Forthcoming as Essay XIV of *Reference and the Rational Mind* (CSLI Publications)

Toward a Naturalistic Theory of Rational Intentionality[1]

I. Preliminaries

This essay some first steps toward the naturalization of what I call rational intentionality or alternatively type II intentionality. By rational or type II intentionality, I mean that full combination of rational powers and content-bearing states that is paradigmatically enjoyed by mature intact human beings. The problem I set myself is to determine the extent to which the only currently extant approach to the naturalization of the intentional that has the singular virtue of not being a non-starter can be aggregated up into an account of rational intentionality. I have in mind a broadly defined family of accounts whose main members are the indicator/information-theoretic approach of Dretske (1988), the asymmetric dependence theory of Jerry Fodor (1987, 1990, 1994) and the teleo-semantics of Ruth Millikan (1984, 1993). Somewhat inaccurately, I will call this family of approaches the information-theoretic family. To be sure, there is only a rough family resemblance among the members of the information-theoretic family. Indeed, several intense quarrels divide the members of that family one from another, but the precise outcome of those internecine struggles is not directly relevant to the aims of this essay.[2] Taken collectively, the information-theoretic family yields a compelling picture of the place of at least a crude form of intentionality -- what I call frog-like or type I intentionality -- in the natural order. Though frog-like or type I intentionality is, I think, a genuine species of intentionality, it may subsist in the absence of rational powers. It is that species of intentionality enjoyed by irritable creatures who, following Brandom (1994)

---

[1] This Essay has been long in the making. It began its life years ago under a different title and with a different argumentative structure. Under that title, versions of it were presented at Stanford University, the University of North Carolina at Chapel Hill, East Carolina University, Virginia Tech, Temple University and The Graduate Center of CUNY. Under its current title, it was presented at the University of New Mexico and at Stanford University. I am grateful to many in those audiences. I have been thinking these issues through for so long, however, that I can't for the life of me remember exactly who said what that led me to make one of the many, many changes this essay has undergone over the years.

[2] It does not, for example, do justice to Millikan's teleosemantics (1984, 1993) to class it as a variety of information-theoretic semantics. Indeed, Millikan has argued at great length for the superiority of teleosemantics over indicator semantics. If I could think of more descriptive name for the rough family, I have in mind, I would use it. Unfortunately, I can't.

might be said to enjoy sentience without sapience. Just because type I intentionality may subsist in the absence of rational powers, some have been tempted to dismiss such low-level intentionality as a mere simulacrum of the real thing. Such thinkers will no doubt find the hope that we might somehow aggregate up from an information-theoretic account of mere frog-like intentionality to an account of rational intentionality a vain hope.

But that assessment misconstrues the significance of frog-like or type I intentionality. Broadly information-theoretic approaches have never been advanced as the complete story about intentionality.[3] Such approaches are best understood not as putatively comprehensive naturalizations of the full panoply of intentional kinds, but as existence proofs of a sort. They show that there does exist in nature something with many of the hallmarks of intentionality -- even if it is only type I intentionality. They thereby reduce our conceptual puzzlement about the very possibility of intentionality as an element of the natural order. To say this is not to deny that we have not yet hunted down the larger quarry. But that, in its way, is entirely unsurprising, since it is really a rather straight-forward consequence of approaches in the broadly information-theoretic family that type I intentionality can subsists even in the absence of rational powers.

My ultimate aim is to show that it takes two things to bridge the gap between type I and type II intentionality: the emergence of what I call logical syntax and the emergence of rational powers. I hold, that is, that type II intentionality results from the marriage of type I intentionality with logical syntax and rational powers and that both logical syntax and rational powers are fully natural. My goal in this essay, however, is not to mount a full scale defense of this ultimate claim. I shall focus primarily on the naturalization of rational powers -- though I shall offer some passing remarks about the nature of logical syntax.[4] And I will close with some remarks about the way in which the marriage of content, rational powers, and logical syntax may be effected in a single natural system.

The bulk of this essay is an exercise in speculative naturalistic psychology. I describe a naturalistic psychological make up which is plausibly our own, which attempts to honor and reconcile three distinct intuitions about rationality and rational powers. The first intuition is what I call the

---

[3]Millikan (1984, 1993, 2000) is the clear exception here. She clearly aims to give a comprehensive naturalistic account of the rational mind, its states and capacities.

[4] For more on the logical syntax of thought see especially Essay VII, Essay I, and Taylor (forthcoming-a)

consequentialist intuition. It is the intuition that rational powers are non-accidentally linked to worldly success. According to the consequentialist intuition, being rational is being such that one's belief-fixation and action-guiding mechanisms reliably and non-accidentally deliver the worldly goods.[5] The proper deployment of rational powers in the generation of action is supposed to guarantee, or at least render it non-accidental, that the rational actor engages in desire-satisfying rather than desire-thwarting actions. The proper deployment of rational powers in the formation of beliefs is supposed to render it non-accidental that our beliefs track the epistemically good, whatever that turns out to be.

The second intuition I seek to honor might be called the internalist intuition.[6] The internalist intuition represents rationality as a matter of the internal coherence of the will and/or intellect. According to the internalist intuition, the proper deployment of rational powers guarantees that the rational will and intellect will not be "self-defeating." The proper deployment of rational powers guarantees that the believer does not simultaneously believe ***p*** and believe not ***p*** and that the desirer does not simultaneously desire ***q*** and desire not ***q***. It thereby guarantees that if we lack worldly success in our conation and cognition then at least the will and intellect themselves will not be at fault. If internal coherence of the will and intellect were *ipso facto* sufficient to guarantee the worldly success of our cognition and conation, then the internalist intuition and the consequentialist intuition would be easily reconciled. But we shall see that the easy reconciliation is not available to us.

The third intuition about rational powers that my account seeks to honor is what I call the Kantian intuition. The Kantian intuition represents reason as the source of its own norms. Now it is fashionable among those of an anti-naturalistic bent to distinguish the order of norms from the order of causes, to distinguish the "must" of causation from the "must" of normativity.[7] Though there is surely something to

---

[5] The consequentialist intuition is, I think, widespread. It underlies pragmatism in epistemology, utilitarianism and other forms of consequentialism in ethics, as well as classical decision theory as well.

[6] There are many doctrines that go under the title 'internalism.' One can be an internalist about content, an internalist about epistemic justification, an internalist about the standing of (practical) reasons as reasons in the sense of Williams (1979). Though what I am calling the internalist intuition is probably compatible with all these varieties of internalism, I doubt it entails any of them.

[7] Contemporary neo-Kantians are often anti-naturalist. Korsgaard (1996) comes quickly to mind, for example, as one who shares something like the Kantian intuition, at least about moral norms, but doubts that normativity can be naturalized. Anti-naturalist Kantians simply fail to appreciate even the possibility

this distinction, I shall argue the reason in its norm-constituting role is not, as Kant and other anti-naturalist imagine, a mysterious other-worldly power with no place in the natural order. The normative power is rather a fully natural power of the mind-brain by which it constitutes for itself and out of itself an order of norms. Further, I shall argue that the Kantian intuition provides the central clue to the proper reconciliation of the consequentialist intuition and the internalist intuition.

II. Toward the Naturalization of Reason

2.1 The Consequentialist Intuition

Beliefs are reasonable in light of evidence and other beliefs; actions are reasonable in light of beliefs and desires; desires are reasonable in light of needs and/or values. But what is a reasonable belief? To a first approximation, a reasonable belief is one licensed by a rational method of belief-fixation. A method of belief-fixation takes evidentiary input, together with "old" beliefs, and yields some (possibly new) beliefs as output. If the reasonableness of a belief is determined by the rationality of the method of belief-fixation that delivered that belief, then, on pain of circularity, the rationality of a method of belief-fixation cannot be determined by the reasonableness of the beliefs it delivers. So we must characterize, in non question-begging, and ultimately naturalistic, terms what rationality for a method of belief-fixation consists in. According to the consequentialist intuition, a rational method of belief-fixation is a method that reliably and non-accidentally delivers "good" beliefs. Similarly, a rational inferential mechanism is an inferential mechanism which is at least goodness preserving. A reasonable belief is a belief, good or bad, licensed by a goodness delivering method of belief-fixation or a goodness preserving inferential mechanism.

Talk of the goodness of a belief is so far devoid of substantive content. We can give it substance by considering the various roles that beliefs occupy in the mental economies of believers. To a first approximation, good beliefs are beliefs such that it is good for the organism whose beliefs they are to have

---

that the norm-making power of the human mind is a brutely psychological capacity. On the other hand, contemporary neo-Humeans, like Gibbard (1992), Blackburn (2000), are typically thoroughly committed naturalists, but their naturalism is often accompanied by too little appreciation of the complex hierarchical organization of human normative psychology. I aim to produce an account of the normative faculty that is as thoroughly naturalistic as the most diehard neo-Humean would want, but as psychologically deep as the committed Kantian could wish for.

(or aim at) them. Beliefs of character □ will be good for an organism to have (or aim at) just in case either having (or aiming at) beliefs of character □ reliably and non-accidentally "pays off" for that organism. The calculus of payoffs cannot be determined a priori by bare reflection on the very concept of a belief, but only by facts about the biological needs, computational and cognitive resources, motor capacities, and environmentally determined opportunities of particular organisms. Speaking conatively, beliefs of character □ are good for an organism ***o*** to have just in case when o has beliefs of character □ and deploys those beliefs in episodes of practical reasoning, ***o*** 's actions lead reliably and non-accidentally to the satisfaction of ***o*** ' s needs and/or wants and/or values. It is always a broadly empirical matter to determine, for particular organisms, just what the relevant □ is.

We must distinguish, at least in concept, the goodness of *possessing* beliefs of character □ from the goodness of *aiming* at beliefs of character □. Imagine a predator for which the having of true beliefs is great good in the sense that whenever its beliefs are true, its actions are reliably successful. Suppose, however, that, as a contingent matter of fact, its method of belief-fixation delivers no belief as true until confidence of truth is high. It is not hard to imagine that for such a creature the cost of aiming at truth may outweigh the benefits of striking the truth. One need only imagine that the cognitive and motor endowments of the predator, together with the speed of the prey, together with the "hostility" of the surrounding environment, conspire to make delay costly indeed. To be sure, even for a system such as this the organism's belief-fixing and action guiding mechanisms might still be so coordinated that *if* its beliefs are true, then, reliably and non-accidentally, its beliefs and desires interact to cause desire-satisfying actions. Moreover, a truth-seeking organism whose belief-fixation mechanism fails to deliver the truth might be better off for aiming at the truth. Compare such an organism to an archer who aims at the bull's eye but is destined mostly to miss. Such an archer might get as near as he can to the bull's eye by aiming at it, even though he is destined mostly to miss.

Might the bare possession of true beliefs fail to be a good? It might if the cognizing agent's belief-fixation mechanisms and action-guiding mechanisms were so coordinated that true beliefs and desires interacted to cause desire-thwarting, rather than desire-satisfying actions. Even so, we might count such an agent rational when its desires were contrary to its needs. For such an agent, it might be that its

own ineptness at desire satisfaction is at the core of its ability to thrive. From a purely consequentialist perspective, there is no obvious reason to deny that status of rationality to such a cognizing agent. Still, it must be admitted that we typically think that in a rational mind, some desires will be rationalized by needs (and perhaps by values as well). It seems doubtful that desires could be *arbitrarily* unrelated to needs in the mental life of a rational cognizing agent.

