Everyone Desires the (Real, Not Apparent) Good:
An Argument on Behalf of Socrates

Imagine I am blocking the door to a room, and a man comes along and says, "I want to go in there." I reply "No, you don't." He pushes me aside, goes in, and eventually emerges bleeding profusely. "You were right," he says, "I really *didn't* want to go in there." Notice what happened here: I told the man what he wanted, and he came to see that I was right, and he was wrong, about what he wanted.

Consider another version of the encounter: when the man approaches I say "There's a lion in there." He says, "I don't believe you." When he emerges, he concedes, "you were right, there really was a lion in there." He wouldn't say, "you were right, I really *did* believe there was a lion in there." Nor did I ever assert this. Why not—why didn't I try to tell him what he believed, in the second story, as I tried to tell him what he desired, in the first one? Is it that we never correct someone else's belief report—that we take what someone says to be the last word on what he believes? No, that's not right. There are occasions on which it is appropriate to say to someone, you don't really believe the thing you say you believe.

For example, I might say that to someone because I think he's lying to me—we can call into question the *sincerity* of a belief-report[1]. We can also call into question the sincerity of someone's desire report: we can say, you are lying about your desires. But my first story seems to suggest another possible way of correcting a desire report, that is, solely on the grounds that the object of desire is *bad*—whereas it does not seem possible to correct a belief report *solely* on the grounds that the belief in question is *false*.

A skeptic about this difference between belief and desire might point out that it's not only when we doubt someone's sincerity that we utter words like "you don't really believe that". There's what we might call the coercive "you don't really believe that," which is meant to pressure you not to believe that—we might analyze it as "*don't* believe it." Closely related, there's the incredulous "you don't really believe that," where I am reporting my amazement at the fact that your belief seems to fly in the face of the obvious—we might translate it as "how *can* you believe p?" And there's the rationalistic "you don't really believe that," where I am saying that p contradicts some other things that you think, that it isn't rational for you to believe p. We can in turn express this by saying, if you were to reflect more carefully, you would be brought to reject p.

If we could interpret my original "you don't want to go in there" as coercive, incredulous, or rationalistic we would dispel the sense of mystery surrounding this purportedly special form of desire-correction—because we would be translating the "you don't desire p" into, respectively, "don't desire p!" "do you desire p?" or "you wouldn't desire p if you were being fully rational." The coercive and the incredulous translations are non-starters: I needn't be

[1] It might seem as though one who grants the existence of unconscious beliefs and desires has another avenue for belief- or desire- report correction. Interestingly enough, that's not the case. If someone says he desires or believes p, to say that he has an unconscious desire or belief that not p is not to contradict the original claim, but simply to add to it, e.g., he believes that p *and* he unconsciously believes that not p. This is connected to Freud's thought that there is no negation in the unconscious, so that unconscious ideas are "exempt from mutual contradiction." (see, e.g., part V of "The Unconscious") Unconscious mental states can serve to correct *negative* belief- or desire- reports ("I don't desire/believe p"), but not positive ones.

trying to get the man not to want to enter; and, I might think there's no way for him to know there's a lion in there, and so have no grounds for incredulousness at his desire to enter. The same problem threatens the rationalistic reading:the man isn't going to be able to deduce the existence of the lion a priori (in the way that someone can figure out that heis beliefs are inconsistent with one another), so it's not irrational for him to want to enter, given his other mental states.

Somehow, it seems that when I claim that he doesn't want to go in there, my claim is resting in an important way on my access to some information that I know he doesn't have, the fact that there is a lion in there. Perhaps what I mean is simply that if he came to have a certain belief, he would (as a matter of fact) no longer have the desire to go in there? The idea here is that there is some belief such that if S came to have it, that would cause him to cease to desire X. But we don't typically ascribe to people every attitude they might come to have—I may know, for instance, that given your current state of post hypnotic suggestion, if you come to believe that I am 34 years old, you will feel an overwhelming thirst for coconut milk. But it would be very odd to insist, *before* telling you that I am 34 years old, that there is any sense at all in which you want coconut milk. You don't, though perhaps you *will*. When I tell the man that he doesn't want to go in there, I seem to be saying something stronger than that he might or could have come to lack this desire—I seem to be saying, at the very least, that it would be appropriate or fitting for him to lack it.

