WHAT IS AN INTENTION?

J.-M. Kuczynski

What is an intention?

J.-M. Kuczynski, PhD

Introduction: Intentions as system-level judgements, values as system-level beliefs

In this monograph, it will be made clear what intentions are. Specifically, it will be proved that intentions are judgments as to how to consolidate one's psychological architecture. It follows, so it will also be established, that *values* are beliefs as to what, in general, is conducive to the integrity of one's psychological architecture.

The meanings of these claims, as well as the justifications for them, are to be understood in terms of Harry Frankfurt's incisive analysis of the concept

of freedom, to which we now turn.

1.0 Frankfurt's analysis of freedom

According to John Locke, you are free to the extent that you can do what you want to do, and you are unfree to the extent that you cannot.

According to Locke, this entails that freedom is a property of actions, not of desires, i.e. whereas actions can be meaningfully describe as free or as unfree, desires cannot be meaningfully be so described. Locke's reasoning:

> *One does not desire to desire to go jogging. Were it one's desire to desire to go jogging, one would* ipso facto *already desire to go jogging. In general, second-order desires (desires to desire) collapse into first-order choices (desires to act). Therefore, given that one is free to the extent that one can do as one desires, it follows that freedom is a property of actions, not of desires* per se.

Freedom to do as one pleases is obviously *a* kind of freedom—a very important kind.

But it isn't the only kind. And contrary to what Locke says, there is *a* sense in which choices can be free or unfree. Harry Frankfurt made this clear. A story will help us begin to understand Frankfurt's insightful analysis of freedom.

Jones is a great writer who is addicted to heroin. Because his addiction is only moderate, he is able to function perfectly well even when he is heroin-deprived. (When heroin-deprived, he doesn't feel good, but he is in no way incapacitated.) He knows that, if he were to shoot up right now, the subsequent lethargy would make it impossible for him to write a single word

for the rest of the day. He also knows that, if he doesn't write 10 pages today, he will lose his book contract. Jones does not want to lose the contract. Plus, he wants to grow as a writer, which he won't do, at least not today, if he shoots up. He thus has a desire to *not* take the heroin—even though he simultaneously has a desire *to* take the heroin.

Jones' desire to take heroin is a first-order desire. Jones' desire *not* to act on this desire is a *second-order desire.* This is because Jones' desire not to take heroin is a desire as to what to do with another desire; it is a desire *not* to cave in to a desire to take heroin.

Suppose that Jones caves in and he takes the heroin. In that case, Jones is a slave to his desires. His will has been compromised. He is not acting freely.

But *now* suppose that Jones does *not* cave in. He does the right thing: instead of taking the heroin, he endures the discomfort of a heroin-free day and manages to get his work done. In *this* context, Jones is not a slave to his desires. His will has not been bent. His conduct expresses a "free will."

In light of considerations such as these, Frankfurt says that somebody has a free will to the extent that his conduct aligns with his *second*-order desires.

But here a question arises. Why are second-order desires so special? Why is it that freedom consists in one's doing the bidding of one's *second*-order desires—as opposed to one's third-order, or fourth-order, or first-order desires? What, in this context, is so special about the number two?

When the heroin-addicted writer decides *not* to take heroin, what is driving him is a rational appraisal of his situation. But when the heroin-addicted writer gives in to his addiction, it is *not* his rationality (or his morality) that is driving him.

Before we can close this argument, we must address Frankfurt's discussion of the concept of a "person." By a "person," Frankfurt does not mean a member of the species *homo sapiens.* He means anything that has the

same basic psychological architecture that we have. What is so special about that structure? First of all, we are self-conscious. Second, we are rational. Third, we are rational not only about how to attain our objectives, *but also about what those objectives should be.* What distinguishes persons from non-persons is that the former, but not the latter, subject their own desires (and, more generally, their own mental states) to critical scrutiny.

Non-persons are often very clever when it comes to figuring out how to attain their objectives; for example, beavers show a lot of intelligence when it comes to building dams. But they are not smart about whether they should be building dams in the first place. The question doesn't arise. They have the desire, and they act on it (or they die trying). In general, persons do, whereas non-persons do not, subject their own desires to rational scrutiny. Non-persons do not ask themselves: "What would the consequences of acting on this desire be? And *should* I act on this desire?"

We persons, on the other hand, *do* subject our objectives to rational assessment. And this is obviously made possible by the fact that we are self-conscious. If I am overtaken by a desire to build a dam, I will ask myself: "how would my acting on this desire help my overall interests?"

We can now close the argument. Whenever someone is driven by a rational assessment of their own desires, they are driven by a desire about a desire. Thus, one is rational only to the extent that it is one's second-order desires that drive one's conduct. There is a sense in which rational conduct is free and in which irrational conduct is not.

To make it clear what this sense is, we must first distinguish between intelligent conduct and rational conduct.

1.1 Rational ≠ intelligent, irrational ≠ unintelligent

An irrational act can involve an enormous amount of intelligence. Smith is a brilliant inventor. He tries to impress his wife by building a giant robot that sings romantic songs in a metallic voice. This alienates her and she divorces him. Moreover, the financial and personal resources he sunk into this project have left him destitute. Smith's act might have been about as intelligent as an act can be. But it was not a rational act. For intelligently carried out though the act was, the immediate object of that act (namely, the construction of a robot that would sing love songs to his wife so that he'd be spared the trouble of doing so) was not intelligently arrived and failed to embody consideration of relevant facts that it was well within Smith's ability to consider.

A rational act is one that embodies due regard for one's various interests. An irrational act is one that does not. An intelligent but impulsive act (e.g., Smith's speedily building this robot to allay his rapidly growing anxieties about his marital situation, all of which are projective but which become legitimate as a result of this endeavor on his part) fails to embody due regard for one's own interests. (Smith isn't taking into account the consequences of his actions; he is misdeploying his vast stores of intelligence. And, granting that there are different kinds of intelligence, he obviously has whatever kind of intelligence one needs to have to know that (supposing for argument's sake that there's no chance that this project would bring him to the attention of any potential benefactors) this project is bound to land him on Skid Row.) So far as one is driven by first order desires, one is neglecting to gratify one's second order desires. So, given that second order desires are clearly more closely hewed to one's middle- and long-term welfare---or, if it isn't one's own welfare that one cares about, one's middle- and long-term situations---it is clearly more rational to act on one's second than on one's first order desires. Also, to make a point that we will soon develop, it is often rational to perform actions in which, were one's first-order desires allowed to discharge

themselves, those desires would be expressed. *Too* much self-restraint is dysfunctional and, in addition to being unpleasurable, seals off possible well springs of insight and creativity. But, setting aside cases of psychopathology, when a person performs one of these enjoyment-oriented actions, he isn't acting on an impulse; he's acting on his *approval* of an impulse. In other words, he's considered the impulse, determined that it lines up with second order desires (or, what I believe to be different, his values), and, on that basis, lifted the initial inhibition on his impulse. So, while the resulting action may be indistinguishable from what he would have done had the impulse itself been the proximal cause of his deeds, his psychological condition, given that his action was *not* directly prompted by that impulse, is radically different from what it would have been had that action been thus prompted.

Having distinguished between rational conduct and intelligent, we may proceed with our discussion of Frankfurt's analysis of freedom.

1.2 Frankfurt's analysis of freedom (continued)

There's more to being free than being able to do what you want to do. Your desires must reflect who you are. If they don't, those desires *belong* to you but aren't *of* you. And in that case, you are alienated from your own mental states, in much the way that a person with Parkinson's is alienated from his body.

But, in a way, the alienation is even worse. Involuntary behavior is unintentional behavior. Intentional behavior is voluntary behavior. Voluntary behavior is desire-driven behavior. (If you don't *want* to raise your arm, but it goes up anyway, you didn't intend for it to happen.) Epileptic fits are not *actions.* Unintended behaviors are not *actions.* Unlike actions, they're *undergone*, not *done.* So estrangement from one's own actions amounts to

estrangement from all of one's own desires; and it's hard to imagine a condition less free than that.

Another story will help expose a problem with the analysis of freedom just described, and it will also clarify the raison d'être for Frankfurt's soon to be discussed emendations of it.

Smith's first-order desires are seriously out of alignment with his second. He values writing poetry; he's good at it; and he enjoys it about as much as it's humanly possible to enjoy it. But he has a voracious craving for drugs. His craving for drugs, especially cocaine, is unusually strong; and his desire to write poetry, though intense, is not as intense as his desire to do cocaine.

Smith's desire to do cocaine is out of alignment with many of his values (second-order desires). Were he to give into this desire, he couldn't write poetry; nor could he maintain friendships or play sports or, with only a few exceptions, do any of the other things he values. For this reason, he doesn't give in. But given how intense his craving is, every day is a struggle for Smith. If, as Frankfurt suggests, one is free to the extent that one acts on one's second-order desires, Smith is as free as a person can be. But given how hard life is for Smith, it's counterintuitive to suppose that he's maximally free. To be sure he has *a kind of* freedom, and he has it in abundance. But there's some other sort of freedom that he doesn't have.