On the other hand, consider a self-assessing predator -- call him Hapless -- whose native endowments and external situation conspire to render his chances for success at predation bleak. Hapless is too slow and dumb, and his prey too fast and smart for Hapless ever to be a good bet to capture his prey. Hapless's psychological makeup is such that were he aware of the mismatch between his endowments and his prey's, he would be overcome with (justifiable) self-doubt; cease his predation; and die. Fortunately, Hapless systematically and imperviously overestimates his own skills and underestimates the abilities of his prey. He has a false but unshakeable belief in his own prowess. He attributes his failures either to random freaks of nature or to the systematic interventions of some malevolent (if unseen) force that is out to thwart him. He is the Ross Perot of predators. Hapless may be accused of either bad faith or self-delusion, but these apparent cognitive "failings" serve him well. His obliviousness to the truth allows him to keep doggedly to his mostly inept, but often enough successful predation. In the short run, he ekes out a meager existence, an existence more fruitful than he would otherwise enjoy. Hapless, therefore, benefits from believing certain sorts of falsehoods and from not believing certain sorts of truths. Indeed, given his peculiar cognitive endowments and psychological makeup, he may well be doing the best that he can. If one is inclined to think in purely consequentialist terms about rationality, it is hard to find a basis for denying that even Hapless's unshakeably false belief as an instance of rational believing, despite its unshakeable falsity.

Our fanciful scenarios have a serious moral. If we take the consequentialist intuition utterly seriously then rationality will turn out to depend not merely on what beliefs a cognizing agent has, nor even on how those beliefs are acquired. If she cannot exploit the truth in pursuing the good, then even a cognizing agent who believes the truth on the basis of truth-delivering methods of belief-fixation may fail to be consequentially rational. Even true beliefs must earn their standing as rational by contributing (non-

accidentally) to the believer's worldly success. Even imperviousness to the truth may be insufficient to rob an episode of believing of its standing as an exercise of rationality.

Some, no doubt, will be tempted to trumpet the intrinsic epistemic goodness of truth and the intrinsic epistemic badness of falsehood. But from our stubbornly consequentialist perspective, both the question whether truth, or any other supposed epistemic or conative good, is worth aiming at and the question whether truth is worth possessing is never a question of intrinsic goodness or badness. Both the epistemically good and the good in the way of action can be determined only relative by calculus defined over the biological needs, computational and cognitive resources, motor capacities, and environmentally determined opportunities of particular organisms. Consequentialism thus suggests a promising strategy for naturalizing rationality. To a first approximation, being rational is a matter of deploying mechanisms that reliably and non-accidentally deliver naturalistically constituted worldly epistemic and conative goods. Reason exists in nature by this line of thinking if such mechanisms exist in nature. If there is no bar to the existence of such mechanisms, there is no reason to deny that rationality may have an echo in nature. We might call rationality so understood consequential rationality.

The consequentialist intuition leads to a kind of relativism. If the epistemic and conative goods are constituted as goods only by facts about particular organisms and their peculiar capacities and opportunities, different methods of belief-fixation may be rational for different organisms. So we are not entitled to suppose a priori that rationality is always and only rationality by our own lights. We may be unable to determine what standards of rationality are appropriate for others and unable to apply those standards as instruments of rational evaluation. That, however, is an epistemological point with no direct metaphysical significance. What goes for creatures unlike ourselves, goes for us as well. Our own current norms of rationality, which may seem inevitable to us, may not be with us to stay. That is, we may someday come to learn to evaluate ourselves relative to different norms of rationality.[8]

The consequentialist embraces the possibility that even for a single organism, there is no single standard of goodness for its beliefs and thus no single measure of the rationality of its method of belief-

[8] Stich (1990), for example, draws just such a conclusion on largely consequentialist grounds. See also the work of evolutionary psychologists like Gigerenzer (2000) with their appeal to what they variously call bounded, situated, or ecological rationality. This too is a species of consequential rationality.

fixation. The goodness of beliefs, and thus the rationality of a method of belief-fixation, is determined by the role or roles that beliefs play in the cognitive and conative economies of an organism. For organisms with sufficiently complex cognitive and conative economies, beliefs may play a multiplicity of roles, each of which gives rise to a different standard of goodness for beliefs. Some organism may have an intrinsic epistemic interest in having true representations of the world, while aiming at beliefs that are "serviceable" whether or not they are true where action is concerned.[9] For any such organism, standards of epistemic goodness may diverge sharply from standards of prudential goodness. If there is such a thing as the overall goodness of a belief or the overall rationality of a method of belief-fixation, that will have to be defined, the consequentilaist will say, in terms of some weighted sum of the various interests relative to which particular forms of goodness and thus rationality are defined. Moreover, the consequentialist will regard it as another and non-trivial matter to spell out the principles in terms of which the relevant weights are to be assigned. This too would be a matter to be decided on a case-by-case basis. It too will depend on facts about the endowments, opportunities, needs, desires, and values of particular organisms.[10]

2.2 Toward the Internalist Intuition

First an obvious point. Short of our achieving god-like infallibility, there is a clear sense in which even methods of belief-fixation which we may be pleased to call rational will never be guaranteed to deliver the aimed at worldly goods. Even "bad" beliefs can sometimes be reasonable. Consider an analogy with bets and betting strategies. In one sense, any winning bet is a good bet and any losing bet is a bad bet. Call the goodness of a winning bet goodness$_1$ and the badness of a losing bet badness$_1$. But now

[9] One sort of pragmatist, of course, simply equates the serviceable – at least the serviceable at the limit of inquiry – with the true. Notice, too, that consequentialism opens up at least the conceptual possibility that we may ameliorate our rational faculties, by coming to deploy them in ways that are more supportive of, as it were, our being in the world.

[10] Consequentialist intuitions carry some bite for the plethora of studies which purport to find pervasive irrationality in human inference strategies.(Nisbett and Ross, 1980, Kahneman, Slovic, and Tversky, 1982). To evaluate the rationality of an inference strategy, the consequentilist will say, we must determine what worldly "goods" that strategy is intended to deliver and then determine whether it delivers, reliably and non-accidentally, the intended worldly goods. We cannot antecedently assume a particular standard of goodness. Only if the calculus of payoffs determines that for us humans truth is goodness for beliefs, and thus that rational inference strategies are those that reliably and nonaccidentally deliver the truth, would it follow even from our pervasive failure to employ truth- preserving inference strategies that we are

consider betting strategies. Suppose that a good betting strategy is any strategy with a positive expected return, one such that, on average, wins are greater than losses. Even if you play a good strategy, you sometimes lose. So there is a sense in which even a losing bet can be good (call this $goodness_2$). A bet, win or lose, is $good_2$ if it is placed in accordance with a good betting strategy. The $goodness_2$ of a bet is thus inherited from the goodness of the strategy in accordance with which it is placed. Just so, a belief is rational ($good_2$), whether true (or more generally $good_1$) or false (or more generally bad_1), just in case it is arrived at by a rational method of belief-fixation.

So far, the consequentialist will be untroubled. She can readily allow that a strategy's goodness is not determined simply by the bare numerical ratio of wins to losses. A strategy that delivers infrequent but large wins may be better than a strategy which delivers many small wins, if the wins it delivers are large enough (and the losses small enough). Similarly, even assuming an organism for which truth is goodness, a method of belief-fixation which delivers few, but highly consequential -- explanatorily deep, serviceable for action -- truths even at the cost of frequent falsehood, may be preferable to a method of belief-fixation that delivers many, but inconsequential truths, even if it seldom delivers bad beliefs. Not all truths are equally relevant to the conative and cognitive strivings of a even a truth-seeking cognizing agent. Truth seekers seek truths that matter. And the consequentialist will say that there must be a calculus of pay- offs, a calculus entirely determined by the worldly needs, resources, capacities, and opportunities of the relevant cognizing-agent.

Just as the soundness of a betting strategy depends on the strategy of one's opponent, so the success of a method of belief-fixation in delivering the goods often depends on the nature of the surrounding environment. Indeed, though environments are not agents, from a consequentialist perspective it is heuristically useful to think of a cognizing agent as playing a competitive game against its environment. To a first approximation, a consequentially rational cognizing agent is one with winning strategies in competitive games played against its environment. Imagine a predator whose prey-whereabouts beliefs are generated by preyish appearances. Suppose that two different animals -- junk food and real food -- have just that appearance. Our predator pursues junk food and real food indifferently and

---

pervasively irrational. Unless and until we independently establish that truth is goodness, no such conclusion is licensed.

faces a variety of possible fates. If the environment is very rich in junk food and very impoverished in real food, then the strategy of pursuing whatever has a certain preyish appearance may often culminate in meals of low nutritional value. If the relatively infrequent captures of real food are high enough in nutritional value, then a strategy of pursuing junk food and real food indifferently will be successful enough. Where junk food is rare enough in comparison with real food that most captures are captures of the real thing, this strategy can succeed even where each quantity of real food is only marginally more nourishing than junk food.

So far, our examples may appear to demand only a more nuanced conception of consequential rationality, in all of its worldliness. We can begin to pump internalist intuitions, however, by considering two predators -- Happy and Sad – who both pursue junk food and real food indifferently. Sad lives in an environment rich in junk food and impoverished in real food, with each quantity of real food being of relatively low nutritional value. Happy lives in an environment rich in highly nutritious prey and impoverished in junk food. Happy is clearly more "fortunate" than Sad. Is Happy "rational?" Is Sad "irrational?" Happy pursues a strategy that tends to deliver the nutritional goods. So Happy has a winning strategy in the competitive game he plays with his environment. Sad, by contrast, lacks a winning strategy. Why not say that Happy is rational, while Sad is not?

Suppose that Happy and Sad are as above. Suppose also that a predator who ingest a quantity of junk food "feels" just as nourished as it would feel if it had ingested a similar quantity of prey. Suppose further that whenever the predator reaches a certain level of "satisfaction" (or pseudo-satisfaction) it ceases, for a period, to pursue further nourishment. Clearly, the strategy of pursuing junk food and real food indifferently may have very disastrous consequences for Sad, even where the consequences for Happy are not so disastrous. Suppose, for example, that due to the luck of the environmental draw, most of Happy's captures happen to be real food, while most of Sad's captures happen to be junk food. Is Sad therefore less rational than Happy?

Suppose that Sad is in no position to discover the difference between genuine nourishment and pseudo-nourishment. By Sad's lights, he is getting all the nourishment he needs. Sad has no rational basis for altering his strategy. If Sad can have no rational basis for altering his strategy, then what basis can we have for condemning Sad's foraging strategy as irrational? Moreover, "from the inside," Happy's current

environment is just the same as Sad's current environment. If Happy were suddenly switched to Sad's environment, he would go on just as he always has, pursuing just the strategy that he always has. That means that Happy might pursue his current strategy even if it failed to deliver the nutritional goods. It would seem, therefore, little more than a lucky accident that Happy's strategy in fact delivers the goods. The difference between Happy and Sad in the case imagined, then, is not a difference of rationality, but of fortune.

If you resist damning Sad as irrational, you are being tugged by the internalist intuition. That intuition has it the rationality of an actor depends entirely on facts internal to that very actor. By internalists lights, rationality isn't directly or essentially a matter of worldly payoffs.[11] To be sure, the internalist may attempt to accommodate the intuitions of her consequentialist friend by allowing that the rationality of a method of belief-fixation depends on whether an intact cognizer would upon reflection endorse that method as likely to produce payoffs. This likelihood is only likelihood by her own lights. Notice that the internalist need not deny the availability of external evaluative standpoints. She can take evaluative notice of the fact that a cognizing agent's cognitive and conative apparatus fails to the deliver the worldly goods. She may even conclude on such external evaluative grounds that a given cognizing agent does not have the method of belief-fixation that she "ought" to have. But unless that inadequacy is detectable, as it were, from the subject's own perspective, deploying the subject's own cognitive apparatus, the internalist will say, then there is no basis for criticizing the subject as irrational. By internalist lights, rationality is a quite peculiar cognitive virtue. The rationality/irrationality of a cognizer is a matter of *its own* cognitive and practical successes and failures. Moreover, the cognizing agent must be able to own those successes and failures as her own from a purely internal or first-personal perspective. If nature has provided a cognizer with discriminatory capacities only this or that sharp or has provided it with a method of belief-fixation that will deliver the goods in only a relatively narrow range of environments, there may be an external standpoint from which we may note and lament the unfortunate mismatch between the

[11] Perhaps the ur-source of this intuition is Kant's Groundwork. For there he claims that the only thing "good without qualification" is a good will. In particular, he holds that intrinsic goodness or badness of the will is entirely independent of its consequences. Apparently, even an utterly ineffective will, a will insufficient onto itself to bring about its aim, can be intrinsically good. Consequences enter not all into the determination of the value of the will.

cognizing agent's capacities and methods, on the one hand, and her goals and opportunities, on the other. Mere misfortune in one's surroundings does not, however, amount to irrationality, the internalist will say.

