent experience and permit doxastic continuity over time, but also set one up for growth, surprise, and revision. Fortunately, as infants we humans start with default trust in our senses,

[26] See Kahneman and Tversky (2000) and especially Lormand (1990).

thoughts, and memory.[27] It is obvious that, without these defaults, there would be no way for infants to learn such trust – for what other than sensation, thought, and memory could furnish their evidence?[28] Thanks to this default trust, infants manage in the fullness of time to become more sensitive to cues for when and when not to trust one's senses, thought, or memory. Eventually, they master the difference between appearance and reality. At the other end of life, as one's eyes, ears, and memory begin to fail, the same expectation-based learning process can alter its own parameters, causing a generalized lowering of default confidence in what one sees, hears, or seems to remember. Hesitancy becomes pervasive where once there was implicit trust.

Back to our animal cousins. The argument just made is quite general, and there is every reason to believe that intelligent animals have a similar sort of default, revisable trust in their eyes, ears, noses, and memories. Especially at the beginning of life, but also throughout life and at a level of which we are often unconscious, humans and animals probably do much of their learning from experience in fundamentally similar ways. Why might it be appropriate to say of humans that they form perceptual beliefs for reasons, while raccoons do not? On pain of regress, we had better not insist on locating our rationality in a special act of endorsement performed between evidence and conclusion. However, we might find it if we looked to the *content* of the evidence and conclusion, and to the larger *mental architecture* in which this content is embedded. **** If humans, but not animals, believe for reasons, we would probably do better not to *here* for So the first part of the answer to the question of how to distinguish rational perceptual belief formation from, say, intelligent experiential learning of the kind found in animals is *don't look* here *to find the distinction.* On pain of regress, one should not insist that learning by "mere conditioning" – a condescending term for trial and error – cannot be a way of forming beliefs on the basis of reasons. **** This sort of "reward learning" operates in the backdrop of everything we do, taking in sensory information, giving it form, and percolating this information through a massive web of associative connections. We would be quite unable to keep up with incoming information without this continuous, "bottom-up" acquisition and consolidation of information. Our more self-aware reflection, deliberation, and evaluation involve these default-governed processes, and their causal sequelae, at every step.

Picture our self-aware mental life, then, as the upper level in a corporate hierarchy, receiving a constant flow of already-processed information from a host of subordinates working

[27] The list of human default expectations is quite long, making possible learning across a wide array of domains. Infants very early on distinguish syntactically-ordered speech from nonsense strings and background noise, distinguish goal-directed from mechanical motion, expect regular cause-and-effect relations and natural kinds, project object constancy, etc.

[28] The environment must concur. Infants growing up in environments where basic trust in adults or in the stability of their environment cannot be established experience long-term developmental deficits.

without one's direct managerial supervision. After all, if the manager spent all her time checking each step of subordinates, there would be no time for her to do anything else—and who would check her? Instead, the manager takes this bottom-up information as initially credible, relying upon it in corporate planning, assessment, and innovation. If difficulties or surprises arise in the execution of these plans, then attention may be redirected to information acquisition and evaluation. Is the policy at fault, or the information upon which it was based, or the implementation, or even the information that it is failing? But such audits and possible shake-ups in organization or procedure must be a transient state, so that the manager can return to managing and the subordinates can return to gathering information and implementing policy.

So the second part of the answer to the question of what is distinctive about rational belief formation is that, in addition to many hard-working subordinate elements in our belief-forming and revising system, we also have the capabilities requisite for this self-conscious, "managerial" level of thought. Moreover, the distinctive nature of beliefs, as opposed to mere representations and expectations, helps make this possible. Nothing about the nature and function of mental representations as such, or the expectations they create, requires them to be *conscious*, or accessible to consciousness. They need only take in information, update, and alter dispositions to think, expect, and do accordingly. Chickens are probably like this. Human belief, by contrast, is often conscious and the experiences or inferences contributing to it are also to some degree accessible to consciousness. One speculation is that consciousness emerged only once the brain became complex enough to get beyond responding to the constant flow of new information and engage in flexible problem-solving and long-term planning[29]—consciousness enhanced these capacities by making various kinds of information "globally" available within the brain, permitting hypothetical, unprecedented scenarios to engage thought, feeling, and action. Moreover, the representations involved in belief are not only often accessible to consciousness, they are richly semantically structured. Such richness, and the inferential possibilities it affords, presumably became possible only with the advent of language, and some theorists speculate that the brain's rapid increase in complexity arose from runaway selection induced by the possibilities suddenly opened up by the first appearances of language. Consciousness and language thus may have co-evolved.

But whatever the truth about our history, once both consciousness and language were present a distinctive kind of epistemic agency could emerge. Just as a manager can carry out an audit, humans can explicitly or implicitly represent their own beliefs and belief-forming processes to themselves, and form and test beliefs or hunches about them—this in turn affecting the rest of

[29] See Crick and Koch (1995).

what we will go on to think and do. Our most striking difference from chickens lies here. At the same time, we share with them a default-governed system that provides, bottom up, most of the evidence and inferential or associative thought-transitions on the basis of which we come to form and test the elaborate, conscious, propositional beliefs that we do. And most of our *reason-based* belief does not come from the distinctive intervention of epistemic agency, even tacit, any more than most of the evidence, calculations, and procedures on the basis of which corporate policies are formed and implemented involve direct managerial oversight. In both cases, this is a very good thing—a person or corporation needs to hear many voices, even voices that notice and say what is contrary to, or even prohibited by, company policy.

Here, then, is a simplified model of belief that brings out the important role of the representations they contain. A belief embodies at least two distinct attitudes, a *degree of confidence* or *trust* in a certain representation, and the *degree of expectation* that this representation elicits.

> **Belief that *R*** (first version)**:**
>
> A *degree of confidence* or *trust* in a representation *R* functions to regulate a *degree of expectation* that things are or will be as *R* portrays them.