Still, there may be a way of combining the reliance on special information with the rationalistic reading: perhaps what I meant was neither that coming to believe that there is a lion in there would brutely cause the elimination of his desire, nor that he could eliminate the desire by pure reflection, without any additional information, but that he'd be rationally required to give up the desire if he reflected on what he knows and wants *and* on the fact that there is a lion in there. The thought is, if he were well-informed *then* he'd be rationally constrained to not want to go in there. We could call this the informed rationalism reading.

On this reading, when I say that "S doesn't really want X," I am saying that

> S wants X, but there is some important information such that, were S to learn it, he would be rationally required to give up his desire for X.

Let us call such a desire (namely, a desire that you'd be rationally required to give up if you were better informed), an irrational desire. On the informed rationalism reading, when I tell the man that he doesn't want to go in the door, I'm telling him that he has an irrational desire. I am going to argue that I can't mean this, because irrational desires are inconceivable—and so the informed rationalism reading must be wrong. The coercive, incredulous, rationalistic, and informed rationalism readings are all ways of denying that I mean exactly what I say when I say that "you don't want to go in there." If none of them work, and given that we are not questioning the man's sincerity, we may have to open ourselves up to the possibility that I did mean what I said; this is the beginning of the defense of the claim that everyone desires not what seems good to him, but what really is good, whether he knows it or not.

******

But I will mostly be occupied, in this talk, with the smaller project of the case against irrational desires. An irrational desire is, as we said, a desire which the agent would abandon, were he fully rational and newly apprised of some fact. An agent described as irrationally desiring is an agent described as susceptible to a certain kind of change, namely, the rational change from having the desire to not having it, as spurred by the recognition of the fact in question. If there can be irrational desires, we must be able to tell a story about this sort of change, or what I'll call desire-correction. Desire corrections are a proper subset of all desire-changes, since presumably sometimes desire-change is either irrational or arational. (I sometimes stop wanting something, for no reason at all: desire fades.)

What I want to point out is that a paradox arises when we try to tell the story of a desire correction. In order for a desire to be corrected, someone must go from desiring X to not desiring X, and do so as a result of a realization that p. The question is: at the moment at which he comes to realize p, does still he desire X? If we say yes, then we have to allow that for one moment he was irrational (that is, for the moment that he knew p that but still wanted X)—but this was supposed to be a story of *rational* desire-change. If we say no, and locate the realization in the period *after* the change then we are, in effect, denying that the realization is what causes the change—for we can't say that the desire is changed by the realization if the desire already has been changed by the time the realization rolls around.

I'll illustrate the paradox with an adaption of a famous example of Bernard Williams': if I want to drink what's in this glass, believing it to be gin, do I stop wanting to drink what's in the glass before or after the moment I learn that it's gasoline? If after, then I was, for at least a moment irrational: wanting to drink what was in the glass because it was gin, despite believing the glass to be filled with gasoline; if before, then why did I even need to learn it, since I'd already lost the desire by the time I learned it.

This paradox about desire-correction is a version of what we might call 'the generic paradox of change'[2] which goes like this: X undergoes a change, losing one of its properties; and this change is due to another thing, Y. When does Y change X? There are three possible answers:

(I.) only while X is A
(II.) *both* while X is A and while X is not A
(III.) only while X is not A

I think these three answers correspond to the three basic *kinds* of change: causal, constitutive and rational. So I will first discuss the causal, constitutive, and rational versions of the change paradox. I will then show that we cannot give any of these three answers in the case of desire-correction. My conclusion will be that we must adopt a fourth option, that in the case of desire, Y changes X neither before, nor during, nor after the change, because there is no change: X was never A.

I. "While X is A": The causal-change paradox

[2] Of course what I call "the generic paradox" itself is only one of a family of paradoxes about change (Heraclitus' paradox, Zeno's paradox, the problem of the substrate, etc.)