2.3 The Kantian Intuiton,

In the last section I allude to failures and successes of the cognizing agent's "own," that count as her own by virtue of her inner capacity to own them as her own. Such talk leads to what I call the Kantian intuition about rationality. The Kantian intuition is the intuition that the order of norms is, a bottom, an order of reason's own constituting. I attempt to honor that intuition by offering a sketch of an account of what I call the ***ownership of norms***.[12] My central claim will be that a cognizing agent ***C*** owns a norm ***N*** if ***C*** would, upon culminated competent reflection, endorse ***N***. I call a norm owned in this way by a cognizer ***C*** a norm of reflection for ***C***. To a first approximation, a norm of reflection is a norm whose binding force over a cognizer is rooted in that cognizer's own "conception of the good" and competent reflective endorsements grounded in such conceptions. Norms that I own are correlative with the space of reasons that are mine. A cognizer who owns a norm of inquiry that directs one to believe the truth thereby makes it the case that she has reason to believe what can be shown to be true. A cognizer who owns a norm of cooperation thereby makes it the case that she has reason to cooperate in circumstances calling for cooperation.

My account of norm-ownership is not an account of the structure and content of abstract norm space. The denizens of abstract norm space are ought-to's. An ought-to articulates what (putatively) ought to be the case or what ought to happen or what ought to be believed. Such directives can be more or less general. They can say what a given agent ought to do or believe at a given time or in a given set of circumstances. Or they can articulate general constraints on action or belief. Think of abstract norm space, on a par with the space of propositions, as a plenum containing every possible ought-to, from the most specific to the most general. Though the structure and contents of this plenum is well worth investigating, it is not my aim to investigate it here.

I distinguish owning a norm from being held to a norm by others. Others may or may not have *their* reasons for holding, or attempting to hold me to norms that they endorse but I do not. Their having

their reasons does not ipso facto give *m*e reason for obeying that norm.[13] Even if I have reasons of my own, rooted in norms that I endorse, for following the dictates of norms endorsed by others, that would not ipso facto make their norms to be norms for me or their reasons to be my reasons. Others may *hold* me to a norm, but they cannot *bind* me to a norm. I am bound only to norms that I own, or would own, through my own culminated competent reflective endorsement. Correlatively, only failures to live up to norms to which I am bound count properly as my failures. Such failures are not necessarily my responsibility in any metaphysically deep sense. The norms that I own need not be freely or autonomously chosen by me. Indeed, I may be causally determined by facts about my contingent psychological makeup to endorse, upon competent reflection, certain norms rather than others. But any such norm would still count as a norm of reflection for me.

Phrases like "culminated competent reflection," "endorsement" and "conception of the good" are intended here as purely psychological or functional role concepts, systematically interdefinable in terms of one another. To a first approximation, a conception of the good is a set of initial (or initiating) convictions and commitments (of whatever strength or intensity) about what is to be, be done, or be believed. Conceptions of the good are initial inputs to reflection. Endorsements, on the other hand, are the outputs of reflection. To a first approximation, a state ***x*** is an endorsement if it is a state of a kind ***K*** such that (a) culminated courses of reflection typically culminate in states of kind ***K*** and (b) states of kind ***K*** typically cause pro-attitudes toward actions, attitudes and states of affairs appropriate to states of kind ***K***. If I endorse Bill Clinton for President that will typically cause me to have a pro-attitude toward any or all of the following: (a) the state of affairs of Bill Clinton's being or becoming president; (b) my own or another's desire to see Bill Clinton become President; and (c) actions taken by me or others that are intended to bring about or sustain a Clinton presidency.

It is not only states of affairs, actions and desires which may be subject to endorsement. A cognizer may also endorse, or fail to endorse, her beliefs and even her methods of belief fixation. To endorse one's belief is, in effect, to represent what one believes as worthy of belief. If one endorses what

---

[12] The account offered here is an expansion and refinement of the account offered in the Essay XIII.
[13] I thus disagree with Korsgaard (1996) in her claim that reasons are *essentially* public. Though reasons can *become* shared, through the give and take of reasons, the do not start out that way. See also Scanlon (2000) for a "universalist" conception of reasons.

one believes one thereby expresses a willingness to stand behind that belief in what I call the contest of reasons. Similarly, if a cognizer endorses her method of belief-fixation, she typically will also endorse beliefs acquired via that method, because they are acquired in that way. Indeed, if a cognizer endorses or would endorse, upon culminated competent reflection, a method of belief-fixation, then she ipso facto has reason to believe any belief generated by that method. By owning a method of belief-fixation, a cognizer thereby constitutes that method as a source of reasons for her.

There is no a priori guarantee that a cognizer will in fact desire or believe that which she would deem, upon culminated competent reflection, worthy of belief or worthy of desire. Nor is there an a priori guarantee that a cognizer would endorse, upon culminated competent reflection, that which she in fact desires or believes. A psychologically well-ordered cognizer may strive to bring it about that she believes only what she deems worthy of belief and desires only what she deems worthy of desiring, but she is not guaranteed of success in that endeavor. Our beliefs and desires are not entirely up to us. Even a psychologically well-ordered cognizer may be causally determined to believe or desire that which, upon culminated competent reflection, she would deem unworthy of believing or desiring. Imagine a cognizer who believes that ***p*** as the result of hypnotic suggestion and lacks any further grounds for believing that ***p***. Imagine that if she were to competently reflect upon hypnosis as a method of belief-fixation, she would not endorse it. Even if our cognizer were to reflectively conclude that her belief is not worthy of belief, she might still be unable to rid herself of that belief.

Endorsements are not created equal. Some actual endorsements are not the outcome of a culminated course of competent reflection -- either because though there was a culminated course of reflection, it was incompetent or because the endorsement was not the product of reflection at all. Only endorsements which are, or would have been, the outcome of culminated courses of competent reflection anchored in an agent's conception of the good play a role in constituting a norm as a norm of reflection for a cognizer. We might call those endorsements which are, or would have been, the outcome of culminated competent reflection anchored in a conception of the good, deep endorsements. Only deep endorsements have any role in constituting a norm as a norm for a cognizer.

Talk of *culminated* reflection has a quasi-normative feel. One is tempted to say that reflection culminates when it reaches an "appropriate" stopping point. But talk of culmination is here intended in an

entirely non-normative manner, without any antecedent standard of appropriateness in mind. To a first approximation, a course of reflection culminates when, given all currently dialectically relevant considerations and a fixed conception of the good, further reflection on those same inputs to reflection would lead to the same endorsement. So reflection culminates when it produces endorsements that are stable under further reflection, given only the currently relevant inputs to reflection and a fixed conception of the good.[14] It is an important feature of this approach that it allows for the possibility that further reflection may yield different endorsements when new inputs are available or even when a new "weights" are assigned to old inputs because of a change in the agent's conception of the good. For that reason the stability that is the hallmark of culmination is a merely local stability. It is important to stress that my claim that reflection culminates when it produces endorsements that are or would be stable under further reflection is not intended to answer a normative or justificatory question. The question I have addressed is not the normative question "when is it appropriate or justified that reflection end?" Is it rather the descriptive question, "When has reflection produced an endorsement that has the agent's full rational backing?" My answer is that an endorsement has an agent's full rational backing when that endorsement would remain standing under any further competent reflection on the currently relevant inputs and given a fixed conception of the good. In answering this question, I am attempting to characterize a basis of deciding *where the agent in fact stands* with respect to the distribution of reasons. We want to know which of the would-be reasons are genuine reasons for the relevant agent.

Even in the absence an episode of culminated competent reflection, there can be a fact of the matter about where the agent stands with respect to the distribution of reasons, as long as her psychology is of the right character. In particular, if the agent has what I call a reflection-determining psychology, there will be true counterfactuals about what norms the agent *would* endorse upon culminated competent reflection. The truth of such counterfactuals will supervene upon the current actual structure and content of

[14] Stability under reflection plays a role in my account analogous to the role played by stable plans and intentions in Bratman (2000). Bratman thinks stable plans and intentions play a decisive role in answering the question of what he calls "agentive" authority. Relatedly, Blackburn (1999) evidently thinks that knowledge is roughly a matter of beliefs that are stable under the pressure of further evidence and inquiry. There is, I think, something right about this thought. Indeed, I defend a similar claim in some work currently in progress. Unlike Blackburn, I see no tension whatsoever between a thoroughgoing realsim and making stability under inquiry be the hallmark of that which we are pleased to honor with the title "knowledge."

the agent's will and the agent's beliefs. When an agent's psychology is reflection-determining, we may say that she is tacitly bound by the relevant norms. The appeal to such counterfactuals is intended to capture the intuition that both the believings and willings of a cognizing agent can be subject to rational criticism, even when she is not explicitly and occurrently guided by the relevant norms. At the same time, the proviso that her psychology must be reflection-determining preserves the internalist intuition that we are bound to no norms except those to which our own psychological make-up determines that we are bound.

Not all cognizing-agents have reflection determining psychologies. An agent may have an incoherent, unstable or indeterminate conception of the good. Incoherence, instability or indeterminacy in the inputs to reflection may lead to instability, incoherence or indeterminacy in the outputs from reflections. To the extent that this is so, there will be no determinate truths about what norms the relevant cognizer would endorse upon competent reflection. In that case, it will be indeterminate where the agent stands with respect to the distribution of reasons. Now even where there is no place that they agent currently stands with respect to the distribution of reasons, it may still be that reflection, if engaged in, would produce some outcome or other. When reflection culminates in endorsement in an agent whose antecedent psychology is not reflection-determining, any such endorsement will determine where the agent *has come* to stand. Her standing there will be a *de novo* act of ownership and self-constitution, one not fully determined by her antecedent psychological make up. So we must distinguish, in effect, reflection that merely elucidates reasons that were, in some sense, there but unacknowledged, from reflection that produces reasons where there were not yet any.

Just as talk of culminated reflection can sound question-beggingly normative, so too with talk of *competent* reflection. But the only normativity involved in my talk of competent reflection is what we might call designlike normativity of well-functioning. Much ink has been spilled over such designlike norms.[15] According to a widely held view, the function of an item ***x*** is an effect of ***x*** for which ***x*** was selected, where the "selection" can either be intentional selection by some agent or some nonintentional process such as natural selection. Function, on this view, has less to do with what an item does, but with

[15]For some discussion see Millikan (1984, 1986, 1989a, 1989b, 1990a, 1990b), Neander, (1991a, 1991b, 1995), Godfrey Smith (1994), Davies (1994), Dretske (1988, 1995), Papineau (1987), Bigelow and Pargetter (1987), Wright (1973, 1976).

what it, or its ancestors, did. For the space of the current argument, the selectionist approach to function suffices. The claim is thus that a certain sort of reflection occupies a certain role in our mental lives --the norm constituting role -- and does so as a consequences of having occupied that very role in the mental lives of our ancestors and having thereby contributed to their reproductive success. Reflection is "competent" when it is a further exercise of that very capacity, and no other. Some have indeed seen selectionist approaches as the foundation of genuine normativity. But selectionist normativity is normativity of a most anemic sort.[16]

My claim about the power of culminated competent reflection to bind a cognizing agent to a norm is not a conceptual-analytic claim about the very concept of normativity. It is a broadly empirical claim about the capacities of rational animals. Rational animals possess a psychological faculty whereby they may take ownership of norms of conduct and inquiry. In exercising this faculty, rational animals bind themselves to norms and in the binding constitute spaces of reasons for themselves. So Kant was right after all. Norms are reason's self-given products. But he was wrong to fear the brutely psychological nature of the normative faculty.