Here is an example to illustrate the difference between these two attitudinal components. Suppose that we might be given two small tiles, one perfectly round and one an irregular shape. Each is marked Heads and Tails on the two sides. Initially presented with them and asked what we think the chance of each is to land Heads on a single toss, we might guess .5 . After all, we have no reason to expect one outcome more than the other. So here we have a degree of expectation, .5 in each case, but our degree of confidence in the representation that supports this expectation, "This tile has probability .5 of landing Heads", is quite low. Imagine now that we examine the round tile carefully, and determine that it is perfectly homogenous and symmetrical, although we have never flipped it. By contrast, we have made no study of the irregular tile, and have no general knowledge of how an object with this peculiar shape will behave when flipped in the air. Now there are two quite different beliefs—a very confident belief that the round tile has equal probability of landing Heads or Tails, and a not very confident belief that the irregular tile has this probability. Several noteworthy features of belief are on display here. First, it is a compound attitude. Second, expectations of future behavior can arise not from trial-and-error alone, but from theoretical inference. Lacking propositional representations, even very intelligent animals cannot carry out such hypothetical, modal "dry-labbing". Third, the dynamics of a belief depends not only on its content, but on the compound attitudes it involves. Thus if the two tiles are flipped three times each, and, surprisingly, all six trials are Heads, the doxastic response we

have will differ. Given our strong confidence that the round tile is "fair", we will likely put the three Heads down to chance—after all, this sequence had probability .125 of happening according to us, not a trivial figure—and alter our expectation concerning the next toss marginally if at all. But given our lack of confidence that the irregular tile is fair, we will likely raise our expectation of Heads from .5, and begin to have give some significant credence to the hypotheisis that it is biased towards Heads.

This seems like a perfectly respectable kind of reasoning—it would be quite unreasonable to ignore the difference in our confidence that the two tiles are "fair" when asking how much to revise our expectations. But note that when we reason thus, our reasoning takes into account not just the content of our belief, but its strength, which is no part of what it represents. That is, we are taking a feature of the belief attitude itself as relevant in determining what we should go on to believe. Put another way, the belief *qua* belief does not "drop out" of our reasoning after all. Might one argue that, really, what figures in the argument is *evidence* that the round tile is fair—that is, the content of our beliefs about its symmetry, etc.? In this case, that might seem plausible. But in the case of many prior beliefs, one will have only a vague idea of what evidence, if any, our particular degree of credence has arisen from. Yet here our familiar default rule operates just as always: if one now believes *p* with degree of confidence *r*, then the default continuing attitude is to believe *p* with confidence *r*. Why revise in the absence of new evidence? The alternative is a belief system with insufficient inertia to make thought or learning possible. [30]

Now we must add one more element to our simple model of belief, a feedback loop that is internal to this state, arising from the nature of trust or confidence. These are *affective* states of

[30] It is sometimes said that we can derive norms for rational belief from the fact that "belief aims at truth". There is much to be said on this score. The point I wish to make here is that the mere fact that *p* is true is *not* a reason to believe that *p*. We can see this from the simplified model of belief. From the mere fact that *R* is true, what *degree of confidence* is it rational to have in *R*? Absent any evidence, the answer cannot be given. And if we have no reason to have any particular confidence in *R*, why would it be irrational to have no particular *degree of expectation* that things will be as *R* portrays them merely because *R* is true? "Believe what is true" is not a norm for belief we can follow. "Proportion your confidence in *R* to the available evidence for *R*" and "Proportion your expectation that things will be as *R* portrays them to the degree of confidence you have in *R* and the degree of probability *R* confers" are both norms we can use for some guidance in belief revision, so long as we interpret them in a way consistent with maintaining the inertia of belief. Neither of these two norms can be "read off" the mind-to-world direction of fit of belief, since degrees of confidence and expectation are (typically) no part of what is represented in the *content* of belief, and since these norms do not automatically govern all attitudes with mind-to-world direction of fit (e.g., hypothesis). Rather, these are relations in which an individual can stand with respect to a representation, or to the world around. Epistemology is concerned with the rational management of these relations, one important desideratum of which is of course reliability. Truth matters, but it does not rule. If it did, we'd have no confidence in anything but tautologies, and have no expectations whatever about the contingent facts of the world.

For an elegant argument that the "guidance function" of belief, rather than an "aim of truth", can explain our norms of belief, see Gibbard (2005).

mind that involve: a representation, presenting to the mind a person, situation, idea, object, or action; an affective guise or gloss, usually positive or negative—the person, situation, etc. is presented as frightening, reassuring, attractive, strange, liked, dear, safe, disgusting, etc.; and an associated set of dispositions to notice, selectively focus, be aroused, associate, etc., and conditional dispositions to infer, act, etc. According to an increasingly influential psychological picture of the mind, an extraordinary amount of our mental processing involves, or is mediated by, affect, often unconscious. Perceptions appear to be coded positive or negative even before they are fully semantically interpreted, and this coding in turn affects the interpretation. The affective salience or valence of a situation immediately influences attention and leads to selective processing of information and selective memory. Associative thought appears to be guided by shared affect. And so on. For example, Jonathan Haidt (2007) writes concerning an emerging "new synthesis" in the psychology of morals:

> … the key factor that catalyzed the new synthesis [in moral psychology] was the "affective revolution" of the 1980's… . [S]ocial 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 (simple feelings of like and dislike) which push us gently (or not so gently) toward approach or avoidance.

Affect is normally liable to the effects of feedback, for example, if a person we have implicitly trusted unqualifiedly turns out to have deceived us, we are no longer as confident in him. In the same way, the degree of confidence present in belief is subject to feedback from experience—if expectations are met, confidence in the representation that elicited them is preserved or enhanced, if they are not met, confidence is lessened. This is part of what differentiates belief from a number of other attitudes with mind-to-world direction of fit. That which we merely hypothesize or imagine, for example, will be true or false depending upon how the world is, just as belief is. But it is not trusted in the way we trust the objects of belief, and thus it does not lead automatically to the sorts of expectations belief does, and resultant dynamical potential for being undermined in the face of negative experience. Our full simple model of belief is now:internal feedback **Belief that *R*** (final version)**:**

> A *degree of confidence* in a representation *R* functions to regulates a *degree of expectation* that things are or will be as *R* portrays them; and this *degree of confidence* in turn is modulated by whether in subsequent experience these expectations are met or violated.[31]

In the case of both of our tiles, the initial degree of confidence in the representation that the tile has probability .5 of landing Heads was weak, but in the case of the round tile, and not the irregular one, it became very strong through examination and inference. Thus the occurrence of three Heads in a row left the probability attribution with considerable confidence in the case of the round tile (three tosses is a small sample, and three Heads had the same degree of expectation as any three toss sequence), whereas it lowered our confidence in the equi-probability attribution to the irregular tile (almost the totality of our evidence consists in these three tiles, which as yet exhibit no tendency to land Tails, so the small trust we had in our initial guess is significantly undermined). Here, then, we see in miniature a significant feature of belief—beliefs differ not only in the degree of expectation they support, but in their resilience in the face of contrary evidence. Often, as in this case, we see this difference in resilience as rational. But notice that the degree of confidence that the round coin is fair is not *perfectly* resilient because our degree of

---

[31] For an interesting demonstration at the neuronal level of the operation of this error signal, see Schultz, *et al.* (1997).