Frankfurt is aware of this problem, and he revises his view to deal with it. According to his revised view, if an act is to be free, it is *necessary* that it be driven by a second-order desire, but it isn't *sufficient.* Another condition must be met: one's first-order desires must be consistent with the second-order desires that are driving that act.

Here's the idea. Person *X* has no self-control. He binges on drugs and food, and he hates himself for doing this; but he can't stop—he's too weak and his urges are too strong. His life has no meaning and he knows it; but he

will not, or cannot, do anything about it. Unlike *X*, Person *Y* has a huge amount of self-control and never caves into temptation. But everything is a struggle, since he's always fighting temptation, like a drowning man trying to stay afloat. (*Y* is like Smith.) Person *Z* cannot be categorized with *X* or with *Y*. *Z* always does what he values, which is composing music. And what he values is what he *wants.* He wants what he wants to want. Because he's doing what he wants, he has the freedom that *X* has that *Y* lacks. Because he's doing what he wants to want, he has the kind of freedom that *Y* has that *X* lacks. And since, of the three, he's the only one who has both kinds of freedom, he's obviously the most free of them all.

Thus, one is free insofar as two conditions are met: (i) one acts on one's second-order desires, as opposed to one's first-order desires; and (ii) one's second-order desires are in alignment with one's first-order desires.

To see the merits, and also the shortcomings, of Frankfurt's brilliant analysis, we must take a moment to say how that analysis relates to compatibilism.

1.3 "*x* is free"= "*x* does what *x* wants to *provided* that what he wants to do is adequately rooted in his personality structure"

We've considered two analyses of freedom:

(i) "x performed act A freely" = "it is because x chose to do A that he did it,"

and

(ii) "performed A freely'' = "A reflects who x really is."

These two views can be combined into the view that:

(iii) "x performed act A freely" = "it is because x chose to A that he did it, provided that this choice of x's reflects who x really is."

This analysis, which is basically Frankfurt's, embodies a profound insight into the nature of human freedom. We will presently see that, although it's a step in the right direction, it isn't correct: for it is possible to warp one's higher-order desires to suit one's first-order desires and, thus, to make one's higher-order desires be mere vehicles for first-order desires.

But right now, I'd like to concentrate on an apparent problem with Frankfurt's analysis that, although, I suspect, some would be inclined to think otherwise, isn't an *actual* problem with it:

> *(FD) Smith has given himself over to drug addiction. He spends days doing drugs. (He has unlimited supplies of money and never has trouble getting them.) Smith has many talents—he's a great water-skier, a great short-story writer, etc., but, because of his addiction, these gifts of his lie fallow.*
>
> *Green, on the other hand, hasn't given himself over to addiction. He spends his days doing things that augment who he is. Green can't indulge every craving that he has. Unlike Smith, he doesn't have the money to do so.*
>
> *You (JMK/Frankfurt) would say that Green is freer than Smith, and your reason would be that, even though he can't gratify as many desires as Smith can, the desires that Green does gratify have the right roots, whereas those that Smith gratifies do not.*

I grant that Smith's condition is worse than Green's, and also that Smith is less free than Green. But, contrary to what you say, that is because, when it comes to desire gratification, Green's batting average is better than Smith's. There are a lot of aspirations that Smith has that will be thwarted as long as he's imprisoned by addiction. Smith wants to do cocaine more than he wants to be a writer, but it doesn't follow that he no longer wants to be a writer. He could very well want both.[454] *Let's suppose he does. In that case, a very strong desire of his is being frustrated. It's true that he's the one who's frustrating it, but that doesn't mean it isn't frustrated.*

Smith will inevitably have many other desires that his cocaine addiction dooms to frustration. In his pre-addiction days, we may suppose, Smith enjoyed many things that he no longer enjoys. We can assume that he desired to do each of these things, and we can also assume that he still desires them, even though he obviously doesn't desire them as strongly as he desires cocaine. Many desires are rooted in personality structures. A desire to play the piano or read Nietzsche doesn't just overtake one; it's rooted in enduring structural facts about one's personality. And as long as the personality structure is there, one's inability to express it in the form of action is, practically by definition, a form of frustration. Smith's personality structure may still support those desires. (And almost certainly will. Addiction doesn't turn people into complete vegetables, at least not right away, and usually not ever.) It could be that Smith thinks it would be greater. But it's not so clear that it would in fact be greater. The phenomenological expression of that frustration would be more intense. But that doesn't necessarily mean that the desire in question is itself more intense. Desires, even short-lived ones, are structures;

they aren't feelings. And the feelings to which a desire gives rise don't necessarily make it clear how intense that desire really is.

1.4 Problems with this analysis

FD is close to the truth, but it falls just short. Another story shows why. Aaronson is a coward who has no self-confidence. So even though he wants to write novels and is extremely talented at it, he doesn't do so. It isn't safe enough. It's too bohemian for his stodgy, blinkered parents and for his equally stodgy friends (who he doesn't even like—he just pretends to like them, since he's supposed to). To make a living, Aaronson sells insurance, which he hates. His life is a waking death.

In living as others want him to, Smith is surrendering his freedom *and* he is also dooming many desires of his to frustration. But those two failings, though related, are not identical. Aaronson is suppressing an important side of himself. The personality structures that he is suppressing *give rise* to many ardent desires, but they aren't *identical* with such desires. His losing his freedom consists in his suppressing these structures. The desire frustration that follows is an *effect* of this loss of freedom and isn't *identical* with it.

So Aaronson's loss of freedom isn't identical with his being unable to do what he wants; his being unable to do what he wants is a by-product of his loss of freedom. To be sure, his inability to do what he wants is *itself* a loss of freedom. But, important though this second kind of freedom is, it's distinct from the first kind and a derivative of it. So being free consists primarily in being oneself, and secondarily in being able to do what one wants. And this is obviously close to what Frankfurt is saying, if it doesn't coincide with it.

1.4.1 Frankfurt’s analysis systemically flawed

Many analyses that philosophers put forth seem not to embody genuine attempts to find the truth. Frankfurt's analysis of freedom *does* embody such an attempt. And that attempt is *almost* successful.

But it falls short. In the next section, we will see why. Here's an outline of what we'll find. Acts that are driven by one's Higher-order desires need not be free; for people can, and often do, warp their higher-order desires, so as to make them fit their first-order desires. Higher-order desires, in other words, can become vehicles for first-order desires. A lazy, weak person can warp his higher-order desires so that they merely rubber-stamp the craven first-order desires of which his actions are compulsive expressions.

We will find that truly free acts are those that are driven, not by higher-order desires, but by *values*. We will find that there is no agency of any kind where there are no selves, and we will find that, where there are no values, there are no selves, though there may be *minds*. I will leave it at that for now, since the entirety of the last part of the present will work will clarify and justify those statements.

2.0 An as of yet unidentified problem with Frankfurt's analysis

According to Frankfurt, a free act is one that is driven by higher-order desires. This is false. A free act is one that is driven by one’s *values*. Values are not desires. (Values may lead to desires, and desires may lead to values. But values aren’t desires.) A story will help us begin to establish these claims.

You hate opera, and you hate the kind of people who hate opera. But those people are chic; they’re the “right” sort of people. You want to be one of them; you want to be chic. But it isn’t because you see their lives as being valuable

that you want to be one of them. In fact, you see them as empty, bad people who act in empty, bad ways. But you're a follower, not a leader. Your desire to enjoy listening to opera has nothing to do with your valuing opera or, indeed, with your valuing anything. That second-order desire of yours is simply a means to end, the end being the fulfillment of some first-order desire, the nature of which we will discuss in a moment. You know that, if you don't come to have a genuine love of opera, you'll inadvertently disclose to the people whose ranks you are trying to join that you don't much care for them. While in your box-seat at the opera-house, you'll look bored and sullen; you'll sound fake when you tell these big people how much you like opera. And you believe that, as a result, they won't let you join them, in consequence of which you'll wither away.

That first-order desire is not a desire to survive: you know that your survival doesn't depend on your being in the good graces of these people. Rather, it is a cowardly desire to fit in. And *that* first-order desire is, we may plausibly suppose, the off-spring of another, equally despicable first-order desire, namely, your desire to avoid the hard, anxiety-ridden work of forging your own identity. So it is laziness and cowardice that underlie your first-order desire to fit in. It is not, I repeat, a desire to survive.

In this context, your behavior is a compulsive response to fear and laziness. It is not a free act. It is a compulsive, pseudo-free act.