It should not be supposed that the constitution of a space of reasons is necessarily an affair of solitary cognizing agents. The capacity to achieve rational solidarity with others through what I call mutual ratification of shared norms is a distinctively human capacity. A norm is mutually ratified by a community of cognizers if: (a) the members of the community one and all endorse or would endorse that norm upon culminated competent reflection; (b) the members of that community mutually recognize that one and all endorse or would endorse that norm; and (c) the members of the community endorse one another's endorsing of the relevant norm. By satisfy condition (a), a community of cognizers makes a norm into a shared norm. By satisfying condition (b), they achieve mutual knowledge of shared norms. And by satisfying condition (c), they make a shared and mutually known norm into a mutually ratified norm. To achieve mutual ratification of shared norms is, in effect, to acknowledge one another as equal partners in the constitution of normative community. It is for me to say to you that the normative authority that I recognize in you is also an authority for me and for you to say the same back to me.

---

[16] Millikan (1984, 1993) is the locus classicus. Godfrey Smith (1995) rightly calls the selectionist brand of normativity weak. I think even 'weak' is too strong a word.

Norms become mutually ratified through what I call the dialectic of ratification. Through the dialectic of ratification, I try to get you to ratify me and my norms as rational sources for you and thereby to make it the case that me and my reasons provide reasons or you. Simultaneously, you try to get me to ratify you and your norms as rational sources for me and thereby to make it the case that you and your reasons provide reasons for me. If we each succeed, we thereby constitute a space of mutually ratified norms and reasons. If we achieve mutual ratification, we each extend the reach of our own rational powers. Through the mediation of mutually ratified norms of inquiry and communication which direct the truth to be sought and told, for example, my having reasons for believing that may give you a reason for believing that f as well. Through the mediation of mutually ratified norms of conduct calling for mutual aid and co-operation, for example, my having a reasons for □'ing may give you a reason either to refrain from interfering with my attempts to □ or perhaps even a reason for aiding me in my attempts to bring it about that □. Mutually ratified norms are thus the rails along which reasons may be transmitted from cognizing agent to cognizing agent.

I do not mean to say that the achievement of normative community is necessarily a two stage process in which agents first endorse certain norms as their own and only then engage in a dialectic of ratification. An agent's initial endorsements may well be simultaneous with her achievement of normative community with others. That is, it may be only in the context of the constitution of a normative community with others, that one constitutes a space of reasons for oneself. But it is important to stress the conceptual distinctness of the two different ratification problems faced by every cognizing agent. First, there is the problem of self-ratification. A cognizing agent's beliefs and desires, for example, present themselves to that very cognizing agent as candidates for her reflective endorsement. Through the reflective endorsement of certain beliefs and desires as worthy of holding, a cognizing agent thereby makes those beliefs and desires truly her own and thereby promotes those beliefs and desires, as well as the sources of those beliefs and desires, into reasons and rational sources *for her*. But there is no logical guarantee that her reasons will count as reasons *for others*. The absence of any such guarantee gives rise to the problem of mutual ratification. A space of reasons that I have constituted as merely my own is not *ipso facto* guaranteed to be

ratified by others as reason-giving *for them*. Nor am I, with all my rational powers, guaranteed by the very nature of rational powers to be owned by others as a source of reasons for them.

Though rational agents and their reasons are not *ipso facto* rational sources for one another, it does not follow that rational agents begin as mere instruments to one another. To cognize another as a rational being is to recognize that she belongs to a different order of nature from the whole of non-rational nature. Non-rational beings are nothing at all either to themselves or for themselves and are at best derivative source of reasons for any rational being. Non-rational beings can indeed be sources of reasons for us, but only in virtue of the rationally optional interests that we happen to take in them. We may esteem non-rational beings as instruments, as objects of wonder and awe, even as objects of a peculiar kind of sympathy or love, but they are not the kinds of beings for which even the possibility of normative community arises. By contrast, each fully reflective rational being recognizes in herself an original, non-derivative source of reasons for herself. To recognize another as a rational being is to recognize that other as also an original and non-derivative source of reasons for that other. The mere recognition of our shared rationality is not yet the achievement of rational solidarity. But in the mere recognition of another as a fellow rational being a question is automatically put: What, if anything, shall we do, be or believe together as fellow rational beings? It is this question that sets for us the problem of mutual ratification. We achieve normative community only when we settle the contest of reasons and achieve mutual ratification of one another's reasons. That we do only when we, as it were, take one another's reasons in. In an original act of ownership, I make your reasons into reasons for me. In a similar act, you make my reasons into reasons for you.

There has never been a time when human beings did not find themselves arrayed in normative communities of varying scopes. Our ancient progenitors were arrayed in normative communities encompassing only small circles drawn around kin, clan or tribe. But the rough general trend of human history has been haltingly toward ever more encompassing circles of rational solidarity. Indeed, the rapid spread, at least from the perspective of evolutionary time, of modern science and technology allows us to conceive of something never dreamt of by our pleistocene progenitors -- the real possibility of a global community of reasons. But one should not suppose that success in the characteristically human struggle to achieve rational solidarity is either antecedently guaranteed or an inescapable commandment of the rational

intellect and will. The contest of reasons is perhaps the deepest, most enduring fact of human social life. Where rational solidarity and normative community emerge from the enduring contest of reason these are typically hard-won, contingent, historical and cultural achievements. To stress the contingency of normative community is not, however, to deny the very possibility of that some norms are universal in the sense that they are bound to be endorsed by any fully reflective and rational cognizing agent. For example, if it is right that each fully reflective rational cognizing agent is present to herself as an original source of reasons, then no rational agent can deeply endorse, at least not coherently, her own enslavement to another. If so, it is not possible for me to achieve rational solidarity with those would use me as a mere instrument. Others may achieve dominion over me through force. I may even endorse my own servitude through incompetent or non-culminating reflection. But that way lies a mere semblance of rational solidarity, not its reality.

2.4 A Kantian Reconciliation

We have already experienced the opposing pulls of internal rationality and consequential rationality. Consequential rationality is focused on success in the pursuit of worldly goods. One is consequentially rational if one's belief-fixation, action-guiding and desire-generating mechanisms reliably and non-accidentally deliver the worldly goods. The consequentialist intuition surely gets at part of our ordinary conception of rational powers. Just as we take it to be non-accidental that the good guy tends to get the girl in the end, we take it to be non-accidental that the good believer tends to believe the true in the end. By contrast, the internalist intuition contains a frank acknowledgement of the worldly limits of rational powers. Where the consequentialist intuition has it that rational powers are powers to reliably and non-accidentally deliver the worldly goods, the internalist intuition has it that rationality consists only in the inner coherence of the intellect and will. In an internally rational mind, beliefs will cohere with one another and with evidence; actions or attempted actions will cohere with beliefs and desires, and desires will cohere with needs and/or values. Conversely, an internally incoherent intellect (or will) is, in a sense, self-negating. For an internally incoherent intellect that affirms ϕ may simultaneously affirm $\neg\phi$. Similarly, an internally incoherent will which desires ϕ may simultaneously desire $\neg\phi$. Such intellects and

wills are at odds with themselves in ways that are intuitively a paradigm of irrationality.[17] It is surely correct that the internalist intuition is deeply rooted in our common sense understanding of rational powers.

If internal coherence did suffice to guarantee consequential rationality, we could have it both ways and there would be no deep tension between the internalist intuition and the consequentialist intuition. Unfortunately, it is fairly easy to demonstrate that, no matter how exactly internal coherence is construed, no form of internal coherence will be sufficient to guarantee consequential rationality. Even when the believings and willings of a cognizing agent are as well-ordered as the exercise of what might be called narrow right reason alone can guarantee, that will not suffice to render her consequentially rational. For even when evidence, beliefs, and desires and intentions, internally cohere one with another, it is not guaranteed that beliefs, desires and intentions will interact, reliably and non-accidentally, to cause desire-satisfying actions. Committed internalist may insist that this fact shows only that it is a mistake to understand rational powers as powers to reliably and non-accidentally deliver the worldly goods. Whatever worldly success we happen to enjoy in our cognition and conation, the internalist may say, will involve an ineliminable element of good fortune. Such a conception of rationality may well be tenable in the end, but if one is at all pulled, as I am, by the consequentialist intuition then one may be moved to seek a reconciliation of the competing pulls of internalism and consequentialism. That is what I attempt to do in this section.

The gap between internal and consequential rationality arises because the narrowly right reasoning mind, though sufficient to guarantee its own inner coherence, is insufficient to guarantee its own external coherence. To a first approximation, one's mental life is externally coherent if there is a metaphysically possible world in which one's beliefs, as widely individuated, are jointly true and one's desires, as widely individuated, are jointly satisfiable. Consider the mental life of Jocasta. By one measure, Jocasta has certain beliefs about her son, which she does not have about her husband. She believes that her son -- call

[17] It is not entirely clear, however, exactly what we praise when we praise internal coherence as one among the cognitive virtues which are definitive of rationality. There are at least two initially plausible candidates: mere logical coherence and baysean coherence. The logically coherent agent believes no explicit logical contradictions, holds no sets of beliefs that are mutually inconsistent, and at least tacitly believes the deductive closure of all that she explicitly believes. The Baysean coherent believer has a coherent, that is, not dutch-bookable, distribution of prior subjective probabilities and updates that distribution on the basis of evidence by conditionalization on the evidence. That is, the baysean coherent agent has a probability

him Tad -- has probably vanished. But she believes that Oedipus is alive and well. Moreover, by one measure, she desires things with respect to Oedipus that she does not desire with respect to Tad. She desires to marry Oedipus, but to avoid marrying Tad. Now Jocasta's mental life is presumably internally coherent. Jocasta's beliefs cohere in the requisite way with her evidence. She plans, in an internally coherent way, a course of action designed to bring about the satisfaction of her two most ardent desires. But despite her internal coherence, Jocasta is in a quite unfortunate predicament, a predicament that is best understood as an epistemic predicament. She has a pair of beliefs that, as long as their wide contents are held fixed, cannot be made jointly true in any metaphysically possible world and a pair of desires which, with their wide contents fixed, are not jointly satisfiable in any metaphysically possible world. Consequently, her internally coherent plans are utterly futile. Jocasta is guaranteed to fail. Any metaphysically possible world in which she gets what she most wants -- marriage to Oedipus -- is ipso facto a world in which she gets what she most wants to avoid -- marriage to Tad.

The very possibility of a mismatch between internal coherence and consequential rationality has its source, I claim, in the existence of two further gaps: (a) the gap between concept and conception and (b) the gap between inner (logical) syntax and outer semantic destiny. Recall our discussion from Essay II and elsewhere of the distinction between concepts and conceptions. A conception, recall, is a kind of mental particular, a labeled, highly structured database of information about the extension of an associated concept. For example, each thinker who capable of thinking <cat> involving thoughts is likely to have stored in his head a database of information (and misinformation) about cats. In English speakers, such a database might be labeled 'CAT'. Such a database may contain a variety of different kinds of information (and possibly misinformation) about cats. It may contain a list of properties that some, many, most, all or typical cats are taken to satisfy. It may contain information about the categorial basis of the concept <cat> -- that is, whether <cat> is a natural kind concept, a functional concept, an artifactual concept. It may contain an image of an exemplary cat, a list of atypical cats, and pointers to sources where more can be found out about cats.

---

distribution that satisfies the axioms of the probability calculus and updates that distribution in accordance with the rule Pn(p) = Po(p/D), where D is a set of data sentences.

But conceptions must be sharply distinguish conceptions from concepts[18]. There are at least two reasons for resisting this conflation:

(1) conceptions do not relate to their extensions in concept-like ways.

(2) conceptions do not "compose" in concept-like ways.