One advantage of this compound model of belief is that it suggests why Bayes' Rule gives so plausible an account of ordinary cases of rational belief revision.

$$\mathbf{pr}(\boldsymbol{H}|\boldsymbol{e}) = \frac{\mathbf{pr}(\boldsymbol{H}) \cdot \mathbf{pr}(\boldsymbol{e}|\boldsymbol{H})}{\mathbf{pr}(\boldsymbol{e})}$$

In our example:

> ***H*** = The probability of Heads on a single toss of coin *c* is .5
> ***e*** = coin *c* lands Heads six times in a row

In the language of our toy model of belief:

> **pr(*H*|*e*)** ≈ the *degree of confidence* in *R* after experience *e*
> **pr(*R*)** ≈ the initial *degree of confidence* in *R*
> **pr(*e*|*R*)** ≈ the *degree of expectation* of a given outcome if things are as *R* portrays them
> **1/pr(*e*)** ≈ a measure of the information value of that occurrence of ***e*** would have as a confirmation or error signal.

The Bayesian picture is less convincing when cases of discontinuous belief revision—e.g., thought re-conceptualizing or re-imagining, and thereby effecting a fundamental change in what *R* can be—takes place.

Default models of reasoning are sometimes contrasted with Bayesian models, but it should be clear that the Bayesian model itself depends essentially upon defaults: the prior probabilities must be treated as having standing in rational belief revision, even when we cannot give ourselves reasons for believing them.

confidence that the probability is .5 is not perfect (i.e., not equal to 1.0). If the relative frequency of Heads remains very high in a larger sample of tosses, our confidence level in the initial estimate of equi-probability will decline correspondingly from its high value.

In other cases, resilience does not seem rational. Our simple model also gives us the beginnings of a unified understanding some of the pathologies of belief. Consider first phobias. An individual phobic about flying, for example, even after gaining vast statistical evidence that makes him highly confident in almost all contexts (including travel planning for his own children) that commercial air travel is safe, still fails to keep his expectations of disaster under control once the plane door shuts and the taxiing begins preparatory to take-off. Instead, bare-knuckled, he finds he stop himself from expecting a crash any second. He thus experiences a momentary "decoupling" of expectation in the absence of any new evidence, which makes it hard for us or for him to say without qualification that he fully *believes* that flying is safe, even though he doubts none of the statistical evidence and does not think closing an airplane door gives him new evidence about the likelihood of disaster. Thus the phobic flyer sees himself as less than fully rational in his attitudes toward flying, and experiences his panicked expectations on the plane as a form of epistemic *akrasia*. Phobias, then, can be seen as cases where the normal regulatory mechanism internal to belief fails, resulting in contextually-specific expectations cannot be "unlearned" or properly controlled.

By contrast, in dogmatism or fanaticism, high confidence in a representation has hardened psychically to become impenetrability to feedback from evidence or argument. ("The fanatic is the man who *can't* change his mind, and *won't* change the subject.") We are inclined, though the dogmatist or fanatic is probably not, to see this as a form of irrationality in belief. It is another failure of the normal regulatory mechanisms internal to the attitude of belief. Our language seems to track the sense in which these are not are mere beliefs—we say, "He doesn't just believe that God exists—he has faith!" Faith indeed often advertises itself as evidence-insensitive and argument-proof. Similarly, we differentiate in ordinary speech between someone who believes that a friend or child is has various good qualities, and someone who has *believe in* this individual. When I tell you I believe in you, I am telling you in part that my attitude towards you is not ordinary belief—but a kind of confidence (and usually also also non-epistemic commitment) that will remain intact despite negative evidence. Things may be very bad now, but I won't give up expecting better from you or for you. We also think these willful attitudes should not play the usual belief role in guiding behavior. Though I believe in someone, I should also contingency plan in ways I would not if this were garden-variety confident belief. Though I have faith that God will see me through, I should buy insurance. Dogmatism and fanaticism

thankfully come in degrees, and a very common, garden variety form of irrational belief occurs whenever this sort of unresponsiveness to evidence or argument creeps into our beliefs. Having a strength of confidence disproportionate to evidence is not inherently irrational—defaults work like this—, but when it sets in as a special barrier to learning, we begin to go astray epistemically.

Finally, since there are many sources or bearers of epistemic authority—one's eyes and ears, one's internal sensations, one's recollections or intuitions, forms of inference, trusted friends, the scientific community, the Church, etc.—the attitude of confidence or trust involved in belief must be able to have very wide scope. We must be able, psychologically, to transfer a degree of trust or confidence in our eyes or memory, or in certain people, procedures, or institutions, into degrees of trust or confidence in specific representations they produce, yielding (e.g.) propositional beliefs. But this wide scope and "affective transfer" can also lead to forms of irrationality. Events in one's life that lessen or raise one's overall self-confidence, such as depression, humiliation, sudden fame, and strong praise, can influence strength of belief, even in areas where the success or failure have no epistemic relevance in the eyes of the individual. An individual who hears on the phone that he has scored well on a vocabulary test becomes more willing to make bets on tomorrow's weather; an individual who has become depressed may lose confidence in her professional opinions, despite being a noted authority; another individual may give more credence to what someone says because she is a famous heiress. So, with nothing one can oneself view as evidence or argument, changes in confidence levels, or forms of social esteem and admiration, can "bleed into" the processes of belief formation and revision, producing a kind of "pseudo-learning" or "pseudo-unlearning" disconnected from normal feedback processes.

These three examples of irrationality testify to a *dynamic* conception of rationality in belief—they are cases where degrees of confidence or expectation resist revision by relevant feedback, or undergo revision without it. In none of the cases need the belief-*contents* in question actually be simply false—even in the case of phobias, acrophobics are right that precipices are dangerous, and aquaphobics are right that water is more dangerous than land. What matters is the way the belief, and the believer, responds to the ebb and flow of experience and reasoning, the loss of control internal to belief, or the exaggeration or interruption of normal belief dynamics by other psychic and social forces.