In this context, it is important to distinguish between fear-driven behavior and cowardly behavior. People are vulnerable. They must take their vulnerabilities into account when deciding what to do. The coward has no compunctions about parting with who he is in his efforts to survive. The fearful, non-coward will avoid parting with his identity *until*, having considered all of the possibilities, even those that would involve the hard work and alienation from ways of living to which he is accustomed, he judges that he has no alternative but to forfeit at least part of his identity in order to survive. It should

be said that many a fearful, non-coward will risk death before forfeiting his identity or even a part thereof.

2.1 Second order desires sometimes are the puppets of first order desires

So, in this context, your higher-order-driven desire is purely instrumental: It is just a means by which you try to gratify your desire to survive. And, though you *want* to survive, you don't *value* it; or, what is probably more accurate, your desire to survive is more basic than your judgment that it would be good to survive, and the latter is an intellectualization, a derivative, of the former. Your desire to survive is a force unto itself. It doesn't await your *approval*. It directly motivates action. And, should you come to approve of it, this judgment of yours has little role, if any, in the instigation of actions on your part that, so you hope, conduce to your survival. Your desire to survive is much more primitive than any judgment of yours as to the extent to which some norm is being complied with. All animals want to survive; but not all animals make judgments. Even though there may well be cases where your desire for such and such is a consequence of your judgment that such and such has value, your desire to survive isn't one of these cases. Thus, second-order desires can be value-free; for they can be nothing more than the techniques for gratifying (valuatively empty) first-order desires.

2.2 Fake freedom: The warping of one's higher-order desires to make them validate one's lower-order desires

What if one *warps* one's higher order desires to validate one's lower order desires? What if one rationalizes?

These questions show that, as it stands, Frankfurt's analysis won't do.

Knowing that doing right by one's existing second-order desires would be difficult, and knowing that doing right by one's first-order desires would be easy and pleasurable, one can convince oneself that one's current second-order desires are the wrong ones and that they ought to be replaced with ones that, as it happens, validate one's current first-order desires. If one is too lazy and weak to do what one believes to be the right thing, one has a choice: (i) one can choose to do the right thing, which, as one knows, involves making much needed, but labor-intensive, changes to one's own character; or, (ii) one can choose to replace one's existing second order desires with ones that---what a surprise!---just happen to validate one's existing first-order desires.

Doing this involves convincing oneself that one's character-defects are actually virtues. So, supposing that Green is somebody who is doing just this, here is what he might tell himself ("GR" is short for "Green's rationalization"):

> *(GR) I'm not "lazy." Unlike the blinkered maniac who works all the time and never takes the time to smell a rose or go to the beach, I see the larger picture. Nor am I "weak." Unlike the so-called "hero" who moronically tries to buck the system, I see the larger picture; I see that despite that system's imperfections, it is one that I am duty-bound to comply with; and, unlike the so-called hero, I am enough of a realist to know that complying with the system involves getting one's hands dirty. The "hero" isn't a hero at all.*
>
> *To be sure, the (so-called) hero's efforts have a sheen of valor. And, in sticking up for the little guy---the one whose work is being plagiarized by the well-entrenched senior professor, the one who is being framed by the oft-decorated police-officer---the hero's conduct might initially appear to be more in keeping with the rules that are constitutive of the system. But that initial appearance is an illusion.*

We need the system. If we don't follow the rules, the system will collapse. But there are two kinds of system-constitutive rules: those that are spoken and those that are not. The hero's behavior is heroic only relative to the first set of rules. But, in this context—the one in which it would be said by a short-sighted person that I am being "weak," the one in which I'm supporting the senior professor's portrayal of events---it is the second set of rules that count. The hero (the odd graduate student or pre-tenure professor who is "telling it like it is" and isn't abetting the efforts of the senior professor) is violating those rules. And he's doing so knowing that, because the rules he is violating must never be spoken, people will be forced to characterize his anti-social, system-undermining behavior as system-compliant. So the hero's behavior is particularly insidious. Therefore, to the extent that his behavior is antithetical to mine, I will take it as a complement when somebody reacts my system-friendly conduct by calling me "a weak, faceless bureaucrat" or "an empty, servile, cringing shell."

To be sure, the system isn't perfect. It often rewards the wicked and punishes the virtuous. It often requires that one speak ill of somebody who one knows is decent---who, though nobody says it, everybody (except for a few imbeciles) knows is decent and who, though nobody says it, everyone (same qualification) knows to be a person of merit---and to laud somebody who one knows (and who everyone knows, etc.) to have no merit.

But, on balance, the system is good, and to keep it going, one must follow the rules---all of them, not just the spoken ones. And if that means taking the "coward's path," so be it.

Deep down, Green doesn't really believe GR. Deep down, he knows that he's a servile, weak person who is doing the pragmatic thing, as opposed to the

right thing. But, at least to the extent that he's convinced by GR, Green has replaced one set of higher-order desires (e.g., a desire to act virtuously, even when, because of circumstances, one wishes to act in a craven and iniquitous manner) with a different second-order desire (a desire to resist the odd temptation one occasionally feels to do the "right" thing, even when it is not to one's practical advantage to do so).

It's pretty clear that this world is replete with people like Green. This is a problem for Frankfurt's analysis. So far as Green is acting on his *new* higher-order desires---the ones that, in telling himself GR, he is installing---Green is letting his own cowardice make his decisions for him and he is thus *not* acting freely. A junky who lets his low cravings make his choices for him; he is acting freely only when he takes the reins and acts in accordance with what *he* wants to do, as opposed to what his *addiction* wants him to do. Green is similarly unfree, since, instead of doing what *he* wants to do, he does what his *cowardice* wants him to do.

So, contrary to what Frankfurt says, an act can be *unfree* even if it expresses a higher-order desire. This happens when the higher-order desire is the puppet of a lower-order desire.

3.0 Values ≠ desires

Given the points just made, Frankfurt's analysis of freedom must be modified or augmented in some way. More specifically, Frankfurt must take one the following four positions (and make any such adjustments to his system as the position chosen requires). The result will, I believe, be a decent analysis of at least one, very important kind of freedom. Here are those positions:

(i) Contrary to what was just said, people don't rationalize; they don't

deceive themselves.

(ii) Although people do rationalize and deceive themselves, their doing so doesn't involve their actually replacing any higher-order values of theirs with other higher-order values. It involves only their falsely believing that they have done so. The original values stay on, this being why rationalizers are tormented by gnawing feelings of guilt (from which they typically attempt to insulate themselves by adding more rationalizations to their existing rationalizations). In other words, rationalization isn't about actually replacing one's values/higher-order desires with new ones. It is about making oneself believe that one has effected such a replacement.

So, in deferring to the senior professor, Green isn't acting according to his second order desires. Rather, he's acting according to what he's deluded himself into believing are his second order desires. (Consequently, the fact that Green's actions are not free, except in the sense that they're voluntary, is consistent with Frankfurt's analysis.)

(iii) Situations such as those described by GR neither constitute cases where one set of higher-order values is replaced with another nor constitute cases where some new set of higher-order values has altogether failed to have been installed. Rather, such situations involve a bifurcation or fracturing of the psyche---a kind of splitting of oneself into two selves, one of which still has the old set of values and the other of which has the new set of values.

So, according to this hypothesis, rationalizations do install new higher-order desires. Contrary to (iii), however, they do so, not by

ousting old higher-order desires, but by adding new ones to them.

Given that the new ones are inconsistent with the old ones, rationalization necessarily involves a degradation of one's psychological integrity. One's thinking must become more illogical than it used to be or one's agency---by which I mean the totality of faculties involved in one's being able to decide what to do---must undergo a sort of splitting.

(iv) Although second-order values can be uninstalled and replaced with others, a distinction must be made between higher-order values, on the one hand, and core values, on the other. All core values are higher-order values, but not all higher-order values are core values. Green's actions are unfree because, although they're consistent with (some of) his higher-order values, they aren't consistent with his core values.

3.1 Free acts = value-driven acts

(i) is out of the question. To say that people don't rationalize is itself to rationalize, given what an obvious empirical fact it is that people rationalize.