Take the case of a window that breaks because a ball hits it. The window changes from whole to broken (that is, not whole) due to the impact of the ball. The causal-change paradox arises when we ask, does the impact occur when the window is whole, or when it is broken? If the impact occurs when the window is already broken, it cannot be what makes the window break. But if the impact occurs when the window is whole, then it looks as if the impact is in some way *compatible* with the wholeness of the window. Does this mean it cannot be the cause of the window's breaking? No—this, in effect, was Hume's insight. We say that the impact was the cause of the window's breaking because we observe a regular association of ball-impacts and window-breakings. This does not mean that the very idea of the impact somehow contradicts the idea of whole, intact glass—there is no internal or logical connection between the cause and its effect. The thought that there is no *necessity* for the effect to follow from its cause is also the thought that there is no *incompatibility* between a cause and the absence (at the time of its occurrence) of its effect. If our story of causation is the story that things of one kind are regularly followed by things of another kind, then there is no problem with saying that the cause does its causing *before* the advent of the effect, that is, while X is still A.

II. "While X is A and While X is not A": the constitutive-change paradox

Consider this parallel to the causal change paradox: the person who gets married goes from being a bachelor to being a husband. But when does his 'getting married' occur? Does it occur when he's a bachelor? But then, at some point, he's a married bachelor! Does it occur when he's a husband? But you can't get married when you're already married, at least in most states. We might try to solve along the lines of the causal change paradox, and say that he gets married when he's a bachelor, and the moment after he's done getting married is the first moment of his being a husband. But if we say this, then it looks as though someone could get married, that is, could *complete* the process of getting married, and yet never be a husband—because, say, the world ends in such a way that the last moment of his getting married is the last moment in which the world exists. I think this violates our linguistic conventions about the phrase "getting married"—that is, if you are a man, and if you got married, then you were (for some time) married, that is, a husband. The answer here is to see that the Y which is the agent of the change from bachelor to husband, namely getting married, is a very different kind of Y from the Y which was the agent of the change in the glass, namely the ball's impact. "Getting married" is (unlike the impact of a ball on glass) an 'infallible' changer, because it is a process that itself includes reference to its change: some process only counts as getting married if the person ended up married at the end of it. We might say Y constitutively brings about X's being A if Y cannot be specified without reference to its bringing about of A. So a rabbi's saying the words "I now pronounce...," the exchange of the rings, the breaking of the glass, etc. only counts as belonging to a process of getting married if, after these things occur, the pair are in fact married. A guest on the way home from the wedding who asks, "at what moment did they get married?" is making a category mistake if he insists on locating a *single* moment. Getting married is something that is located on both sides of the "bachelor" and "husband" divide: it will, with respect to that distinction, never have the unity of a *single* moment.

Note: I am not denying that there is some first moment of being married, and hence, immediately prior to it, a last moment of being a bachelor. Nor am I denying that there are *causal* relations (of the type discussed above in I) between the two moments, relations where the cause on the bachelor side can be understood fully independently of the effect on the husband side. What I am denying is that anything which can be described as "getting married" can be understood as such a type-I change, that is, without reference to both sides of the divide.

III. "While X is not A ": The belief-change paradox

Consider a third variant: S believes that $\neg q$ and comes to believe that q by realizing that $\neg q$ is false. *When* does his realization that $\neg q$ is false occur? Does he come to realize that $\neg q$ is false while still believing that $\neg q$? Then for a little while he believes a contradiction—but that's impossible! Does he come to realize that $\neg q$ is false when already believing that $\neg q$? Then the realization cannot be the explanation of the rational change in belief.

The belief-change case cannot be understood in the mode of the causal change case, because one's realization that $\neg q$ is false is incompatible with continued belief in $\neg q$. But it also cannot be understood along the lines of the constitutive change case, because the realization, unlike "getting married," must have the unity of a point. That is, it cannot consist of two distinct "points" occurring in temporal sequence: a moment of believing that $\neg q$ is false, followed by a moment of believing that q. The constitutive analysis of belief-change would require us to say "a belief that $\neg q$ is false, in order for it to be a belief that $\neg q$ is false, must be *followed* by a belief that q." What wrong with this is that the belief that $\neg q$ is false, in order for it to be a belief that $\neg q$ is false, must already *be* the belief that q. And this is going to hold even if we make the transition from holding to giving up a belief much more complicated. Say I believe that $\neg q$, and I come to believe that p, and that $p \rightarrow r$, and that $r \rightarrow s$ and that $s \rightarrow q$, and all this is what gets me to change my mind about q. I might come to believe these things severally and not change my mind about q—for I might not see that all these new beliefs require me to give up $\neg q$. So my realization must be more than an acquisition of these new beliefs, and it must be more than an acquisition of these beliefs followed by the belief that q. My realization must be a synthesis: I must see that p and $p \rightarrow r$ and $r \rightarrow s$ and $s \rightarrow q$, taken together[3], imply q. But as soon as I count as having taken them together, I must see that q—for seeing that q is just what it is to take them together. In a description of the realization like this one: "p and $p \rightarrow r$ and $r \rightarrow s$ and $s \rightarrow q$ and therefore q," the "ands" and "therefores" don't represent a temporal sequence, they represent a sequence of reasons within a single thought.