These are also examples of reduced epistemic autonomy. As the case of dogmatism or fanaticism shows, the road to autonomy is not always a matter of distancing oneself from experience and "taking charge" of what to make of it, "making up one's own mind" about what counts as a reason for belief. This can simply be an invitation to *ad hoc* rationalizations. Real

autonomy is possible because of the multiple ways we can receive and respond to information about ourselves and our world, and thus the many chances we have to self-correct so long as we keep our mind open and cultivate ways attitudes that keep our belief dynamics labile and functioning normally. In a given case, autonomy may require a "top down" ("managerial") intervention into the "normal science" processes of updating, to revisit and revise default attitudes, or throw into doubt a large range of beliefs (as when one discovers oneself to have a prejudice). In other cases, what's needed to protect autonomy is an openness to a cold dose of reality to awaken one from one's dogmatic slumbers. Such awakening can come not only "top down", from reading Hume in one famous case, but also, as Hume would insist, from the "bottom up", as re-entering society and playing a game of backgammon dispel the skepticism induced by too much closeted, "top down" thinking. This enables us to see most clearly the multiple capacities—some agential and some receptive—that account for why we are neither slaves to the onslaught of experience, nor prisoners of our own firm convictions and principles. We need not "wait" for conditioning to support, update, or revise our expectations, and we need not "wait" for natural selection or other chance process to acquire, revise, or revolutionize the default expectations and regulative principles to guide us in belief. We can convince ourselves with arguments and rationalizations, but we can also find our confidence in these eroded by unwanted experience. So while we congratulate ourselves on our distinctive forms of epistemic agency, we should be thankful that this agency inherits a large, complex, intelligent system of representation- and expectation-forming and revising processes that operate by default, and which make learning and inference possible, and may cue us when reflection and agency are needed.[32]

5. How to be rational in desire: default, defeasible reasons for action

As we left off discussing reasons for action, we had concluded that desires as such are merely psychological states, not reasons. Some form of assent to, endorsement of, or identification with

[32] The role of self-conscious thought in regulating our lives may be much less than we think, and our fate much more in the hands of non-conscious but intelligent processes. See Hassin, *et al.* (2006). The most spectacular human cognitive errors may come not from excessive trust in piecemeal "bottom up" information, but from our capacity to form abstract, general representations of how things are. If our theorizing can save us from the tyranny of sensation and reinforcement learning, what except the direct and difficult-to-ignore assault of contrary experience can save us from the tyranny of our theorizing? In a famous series of experiments, Antonio Damasio and colleagues showed that subpersonal reinforcement-based learning of the probabilities of chance outcomes goes on even in the presence of higher-order declarative belief formation concerning the same chance process. This subpersonal process can be much more rapid and accurate, can shape behavior directly, and can "cue" higher-order cognition that is slower to catch on. See Bechara, *et al.* (1994)

the desire on the part of the individual—some form of practical agency in which an agent's normative construal of her situation results in her taking a given desire (or, more likely, its object) as a reason for action. It thereby becomes *a* reason and *her* reason in acting, not just a causal factor of some degree of motive force. As in the case of belief, let us allow that this "top down" exercise of agency be tacit and lightning quick. But its presence or absence would nonetheless mark the fundamental distinction between behavior that is properly my own action, and motivated behavior that simply happens in me or to me. "Desires propose; *you* dispose."

And so now we must ask: What guides this disposing? In tiny, transient matters, it may not matter. Just pick. But in life's larger questions, it must be of the right kind to have some normative authority for the agent. I cannot just pick what will have normative meaning or importance for me. Indeed, I could not, even if I tried, simply *declare* that something will or will not matter deeply to me in its own right, and make it so. Moreover, despite their normative "flavor",[33] endorsing, assenting to, siding with, or identifying with a desire are all in themselves nothing more than mental acts or states—and the argument has just been given that mere mental states and operations are not *a priori* necessarily reason-giving.

In order that one's disposings not be arbitrary, they would have to be governed by something non-arbitrary. As acts, they must be governed by beliefs and aims. Very well, let us say that they are to be governed, not by mere desires that assault us willy-nilly, but by our convictions and our values. That is by what we most strongly believe and what matters to us most—what we trust, value, admire, love, prize, respect, and so forth. However, these guiding attitudes, too, are but psychological states. Given the many influences of society and circumstance upon what or whom we trust or distrust, value or despise, respect or disrespect, identify with or see as "other", allowing these attitudes to govern what we endorse—which will, after all, affect the whole course of our lives and the lives of those we care about—without further qualification would be allowing ourselves to be "dragged around" by whatever beliefs and attitudes happened to be present in our minds at the time. This could include: whatever personal affections or animosities I might have developed or chanced to fall into, whatever values or aspirations I might have internalized from my culture, whatever beliefs and principles were drubbed into me, whatever attitudes toward child-rearing I grew up with, etc. What, then, might speak on behalf of the relevant authority of these beliefs and attitudes—or who?

[33] A view commonly encountered nowadays holds that, say, to value *X just is* to take there to be a reason for bringing *X* about. So from valuings, reasons for action follow naturally. But actually, what follows is *taking oneself* to have a reason for action, and this entails nothing about whether one has such a reason.

Surely, in the end it must be the agent himself if the endorsement is to be autonomous. If it is to be my normative voice that speaks throughout, as a self-governing, self-aware agent—rather than "the dead hand of the past, weighing like a nightmare upon the brain of the present"—then these beliefs and attitudes, too, will need my endorsement. I must, at least tacitly, take these to be reasons for endorsing or identifying with some aims in life rather than others. Of course, we have just been told how such *taking to be a reason* occurs: it is a special kind of agency, an initiative of mind in which a normative construal on my part regulates the weight the consideration has in my deliberations. Very well, now we only need to know what will guide this initiative.

But we have been here before. It is clear that we are once again in a regress. However rapid or tacit the agent's "taking as a reason" might be, so long as this is a something rather than a nothing, he will never reach the point of action itself.

One response would be, as discussed in the case of belief, to take this argument as a *reductio* of our notion of practical autonomy or acting for a reason. The only thing that would fill the bill as autonomous rational action involves an impossibility—rather like the view that only agent-causation could be freedom of the will, conjoined with the belief that agent-causation is incoherent. Once again, I think we owe it to ourselves to try an alternative before throwing in the towel. The argument here will also be conditional, and demi-transcendental. *If* (only if?) an account of the sort given below of desires as reasons for action works, then rational action will at least be possible. If it does not work? As of now, I don't myself see another way.

Let's begin with the general point first. We got launched on a regress because we could not accept a "bare" pro-attitude—be it an endorsing, an identifying with, a valuing, an admiring, a respecting, or a desiring—as in itself a reason for action, without further condition. I want to keep *desiring* on this list, since we once we have agreed to consider whether "bare" psychological states could be reasons, we have opened the door previously shut on desire. It won't do any good at this point to appeal to the fact that *valuing* or *identifying with* have a more normative air about them—because we haven't really investigated whether *desiring* might not have a similar, perhaps less visible, normative character. Moreover, it certainly won't do to appeal to autonomy. What individuals value or disvalue, identify with or against, etc. is certainly no less likely to be the result of socialization, narrowness of experience, or contingencies of life course than what they desire. Indeed, looking at the history of stigmatized groups struggling for recognition or equal standing, it often was the fact that individuals could come through experience to *conceive* and *desire* things that were prohibited them not only by the dominant culture, but even by culturally-

laden attitudes they themselves had internalized, that led them to become, often indirectly, agents of change.