Also, the essence of Frankfurt's position is that, if one is to be free, one must not be at war with oneself: intrapsychic conflict is the enemy of freedom. Frankfurt is aware that there are different ways of eliminating intrapsychic conflict and that some of them not only fail to bring one freedom but strip one of it. For example, given somebody who is torn between shooting heroin and kicking his addiction, that person can cease to be at war with himself by relinquishing his desire to kick his addiction and, thus, to surrender to it. But, as Frankfurt himself clearly states, by resolving his internal conflict in *that* way,

the junky is forfeiting his freedom. Nonetheless, Frankfurt holds that, in cases where internal conflict is so intense that no one party has uncontested control over the reigns of agency, one cannot possibly become free without ceasing to be internally conflicted. Knowing that some ways of resolving internal conflicts fail to make one free, Frankfurt holds that the way in which a person resolves a case of internal conflict will bring that person freedom if and only if two conditions are met, to wit:

> *(a) The resolution of the conflict involves his higher-order desires prevailing over one's lower desires;*
>
> *(b) The way in which that person's higher-order desires prevail over his lower-order desires doesn't involve the latter being unduly frustrated and, consequently, requires that the victorious higher-order desires so align with the lower-order desires over which they've triumphed that the latter can discharge themselves in a way that doesn't threaten the hegemony of their higher-order masters.*

This position is a near duplicate of a position that Freud explicitly advocated, in various forms, throughout his career.[459] It may be possible in some attenuated, strictly logical sense for there to be Intrapsychic conflict without self-deception or rationalization. But, human psychology being what it is, it seems that self-deception is essential to the perpetuation, if not the initiation, of such conflict. In any case, (i) is *prima facie* discrepant with empirical facts; and, although many philosophers and psychologists virulently deny (i), they do so on *a priori* grounds that, as we'll see, are made of straw *or*, what is much more common, they don't bother to defend their position.

(ii) is inconsistent with empirical facts. People's values do change. As a 20

year old, Jim wants to save the world. He works for the Peace Corps. He spends his weekends working at soup kitchens. As a 40 year old, Jim is an investment banker who lives to accumulate wealth and wouldn't dream of not having a live-in maid, butler, etc. One *could* say that, even as a 20 year old, Jim's values were *really* those of a greedy banker *or* that, as a 40 year old, his values are really those of a hippie do-gooder. But such a position would be *ad hoc* and counter-empirical.

To be sure, there are many cases where somebody who is *in fact* a greedy sociopath works at a soup-kitchen (etc.), so as to bolster his resume in the hopes of making a lot of money on Wall St. And there are people who make a lot of money on Wall St. who wish that they were doing good for the world and whose consequent grief all the gin in the world cannot drown out. But there are also people whose values truly change.

(iii) describes something that, I suspect, really does happen. But, so I also suspect, it does not describe what is *typically* happening in cases of rationalization. In my personal experience, and in the clinical experience of many an authority, what *typically* happens is that some *one* set of higher-order desires becomes corrupt: there is a mixing of the old (pure) higher-order desires with new (corrupt, rationalization-driven) higher-order desires. There is still just one agent. There aren't two agents—a Dr. Jekyll and a Mr. Hyde—inhabiting one body. There is but one agent. Some of his higher-order desires are wholesome and some of them are unwholesome; and in some cases a single higher-order desire may be wholesome in some respects and unwholesome in others.

(iv) is the only remaining option. In the pages to come, we'll find much reason to accept (iv) and no reason to reject it. Incidentally, this analysis, though clearly similar to Frankfurt's analysis, does differ from it in a non-trivial way. Frankfurt seems to hold that one's higher-order desires *ipso facto* are one's core

desires. But that is precisely what the present hypothesis denies.

3.2 Some general facts relating to the nature of values.

In the next section, we'll answer the question: "what are values?" Having answered this question, we'll be able to answer the question: "what is it for an act to be free?" We'll answer the latter question in the next section.

(1) Values must be distinguished from things that have value. A great novel has value, but it is not itself a value.

(2) "Value" is a relational term. What has value for Smith may not have value for Jones.

(3) (2) doesn't entail that no value is better than any other value; nor does it entail that values are in any way 'subjective' or 'unreal.' The term 'three miles from' is relational; but whether Smith is three miles away from Jones has nothing to do with anyone's feelings or opinions about anything. The same thing is true of values, as we'll see. Given that values are judgments, values can be correct or incorrect.

Given that values are relational judgments, two people who have different values needn't disagree about anything. I value playing the piano and I don't value playing scrabble. That means that, in my judgment, I bear a certain relation to the first activity that I don't bear to the second. Smith, let us suppose, values playing scrabble more than he values playing the piano; this means that, in his judgment, he

bears a certain relation to the one activity that he does not bear to the other. Those two judgments are perfectly compatible. So even though my values are antithetical to Smith's, there is nothing that we disagree about.

(4) Liking something is not the same thing as valuing it. People value activities (e.g., helping somebody in need) that they don't enjoy and enjoy activities (e.g., doing cocaine) that they don't value.

One tends to enjoy the activities that one values. I value playing the piano, and I also enjoy it; and these facts are obviously closely linked. But the relation I bear to that activity by virtue of enjoying it is distinct from the relation that I bear to it by virtue of valuing it. By virtue of the fact that I value playing the piano, I bear a cognitive relation to that activity; in other words, my valuing it consists in my having made a judgment about it (viz., I see it as having value and therefore---for reasons that we'll soon identify---as being agential). By virtue of the fact that I enjoy playing the piano, I bear a purely affective relation to it (viz., it makes me feel good).

3.3 What are values?

A value is a judgment on the part of some person (or sentient creature) to the effect that a certain course of action, or a certain outlook, will increase the scope of its *agency*. A moment ago I was reading a book (*The Scientific Image*, by Bas Van Fraassen). I then had a sudden urge to read a different book (*The Will to Power*, by Nietzsche). I was driven by a whim. Van Fraassen's book was illuminating; I was not wasting my time by reading it. And, although, for personal and professional reasons, I do at some point have to read Nietzsche's

book (which, judging from the fragments I've read, is extremely good), I would, in giving in to my whim to ditch Van Fraassen, be undermining a resolution of mine, and doing so for no good reason.

The result of my ditching Van Fraassen under these circumstances would have been a degradation of psychological integrity. I would have been allowing my decision-making process to give undue weight to sudden fancies. I would have been sending a message to myself that it's ok *not* to carry out an intention that I had formed in a rational and informed manner and that I knew that, if successfully carried out, would make me a better and more powerful person. This illustrates, and also supports, though only to a small degree, our contention that value is a judgment to the effect that a certain course of action or way of living increases the scope and power of one's agency.

I value carrying out rationally formed intentions. But why? Because I would be diminishing myself if I indulged every passing whim. So far as the fickle winds of fancy determined by life-path, *I* would not be calling the shots; *I* would not be the captain of the U.S.S. Mordecai Einstein. So far as impulses, rather than rationally formed intentions, determined my conduct, I would not be an *agent.*

Even though I may have *desired* to read *The Will to Power* more than I *desired* to read *The Scientific Image*, I *valued* reading the latter at that particular juncture more than I *valued* reading the former. My reason: under the circumstances, my reading *The Will to Power* would have constituted a surrender of agency to sub-agential whims.

3.3.1 An objection to what we just said

There seem to be counterexamples to this contention. For there seem to be cases where, driven by his values, somebody martyrs himself or otherwise

diminishes himself. There are cases where a person correctly characterizes an act as “valuable” even though that same act, if carried out, would diminish his agency and possibly destroy it. But that isn’t because our thesis is false. It’s because the word “value” is ambiguous. Sometimes it means “belief that a given way of acting (or being) enhances *one’s own* agency.” Sometimes it means “belief that a given way of acting (or being) enhances *agency in general*---not necessarily one’s own, but net amount of agency had by people overall.” This second usage of the term “value” is derivative of the first, even though one’s initial impression would probably be that it’s the other way around.

Most ethical systems use the word “value” (or its analogues) in the just-mentioned derivative manner. The one exception to this is *ethical egoism*, which was advocated by Nietzsche (and not---at least not to my limited knowledge----by very many other thinkers of substance). We’ll find at the end of this book that ethical egoism, though not ultimately a viable moral system, has some hidden virtues.

3.3.2 What are values? (revisited)

There are times when I *desire* to watch TV more than I desire to read *Non-Standard Analysis* by Abraham Robinson. But, in general, I *value* doing the latter more than I value doing the former. (Why “in general” as opposed to “always”? There are times when my reading *Non-Standard Analysis* would represent a compulsive, pseudo-agential act; and on many of those occasions, my watching TV would be a more “integrated” and rational course of action than would my reading some onerous treatise. In general, many is the time when what would ordinarily be characterized as the “lazy” course of action is the rational, non-lazy one, the reason being that what is in fact an agency-diminishing impulse cloaks itself as an intention to perform some action that

would ordinarily be seen as agency-promoting, as when one's intention to cut one's break short so as to go for a five mile run (even though one's bursitis makes it a bad idea to do so) or to continue reading some treatise (even though, owing to one's fatigue and consequent vulnerability to performance diminishing influences, one would ultimately be a better treatise-reader (and treatise–writer) were one, at this particular juncture, to take a break from treatise reading---we'll revisit this obscure parenthetical point). I *value* reading *Non-Standard Analysis* even though, at that particular juncture, I do not *desire* to do so, at least as much as I desire to watch TV, because I judge that reading that book would consolidate my agency and expand its jurisdiction, whereas watching TV would fail to have such consequences and would indeed lead to a withering of agency.