Like a concept, a conception may have an "extension" -- the set of things of which it is a conception. But nothing belongs to the extension of a concept except instances of that concept. All and only cats belong to the extension of the concept <cat>. But a 'cat' conception may fail to be true of any actual cat and may be true of many non-cats. This is largely a consequence of the fact that conceptions, unlike concepts, are structures of belief. Though the concept <cat> has all and only cats within its extension, that concept may be misapplied to non-cats in the context of one or more false beliefs. One may falsely believe Nikola the dog to be a cat. And this false belief may well infect both one's 'cat' conception and one's 'Nikola' conception. As a consequence, one may come to have both a 'cat' conception and a 'Nikola' conception each of which applies to something of which it is not true and fails to apply to something of which it is true.

Now consider an arbitrary speaker's conception of gray cats. What constraints are there on the relations among her 'gray cat' conception, her 'cat' conception and her 'gray' conception? Evidently, there is no way these conceptions must be related. Suppose that each of these conceptions includes a list of exceptions. There is no way in particular that the list of exceptions for <gray cats> must be related to the list of exceptions for <gray> and the list of exceptions for <cat>. An exceptional gray cat may fail to be either an exceptional cat or an exceptional gray thing. Much the same seems to be true for any aspect of the relation between a "complex" conception and its simpler constituents one cares to name. Prototypical gray cats need be neither prototypically gray nor prototypically catlike. Cat x may be a gray cat exemplar while failing to be either a cat exemplar or an exemplar of gray. One may have a "false" (true) conception of gray cats, while having a "true" (false) conception of gray things and a "true" (false) conception of cats. Facts about "complex" conceptions are simply unconstrained by facts about the simpler conceptions out of which they are "composed." This is so because conceptions are structures of belief, not *constituents* of belief. What one believes about gray cats is no straight-forward function of what one believes about cats

[18] Fodor (1998) goes to great lengths to argue that conception-like structures (prototypes, stereotypes) do not compose in any systematic way. Prinz (2002) attempts to show, contrary Fodor, that conceptions do

and what one believes about gray things. Since beliefs about gray cats are not built out of beliefs about gray things and beliefs about cats the failure of conceptions to "compose" in any straight-forward way is precisely to be expected on the assumption that conceptions are structures of beliefs rather than constituents of beliefs.

With concepts, matters are more systematic. The inference from 'x is a gray cat' to 'x is gray' and 'x is a cat' is clearly compelling. Whence the source of its compelling nature? A natural seeming answer is that the language of thought -- which Frege appropriately called the "concept writing" -- has a combinatorial syntax and a compositional semantics. The thought is that the mental representation that expresses the concept <gray cat> is literally "built out of" the mental representation that expresses the concept <cat> and the mental representation that expresses the concept <gray>. Moreover, the intension of the concept <gray cat> is the set theoretic intersection of the intensions of the concept <gray> and the concept <cat>. It is crucial here that there exist a general and systematic story about how "simpler" concepts compose to yield more "complex" concepts and a systematic story about how what the more complex concept expresses is fully determined by what its constituents express.

For our current purposes, it is crucial that cognizers who share a concept may have quite diverse conceptions of the extensions that fall under their shared concepts. For example, the 'water' conception of the expert is likely to include representations of the molecular structure of water, while the 'water' conception of the lay person is likely to include only representations of the superficial look, feel and taste of water. But both conceptions are, nonetheless, conceptions of water. What can happen in two minds, can also happen in one. Just as two cognizers can have distinct conceptions of the same extension, so a single cognizer can have two distinct conceptions of the same extension. Jocasta, for example, had two distinct conceptions of Oedipus -- one labeled 'Oedipus', the other labeled 'Tad'. And many ancient Babylonians apparently had at least two distinct conceptions of the one planet Venus.

Many will concede both that diverse cognizers and that a single cognizer can have diverse conceptions of the same extension. Now I argued at length in Essay II that diversity of conception neither constitutes nor presupposes diversity of concept. Cognizers may share a concept even when they have radically different conceptions of the extension of that concept. Indeed, I claim that a single cognizer may

---

compose.

possess the same concept twice by having two distinct inner representations that express that concept and two dynamically unlinked conceptions which independently mediate deployments of that concept in thought episodes. When a cognizer has two dynamically unlinked conceptions that independently mediate deployments of the same concept in thought episodes, her mental life is quite liable to external incoherence. Jocasta is a case in point. She has two distinct conceptions of the selfsame person. One conception is labeled 'Oedipus'; the other is labeled 'Tad'. The conception labeled 'Oedipus', we may presume, mediates such deployments in thought episodes of the individual concept <Oedipus> as are constituted by tokenings of some inner mental symbol which we may name *Oedipus*. The conception labeled 'Tad' mediates such deployments in thought episodes of that very same individual concept as are constituted by tokenings of some distinct inner mental symbol which we may name *Tad*. Jocasta believes, via tokenings of *Oedipus*, the proposition <<Oedipus> <alive>>. But she also believes, via tokenings of *Tad* the proposition <<Oedipus <dead>>. Moreover, she desire true, via tokenings of *Oedipus*, the proposition <<Jocasta><marry><Oedipus>>. She simultaneously desires true, but via tokenings of *Tad* <not<<Jocasta><marry><Oedipus>>>. Unfortunately for Jocasta, there is no metaphysically possible world in which Jocasta's beliefs are simultaneously true and no worlds in which her desires are jointly satisfied. And that is why she is externally incoherent.

Consider a case of a different kind. Like Jocasta, Lex Luthor has two conceptions of the selfsame extension, one conception labeled 'Clark Kent', the other labeled 'Superman'. Luthor's most ardent desire is to put an end to Superman. To that end, he engages in internally coherent planning by which he devises an elaborate scheme for bringing about Superman's demise. At this very moment, Superman is standing right there, presenting Luthor with a golden opportunity. Yet Luthor lets the opportunity slip by. Unlike Jocasta, Luthor is not guaranteed to fail. He has it within his power, at this very moment, to flip the switch that releases the trap that does in his nemesis once and for all. But Superman is present under the wrong guise -- he is present under his Clark Kent guise, rather than his Superman guise. Superman's presence under his Clark Kent guise "activates" the file in Luthor's head labeled 'Clark Kent'. This file mediates such deployments of the individual concept <Superman> -- a concept which is identical in "wide content" to the individual concept <Clark Kent> -- in thought episodes as are constituted by tokenings of some inner mental symbol *Clark Kent*. And Luthor comes to believe, via a tokening of this symbol, the

proposition <<Clark Kent > <standing at p>>. Luthor has, in addition, a file in his head labelled 'Superman'. This file mediates such deployments of the individual concept <Superman> = <Clark Kent> as are constituted by tokenings of some inner symbol *Superman*. Luthor believes, via a tokening of *Superman*, the proposition <not <Clark Kent> <standing at p>> (= the proposition <not <Superman> <standing at p>>). In addition, Luthor desires true, via tokenings of *Superman*, the proposition <<Clark Kent> <meet demise at p>>. Unfortunately for Luthor, his *Clark Kent*-tokening constituted belief and his *Superman*-tokening constituted desire co-exist in splendid isolation from each other. And that is what costs him the opportunity to satisfy his most ardent desire.

The cognitive predicaments of Luthor and Jocasta are the stuff of which tragedy and missed opportunities are made. Unfortunately, it is just a fact of cognitive lives like ours that even within an internally coherent mind, there occur dynamically unlinked concepts of the selfsame extension. As a consequence we run the constant epistemic risk of external incoherence. And when we are externally coherent, then narrow right reasoning alone cannot guarantee that our beliefs, desire, intentions, and actions, will relate to one another in ways which are characteristic of consequential rationality.

What are we to say, in the end, about Jocasta and her rationality. Is she a rational believer or is she not? If we are deeply committed to the consequentialist intuition, with its focus on worldly success, then we might be tempted to declare Jocasta irrational. To do so would be to elevate external coherence, a kind of coherence which seems to require cooperation from the world beyond the exercise of narrow right reasoning, into a necessary condition on being a rational mind. Alternatively, if we are deeply committed to the internalist intuition, with its emphasis on rationality as the power to avoid self-negation, then we may be tempted to abandon the consequentialist intuition and with it the requirement of reliability in the attainment of worldly goods. What matters for rationality, from this perspective, is not actual worldly success, but that we have done all we can which is contributory to success, given our worldly situation, and that we always have "clean epistemic hands" in whatever failures happen, nonetheless, to come our way.

I do not think that we should altogether abandon the internalist intuition. It is not plausible that external coherence is constitutive of rationality. If external coherence were constitutive of rationality, rational minds would have to know *all* the true identities upon which the success of their behavior is

contingent.[19] That requires too much. At the same time, there is a clear sense in which mere internal coherence demands too little. Internal coherence is the merely negative syntactic virtue, displayed by a will and/or intellect that are merely not syntactically self-negating. It is not a virtue that makes an intellect

[19] Millikan (1993) recognizes the relevance of something like external coherence to rationality, but she goes too far. Consider the following:

> Because rationality pivots on Kontent [roughly, broad content] and does not reside in some inner, safer realm, it is also true that no manipulation of modes without regard to whether or not these in fact have kontent, and without regard to whether they have, perhaps multiply ambiguous kontents, could possibly be a manifestation of rationality. Having an automated formal system unfolding inside one's head is not being a rational creature if the system has no interpretation. Nor is it being a rational creature if each symbol ambiguously means, or is undifferentiated among meaning several different things. Imagine a head full of descriptions whose component terms are empty. Imagine a head full of a thousand descriptions and (indexical) names all of the same object but without the head's knowing this. This would be rationality? (p. 349.)

Millikan is right to suggest that a cognizer who cognizes the same object twice, but without recognizing that she has done so, is no paragon of cognitive virtue. Such a mind will be externally incoherent, its internal coherence not withstanding. But to elevate external coherence, as Millikan does, into a requirement on rationality is really to insist that only minds which are "nicely" embedded in their environments count as rational. It is hard to see how this approach can distinguish rationality from mere cognitive good fortune.

Fodor (1994) too apparently thinks that something like external coherence is partly constitutive of rationality. For note the following:

> Rational behavior is, generally, pretty successful as a matter of fact; a lot more successful, anyhow, than behavior that is simply crazy. No remotely acceptable intentional psychology could count this fact as accidental. But it *would* be accidental -- it looks like it would be unintelligible -- unless generally speaking, people know and respect the facts that the outcomes of their actions depend upon; including, in particular, the facts about what is identical to what. The *only* means that a belief/desire psychology *has* to insure that, in typical situations where a = b, Smith will assign appreciably similar utilities to Fa and Fb, is to insure that Smith *believes* that a = b and that he makes the relevant inferences.

Again, I agree with Fodor that our plans and actions will often go astray and that we will often miss opportunities if we do not know relevant true identities. Moreover, we can be brought by reflection to know this about ourselves. Because we can know this about ourselves, we endorse external coherence as a regulative ideal on our cognitive strivings. But it demands too much to require that a rational mind know all true identities. Nothing merely internal to our own heads, taken either severally or collectively, can guarantee that when we are presented with the same object again, we will recognize that we are. No cognitive policy or practice and no bit of "social engineering" can guarantee such a happy outcome. Indeed, assume that Jones and Smith adopt exactly similar cognitive policies and practices and are subject to a similar social division of cognitive labor. Still, it is possible that, due to the luck of the environmental draw, Jones is confronted with a plethora of Frege cases and Smith is confronted with few. By Fodor's lights apparently (and apparently also by Millikan's) Jones may count as irrational, while Smith counts as rational. So again, rationality turns out to be partly a matter of the luck of the environmental draw. But again, though a commitment to try, ceteris paribus, to reduce the dependency of our cognitive and practical success on the luck of the draw does seem constitutive of rationality, it seems wrong to elevate actual achieved external coherence into a constituent of rationality. Rational minds ought to strive to be externally coherent. They ought, ceteris paribus, to prefer being externally coherent to being externally incoherent. But no merely finite mind can *guarantee* its own external coherence.

sufficient for cognizing the epistemically good nor a will sufficient for achieving the conatively good because nothing in a merely internally coherent mind can guarantee that inner representations are bound down to external existents in an external coherence making way. It is fair to wonder whether there is middle path, a path that is still internalist, but accommodates from within a still internalist perspective the normative pull of external coherence. I devote the remainder of this section to sketching one such path.