So where is this single thought, this unified realization, located?—or rather, *when* is it located? We can see what the answer must be if we consider the fact that there is no way to talk about the realization as cause except to talk about the realization as effect—that is, the cause just is the effect. This is much stronger than what we said about the constitutive case, for there we said only that we must *make reference* to the effect in characterizing the cause. The solution I am proposing is that the realization must be squarely on the right hand side of the change: it is the first thought of the period in which he believes q rather than the last thought of the period in which he believes $\neg q$.

[3] And of course the 'realization' may piggyback on having done some of the unification already: perhaps I realize that $p \rightarrow s$ at t_1 and at t_2 make use of this to conclude that $p \rightarrow q$, and only at t_3 bring this together with p to conclude q. In this case the realization in question would be the one at t_3, which would be predicated on the earlier realizations.

I suggest that here we should not be chagrined to accept the second horn of the dilemma, that the realization does not explain the change: what's rational about rational belief change is the new belief (one has good reasons for it, better reasons than for holding the opposite belief) rather than the change *per se*. If we are asked "why did you change your mind about q" we will give our reasons, reasons which still hold even now—the answer to that question does not make reference to *an event* occurring just before we started believing what we now believe. Of course there may have been a precipitating event, say, a neuron firing in a brain, which occurred at the very last moment of believing ¬q—and perhaps that event was linked to the one following it, the first moment of believing q, by inexorable causal laws—nonetheless, the rational story of belief change doesn't ask us to make reference to this neuronal firing. This event figures crucially in a *causal* story of change from one brain-state to another, and I have nothing to say about the relation between this story and the story of rational belief-change. What I want to insist is that we be very clear on which story we are trying to tell, because they are stories of different kinds of change.

IV. The desire-change paradox

Can we use any of I-III to model desire correction? Let's try.

I. The causal case won't work because it won't allow us to pick out specifically *rational* desire changes. We are trying to explain cases where desire change is not just a matter of moving from one state to another, but of having one's desires *corrected*. If there is such a thing as a species of reasoned or rational desire-change, then this cannot merely be a case of one mental state's being *followed* causally by another. Recall that in the story of causal change, the cause and the effect have, in some sense, nothing to do with each other. The effect is just something that tends to follow the cause—this doesn't give us the resources to capture the idea of desire correction. We need to find a way to make room for the fact that given an agent who desires X, and comes to believe p, what not only does but somehow must come next is the elimination of the desire that X. (Again, this is not to deny that we can *also* tell a causal story about, say, the events that underlie the desire change—the point is just that the causal story can no more be a story of desire correction than it can be a story of belief correction).

II. In the case of a constitutive change, the cause of the change ("getting married") is logically posterior to the change itself—getting married explains one's going from bachelor to husband but only because getting married is itself understood as (for a man) the transition from bachelorhood to husbandhood. This ensures the 'infallibility' of a constitutive changer: it never fails to achieve its effect, because it has already been specified in terms of that effect. None of this seems to hold of the realization that occasions a desire change. Realizing that this is gasoline is not logically posterior to moving from wanting to drink what's in the glass to not wanting to drink it; nor does the realization in any way logically entail that the change will take place. The entailment is supposed to be *on pain of irrationality*.

III. That's why the serious contender here is III, the belief change model. Why can't we solve the desire-paradox along the lines of our solution to the belief-paradox? Why can't we say that the first moment of my having a new desire *is* the realization that, say, this is gin? This answer would require us to be able to unify, into a single representation or thought or mental state, the belief that this is gin and the desire not to drink this. What I will argue is that such a unification is impossible: the resultant thought would have to be *either* a belief or a desire, in which case one of the two thoughts would be left out of the unity. My not wanting to drink this cannot *be* or *contain* my belief that this is not gin. Why not?