To be sure, once I turn off the TV and start reading *Non-Standard Analysis*, I not only experience gratification, but do so on a number of levels. First of all, I am proud of myself for doing the right thing. Second, and much more importantly, in reading *Non-Standard Analysis*, I experience what I will refer to as *agential enjoyment*, whereas in lazily watching TV I am experience non-agential enjoyment. (We tend to use the word "pleasure" to refer to non-agential enjoyment: one experiences "pleasure" when one enters a warm bath. But, although one *enjoys* reading difficult treatises or working on one's backhand, and although one *does* experience *a kind of pleasure* in doing so, that pleasure is derivative of a certain kind of enjoyment (for lack of a better word) that comes from exercising one's agency.)

3.3.4 What are values? (continued)

I *have* impulses. Impulses are *in* me. But they are not *of* me. *I* am not my impulses. There is a sense in which, so far as my behaviors are impulse driven, *I* am not acting. To put it another way, so far as I *am* acting in such

circumstances, my actions are but vehicles for sub-agential impulses. Consequently, my behavior in such circumstances is ultimately only pseudo-agential. In such cases, that is to say, my agency is involved, but only as a way of discharging impulses that are sub-agential and, in being acted on, diminish my agency.

Here it is important that we make a distinction between two kinds of impulses: those that one has *by virtue* of being an agent and those that one has before one's mind becomes agential. My impulse to read *The Will to Power* is obviously not in the same category as my impulse to eat tubs of ice-cream or to engage in certain carnal acts. The former impulse---though one that, in the previously described circumstances, ought to be suppressed---is obviously derivative of the development in me of values and aptitudes that jointly constitute agency. For that reason, were I, on a whim, to stop reading Van Fraassen's book and start reading *The Will to Power*, I would rightly not judge myself as harshly as I would if I stopped reading Van Fraassen's book and punched my annoying next door neighbor or ate a tub of ice-cream. Giving into impulses---be those impulses agency-derivative or agency-independent---is a bad thing. But giving into agency-independent impulses is worse.

"But by *never* giving into one's impulses," it must be asked, "doesn't one become a robot? Doesn't one become a brittle, neurotic square who can't live, love, or laugh?"

How that question is to be answered depends on what, in this context, it means to "give in" to an impulse. A healthy person ought indeed to discharge his base impulses and---what sometimes coincides with so acting---one ought to act spontaneously. But when one scrutinizes a non-maladaptive case of "giving into one's impulses" or of "acting on the fly," one finds that the decision to do so was a considered one; one finds that the thought-process behind the decision was to the effect that, since one damages oneself and deprives oneself of

strength by alienating oneself from one's base impulses---of which one's higher values are rarified derivatives---one weakens oneself: one turns oneself into an emotional cripple, a neurotic nerd, who, having drained his agency of a source of power, is less agential than one would otherwise be. So what *appear* to be cases of acting "on a whim" are cases in which one non-whimsically (though perhaps almost instantaneously) realizes that it is to the advantage of one's agency to act in a way that *seems* to be whimsical. *Genuine* cases of whimsical behavior are the prerogative of schizophrenics, whose minds are so disorganized that impulses don't have to go through the usual vetting process to be converted into action, and of psychopaths, whose vetting process operates much more intermittently than that of a normal person and is governed by different operating principles. (Though it may not be apparent, these points about psychopathy are in the process of being substantiated.)

3.4.0 The question "what are values?" considered in light of Mill's analysis of moral value

According to John Stuart Mill, happiness is the one thing that truly has value.[460] So far as anything other than happiness has value, Mill says, it's only to the extent that it leads to happiness. Right now we'll examine this thesis of his, along with his argument for it. In so doing, we'll find support for our contention that to value *x* is to judge *x* to increase the scope or the strength of one's agency.

There's no denying that writing novels of the same level of merit as *War and Peace* is more valuable than doing cocaine. Mill admitted this. Mill also admitted that, not having been duly educated or endowed with the requisite cognitive faculties, one might fail to know the delights of reading *Hamlet* and would prefer drinking vodka to reading Shakespeare. But, Mill insisted, so far

as reading *Hamlet* is more valuable than drinking vodka, it is *only* to the extent that the former brings more happiness than the latter.

One problem with what Mill says is that happiness *presupposes* a value-system. Happiness involves believing oneself to have measured up to certain benchmarks and thus to embody certain *values.* (Mice can probably feel good. But they can't be *happy*, they can't make judgments of the relevant kind.) A consequence is that happiness *presupposes* that there are things other than happiness that are of value.

Thus, Mill's contention that happiness is the one and only non-instrumental good is false. In fact, it's tautologously false, given that happiness consists in judging oneself to be good in some way or other. What is good for one person may not be good for another.

3.4.1 The question “what are values?” considered in light of Mill’s analysis of moral value (continued)

Also, Mill claims that, if a person fails to prefer Tolstoy to grain alcohol, it's because he's defective in some way. But in making that claim, Mill is saying that Tolstoy is valuable in a way that grain alcohol is not and that one must be deficient not to see that and, therefore, not to derive more pleasure from reading good literature than from drinking cheap liquor. So Mill is saying that *there are* values other than happiness and that our happiness is a *consequence*, and thus not the essence, of our receptiveness to these values.

Of course, Mill could claim that one can fail to derive any pleasure at all from Mozart's music, etc. and yet *not* suffer from any defect. But in addition to being totally *ad hoc*---not to mention hypocritical, at least when put forth by the likes of a cultured person such as Mill, who spent his days reading “good” literature and listening to “good” music---such a position is brazenly false. A

tone-deaf person who can't distinguish Mozart's music from the noise made by a dump-truck clearly is deficient in some way: there is something to which he is blind (or deaf, rather). Setting aside people who are rightly held in contempt for their musical deficiencies, nobody really believes that any given person is as good a composer as Mozart or, equivalently, that Mozart is as bad a composer as any given person. At least some of Mozart's works are better than at least some of those of at least some other people. This is a conservative claim. It's hard to see how one could sincerely deny it.

This is not to say that, given *any* two musical works, one is better than other. Two musical works may not be comparable. Each may be good in its own right and it would be artificial to say that one was better than the other. But there are aesthetic truths. I wrote a fugue once. It wasn't as good as any of the fugues that Bach wrote. The statement "it's *not* all a matter of how we feel about a given work of music: some musical works really have merits that others lack" doesn't entail "there is some *one* relation *R* such that musical work *x* bears *R* to *y* iff *y* is better than *x* and such that, given any two musical works, one bears *R* to the other." (Aesthetic properties are not unusual in this respect. Outside of the narrow horizons within which we operate, the question ‹is *x* hotter than *y*?› ceases to have a determinate meaning or therefore a single answer. But surely an object's thermal properties are not "subjective" or "unreal.")

Of course, these days, many *claim* to reject this sort of non-relativism, and to hold its advocates in contempt. But that can very easily be explained without supposing either that it's true or that anyone believes it. If it's all just a matter of taste, then—good news!---your sonatas (which even your mother describes as "execrable") are just as good as Mozart's. We're all geniuses! Break out the party hats!

If there is no aesthetic right or wrong---if a person's view as to the level of value of a given musical work is a projection of his emotional (and, so it is

assumed to follow, unfounded) response to it, instead of his emotional response to it being a consequence of his belief as to its degree of merit then, at least in the arena of musical composition, nobody need fear censure. And given how obviously false that view---or, at least, how thoroughly at variance it is with people's behavior and sentiments---and given, therefore, how much self-deception is involved in making oneself believe it, that view makes it possible to accept other positions that absolve one of one's deficits. No longer is one stupid; one just has a different "but equally valid" opinion.

Mill could rid his position of its vicious circularity by replacing the term "happiness" with the term "pleasure," in which case his thesis becomes "a thing is valuable to the extent, and *only* to the extent, that it promotes pleasure." An immediate consequence of this view is that the life of a tapeworm that experiences a large amount of pleasure is more valuable than that of an Einstein who does not experience quite as much pleasure. Mill is aware of this problem with his view. He deals with it by saying that listening to Mozart, reading Shakespeare, etc. brings one more pleasure than does shooting heroin. But while it can be said that reading *Timon of Athens* brings one more *happiness* than does shooting heroin, it cannot plausibly be said that it brings one more *pleasure.*

3.5. An objection to our position concerning the nature of values

People sometimes see things that diminish them---that abridge the scope of their agency and possibly even kill them---as being "valuable." In some cases, this is simply the result of error. (I value doing push-ups all day because I foolishly think that, as long as I do so, I'm in control of my life.[461]) But in many cases no error is involved. Smith *rightly* sees his self-diminishing activities as a UNICEF-volunteer as having "value." In fact, practically every system of ethics *demands* a reduction of oneself. [462]

This objection is not hard to neutralize. Let VS be a value-system according to which one's primary ethical obligation is to augment the scope of other people's agency. One has no moral obligation to promote what has no moral value. So, one has no moral obligation to promote welfare (agency) unless agency has moral value. *A fortiori* one's primary moral obligation cannot be to promote other people's welfare unless agency is the most valuable thing. But agency is, of its very nature, something that has only a *relative* value: *relative to me*, my agency may well be the most valuable thing. (In fact, I have no doubt that it is.) But *relative to you*, my agency has no value at all--except in an irrelevant, purely instrumental sense. (All references to "value" in this context are to intrinsic, not instrumental, value.) So VS is saying that, ethically, what is worthless to a given individual is the very thing that has the most worth. And I do not see how such a value system could be a coherent one.