However this may be, the project of showing how valuing, endorsing, desiring, etc. could be reasons for action might seem doomed from the outset.

First, we face the problem of *bootstrapping*. From the fact that I have an attitude of endorsement or desire for *X*, I suddenly would have a reason to encourage *X* or act to promote it. But part of what endorsing and desiring involve as attitudes are dispositions to encourage or pursue. So now my attitude-generated reason has come around and patted my original attitude on the back—something for nothing. As before, we will respond to this objection only a bit later, noting for now that it may be better to boot-strap up than to boot-strap down.

Second, we face the problem of "direction of fit". In the standard neo-Humean way of looking at things, attitudes like valuing, endorsing, admiring, desiring, etc. lack a mind-to-world direction of fit, and hence are not capable of truth or falsity. Therefore we cannot reason with them, give evidence for them, etc. in any straightforward way.[34] This sort of claim, however, depends a great deal on 'straightforward'—usually understood as the sort of truth-functional inference or confirmation procedures found in theoretical reason. Suppose there were some "non-straightforward" way of reasoning with or learning about valuing, desiring, etc. There is only one way to answer this question, and that is by looking beyond a blunt picture of attitudes that divides them into two camps, mind-to-world and world-to-mind.

I would like to do this in the case of desire. I think I see ways to transfer this conception of things to other pro- (or con-) attitudes, but I'll settle for making the case for desire, especially since, initially, it seems the *least* promising of the lot, the most "non-cognitive" or "judgment-insensitive".

Some of the resistance to seeing desires as reasons or originating sources of reasons arises from an underestimation of what desiring is, which has been abetted by the unfortunate tendency to treat it as a catch-all for any form of motivation—an urge, a drive, a whim, a craving, an addiction, an ambition, an aspiration, an ideal.[35] Here I will be developing the view that within this grab-bag there is a least one attitude worthy of the name *desire*, and able to earn the name *rational*.

---

[34] One alternative response to this worry is to claim that valuing *X*, for example, is a special sort of motivating belief. It would therefore have mind-to-world as well as world-to-mind direction of fit. I won't be pursuing that alternative here. But I will be considering a view of desire that makes it clear, I think, why there is some attraction in this sort of view.

[35] There are notable exceptions. See Schroeder (2004), Scheuler (1995).

Tell someone other than an analytic philosopher that you're studying desire, and one thought pops into the head: sexual desire. So let's begin there. Consider the difference between a sex drive, such as that of an animal in heat or rut, and an erotic desire. Both involve bodily arousal and directed motivation to "have sex", in some sense. But the former can be the work of exclusively of hormones, usually responds to species-specific biological cues (certain pheromones, swellings, etc.), and typically orients the individual toward performing stereotyped, autonomic courting and mating behavior. The latter is the work of images, narratives, performances, gestures, and music. It ranges across a vast array of cultural differences and differences in mode of presentation, including some of the most indirect or symbolic nature, and orients individuals toward an extraordinary range of actions as well as behaviors, some entirely unprecedented and many purely imaginary or sublimated. This is not to deny that erotic desire involves "mere" biology or "raw" sexual drive (although in some cases it clearly does not), but rather to emphasize that the operation of the erotic works its distinctive wiles through the *mediation* of images, ideas, narratives, and fantasies that seduce us.

The essence of desire, as I will understand that state here, is motivation that is not a bare urge, drive, or appetite, but motivation that is elicited by, and under the guidance of, an idea or image to which we are drawn. Here's an example. A review of a Japanese restaurant in the newspaper presents me with a vivid, intriguing description of a dish I have never tried. The dish is made with a fish I do not know, simmered in sauce made with a fruit I have never tasted, rambutan. This dish is said to concentrate the essence of the mysterious "fifth flavor", *umami*—a taste I have heard spoken of, but never to my knowledge experienced. Suppose that I am someone who has an intrinsic interest in exploring the aesthetic possibilities of the senses and cultural variations. I want to expand my palate to include this taste, and expand my gustatory repertoire to give it a role. Of course I care about gustatory pleasure, but I do not eat simply for pleasant sensations—but also to explore the aesthetic terrain and experience its themes and variations. I am smugly pleased with myself for my catholicity of taste. The review makes me, and presumably thousands of other readers, very attracted to the idea of trying this dish.

I now have a desire, which points me toward the steps I need to take—call the restaurant, make a reservation, etc.—and motivates me to do so quickly. Not because I am hungry, because I'm sure the place will soon be booked solid. My interest in this case is gustatory, but non-appetitive. A busy signal when I call does not deter me, and I keep dialing until I get through. I am delighted to find I can make a reservation. Here I haven't tasted a bite and *already* I feel some satisfaction as a result of the recently-acquired motive. Not because I enjoyed talking with the harried headwaiter, but because I have made the alluring goal one step closer. This, too, is an

image—I see this small accomplishment as part of a narrative laid out before me by bringing my practical understanding into the consideration of this desire. This narrative has come to occupy a place in my life on an otherwise ordinary morning thanks to my "governing" attraction to its *dénouement*. As a result, I now experience the minor but real gratification of, competently, taking a step forward.

The night of the reservation arrives, I collect my companion—who shares my gustatory aspirations and ambitions—, and we endure an excruciatingly slow trip in a bus in cross-town traffic in bad weather. Why excruciating? No one is actually torturing us and we're not out in the snow and blasting wind. We're anxiously looking at our watches because we begin to see a different narrative unrolling, in which we lose our place at the restaurant and go home empty-handed or wait forever for a later seating. When finally we arrive, a quarter hour late, our worry intensifies as we see that the place and the small anteroom are packed, with no headwaiter in sight. But the headwaiter finally appears, and, miraculously (so it feels) he has saved our table despite the many still waiting. we are pleased and very grateful that our reservation has been held, and we tip him handsomely. More satisfaction awaits us when we're told that the special dish is still available that night, despite the crowd. We're pleased to order, and to wait in the hot, noisy restaurant, crammed into a corner. All of this effort, all these bits of gratification or annoyance along the way—all on behalf of, under guidance by, an idea of something we do not know. Still, this none of this pathological. On the contrary, it is human desire functioning in its normal, distinctive way: what we want is being regulated and guided by what we favorably represent, even if the object of this representation is has as yet only the reality of an idea in our minds. Think of desire, then, as a compound, articulated state, similar to belief, in which a degree of affect (in this case, more akin to liking or being attracted to than trust) toward a representation regulates and directs a basic, future-oriented state (in this case, wanting rather than expecting):