Here is the argument: My recognition of my reason to believe p *is* my belief that p, but my recognition of my reason to desire ϕ is not my desire to ϕ. Even if we want to say that the person who recognizes she has every reason to desire to ϕ but doesn't is *irrational*, we cannot say that she is, like her belief-counterpart, impossible. There is simply a constitutive connection between seeing oneself as having decisive reason to believe p and believing p, but there is at most a non-constitutively rational connection between the perception that this is gin and the desire not to drink it. And if the recognition of a reason to desire to ϕ is not identical to the desire to ϕ, then the belief (that there is a lion, that this is gasoline) will never be identical to the desire (not to enter, not to drink). They cannot constitute one unified mental state.

Let us imagine the following counterargument, offered by someone I will call the hyperrationalist: "the desire to ϕ *just is* a belief[4]„ namely, the belief that ϕ-ing is good." The hyperrationalist might say that the realization, in the desire-change case, is the (old) belief that drinking gasoline is bad, brought together with the new belief that this is gasoline—to bring these two beliefs together just is to believe that drinking this is bad, and that just is what it is to desire not to drink this. Certainly the hyperrationalist seems right that if desire is belief, desire-change can be modeled as belief-change.

The hyperrationalist gives us a translation formula for turning desires into beliefs. It is this: I desire to ϕ, or I desire ϕ-ing, or I desire that I ϕ (whatever your preferred way of representing the logical structure of desire) just is the belief that ϕ-ing is good (or, alternatively, the belief that [that I ϕ] is good. My question is: how will the desire-as-belief theorist translate my desire *not* to do something? He needs a way of handling negative desire if his reduction of desires to beliefs is to be complete. Consider the following sequence:

[4] For the first (there have been many by now) argument against the desire as belief thesis, see David Lewis ("Desire as Belief" Mind 1988). Interestingly, Lewis' argument centers on the impossibility of the desire-as-belief theorist's modelling a rational change of mind. On my view, however, the desire-as-non-belief theorist is subject to just the same problem! Since Lewis' paper, a number of arguments have been piled on to his (John Collins, "Belief, Desire and Revision" *Mind* 1988, Costa/Horacio/Collins/Levi "Desire-as-belief implies opinionation or indifference" *Analysis* 1995, Byrne/Hajek "David Hume, David Lewis and Decision Theory" *Mind* 1996, and Lewis himself again "Desire as Belief II" *Mind* 1996). In some way, the best argument for my claim that desire change cannot be understood on the model of belief change is to be found in Petit and Hajek's ("Desire Beyond Belief" *Australasian Journal of Phil.* 2004) argument *for* the desire as belief thesis. Exploiting a loophole in Lewis' argument, they show that DAB can be resuscitated if indexicalized, but the resultant desire-as-indexical-belief thesis turns desire into precisely the kind of belief which doesn't allow for stripping of the operator to combine its contents with that of another belief. That is, the beliefs that desires are, if Pettit and Hajek are right, are exactly ones whose changes cannot be understood on the model of III! (e.g. my desire to φ is my belief that "φ-ing is good" where that belief is understood as meaning "I have an attitude of approval toward φ") My argument here shows exactly why Lewis was right to claim (1996, p.312) that the indexicalizing version of the thesis (or, "the inconstancy thesis", as he called it) that desire is belief also trivializes it. This point about Pettit and Hajek's paper is also the point made by the second horn of the dilemma in the argument against the hyperrationalist.

(1) I desire (not to ϕ)[5] *or* I desire (that I not ϕ) *or* I desire (not ϕ-ing) (all = acc. to DAB)
(2) I believe that (not ϕ-ing) is good
(3) I believe that ϕ-ing is not good