This suggests that one of the following propositions holds:

> *(i) What is of moral value may not be of the slightest value to any given individual, in which case anyone who acts morally is ipso facto acting irrationally. (In fact, even in cases where what is moral happens to coincide with what is valuable to a given agent, it is not as somebody who is acting morally that he is acting rationally. His behavior, if rational, merely coincides with that of somebody who is genuinely moral. It does not itself embody any morality, being nothing more than a mere simulation of it.)*
>
> *(ii) What is of moral value is what it is in one's own interest to do.*

So a certain incoherence is internal to *public* value systems. But such systems *must* be incoherent in the just-described ways *if* they are to do their job.

Public value systems exist to ensure that, on balance, people's rights aren't horribly violated. The only value systems that can accomplish this are ones that, from a logical viewpoint, are incoherent. (There is nothing incoherent in the supposition that a value system must be incoherent if it is to have the causal properties that a public value system ought to have. This is because how much merit a value-system has is to be judged by the effects on people's lives through implementing it, not by its coherence. Value systems are not scientific doctrines. Thus, a welfare-maximizing viewpoint may be incoherent.) Setting aside people who are simply deluded (e.g., people who believe that joining cult *X*, a precondition for which is wire-transferring their life-savings to the Master's Swiss Bank account, is the royal road to self-actualization), when people say that performing self-diminishing acts (e.g., working at a soup-kitchen) has "value," they may be speaking the truth; for what they are saying is that performing such acts, though not *intrinsically* valuable, is valuable *in that* it consolidates value-systems that are instrumental in preserving and augmenting *agency in general.*

4.0 Minds ≠ Selves

In this section I will make some points about the difference between *minds* and *selves.* What I say in this section will be obscure, and I will provide little or no argumentation for it. The contents of this section will be both clarified and justified in the sections following it.

All selves are minds, but not all minds are selves. A self is a mind that is an *agent.* An agent is a mind that *acts*, and doesn't just *react.* In order for a mind to be capable of acting, instead of just reacting, that mind must do more than simply facilitate the gratification of primal urges; it must do more than provide the organism with information concerning the external world that makes it clear

how, without being killed or injured, that organism can eat, sleep, etc. A mind whose only objectives are to gratify such urges, and in which the sole purpose of knowledge is to enable such urges to be gratified, merely *reacts*, in light of the knowledge that it has, to the directives of instinct. Such a mind does not act. It does not itself do anything. It merely reacts to instinctual pressures.

These reactions are not blind. The exact nature of these reactions is conditioned by the knowledge stored in that mind. But non-blind reactions are not actions; and given an organism of the just-mentioned kind, that organism's *self* does nothing. It has no self, as we'll see; and its instincts, not its non-existent self, are the real agents. The behaviors in which the organism engages are compulsive responses to instinct. They are therefore not *acts*. Put another way, so far as those behaviors *are* acts, they are not acts on the organism's part, but rather on the part of its instincts.

The organism's mind doesn't act. Its only function is to determine, in light of its knowledge of the external world, what exact form the organism's instinct-directed behaviors must take if those behaviors are not to lead to the organism's being injured or killed. Such a mind's function is not to act, but only to ensure that the organism's *reactions* to its instincts don't run afoul of the forces at work in the external world.

4.1 Minds ≠ Selves (continued)

A *self* is a mind that has values; it is a self whose behaviors are driven beliefs as to what is of value. Value-driven acts often involve the gratification of instinctual urges; indeed, if value-driven acts systematically make it impossible for such urges to be gratified, the value-system underlying those acts is not a viable one. Nonetheless, a self is a mind that permits the organism housing it to behave in a given way only if, in its judgment, that behavior is in compliance

with some value. Of course, its judgments as to what is of value are, and should be, heavily influenced by its knowledge of what will permit the gratification of primal urges. But what is relevant is that, in an organism that has a self, as opposed to a mere mind, it is *judgments* as to what is of value that lead to behavior; and it is not, except in the indirect sense just described, urges *per se* that do so.

It follows that selves *act*. Of course, creatures that are selves sometimes merely react. But it isn't *as* a self a creature reacts. (I am a tennis player. But it isn't *as* a tennis player that I am writing this book.)

4.2 Minds ≠ Selves (continued)

A story will help us move forward. Smith is a sapient being. (It doesn't matter whether Smith is human or not.) Smith can make judgments as to how it must act in order to discharge its instinctual energies. Like all judgments, these judgments involve knowledge of principles. So there must be principles that Smith is aware of and that guide Smith's thinking. But given an arbitrary one of these principles, Smith doesn't know that he has knowledge of that principle.

One day, however, Smith becomes aware of at least some of the principles that guide his thought. (*How* Smith could come to have this second-order knowledge is discussed in Sections 5.0-6.2. Let "*P*" stand for an arbitrary one of the principles that Smith thereby comes to know himself to know. Let us assume that Smith has these realizations at time *t*. (Though they're experienced as mere feelings, embodied in these responses are rational judgments as to what to do, given what Smith's objectives are (instinct-gratification) and given what he knows about the circumstances. We must distinguish the instinctual response *per se* from the conscious experiences

that merely represent it, much as the icons on your desktop must be distinguished from the patterns of electrical activity to which they correspond.) Before he knew that he had knowledge of *P*, Smith obviously couldn't make judgments of the form: my judgment that such and such is consistent with my knowledge of *P*. Prior to time *t*, Smith's judgments were indeed guided by a knowledge of *P*; but he obviously could not, as of yet, make judgments as to whether his thinking was being *properly* guided by his knowledge of *P*. (One's thinking can be *mis*directed by one's knowledge of a legitimate principle. This may happen either because one's auxiliary assumptions are wrong (e.g., knowing that gross disparities in wealth lead to injustice and are themselves unjust, but not knowing that certain ways of eliminating such disparities may lead to even greater injustices, President Jones destroys the economy in his attempt to redistribute wealth) *or* because one's ability to deploy one's knowledge of the principle in question has been blunted by garden-variety performance-inhibiting factors, such as fatigue and lack of memories.) But, as of time *t*, Smith can form judgments as to whether the manner in which he arrives at judgments is consistent with *P*. He is thus aware of a *norm* from which his thinking mustn't deviate; and, when such deviations occur, he can make judgments as to their nature and extent. He can thus evaluate his own judgments. And his evaluation of them is positive (negative) to the extent that, to his knowledge, they (fail to) comply with *P*.

Integral to the argument just put forth is a delicate but important distinction. There is a big difference between being aware of a principle and being aware that one is aware of it. A talented athlete's behavior embodies awareness of principles that the athlete doesn't know himself to be aware of. The point being made here is not that his behavior *accords* with some principle of which, at every cognitive level, he is ignorant. It is that, although a knowledge of a given principle may be what is guiding that behavior, the

athlete in question may not be aware of that fact and, indeed, is exceedingly unlikely to be.

Philosophers and psychologists often regard an act as being entirely "mechanical" if it isn't guided by knowledge that the agent can articulate. But it is only when one has second-order knowledge----knowledge as to what it is that one knows---that one can articulate what one knows; and philosophers are therefore likely to regard behaviors that fail to embody second-order knowledge as failing to embody *any* knowledge and, therefore, as being entirely unprincipled. In fact, one needs a certain amount of verbal dexterity *in addition* to such second-order knowledge if one is to be able to articulate the grounds on which they make practically any one of the decisions that they make.

Of course, in saying "snow is white," I am not affirming my knowledge that I *know* that snow is white; I am talking about snow, not about my own mind. But, in producing that utterance, I am *expressing* the fact that I know that snow is white. And one could very easily know that snow is white without knowing that one knows it. One can know that snow is white without having the concept of knowledge. But one must have that concept to know that one knows that snow is white. So knowledge of *P* is different from, and needn't involve, knowledge that one has knowledge of *P*. Also, by virtue of having a concept of knowledge, one has various concepts---e.g., belief, truth, mental activity---not had by an otherwise comparable creature that doesn't have the concept of knowledge.