> **Desire that *R*** (first version)**:**
> A *degree of positive affect toward* a representation *R* functions to regulate a *degree of positive motivation* toward bringing about the state of affairs that *R* portrays.[36]

[36] This idea is similar to Kant's:

> The **faculty of desire** is the faculty to be, by means of one's representations, the cause of the objects of those representations. … That pleasure which is necessarily connected with desire (for an object whose representation affects feeling in this way) can be called a **practical pleasure** whether it is the cause or the effect of the desire. [*MM* 212]

As in the case of belief, moreover, there is an internal feedback loop. It is of the nature of such positive affect that it tends to be undermined by negative experience (an "error signal"—experience worse than represented and thus expected), and strengthened by positive experience (experience better than represented). Desires in this way differ from other states with world-to-mind direction of fit—the failure of experience to match our positive images does not similarly discourage wishing. Finally, when in desire the experience matches the representation, there is "no news", and the default is for the degree of positive affect to stay constant.[37] The result is:

> **Desire that *R*** (final version)**:**
>
> A *degree of positive affect toward* a representation *R* functions to regulate a *degree of positive motivation* toward bringing about the state of affairs that *R* portrays; and this *degree of affect* is subsequently modulated by whether actual experience of moving toward or realizing *R* is better, worse, or in conformity with expectations arising from the affective representation.

So let's return to the restaurant. When the long-awaited meal arrives, it is beautiful and pungent. A first taste—incredibly intense and wonderfully complex!—though with a bit of aftertaste that's puzzling. Another taste and it has become a bit cloying, with the peculiar, unwelcome aftertaste emerging as the dominant note. A third taste and I realize that I will have trouble eating this—I'm almost gagging. A few more tentative bites, mixed with rice, to check, and that's all I think I'll ever want. My aesthetic pride is somewhat dented—the boundary of my gustatory cosmopolitanism will not be extended in *this* direction. My companion, by contrast, is very pleased—"I really had no idea what to expect, but this is something else. Not pleasant—not something I'd want to eat every night, but fascinating and full of possibilities". Because desire is

---

A similar picture could be given of the negative forms of desire, such as hating, disliking, being averse, etc., with the affect and motivation both flipped to negative.

As we noted above, there is no difficulty about making empirical sense of the idea of an affectively-coded representation—this is in fact a predominant mechanism in the brain, shaped by long evolution to unite in the mind representation and expected value. As Zajonc argued in the paper that began the affective revolution in cognitive social psychology, "preference needs no inference". See Zajonc (1980) and Bargh and Chartrand (1999).

It is sometimes said that we can derive norms for desire from the fact that "desire aims at the good". There seem to be many counterexamples toward this, unless we understand the slogan in something like the following sense: "desire is a mode of presentation in which the object of desire is presented in a positive light". That is the view here, where this is understood not as a *judgment* about the object, but an affective attitude toward it or the presentation of it. When Lucifer says "Evil be thou my good" he is not making a merely functional claim, he is expressing his fundamental, evil nature: he loves the very idea of evil, and, of doing evil because it is evil. In consequence, he is motivated to pursue evil for its own sake—not for the sake of the good. Many mortals have also had episodic bouts of a similar passion, it seems.

[37] For a discussion of the neurological mechanisms here, again, see Schultz, *et al.* (1997).

functioning normally in both of us, the result will be that I no longer am at all attracted to the idea of eating this dish or experiencing this flavor, and, as a result of that, no longer want to do so. She, by contrast, now has an interest in it that is not merely experimental—she'd has revised her image to one for which she has a distinctive degree of liking, and, as a result, a qualified degree of motivation toward pursuing it in the future.

What can we say about direction of fit? In one sense, I have brought the state of the world to fit the object of the idea in my mind—myself, eating this dish. But the result turned quickly from satisfaction to disappointment as I came to realize that my favorable representation of what this states of affairs would be like was in error—I did not expect a *pleasant* taste, but I expected an inherently rewarding, open-ended aesthetic experience and engagement. So in this other sense, the world did *not* come in line with the idea in my mind, to my chagrin. This altered my future-directed wants. My companion, too, was motivated to sample the dish for aesthetic reasons, but had a rather contentless favorable representation of what tasting this dish would be like. She, too, brought the world in line with the idea in her head—herself, eating this dish—but as a result her favorable representation suddenly became wonderfully richer and more definite, exceeding expectations. This altered her future-directed wants, too, but in the opposite way.

It seems natural to say this has been a *learning experience* for both of us. In each case, there was an appearance/reality gap that has been overcome, and an apt, forward-looking response. Via a feedback mechanism very similar to belief revision, a "prior" degree of affect has been revised into a "posterior" of different strength, thereby reshaping what one is disposed to pay attention to, or go on to think and do.

I picked a very simple example, involving an intrinsic aesthetic interest. But we can speak of learning in desire from experiencing "what it really is like" in the case of much more significant desires—desires, say, for a career in politics, or to be rich. Here, too, there will be all the satisfactions and frustrations along the way as the mileposts of the narrative laid out by these long-term desires are set and passed, or obstacles and delays confronted and, if successful, overcome. But, here, too, the proof of the pudding will be in the eating—the lived experience of politicking or having wealth may fail to correspond to the favorable representation that has so long provided an organizing idea for a life and inspiration for motivation. One may learn, perhaps after many years, that "I wanted to be politician, but the real world of politics totally turned me off" or "As a kid, I saw people with wealth and I thought, 'That's what I want to be'. Now I find that money brings no magic transformation of life—I'm the same person, with the same kinds of cares, and the good things in my life seem mostly to have nothing much to do with my money." These people will see themselves as having learned that they were mistaken, even

though they in one clear sense brought the world to fit their aim, and "got what they wanted." Not all desires are personal. One can learn in a similar way about many impersonal desires. For example, someone who holds an ideal of group or social organization, and who has the experience of bringing about or participating in a group or society patterned on this ideal—one thinks of the utopian experiments, communes, religious communities, Soviet or Maoist communism, various corporations, clubs, and voluntary associations, but also immigration into a different cultural or economic system. From the experience of living within such a society, or belonging to such an organization, one can *unlearn*, perhaps *first* through frustrated desires or hopes on a small-scale, daily, interpersonal level, one's passion to see this ideal realized. In fact, this is a familiar story in the explanation of the decay or decline of such social collectives.