(2) follows from any of the variants of (1), but (3) does not follow, logically, from (2). For one thing, there is a question whether the *content* of the belief in (2) licenses the inference to the *content* of the belief in (3). For any proposition p, and predicate F, whether it is the case that $\neg$p is F $\leftrightarrow$ p is $\neg$F depends on the predicate in question. For instance, if F is the predicate *true*, then the biconditional holds.[6] But most of the things we might say about a proposition are not like 'is true': consider "is profound" or "is being entertained in the mind of someone in this room right now" or "is grammatical" or "speaks to the issue of whether or not it is raining." These are predicates which can apply to both, or neither, of a proposition and its negation. To take one example: from the fact that the proposition "it is not the case that it is raining" speaks to the issue of whether or not it is raining, it does not follow that "it is raining" fails to speak to the issue of whether or not it is raining. We might describe a predicate F of which it is the case that, for any proposition p
$\neg$p is F $\leftrightarrow$ p is $\neg$F as 'convertible with respect to negation', *or negatively convertible.* Truth is negatively convertible, but we can say more than that about it: truth is transparently negatively convertible, because we can substitute the one side for the other even in an intensional context. That is, we can say S believes $\neg$p is true $\leftrightarrow$ S believes p is $\neg$true. This is, in fact, precisely the point we relied upon at the crucial moment of the argument of (III) above, when we said the belief that $\neg$q is false, in order for it to be a belief that $\neg$q is false, must already in some sense *be* the belief that q is true. In order to infer from (2) to (3) we would need 'is good' to be negatively convertible, and to be transparently so. Is it? Is good truelike?

This is the hyperrationalist's dilemma. Either way, he will encounter a problem: if he says good is truelike, it will not be the kind of thing, belief about which could be identical to a desire; if he says good is not truelike, he will make it impossible to use the apparatus of III to explain desire-change. That is, he either gives up modeling desire on belief, or he gives up modeling desire-change on belief-change.

Horn #1: Good is true-like. On this understanding of "good," the desire not to ϕ entails the belief that ϕ-ing is not good. On every understanding, the desire to ϕ entailed the belief that ϕ-ing is good (that just is the desire as belief thesis). So on the true-like reading, the person who desires to ϕ but also desires not to ϕ has contradictory beliefs. But it is impossible to (occurrently, self-awaredly) believe a contradiction[7]. So the hyperrationalist who takes the

[5] The "desire to" way of representing the structure of desire does not lend itself to belief-translation, and therefore will itself have to be pre-translated into either the "that" or the "–ing" formulation before that in turn can get translated into the corresponding belief.

[6] To deny the inference from $\neg$p is true to the conclusion that p is not true is to accept the possibility of true contradictions (dialetheism); to deny the inference in the other direction is to deny bivalence.

[7] What about the hyperrationalist who wants to assert that desire-beliefs are special, in that it is possible to desire-believe a contradiction? They are handled by the second horn of my dilemma as well. That is, insofaras the second horn exempts desire-beliefs from being incompatible with contradictory desire-beliefs, exactly so far does it make it impossible to identify the desire-belief that p with the desire-belief that it's not the case that not p. This move is the parallel to the move in footnote 8—wavering between the two answers is not a way out of the dilemma.

first horn of the dilemma is denying that I can desire to ϕ but also, at the same time, desire not to ϕ; or, equivalently, he denies that I can desire to ϕ but at the same time believe that it is not good to ϕ. The equivalence is evident if we note that

(4) I don't believe that (ϕ-ing is good)

follows from (3). But I can be motivationally conflicted, that is just a fact of life. I want to eat it because its tasty, but I don't want to eat it, or simply believe it's good to not eat it, because its unhealthy. If seeing desire as belief requires us to deny that I can want things I think are bad for me, it's a mistake to see desire as belief.

Horn #2: Let's try the other horn. To insist that good is not truelike is to say that there is no logical relationship between believing that not doing something is good and the believing that doing it is not good. So, for instance, if I discover that this is gasoline, and couple that with my desire not to drink gasoline (that is, my belief that not drinking gasoline is good) I am rationally bound to conclude, by the norms governing belief, that I desire not to drink this (that is that I believe that not drinking this is good). But, for all that, I can still believe that drinking this is good—nothing has touched that belief. But that is just to say that my desire to drink this is untouched by anything else that I desire.

And every desire will always be untouched, on the non-truelike horn. That is because on such a view there is no desire equivalent for the either the belief, or the absence of the belief, that drinking gasoline is not good—there is no way to desideratively represent something's not being good. But this would be the only kind of representation that would require me to give up a desire, because this is the only belief which is incompatible with either failing to believe, or believing that ϕ-ing is good (that is, failing to desire, or desiring, to ϕ). So given that I desire to ϕ, there is nothing that I could learn about the world (i.e., that this is gasoline, that this will kill me, etc.) that could rationally require me to change my desire. All that I could be required to do is *also* develop the desire not to ϕ[8]. So the nontruelike horn cannot explain precisely what we are trying to explain, namely, being rationally required to give up a desire.