4.3 Minds ≠ Selves (continued)

Now endowed with knowledge, not just of *P*, but of the fact that he has knowledge of *P*, Smith doesn't react to situations; he reacts to his reactions to

situations. Given a situation, he instinctively makes a judgment. Aware that his judgment must comply with *P*, he makes a *second* judgment. This second judgment doesn't concern the situation in question, at least not directly; it concerns his initial judgment concerning that situation. It concerns the extent to which his initial judgment is consistent with the principles his knowledge of which eventuated in it.

Assuming there to be any cognitive dimension to it at all, one's response to any given situation is *ipso facto* principle-driven. But that, by itself, isn't enough for that response to be an action, as opposed to a reaction. Actions are behaviors that result from judgments about judgments. Confronted with a situation, one makes an initial judgment. One may then make a judgment as to whether that initial judgment was consistent with the principles guiding it. In other words, having made an initial judgment, one may *evaluate* that initial judgment. One is acting, as opposed to reacting, if one's behavior expresses such an evaluation. One is reacting, as opposed to acting, if one's behavior expresses one's initial judgment.

There is no self where there is no ability to act---where there is no agency. There is no agency, where there is no self. There is no self where there is no self-evaluation.

To put this last point in a way that, in addition to being more precise, doesn't make it seem viciously circular: there is no self where there are no judgments as to the integrity of other, co-corporeal judgments.

Where there are such judgments, there are values. To approve of (criticize) a judgment (or anything else) is to judge it to have (lack) some value that it ought to have.

What *directly* leads to a *bona fide*action is never a desire and is always an attitude of *approval* towards a desire. Where there are no values, there is no approval or disapproval. So where there are no values, there is no agency.

5.0 Co-corporeal, informationally integrated mental states enough for a mind, but not for a self

Anything that has the cognitive wherewithal to have a series of mental states *ipso facto* has a mind. But there is more to having a self than there is to being capable of having a series of mental states. Consider Smith as he is *before* time *t*. Each of his mental states could be understood on its own terms: as a reaction to immediate stimuli. Each of those reactions may have been intelligent. (In other words, each resulted from cogent reasoning as to how to discharge the relevant instincts, given the available information about the environment.) And, supposing that they were intelligent, each was more likely than not to incorporate the information encoded in previously occurring, co-corporeal mental states. So, if we think of Smith's pre-*t* mind as a series of links in a chain of events (or conditions), the links in that chain are bound together not merely by virtue of being co-corporeal and not merely, therefore, by virtue of the earlier being partly causally responsible for the existence of the later ones, but by virtue of the fact that (in at least some cases) the former ones bequeathed the information encoded to them to the later ones and, therefore, that it was *as* mental entities that the earlier ones (in some cases) had a hand in determining the contents of the later ones.

5.1 Co-corporeal, informationally integrated mental states enough for a mind, but not for a self (continued)

Still, co-corporeality, even when coupled with content-transmission, isn't enough for self-hood: supposing that $M_1 \ldots M_n$ is a series of co-corporeal mental states such that, for each i and j, $(1 \leq i < j \leq n)$, for $M_1 \ldots M_n$ to form a

self, as opposed to a mere mind, it is not enough that M_i be corporeal with M_j; nor it is enough that M_i help bring about the occurrence of M_j; nor, most importantly, is it enough that M_i do this in the distinctively psychological manner just described.

A story will help make this clear. There is a group of around 20,000 people, all of whom live on the same land-mass, which we'll call "*LM*." (*LM* is big enough that nobody feels crowded but small enough that nobody feels lonely and, more importantly, interactions among its occupants are frequent.) *LM*-occupants all speak the same language, and they often exchange information with one another. They also often exchange goods with, and provide services for, their fellow *LM*-occupants.

Being reasonably intelligent and well-informed people, these people tend to have accurate beliefs; and being reasonably honest people, they tend to provide one another with accurate information. Also, most of the time, they're reasonably honest not just in their communications with their fellows, but in their transactions with them. They tend not to defraud one another.

In at least some cases, the ties that *LM* occupants have to one another are deep ones. (For the moment, we'll set aside cases where *LM*-occupants have no significant ties to their brethren.) Those ties are not purely geographical: it isn't just that the parties involved live in the same place. Nor are those ties purely causal. (It isn't just that *LM*-occupants accidentally bump into other *LM*-occupants, thereby inadvertently changing the latter's lives.) The ties in question involve:

> *(i) the fact that one person intentionally transmits information that is arrived in a principled manner to some other person who makes principled inferences from that information,*

and:

> *(ii) the fact that two people may (for prudential, moral, or even libidinal reasons) benefit each other by honoring an agreement to exchange goods or services.*

One fact about *LM*'s occupants must be made explicit. There is nothing *regulating* their exchanges with one another. Should one *LM*-occupant defraud another, there would be no penalties. Given that there are no laws in *LM*, the victim would have no legal recourse and he'd have to take it upon himself to see that justice be done. An *LM*-occupant who'd been misinformed by another would be in the same situation *mutatis mutandis*. It may be that, because *LM*'s occupants are so honest with one another, they're better off for the fact that their lives aren't weighed down by the presence of the regulatory entities that would be needed for there to be any formal penalties for fraud.

But what is relevant here is that, precisely because their lives are not thus "weighed down," the occupants of *LM* do not jointly constitute a *nation*, at least not in the sense in which France constitutes a nation. There is nothing under whose jurisdiction the occupants of *LM* fall.[463]

5.2 Minds that are selves and minds that are not

Bearing these points in mind, let us turn our attention back to $M_1 \ldots M_n$, and let us make some further assumptions about that series. The information encoded in M_i $(1 \leq i \leq n)$ is likely to be accurate. That information, if transmitted to M_j, $(1 \leq i < j \leq n)$, is likely to be transmitted in an accurate

form. Moreover, the synthesizing of that information with other bodies of information (e.g., information currently being uploaded through the senses as to the nature of environment in Smith's vicinity) is likely to consist of inferences in which correctly understood correct rules of inferences are correctly applied.

But----and this is the final, and most important, assumption that we'll make about $M_1 \ldots M_n$-----*none of the members of that series constitute judgments concerning the merits of either the information-exchanges occurring among other such members or of the manner in which the beliefs encoded therein are generated*. Those exchanges and syntheses of information are not policed. It may be that each of the exchanges and inferences in question is done impeccably. But, if so, that isn't because there is anything overseeing those operations; it is simply because, despite the absence of any supervisory agency, the mental states involved in those operations happen to satisfy the relevant desiderata. By the same token, if the ratiocinative activity that generates $M_1 \ldots M_n$ is inconsistent, even systematically so, with the very principles embodied in that activity, no *internal* mechanisms exist to ensure that the needed course-corrections occur. The spuriousness of the reasoning involved in $M_1 \ldots M_n$ may well ensure that Smith fails in his attempts to gratify his desires.

But there isn't any internal agency that *prevents* Smith from making such mistakes or, therefore, that protects him from the damage to his person that such mistakes are likely to entail. There is, in other words, nothing under whose jurisdiction the various members of $M_1 \ldots M_n$ fall. Nothing polices either the manner in which members of that series exchange or produce information or the manner in which that series generates new members. Thus, $M_1 \ldots M_n$, though constitutive of a *mind* and perhaps of something that has an

enduring identity in the sense in which a rock does, doesn't represent anything that has an enduring identity in the distinctive sense in which a human being does.

6.0 When minds become selves

Let's revisit, and duly extend, what we said earlier about Smith/$M_1 \ldots M_n$. At time *t*, it will be recalled, that series generates a mental state that embodies a judgment as to the fitness of (some of) the activities by which that series has thus far arrived at judgments. An internal regulatory agency is thereby born.

That agency will inhibit the conversion of impulses into actions. This does not mean that it will simply thwart those impulses. It means that if, once that agency scrutinizes a given impulse, that agency judges it to be one worth acting upon, then it will lift the inhibition. But, for the reasons given in Section 3.3., the aforementioned impulse isn't the proximal cause of the body-movements that result from the lifting of such an inhibition.

Even before $M_1 \ldots M_n$ became to any degree self-policing, the activity constituting $M_1 \ldots M_n$ embodied knowledge, not of some one principle, but of many. So for that activity to become *systematically* self-policing, it must be guided by a knowledge of many different second-order principles of the sort previously described. But it's a distinct possibility that not all of those second-order principles will be consistent with one another. As a result, an analogue of the process that led to Smith's acquiring knowledge of those second-order principles will lead him to acquire knowledge of third order principles, the purpose of which knowledge being to ensure that the totality of second–order principles guiding Smith's mentation be a coherent one---that those principles not be such that, when complied with, the result is that it

is judged *de rigueur* to lift the inhibition on some impulse and also *de rigueur* not to lift that inhibition. Second-order principles that don't fit in with the rest are weeded out; they are no longer countenanced. For the same reason, second-order principles *are* countenanced that were not previously so. So far as this enterprise is successful, the result is an integrated body of principles, any two of which are consistent with each other, under whose jurisdiction falls all of the activity constitutive of Smith's decisions as to how to act. So, given the reasonable presumption that, at this stage, none of Smith's mental activity is purely theoretical, all of it being oriented towards instinct-gratification, this means that this enterprise, so far as it's successful, yields a single, cohesive regulatory agency under whose jurisdiction *all* of Smith's mental activity falls.