The middle path takes external coherence to be a self-given regulative ideal on our cognitive strivings. When we reflectively endorse external coherence as a regulative ideal, we commit ourselves to striving, ceteris paribus, so to arrange our mental lives that representations that share an external semantic destiny also come to share a common syntactic life in episodes of narrow right reasoning. In endorsing external coherence as a regulative ideal, one thereby takes the plenum of one's inner representations to be more than a closed Cartesian circle, subject only to constraints of syntactic well-orderedness. In aiming to be externally coherent, one aims to bring it about that the inner syntactic lives of one's inner representations mirrors their outer semantic destinies. One strives to make it the case that, ceteris paribus, one does not have two dynamically unlinked ways of thinking the same thought again, of deploying the same concept again.

It is only because external coherence is one among the defining ideals of a rational mental life that such a mental life present itself to us as a mental life worth pursuing. After all, it is only when the inner syntactic lives of our inner representations track nicely with their outer semantic destinies that our cognition and conation will exhibit the hallmarks of consequential rationality. In the absence of external coherence, the merely negative virtue of internal rationality leaves us still dangerously liable to tragedy and missed opportunities just because it leaves us with the possibility that representations which are semantically connected will persist in splendid syntactic isolation one from another. In cognizing the worldly limits of our rational powers while, nonetheless, elevating external coherence to a regulative ideal on our cognitive strivings, we commit ourselves to the unending endeavor to free ourselves from the whims of mere epistemic fortune. It is hard to see how a reflective mind that would not be free of the whims of fortune could reasonably count as a fully rational mind.

To be sure, possible worlds there are in which, though Jocasta's internal mental life is, in one sense, just the same as it actually is, she is so embedded in the world that her inner symbol *Tad* and her inner symbol *Oedipus* are "anchored" to two different individuals. And this will suggest to many that the so-called "narrow" or "intrinsic" contents of Jocasta's head might remain just as they are even in a world in which Jocasta's *Tad* and Jocasta's *Oedipus* fail to co-refer. In any such world, Jocasta's *Tad* involving thoughts and her *Oedipus* involving thoughts will indeed be thoughts about different objects and also, presumably, thoughts involving different (wide) concepts. In some such worlds, Jocasta marries the referent of *Oedipus* without marrying the referent of *Tad*. In such worlds, Jocasta's mental life is externally coherent and she enjoys, therefore, the possibility of worldly success. Unfortunately for Jocasta, as for us all, her intrinsic powers cannot guarantee that she is embedded in the world in an external coherence making way. These considerations will suggest to some that rationality should be defined not over broad contents, which depend partly on worldly relations between representations and representeds, but only over narrow contents, which supervene only on what's in the individual cognizer's heads. Viewed in this light, Jocasta's "predicament" isn't a predicament of reason at all. Her problem lies not in her representations or in her rational manipulation of those representations but in her "extra-rational" connections with the world.[20]

The lack of any inner guarantee of outer coherence makes it implausible that external coherence is partly constitutive of rationality and that makes it implausible that rationality could just be consequential rationality. Still, I think it would be a mistake to entirely deny the normative pull of the considerations of metaphysical possibility and necessity on which the notion of external coherence depends. By treating external coherence as a self-given regulative ideal, we give due deference to such considerations both without abandoning internalism and without reducing rationality to a merely negative, and as it were, syntactic virtue. External coherence is a self-given project of reason. Moreover, it is not just that the rational mind *happens* to give itself the project of achieving external coherence. Rather, it is a project that reasons gives itself from, as it were, the depths of its self-conception. To be sure, reason strives to carry out this self-given project without any internal guarantee of success, But the absence of a guarantee of success is not the absence of the possibility of success. The directive that reason gives to itself in endorsing

[20] See Essays VII, XI- XII for further discussion of narrow content.

external coherence does not command the impossible. Rather, it commands pursuit of the possible and an attitude of non- complacence in mere internal coherence.

That we who have the inner power to guarantee only our own internal coherence, nonetheless, constitutively endorse external coherence as a regulative ideal says a great deal both about the nature of our minds and about our self-conception of ourselves as beings in the world. First, the standing of external coherence as a self-given regulative ideal is an expression of the syntactic plasticity of our minds. A syntactically plastic mind contains a plenum of inferentially articulated representations, together with the power to rearticulate its representations by laying down new and extinguishing old inferential links among its representations on the basis of its experience in ways which enhance both their epistemic and non-epistemic worldly goodness. In such a mind, reason's giving itself the regulative ideal of external coherence amounts to reason's giving itself the project of constraining the rearticulation of its syntactic landscape by an external world not of its own constituting. It is as if the syntactic landscape is present to reason as simultaneously subject to reason's own rearticulation but also as not *directly* constrained by the very world that reason takes to be external to itself. In giving itself the regulative ideal of external coherence, reason introduces, on its own authority, a constraint from the world, while retaining full awareness that it cannot be guaranteed to satisfy the constraint merely through the absence of syntactic self-negation.

Syntactic plasticity has many expressions. It is expressed as a capacity to engage in the sort of theory-mediated, non-demonstrative inferencing via which low level irradiations at our sensory surfaces can come to be inferentially linked to representations of things remote from immediate sensation. It is also expressed as capacity to engage in the sort of nonmonotonic reasoning which enables us to withdraw a conclusion based on increased information, even when we continue to endorse the premises from which the now dropped conclusion earlier followed. Yet another expression syntactic plasticity is our ability to alter our action plans on the basis of our ever changing beliefs, desires and opportunities via the practical syllogism. Such powers of rearticulation enable the syntactically plastic mind to construct ever better theories of the world, to learn new, deeper, more explanatory regularities, to come to see that which is merely locally and contingently reliable as merely locally and contingently reliable and to acquire ever more powerful practical abilities to manipulate the world for its own ends.

Some intentional systems inhabit environments in which the presence of the good can be reliably enough discerned in some uniform, low cost way. Such intentional systems can afford to be prisoners of fixed ways of locating and tracking the good. But for systems that inhabit more unanticipatable environments -- environments that are variable in ways relevant to the system's pursuit of the good, in which the presence of the good cannot be reliably discerned or achieved in any uniform way -- syntactic plasticity may be of great value. Only cognizing agents with some degree of syntactic plasticity can face an unanticipatable world with promising prospects of success, relatively independently of the luck of the environmental draw. To be sure, no cognizer with a finite brain is likely to be indefinitely syntactic plastic.[21] But a modicum of syntactic plasticity is a sine qua non of rationality. Nothing is a rational mind that entirely lacks the power to rearticulate its syntactic/inferential landscape on the basis of its experiences in ways which enhance epistemic and conative goodness.

I have not told the entire story about rationality. I hope I have said enough to make the claim that rationality is a fully natural power of the mind-brain both plausible and attractive. To build a rational mind, one endows a mind with a plenum of representations with inner syntactic lives and with outer semantic destinies. One links those representations in ways that subserves the cognizing agent's worldly pursuit of naturalistically constituted epistemic and practical goods. One endows that mind with the power to reorder its syntactic landscape, in goodness enhancing ways, on the basis of its experience in the world and its actions on the world. And one endows that mind with a power of reflection by which it takes ownership of norms of conduct and inquiry and thereby constitutes a space of reasons as its own. I can see no deep or principled reason to deny that such a mind may be built out of a kit of entirely naturalistic

---

[21] In holding that relatively implastic cognizers often succeed, if they succeed, merely because of the luck of the environmental draw, I need not deny that, for example, naturally selected cognizers typically are embedded via natural selection only in environments in which their cognitive strategies are at least reasonably adaptive. Nor can we necessarily say that the cognitive strategies of relatively implastic cognizer's succeed merely because of the luck of the environmental draw, at least not if we view the organism-environment relation from the perspective of evolutionary time. From the perspective of evolutionary time, it's in one sense non-accidental that cognizers tend to be embedded in environments in which their cognitive strategies represent winning strategies in the competitive game the organism plays with its environment. But this outcome results from the fact that over the course of evolutionary time the cognizers that are left standing are the one's epistemically lucky enough to find themselves embedded in environments in which their cognitive strategy is a winning one.

materials. The kit will include such relatively familiar items as inner representations in a language of thought, an evolutionarily instilled functional role psychology, second-order mental states within that functional role psychology, and a version of the causal/informational theory of conceptual content. To be sure, the evolutionarily instilled functional psychology I envision includes states such as culminated competent reflection, conceptions of the good, convictions, commitments, reflective endorsement, and the like, that are not explicitly countenanced in standard philosophical glosses of common sense belief -desire psychology. But these standard glosses have never been plausible as the full functionalist story about our psychological make up. I am doing only what has always needed doing (and what many others have already begun to do) -- introducing into the functionalist psychological story some of the complexities that would render such approaches more nearly adequate accounts of the mental states and capacities of real human beings. To be sure, I have only begun to sketch the additional layers of complexity. But I hope I have done enough to shift the burden of proof to the anti-naturalist who purports to hear no echo of reason in nature.

III. The Great Chain of Intentionality: From Sentience to Sapience

3.1 Of Pouncers and Frogs.

So far, my argument has focused on the nature of rational powers. I turn now to consider intentionality. I turn away from speculative naturalistic psychology and undertake an exercise in speculative natural history. My aim is to locate rational or type II intentionality on the great chain of intentionality that stretches from the crudest intentional systems, altogether lacking the sorts of transformative and reflective powers that are a sine qua non of rational powers, to creatures like ourselves who enjoy the full panoply of rational powers and intentional states. My arguments are designed to show that it is plausible that the emergence of rational intentionality was not the first appearance of intentionality in nature but a refining and deepening of it. I claim that we gain a deeper understanding of rational intentionality by understanding its place in the natural chain on which it is but a single link.

I begin with a somewhat fanciful example designed to pump the intuition that something with at least some of the hallmarks of intentionality can subsist even in the absence of rational powers. Imagine an organism $Pouncer_1$ with a quasi-perceptual control device, called DETECT. DETECT has on state and an off state. Ceteris paribus, DETECT turns on when X appears, and turns off when X disappears. Now suppose that $Pouncer_1$ needs certain amino acids. Though $Pouncer_1$ *needs* the relevant amino acids, it does not explicitly *want* them. But neither does it explicitly want something other than those amino acids. Needs are cheap. Even systems plausibly lacking intentionality -- like amoeba -- have needs. Wherever there is homeostasis there is need. Explicit wants are rarer. They require, I claim, that one be related to inner representations in a certain way. $Pouncer_1$ is an utterly simple homeostatic system with quite limited representational powers. It has only the utterly simple representational capacity to represent X as present or not present. Still, if things go right, $Pouncer_1$ does a decent job of getting the amino acids it needs. For in $Pouncer_1$'s home environment E (and only in E) the presence of X is fairly (though not perfectly) reliably correlated with the presence of the relevant amino acids. DETECT is so wired to $Pouncer_1$'s motor effectors that when DETECT goes into its on state, $Pouncer_1$ emits behavior P (POUNCE) . Fortunately for $Pouncer_1$, POUNCE typically culminates in E in the ingestion of the relevant amino acids.

Some will balk at the intuition that $Poncer_1$'s crude X dectector is a bona fide representational system. Such reticence is perhaps understandable. After all, $Pouncer_1$'s very crude control structure means the its crude intentional states function very differently from what we take to be paradigmatic intentional states, viz., our beliefs and desires. $Pouncer_1$'s crude DETECT device can contain nothing like the plenum of inferentially articulated possible contents that our own "belief box" can contain. Second, unlike either our beliefs or our desires, $Pouncer_1$'s detector states are in direct and inflexible control of its behavior. And unlike our desires, $Pouncer_1$'s detector states are, it would appear, representations of how the world stands, not representations of a possibly non-obtaining but "preferred" state of the world. I concede that $Pouncer_1$ may enjoy neither beliefs nor desires. It may even be right, as Davidson, Dennett, and others allege, that there is some sort of necessity that no intentional system entirely lacking in rational powers could enjoy beliefs or desires. Though I do not doubt that the marriage of rational powers with intentional contents was one of nature's great advances in the design of minds, my

own intuitions is that there are more contentful states on the great chain of intentionality than those that are initially present in only rational minds. For those with harder intuitions, I offer the following as a softening agent.