The psychologist Daniel Gilbert has coined the useful phrase "miswanting" for desires in which a favorable representation does not correspond to the actual experience of bringing about the object of that representation.[38] It is a natural and pervasive part of life that we are constantly forming new desires, discovering some to be miswantings, and moving on—wised up. Some desires are harder to unlearn than others—some people, for example, complain that the only mates they are attracted to are ones they should avoid. To them, this seems irrational—"Why can't I ever learn?"

As in the case of belief, then, we can think of rationality in desire in a dynamical way. Desires may come to us from many sources, internal and external—of one has no accurate idea of where they come from. So we find ourselves with many motivational "priors". As we saw, it is impossible to demand that such attitudes be endorsed before they can play a role in shaping what we seek. Our rationality in response to desire instead consists in how we and they evolve through the course of experience, and how open we are to new experiences, information, and ways of thinking or being, and consequent ways of relearning what we really want and what really matters. Our lives can be, to a greater or lesser degree, Millian "experiments in living", and can, to a greater or lesser degree, give us the depth and breadth of understanding to see what various different ways of being would be like, and form resilient preferences among them. Such a view helps us understand some of the influential examples of irrationality in desire that have contributed to the view that desires, as such, cannot be reasons for acting.

An unwilling addict, who sees the ruination of his life and no longer derives pleasure from the drug to which he is bound, and hates the grueling daily task of finding a fix, can

[38] Gilbert and Wilson (2000). This also corresponds to the first respect in which Hume thought a desire could be called mistaken: the luscious-looking pear in view was not of "excellent relish" after all. Hume of course saw talk of a mistake in desire as a misleading idiom. His notion of sentiment is much closer to the articulated picture of desire presented here.

certainly be said to *want* this next fix. An obsessive-compulsive who has become aware of how her obsession with cleanliness is cutting her off from her family, her career, her friends, and even the small pleasures in life, can still be said to *want* to escape the touch of her own child's hand, which she cannot help visualizing and fearing as germ-laden, even though she knows that this is unrealistic and that there is no real danger. A 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, [39] can nonetheless be said to *want* to turn on the radio he spies next his unfriendly neighbor, who is sleeping next door on the patio. Despite these wants, the argument goes, these are cases in which the individual has *no* reason carry out the wanted behavior. [40]

What is striking to me about these cases is that they all seem to be *pathologies* or *random quirks* of our motivational system. Addiction and obsessive-compulsion figure in the manual of mental disorders because they are ways in which our motivational system can go awry, become "disordered". Consider a comparison with the "disorder" we saw in the case of phobic belief, in the way in which the elements of belief that normally work together have become "decoupled", as the regulative process internal to belief fails to control expectation. As a result, the phobic has expectations he does not "want" or find credible, and yet which are impervious to epistemic control or unlearning.

The unwilling drug addict and the obsessive-compulsive represent a similar failure, in these cases extreme, of the internal regulative processes of desire. There is a "disorder", an "uncoupling" in which acts or states of affairs that are no longer represented in a way that attracts positive affect—for which indeed there is strong negative affect—remain stubbornly, overpoweringly wanted. In the case of addiction, certain natural substances for which the brain has especially sensitive receptors can produce altered experiences that are initially exhilarating, enjoyable, calming, etc. But with repeated exposure the brain becomes "sensitized", and the individual requires larger and regular dosages simply in order not to feel bad, or (much) worse. Kent Berridge and colleagues have analyzed this sort of addiction as, in effect, a chemical hijacking of a portion of the motivational system of the brain (what he calls the "wanting" system), creating a state in which there is an insistent, unmodulated demand for the substance that continues to operate irrespective of whether the individual is experiencing any pleasure or

[39] The example is Warren Quinn's. See Quinn (1993).

[40] Here we need to make a distinction. We should set aside reasons generated by desires indirectly. For example, any strong desire could be a discomfort, and until satisfied a source of frustration, so that the bare fact of the desire gives *some* one reason to satisfy it, simply to avoid these unwanted experiences. But this reason does not speak at all to the question whether there is a reason in favor of attaining the *object* of desire. Thus an equally good way to address this indirect reason would be find a way to kill off the desire.

positive experience from its use, or sees it positively (what he calls the "liking" system).[41] A forceful "wanting" thus is no longer regulated by any active "liking", and the individual experiences an incessant motivational demand that can monopolize his life. Yet this demand is insensitive to reasoning or learning, to changes in the intrinsic quality of the experience, or to competing desires. Obsessive-compulsives, for other reasons, have dispositions to act that they cannot control, and cannot alter by learning or reasoning even after they have stopped gaining any satisfaction from or seeing any point in the objects of those acts. The "radio man" seems to have a (usually) harmless compulsion, but it is nonetheless a compulsion—a wanting that is alien to him, unassociated with any favorable representation that might explain its force and object, or any enjoyment that might make its pursuit rewarding. Just as it was difficult in the case of the phobic flyer without qualification that he fully *believes* flying unsafe, it is difficult in the present examples to say without qualification that these individuals truly *desire* the object of their "uncoupled" wants. It seems much more plausible to describe them as being in the grip of an *urge*, and irrational in desire—and very likely they will agree.

These are cases where the dynamics of feedback and regulation that figure in normal desire have broken down, and where learning and unlearning have become impossible. Much more mundane examples of irrationality in desire operate somewhat similarly. The dieter who has a strongly favorable representation of being thinner and more self-disciplined, but who is confronted with a very salient, alluring dessert, finds that his want to realize what he most favorably represents is swamped by a more insistent want that even he sees as on the whole unfortunate.[42] An individual who loses overall self-confidence or becomes depressed, this flattening of affect can "bleeds into" desire and sap motivation. She may find that she cannot sustain a significant level of positive affect even for a social cause that has previously been vital to her. It seems to have become unimportant, yet she has learned nothing—even by her own lights—that would would suggest this. After all, her representation of the cause (say, of decent health care for the poor), was never as important because it mattered to *her*.

In normal desire, learning goes on constantly at the subpersonal level, as positive and negative experience feeds back into how favorably we see various ends or objects. All this is part of the general "affect regulation" that predominates in the brain—in animals as well as humans.

---

[41]Wyvell and Berridge (2000, 2001). They provide a range of behavioral and physiological evidence of the separateness and possible decoupling of the "liking" and "wanting" systems.

[42] Though in this non-compulsive sort of *akrasia*, it is worth noting that the want for the "tempting" item is usually mediated by a tempting image—it is not the actual taste of the fancy dessert that elicits the otherwise unwelcome urge to go ahead and order it, but the alluring mental representation or visual presentation of the dessert. This is why the subsequent behavior makes sense to the individual—he does seem something positive in it.