6.1 Summary

Just as group of people who live in the same place don't become a state until their activities come to fall under the jurisdiction of some regulatory agency, so a group of co-corporeal, informationally integrated thoughts don't necessarily constitute a self until *they* all fall under the jurisdiction of a regulatory agency. Before that happens, there are no actions, only reactions; after it happens, there are actions. We finally have an *agent* and, therefore, a *self*.

7.0 The concept of intrinsic value in relation to the concept of selfhood

With the inception of selfhood, values replace desires. Under the growing power of a supervisory structure of the just-discussed kind, the various constituents of that mind become increasingly unified: the structure they form

becomes more and more of a self. (Incidentally, a consequence of this last statement, and therefore of our analysis, is that there are degrees of self-hood. This redounds to the credit of our analysis, since selfhood *does* come in degrees. Though clearly on independent grounds, a justification for it is given later.) Certainly, the *initial* purpose of this supervisory superstructure was to maximize the amount of instinctual gratification enjoyed by the mind falling within its power. And, surely, that always remains *one* of its purposes. Indeed, it arguably always remains its *primary* purpose (though, for reasons I'm about to state, I personally don't think this).

But that superstructure has needs of its own; and, as that superstructure develops, it becomes increasingly difficult to find an instinctual basis for those needs and increasingly necessary to understand them on their own terms.

7.1 Some preliminaries

Before we can see what exactly this claim amounts to, or, therefore, why it's true, we must make preliminary points of a general nature as to the nature of the supervisory agency in which we have claimed the essence of selfhood to lie. There are some impulses that, if acted on, would weaken the integrity of this supervisory agency. Tautologously, this agency, so far as it's in good working order, will inhibit such impulses. But little harm will come to that agency if it's *too* good at its job---if it inhibits impulses that it could, without in any way jeopardizing its own authority, allow to be expressed in action. If that agency is too good at its job---that is, if it's too conservative, if it inhibits impulses that, erring on the side of caution, it needn't inhibit—the result will be merely a forfeiture of pleasure; it will not, at least not in the short term (or even the medium term), do any damage to itself or to the structures

responsible for the mentation falling in its jurisdiction.

As a result, this supervisory agency comes to be overly restrictive: it disapproves of the disinhibition of impulses that ought to be disinhibited. So far as those inhibited impulses are permitted gratification, it is only on the condition that they assume such forms that, if acted on, consolidate this agency's power over the course of mental activity falling into its jurisdiction. In some cases, it is possible for those impulses to assume such forms. (This is what Freud referred to as "sublimation." My aggressive instincts are discharged by my writing treatises that aggressively critique the views of my colleagues.) In other cases, this is not possible. This leads to what Freud calls "neurosis": an inhibition of impulses that cannot be discharged in ways that are in keeping with one's values but that, because they are so strong, must come out in some way or other and are forced to be expressed as *symptoms*. Since symptoms are unintentional, neurosis involves an abridgment of agency. One would have been better off yielding to the impulse and discharging it agentially, instead of forcing to find non-agential ways of discharging itself. (Freud makes this point in his (1926) book *Symptoms, Inhibitions, and Anxiety.* As he puts it: somebody riding a horse must sometimes let the horse go where it wants to go, lest the horse cease to let the rider have any control over it.)

7.2 Two Illustrations of Our Thesis that what is Valuable is what consolidates one's Agency

Illustration #1: We see Bach's music as being more valuable than Eminem's. (Eminem is a pop-star.) Why? First of all, the enjoyment of Eminem's music is easily traced back, by way of multiple pathways, to enjoyment of the sexual act, it being largely for that reason that we enjoy

Eminem's music. By contrast, it is not because of any obvious connection that Bach's music has to the sexual act that we enjoy it. The pleasure Bach's music gives us is more akin to the pleasure given us by a philosophical or mathematical insight than it is to the (sub-agential) pleasures given to us by the sexual act. Implicated in the experience of listening to Bach's music, and largely constitutive of the pleasure it brings us, are ratiocinative abilities that have no obvious connection to sexual gratification, or to any other form of instinctual gratification, and that *do* have obvious connections to the rational operations on which the supervisory agency in question depends to consolidate and extend the hegemony it has over the course of mental activity. Bach's music is studied in conservatories; it is performed in places (e.g., concert-halls) where the operative norms of conduct require tremendous self-restraint---where one has to sit still for hours on end, where one must speak to others in a very controlled manner, and where one must display all of the asceticism involved in the transformation of instinct-driven proto-humans into instinct-suppressing actual humans.

When people say that Eminem's music is as highly "valued" as Bach's, so far as what they are saying is true, they are saying in a distorted way that Eminem's music is *liked* as much as Bach's. And there's no doubt that (at this juncture in history) there are more Eminem fans than there are Bach fans (that said, pop-stars are seldom liked *solely* for the merits of their music); and it may be that, during its brief recess from oblivion, Eminem's music is a source of more pleasure (not pleasure *per capita*, but aggregate pleasure) than Bach's music. But, in saying that Eminem's music is more highly *valued* than Bach's, one is confusing the concept of pleasure, and the related concept of desire, with the very different concept of value.

Illustration #2: People enjoy reading (and writing) philosophical treatises. To become capable of this sort of enjoyment, one must alter one's

psychological structure in non-trivial ways. In order to pull off the remodeling of one's intellect needed to become at all adept at this discipline, or therefore to derive any enjoyment from it, one must suspend gratification of urges that one would otherwise gratify, and one must do so without having any firm assurance that, in due course, one will be able to enjoy gratifications as intense as those that one has foregone. Because it takes a long time to acquire the philosophical acumen needed to make the reading, let alone the writing, of philosophical works more pleasurable than painful, and because during that long period otherwise fecund sources of instinctual gratification are off-limits, one must, if one has chosen to acquire that degree of acumen, develop a certain asceticism. One must come to de*value* (some forms of) instinctual gratification and to hyper-value the self-control that comes with becoming capable of suspending instinctual gratification.

This is not to say that the enjoyment of philosophical discourse has *no* instinctual foundation. Like Nietzsche, I believe that the enjoyment one derives from the exercise of one's intellect is a form, or at least a derivative, of the instinctual pleasure one receives from having power over others and acting aggressively towards them. But, supposing this correct, the way in which, by virtue of engaging in philosophical discourse, one discharges these primitive impulses is contingent on one's exercising faculties that, being purely ratiocinative, are much less directly connected with instinct-gratification than they are with the inhibitions of instinct so essential to the existence, and to the consolidation, of the supervisory agency in which the essence of selfhood lies.

The same point *mutatis mutandis* holds of the pleasure one derives from very sophisticated forms of music. Bach's music is passionate, and the passion embodied in it arguably has *a* sexual basis. But Bach's music expresses that passion in an extremely controlled and ratiocination-heavy

manner; and the faculty underlying the ratiocinative component of Bach's music is one and the same with the faculty underlying the instinct-inhibitory ratiocinations constitutive of the regulatory agency that is one's self.

7.3 Selfhood a matter of degree

Some minds are more intact than others. In this respect, minds are like nations. The official government of Pakistan has uncontested control over some areas and only little or no control over others. Thus, the landmass coincident with the nation of Pakistan is *less* unified than it would be if the loyalties of its occupants weren't fragmented by the para-governmental agencies that occupy it. By contrast, the landmass coincident with the continental United States *can* be said to fall under the more or less uncontested jurisdiction of some one government, the same being true of the discontinuous landmass that is formed when Hawaii and Alaska are added. Therefore, the landmass coincident with the U.S. is *more* of a nation than the landmass coincident with Pakistan. Thus, nationhood comes in degrees. *x* is more of a nation than *y* if *x*'s government does a better job than *y*'s government of unifying the occupants of the landmass falling in its jurisdiction.

It is for similar reasons that one mind can be more of a self than another. If the regulatory agency governing mind *x* has more scope and power than its counterpart in mind *y*, then *x* is more of a self than *y*.

That our analysis has the consequences that self-hood isn't equally distributed among minds is to its credit. People's selves disintegrate. And there are degrees of integratedness (i.e., of non-*dis*integratedness), as everyday experience makes obvious and as clinical observation of the mentally ill makes even more obvious.