Imagine a somewhat more complex Pouncer -- $Pouncer_2$ -- with a control structure rather more like our own, but still devoid of anything like rational powers. $Pouncer_2$ is like $Pouncer_1$ except that in addition to a DETECT control device, $Pouncer_2$ also has a WANT control device. $Pouncer_2$ can perform two behaviors -- SEEK and POUNCE. Like DETECT, WANT can be in an on state and an off state. Unlike DETECT, the states of WANT are not directly controlled by the presence or absence of X *in the environment*. Rather, $Pouncer_2$ is so designed that on some fairly regular schedule WANT is caused to be activated by some yet to be specified condition. We might imagine, for example, that state of WANT is causally correlated with the intensity of some "inner" signal ***S*** such that WANT is on when and only when S is above a certain level of intensity, and WANT is off when and only when ***S*** is below a certain level of intensity. Assume that the intensity of ***S*** is either statistically or causally fairly well negatively correlated with state of $Pouncer_2$'s store of amino acids such that when the store is low the intensity of ***S*** is high and when the store is high the intensity of ***S*** is low. Like $Pouncer_1$, $Pouncer_2$ needs amino acids but does not want them, at least not under the aspect of amino acids. Moreover, $Pouncer_2$ is entirely unable to detect the amino acids that it needs.

Now imagine that $Pouncer_2$ is so wired that whenever ***S*** causes WANT to be activated, $Pouncer_2$ emits SEEK and keeps emitting SEEK until DETECT is caused to go on (typically, but not infallibly, by the presence of ***X***). Once on, if WANT also remains activated, DETECT seizes control from WANT, causing $Pouncer_2$ to emit POUNCE. Several different things are liable to happen once ***X*** is detected. ***X*** may disappear again from detectable range. If this happens, DETECT goes off. In that case, if ***S*** remains intense and WANT remains on, WANT regains control, causing SEEK. If ***S*** has low intensity, causing WANT to deactivate, $Pouncer_2$ remains quiescence until WANT goes on again. Alternatively, ***X*** sometimes remains within detectable range, but just beyond catchable range. In such cases, DETECT remains in control, causing $Pouncer_2$ to emit POUNCE again and again. Now as things happen, $Pouncer_2$ never succeeds (or at least hardly ever) in capturing ***X*** by emitting POUNCE -- ***X*** is just too fast. In a sense, $Pouncer_2$ is frustrated every time WANT and DETECT jointly cause it to SEEK and POUNCE.

But Pouncer$_2$ is a quite fortunate creature. For ***X*** hops about from one leaf to the next. Each leaf swarms with tiny amino acid rich bugs, bugs which Pouncer$_2$ neither wants nor detects. But it just so happens that every time Pouncer$_2$ SEEKS and POUNCES he ingests a large quantity of bugs and thus a large quantity of the needed amino acids. Hence, though Pouncer$_2$ can't always get what it wants, when it tries, it sometimes gets what it needs.

I hope it is intuitively plausible that Pouncers have representational states with determinate enough intentional contents which play dynamic roles in Pouncer cognitive economies, such as they are. But Pouncers are clearly possessed of no rational powers. A pouncer's crude behavior, crude perception analogs and crude desire analogs are not related in the way characteristic of either consequential, internal or kantian rationality. Even when a Pouncer "acts" on the basis of such of its crude perceptual belieflike states that might be called true, its actions fail to lead with any degree of reliability, to the satisfaction of its desires. Lacking appropriately general and conditional beliefs, its actions are not linked to its beliefs and desires via the practical syllogism and do not exhibit the syntactic plasticity characteristic of rational action. And Pouncers are certainly not original sources of any self-given rational projects or directives.

Intentional attributions can, of course, be cheap. And even after the application of our softening agents, some will not be able to get beyond the intuition that our Pouncers enjoy at most a gratuitous "as if" intentionality. If we are to combat this response with an argument which goes beyond the power of mere intuition pumps, we will need some principled basis for deciding when we've located bona fide (but arational intentionality) and not merely a simulacrum of intentionality. I suggest that we will have succeeded if we can show that there is a principled basis for assigning at least reasonably determinate contents to the states of some pouncer-like creatures and if assigning contents in accordance with that principle itself subserves some deep explanatory purpose. It is to that task that I now turn.

Consider first the lowly frog and its representational capacities. Frogs have certain neurons that, to first appearances, indicate the presence or absence of flies. These bug indicating neurons (indicator I) are eventually wired to motor effectors that control the snapping of the frog's tongue (effector E) in such a way that, at least in the frog's home environment H, flies will often enough be present when the frog snaps and frogs will often enough snap when flies are present. Natural selection (together with mechanisms of trait inheritance) explains why indicator I is wired to effector E by appealing to the selective advantage

conferred by that wiring on the (ancestors of) the frog. Ancestral frogs so wired stood in H an enhanced-enough-for-reproductive-success chance of getting their tongues to where the good stuff was and an enhanced-enough-for-reproductive-success chance of not wasting excess calories snapping in the absence of the good stuff.

Now suppose that indicator I has many different indicational properties. Imagine, for example, that the states of I are as reliably correlated, at least in ***H***, with the presence of flyish small dark moving things as they are with the presence of non-flyish small dark moving things. It is sometimes claimed that natural selection can bestow upon ***I*** the function of indicating one rather than the other of these, thereby promoting a mere indicator into a bona fide representation with a relatively fine-grained representational content.[22] The motivating thought is that some but not the other of ***I***'s indicational properties will be relevant to explaining why in ***H*** frogs in which ***I*** was wired to ***E*** stood an enhanced-enough-for-reproductive-success chance of getting their tongues to where the good stuff was. It is arguable, for example, that while the fact that in ***H***, ***I*** indicates a smallish patch of darkness moving across the frog's visual field is irrelevant to explaining the selective advantage enjoyed in ***H*** by frogs in which ***I*** was wired to ***E***, the fact that in ***H***, ***I*** indicates the presence of flies is, by contrast, highly relevant to explaining that selective advantage. The further thought is that that such a difference in explanatory relevance between ***I*** 's property of indicating flies and ***I***' s property of indicating small dark moving things can make it the case that ***I*** has the function of indicating the former, but not the function of indicating the latter. Once I acquires a certain indicational function, we can make sense of the idea that ***I*** might sometimes fail to perform its indicational function. Misrepresentation happens when the states of ***I*** fail to indicate what they have the function of indicating.

Natural selection is about past history. This indicator is wired to this effector not because of what it indicates here and now in the life of this very organism, but because of what it once indicated, there and then, in the lives of some ancestors of this organism. Fred Dretske (1988) has argued that what natural selection accomplishes for a species, learning can accomplish for an individual. Beginning with an

22 Dretske (1988) is the leading proponent of indicational functions as the key to semantic content. But see also Millikan (1984, 1993), Papineau (1987) for accounts in which function plays a role in constituting content.

organism ***O***, the control structure of which has a certain degree of plasticity, a proper scheduling of reinforcement can bring it about, Dretske claims, that ***O*** emits response ***R*** in, and only in, condition ***C***. That can occur if some indicator ***I*** that indicates ***C*** is recruited as one among the controllers of ***R***. Again, the idea is not just that learning recruits ***I*** as a controller of ***R***, but of learning recruiting ***I*** *because* it indicates ***C***. When that happens, Dretske claims, it thereby becomes the function of ***I*** to indicate ***C***. And again, if the states of ***I*** are reliably correlated with both ***C*** and ***C***', it's indicating ***C***, but not it's indicating ***C***' can explain its recruitment by learning as a controller of ***R***. Thus learning promotes a mere indicator, with a rather coarse-grain indicational content, into a representation, with a more fine-grained representational content.[23]

My aim is not to show that some version or other of selectionism is correct. Such theories admittedly face difficulties that are not obviously resolvable. They require, in particular, that the content making function be determined by the distinguished role of a certain (indicational) property in explaining (the recruitment of some indicator as the controller of) some successful behavior. But the success of an organism in an environment is typically the product of the interaction of many factors. There will often be a variety of competing ways to divide the labor among the interacting factors in explaining that success. Putative representational contents will be different depending on how the labor is divided among the contributing factors.

---

[23] There is an important difference between so called indicator semanticists like Dretske and so-called teleosemanticists like Millikan that deserves brief mention. Though Dretske and Millikan both suppose that the content-determining facts are those facts that determine function, they have rather different views about the nature of the content-determining functions. Millikan assigns little role to learning in the determination of function. She denies that *indication* functions are the content-determining functions. For she thinks that there can be representation without indication. Her view is more "success-linked." Suppose that the frog could not "detect," in Dretske's sense, the presence of flies, but could detect something else -- **X**, say -- which, at least in **H** -- the frog's home environment -- is well-enough correlated with the presence of flies so that at least in **H** emitting behavior **B** when **X** is present culminates reliably-enough-for-reproductive-success in the capture of flies. Millikan's thought is that since it is flies that the frog is "after" then if emitting **B** in the presence of **X** contributes to the frog's chances for reproductive success *because* **B** culminates in the capture of flies when it is emitted in the presence of **X** then the content of the representational states which "control" **B** will be to represent flies, and not **X.** And this is so, Millikan claims, even if nothing in the frog's representational system indicates, in Dretske's sense, the presence of flies. Though a number of important issues are at stake between Dretske and Millikan, I will not need to be careful of the distinction between Dretske's indicator approach and Millikan's more success-linked approach. So I will treat them both indifferently as what I have called selectionist theories.

Consider what I call the "high content" explanation of the contribution of the frog's representational powers to the success of the frog's fly-directed behavior. The high content explanation has it that selection has recruited a certain indicator as the controller of the frog's fly directed behavior because the recruited indicator is (or was) a reliable indicator of flyish things, and not because it is (or was) a reliable indicator of small dark moving things. Being controlled by a reliable enough indicator of flyish things partly explains, according to the high content approach, why the frog's fly-directed behavior culminates often enough in the ingestion of a fly. Since it is ingesting flies that matters to the frog and since the indicator's property of indicating flyish things helps to explain why behavior controlled by that indicator leads to the ingestion of flies, the indicator will come to have the function of indicating flyish things, but not the function of indicating small dark moving things. That is, once the indicator is recruited to control the frog's fly-directed behavior, the indicator is now supposed to indicate flies. That is what it did that caused it to be recruited (and replicated). This is so even if the recruited indicator remains, post recruitment, a more reliable indicator of small dark moving things than it is of flies. This last fact explains how misrepresentation is possible. Misrepresentation of ***x*** as a fly happens, for example, when an indicator which is recruited because it is a reliable enough indicator of flies turns out to be an even more reliable indicator of small dark moving things and is tokened in the presence of a small dark moving thing which is not a fly. In that case, the indicator misrepresents a non-flyish small dark moving thing as a fly.

Contrast the high content approach with what I call the low content approach. The low content approach says that the relevant indicator is recruited to control the frog's fly-directed behavior because: (a) it indicates small dark moving things -- flyish and non-flyish alike -- and (b) enough of the small dark moving things are flies that snapping at them eventuates regularly enough in the ingestion of a fly. The low content approach says, in effect, that the frog's fly-directed behavior is successful in part because it is under the control of a reliable indicator of small dark moving things. And it adds that being under the control of a reliable indicator of small dark moving things leads to success in the pursuit of flies because in the environment in which the frog does its foraging enough of the small dark moving things are flies that snapping in the presence of whatever is small, dark and moving eventuates often enough in the ingestion of a fly. On the low content approach, the frog does not ipso facto misrepresent a non-flyish small dark moving thing as a fly when it tokens the relevant indicator in the presence of a non-fly. But

misrepresentation is possible as long as it is possible for the frog to token the relevant indicator in the presence of something which is not small dark and moving. For example, perceptually abnormal frogs may token the relevant indicator in the presence of something that is not small dark and moving and may thereby be caused to snap. Such a frog would misrepresent that which is not small dark and moving as small, dark and moving.