Distinctly human desire is possible because this normally subconscious affective coding can be brought to the surface, and self-consciously thought about. The semantic character of the representation that mediates motivation in desire makes it possible for inference and other forms of reason to work with desire, extending the reach or generality of desire, locating means, bringing to bear evidence concerning the accuracy of the affective representation. Notice that in such cases, what appropriately figures in our deliberating is not just our understanding of the content or object of a desire, but also (as we saw in the case of belief) its strength, intensity, or resilience. It is perfectly sensible that in many of life's choices a key reason to act one way rather than another is a matter of deciding whom or what one likes most or is most attracted to, whom or what one cares most strongly about, which activities, relationships, careers, etc. one feels most enduringly or passionately about.[43] These are not usually value judgments—who is the better person, more deserving, or which is the better place to live or post-retirement avocation to pursue—but serious attempts to understand something requiring real effort to understand: what makes one tick, what one really wants, what one truly finds interesting, whom one really cares for or would hate to lose, and how frustrated one would be to give something up.

Again, I certainly do not mean to suggest that all reasons are of this kind. Moreover, abstract values can themselves be objects of the most intense desire, and when we act on their behalf it isn't a question of "satisfying a desire". Napoleon, who knew good deal about how to elicit motivation through representation, may be famous for having said that an army marches on its stomach. But he did not expect his troops to march into canon-fire for the sake of a hot meal. He knew the importance of inspiring the love of ideals and glory, the passion of regimental honor or liberation from tyrants, and devotion to comrades-in-arms. In the end, he wrote, "Imagination rules the world."

Desire, our human, imaginative way of being motivated, is thus different from animal appetite in the same ways that belief is different from animal expectancy—in both cases, the crux of the matter is that an affective representation that gives an idea a key regulative position in our lives—what we want in life, what we expect—also serves it up for reasoning and experience to be brought to bear. In the case of desire, positive affect provides a semantic representation from which to deliberate—for example, in thinking through what a satisfying a desire might lead to, or require, or in assessing one's progress and prospects in fulfilling it. Mechanisms of "affective transfer" can then make something which we can represent as a necessary means, and which, in itself, we would simply dread and avoid, into something we actually *want*, and will expend

[43] Here I differ with Pettit and Smith (1990), who claim that, except in a few cases such as craving for a cigarette, one's desires are always "back-grounded" in practical deliberation.

mental and physical resources in pursuing, with attendant satisfactions and frustrations. We find ourselves *wanting* to get to an exam or meeting we otherwise would hate. In desire, then, as in belief, there is an enormous gain in flexibility and learning harnessing a powerful "animal" system for regulating behavior—expectancy or motivation—to the power of ideas, thought, and norms.

The distinctive features of human belief and desire create: (1) potential roles for abstract knowledge, deliberation, norms, imagination, and communication in the *forward* dynamics or regulation of representational and motivational states, and (2) the possibilities for knowingly, deliberatively, normatively, imaginatively, and collectively combining representations and motivations to form intentions of unprecedented kinds—for novel aims or courses of action, for the attainment of distant or complex goals, for shared action on almost any scale, and for the realization of abstract ideals.

In neither case should we think that it is only at the "high" or cognitive end that we find responsiveness to reasons or normative authority. There is ample evidence that our marvelous cognitive, deliberative, and imaginative capacities can lead us down the garden path, or into insane fixations, or into moral horror. These features make desire infinitely more flexible and suggestible than drive or appetite, no doubt partly explaining how humans have managed, unlike any of our ape ancestors, to move into entirely unprecedented habitats. Unfortunately, it also explains how desires can lead us, individually or collectively, very badly astray.

6. Conclusion

A plausible approach to rationality in belief holds that no beliefs are *intrinsically* rational. 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 an authority one has reason to suspect. What makes for rational belief, then? Not whether an individual simply *has* a certain belief, but how she has come to have it and how that belief has since evolved and can evolve in response to new evidence or argument. Direct perception, without deliberation or confirmation, can yield perceptual beliefs we have good reason to have. Immediate transitions in thought that follow patterns of logical inference can likewise yield rational belief. And so, of course, can reflective assessments of one's beliefs in light of plausible norms or considerations of mutual coherence. All of these ways of believing

work together in rational believers. Rational believers moreover maintain open minds and thus typically accord varying degrees of confidence to conflicting objects of belief (e.g., *p* and not-*p*). This gives them the "seeds" from which to grow changes in belief in response to trial-and-error or sheer inspiration. At the same time, rational believers feel some pressure toward consistency and probabilistic coherence. The rational believer takes the fact of believing that *p* (to some degree) as a *default, defeasible reason* to expect *p* (with proportionate probability) and to take *p* as a premise in inference. Without this default she would be trapped in a reason-seeking regress. But at the same time, she shows her autonomy in being open to many ways of *defeating* or *revising* that default through experience, new ideas or arguments, and communication with others.

A similar picture, I would argue also plausible, can be given of rationality in desire. No desire is *inherently* rational or irrational. A desire for a genuine good that is exaggerated or obsessive, or would be impossible to unlearn even through contrary experience, can be irrational, and a desire for something not at all good that has arisen from first-hand, positive experience, or from inference involving false but justified beliefs, can be rational. Desirers and desires display rationality by being dynamically sensitive to changes in experience and belief, and by seeking to enlarge their sphere of experience and understanding. In such a setting, even desires that are effectively mutually antagonistic (as much so as love and hate) can be rational. Rational desirers maintain open minds and thus typically feel varying degrees of affect toward conflicting objects of desire, and this gives them the "seeds" from which to grow changes in desire on the basis of new experience or inspiration. At the same time, rational desirers feel some pressure toward forming their desires into coherent aims and lives into coherent plans. The rational desirer takes the fact of desiring that *R* or liking the idea of *R* (with a certain intensity) as a *default*, *defeasible* reason to want (with proportionate intensity) to bring *R* about. But at the same time rational desirers show their autonomy through the many ways they are open to revising or defeating these defaults in response to new experience, ideas, and people.

If something like this picture of belief and desire alike is right, we can see how practical reason—which is very much concerned with the formation and revision of desire—can solve the "Problem of Significant Digits" for rational action mentioned at the outset. Practical reason need not magically suspend the Rule of Significant Digits—but only apply it. For we can begin to see how we might fill in the blank in (2), now yielding (3), not by theft, but by the honest toil of learning from life:

rational belief + **rational desire** + **rational deliberation** →
rational intention → **rational action**

References

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.

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.

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. (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-