I do not propose to solve the determinacy of function problem here. For our purposes, it matters only that either the high content approach or the low content approach would provide a principled enough basis for the assignment of reasonably determinate intentional contents to a frog's representational states. Still, I want to suggest that both high content and low content approaches contain important grains of truth -- grains of truth that should be reflected in any future solution to the determinacy of function problem.[24] And seeing the grain of truth contained in each will help us to see the explanatory point of ascribing intentional contents to the frog's representational state.

I begin by distinguishing stimulus from target . The stimulus for the frog's snapping behavior is what causes that behavior to be initiated; the target of the behavior is what defines "success" for the behavior. Anything sufficiently small, dark and moving will cause snapping behavior. So we may take small, dark, moving things to be potential stimuli for the frog's snapping behavior. Suppose, further, that there is a state of the frog's nervous system that "encodes" the property <smallish dark moving thing> -- where "encodes" is to be understood in broadly information-theoretic terms. We may think of such a state as an inner representation, the stimulus content of which is determined by the property that it was selected to information-theoretically encode. That stimulus content is determined by the property an inner representation was selected to information-theoretically encodes is the grain of truth in the low content approach.

Now imagine a frog whose snapping is under the control of an inner representation which was selected to encode the property <small dark moving thing>. Such a frog will snap indifferently at flyish and non-flyish small dark moving things. Does a frog that snaps at a non-flyish small dark moving thing thereby misrepresent? If we consider just the stimulus content of the frog's inner representation then,

[24] Neander (1993) is a particularly helpful discussion of the determinacy of function problem from within a selectionist perspective.

strictly speaking, the answer is no. The frog represents a certain region of frog-centric space as filled with a small dark moving thing. That region of frog-centric space is, in fact, filled by something small dark and moving. But distinguishing the stimulus content of the frog's snap-controlling representation from the target of the frog's snapping behavior reveals a more complex story about the frog. Even when the frog snaps at a small dark moving non-fly, we can see that the (evolutionary) point of snapping remains to deliver fly-shaped packets of amino acid to the frog's digestive system. A snap emitted in the absence of such fly-shaped packets of amino acids is like an arrow which fails to hit the target at which it is aimed. Not hitting the target is not a way of not being aimed at the target. That the target of the frog's snapping behavior is not determined by the stimulus content of the frog's inner state is the grain of truth in the high content approach.

If this is right, strictly speaking, a frog, the snapping behavior of which is under the control of an inner representation which encodes the property <small dark moving thing>, makes no representational error in snapping at a non-fly, just as the low-content approach says. Nonetheless, such a snapping misses the real evolutionary point, just as the high content approach would have it. Call such pointless snaps "wild" snaps. That the frog will emit wild snaps in certain circumstances is a direct consequence of the nature of the frog's evolutionarily instilled "control structure." The gap between the stimulus content of the frog's snap-controlling representation and the target of its representation-controlled snapping behavior makes it the case that in certain foraging environments the frog is bound to emit wild snaps.

Nonetheless, the frog is well-designed by evolution for foraging in its intended environment. Indeed, the frog's control structure is a simple, but effective solution to the following evolutionary challenge. Given a syntactically expensive to encode target property ***t*** and a foraging environment ***E***, evolve a triple consisting of a relatively syntactically cheap to encode stimulus ***s*** (or set of stimuli), an inner representation ***r*** (or set of inner representations), and a behavior ***b*** (or set of behaviors) such that: (a) in ***E*** , s reliably enough indicates the presence of instances of expensive to encode ***t*** ; (b) ***r*** information-theoretically encodes ***s*** ; (c) ***b*** is controlled by ***r*** ; and (d) in ***E***, ***b*** reliably enough culminates in striking of instances of ***t*** . Possible solutions abound. The frog's "control system" occupies just one point in that solution space. In the frog, syntactically cheap to represent properties of shadow and

shape are put in control of behavior that aims at syntactically expensive to represent properties of molecular structure. Because the syntactically cheap to represent properties and the syntactically expensive to represent properties are reliably enough correlated in the frog's foraging environment, the frog thrives, despite, as it were, it's representational insensitivity to the amino acids at which its snapping behavior is aimed.

The frog's inner representation really have two distinguishable functions -- the representational function of encoding the property <small dark moving thing> and the executive function of driving the frog's snapping behavior. Moreover, snapping behavior in the frog may itself be said to have a distinguishable function -- the function of delivering fly-shaped amino acid bundles to the frog's digestive system. Now we need not suppose that the representation that drives a bit of targeted behavior inherits its representational content from the target of the behavior that the representation drives. What frogs represent are regions of frog-centric space as filled with small dark moving things. What frogs pursue are flies. The mismatch between content and target will sometimes causes wild snaps. But fogs lucky enough to spend their lives foraging only in the environment for which they were "designed" will seldom be the victims of this design limitation.

Nature is, I suspect, replete with frog-like intentional systems. For it is likely that in "designing" intentional systems, natural selection has obeyed the constraint of representational minimalism: ceteris paribus, prefer syntactically "cheap" representational contents to syntactically expensive ones. Nature is unlikely to have endowed an organism with a reliable, but syntactically costly amino acid detector, capable of reliably detecting the presence of certain amino acids in a variety of guises, via a variety of different evidential pathways, where, without selection disadvantage to the organism, it could have endowed that organism with a syntactically low cost detector of a pattern of shadow and shape. The principle of representational minimalism provides a basis for generating hypotheses about the representational contents of evolved natural systems by licensing the defeasible assumption that nature "prefers" to spend fewer rather than more syntactic/inferential resources in designing organisms. The calculus of costs and benefits cannot be an a priori one. Costs must be relativized to facts about the opportunities, challenges, and ancestral endowments of particular species over the long course of evolutionary time. Indeed, sometimes when the payoff in increased fitness is high, nature will pay the costs. What is crucial for our

current purposes is that we can frame a calculus of cost and benefits of different styles of representation without imputing either logical syntax or rational powers to the relevant organism. Indeed, it will often be the case that we can determine just what point in solution space is occupied by the control structure of a given organism only by appreciating how very unlike a fully rational intentional system, with the fully panoply of rational and representational powers, it is.

3.2 Aggregating up: A Sketch

It remains to determine whether frog-like or type I intentionaliy can be plausibly aggregated up to yield full blown rational or type II intentionality. The key to seeing that it can, I claim, is to execute something of a gestalt shift. Suppose, in particular, that we ask whether it is plausible that the advent of rational powers and logical syntax was an adaptive solution to a problem faced by our non-rational progenitors. If it is true that nature obeys a principle of representational minimalism in the design of intentional systems, then our pre-rational and syntactically less articulated progenitors will often have exhibited a mismatch between some or all of their behavior guiding representational contents and the targets of their representation driven behaviors. To that extent our pre-rational progenitors will often have been prisoners of idee fixe, of fixed ways of detecting and pursuing the good. Consequently, their worldly success, such as it was, was likely to have been highly dependent on the luck of the environmental draw. Against such a background, it is not difficult to imagine the relative advantages enjoyed by that fortunate creature in which the full range of rational powers and logical syntax first emerged. Its mind contained a plenum of syntactically articulated representations and enjoyed sufficient syntactic plasticity to enable it to link and unlink, in goodness-enhancing ways, these representations to one another and to action on the basis of its experience in and actions on the world. Such syntactic plasticity would have endowed it with the capacity to build ever better theories of the world, to learn new, deeper, more explanatory regularities, to come to see that which was merely locally and contingently reliable as merely locally and contingently reliable and to acquire ever more powerful practical abilities to manipulate the world for its own ends. A mind endowed with such powers would no longer be hard-wired to pursue a given target solely on the basis of a given stimuli. Such powers would have been the foundation of brand new abilities to manage the ancient gap between stimulus and target. Equipped with its representational plenum and its ability to

manipulate the plenum in goodness-enhancing was, this newly minted mind would have possessed the capacity to travel diverse evidentiary pathways to a given target, to acquire new evidentiary routes to a given target and to extinguish old routes to that target. It would no longer, that is, be the prisoner of fixed ideas and would be far less dependent for its worldly success on the luck of the environmental draw. To say this is to not say that this new mind would have been indefinitely syntactically plastic. There are likely to be real limits on the power of any finite mind to reorder its representations in goodness-enhancing ways. Moreover, though I have heretofore left to one side the subject of modularity here, it is very likely that the mind's rational powers sit on top of fixed array of modules. The presence of some more or less extensive array of fixed modules would no doubt set real limits on the capacity of even a mind newly endowed with rational powers to alter its representational landscape in goodness enhancing ways. In addition, though the first emergence of rational powers would have brought with it new prospects for worldly success, it would also have given rise to cognitive challenges never before encountered. For once it became possible for a single mind both to represent the same thing twice from distinct but simultaneous perspectives and to be aware of is own potential to do so, a brand new coordination problem arises: the problem of coordinating the inner syntactic lives of the mind's inner representations with the outer semantic destinies of those representations. Eventually, for example, there came to be minds possessed of inner, syntactic marks of same-purporting thoughts, minds whose inner representations achieved the character of objectuality.[25] Only when the inner syntax of thought became sufficiently articulated and conjoined with rational powers, did the causal impact of mind on world suffice for situated rational intentionality.

The foregoing reflections do not amount to a knock down argument in favor of the claim that the conjunction of frog-like or type I intentionality with rational powers and logical syntax yields type II or rational intentionality. They are intended only to render it plausible that although the advent of rational powers and logical syntax in an already contentful mind represents a great advance in mind design nonetheless that advance may well have occurred as an adaptive response to both pervasive problems faced by living things in general and the peculiar problems that come with Type I intentionality. All life faces certain evolutionary challenges generated by the famous four f's: living things must feed, fight, flee,

[25] For a discussion of same-purporting thoughts see Essay VII. For a discussion of objectuality see VII and also VI

and reproduce. If an organism is to thrive in the environment in which it finds itself, it must embody some strategy for getting to where the food is, for getting away from where the danger is, for overcoming its foes, and for getting more of its kind into existence. What divides intentional systems from non-intentional systems is not the broad evolutionary challenges they face, but the range of strategies for meeting those challenges that they embody. Intentional systems are the ones that have achieved representational capacity as a means of solving their ecologically determined problems. They are the ones that perceive their mates, their food, their foes; the ones whose actions on and in the world are guided by what they perceive, where there perceive it, and what they "want" with respect to the things they perceive. Not all organisms embody such strategies. The lowly but lovely amoeba has solved its evolutionary challenges in a radically different way. It reproduces by dividing in half and thus needs neither to find nor attract a mate; it is semi-permeable and swims about in a rich nutrient bath and thus has no need to encode a signal for its target nutrients; and it is not surrounded by too much danger in proportion to its number, so that it needs no elaborate strategy for escaping or defeating its foes. The amoeba's problems are all solved without it having to represent anything at all in the world. Some number of evolutionary steps removed from the amoeba are type I intentional systems. Such systems have representational powers finely tuned to their peculiar evolutionary niches. Their powers respect the principle of representational minimalism, but they lack the transformative and reflective powers that are a sine qua non of rational intentionality. Such transformative and reflective powers are, I have been suggesting, relatively recent evolutionary innovations. The emergence of such powers was not the first appearance of intentionality, but a refining and deepening of it. Organisms with such powers will have capacities not shared by implastic, unarticulated, and unreflective organisms -- abilities to deliberate, to learn, to theorize, to infer, to constitute spaces of shared reasons. But lacking such powers is not a way of lacking representational powers altogether. Indeed, we have gained a deeper understanding of rational intentionality by understanding its place in the natural order of which it is a part and by appreciating what, besides mere intentionality, it takes to build a rational mind.