A table that might be helpful for comparing the various translations: [[I won't read this]]

belief	Horn #1: Truelike translation	Horn #2: Non-truelike translation
I believe that ϕ-ing is good	I desire to ϕ = I don't desire not to ϕ	I desire to ϕ
I don't believe that ϕ-ing is good	I don't desire to ϕ = I desire not to ϕ	I don't desire to ϕ
I believe that ϕ-ing is not good	I desire not to ϕ = I don't desire to ϕ	*undefined*
I don't believe that ϕ-ing is not good	I don't desire not to ϕ = I desire to ϕ	*undefined*
I believe that not-ϕing is good	I desire not to ϕ = I don't desire to ϕ	I desire not to ϕ
I don't believe that not-ϕ-ing is good	I don't desire not to ϕ = I desire to ϕ	I don't desire not to ϕ

[[The following paragraph might get cut when reading, given time constraints]]

[8] Of course there is one way 'out' of the desire to drink this: coming to believe that this is not good. The point is that I can't get to such a belief from any of the attachments encoded in my desires. And insofar as I invoke such a thing, in the form of, e.g., a belief that this is unhealthy, or immoral, I am back with a version of problem #1: because it 'eliminates' my desire too easily, too logically, and thus seems to preclude conflict.

I think the underlying problem with trying to understand desire-change on model III is this: it requires us to unify a belief and a desire, into a single mental state. This entails combining the contents of the two states into the contents a single state. For instance, when I unify my belief that p with my belief that p→q, what I do is strip off the belief operators from both of those mental states and conjoin their contents, which are combinable into the conclusion that q. But heterogeneous mental states, such as beliefs and desire, are heterogeneous precisely because they do not have the same rules for the combination of their contents. (We can already see this if we notice that we must straightjacket a desire to get its contents into belief's favored "that" formulation.) Consider what happens if I try to simplemindedly 'combine' the contents of a belief and desire that often come together: the belief that it is not the case that I am ϕ-ing and the desire to ϕ. If I try to combine them into one belief, I get a contradiction (I am not ϕ-ing and I am ϕ-ing), and if I try to combine them into one desire, I get a state of conflictedness (I want not to ϕ and I want to ϕ). But the belief that I'm not ϕ-ing and the desire to ϕ are not supposed (rationally) to give rise to a state of conflictedness, or a state of self-contradiction, but rather a state of being motivated to act. What we mean by the unity of a mental state is the unification of the contents of that state, but desire and belief have different unification rules, and so you can't unify across these mental states. The problem here is a variant of the problem of opacity: desires and beliefs set up something like indissolubly different intensional contexts, but here the boundary line is *within a single mind*, at a single time.[9] To identify desire with a form of belief, or to try to get it working along the lines of belief, is therefore to abolish the very category of desire.

There is a fundamental incoherence in the story of so-called desire-correction: the agent is supposed be brought from desiring to ϕ to not desiring it by a realization which cannot be located on either side of the divide, or across the divide (as in the constitutive case[10]). The hyperrationalist who locates the realization that occasions desire-change on the right hand side effectively eliminates desire in favor of belief; the constitutivist who locates the realization across the change effectively eliminates belief in favor of desire, defining the realization in terms of the desire change it is to cause. And, finally, to locate the realization on the left side of the change is simply a straightforward denial of the possibility of desire correction, since assimilating desire-change to causal change eliminates the basis for a distinction between rational and nonrational desire change.

****

My conclusion is that desires are incorrigible: there are no rules which ever tell us to change our desires so that they might better live up to the standards set by what it is to be a desire. There is simply no such thing as 'rational desire change.' Let me end by saying what options

[9] If I'm right about this, it also serves as an argument that the conjunction of desire and belief cannot be another mental state, such as an intention. It must, as Aristotle thought, be a different kind of thing altogether, such as an action.

[10] I know I need to say more about why the constitutive model can't work as a model for desire. I'm tempted to say "it combines the vices of the other two!" but I know I need to say more than that. That is, I need to explain why the impossibility of a constitutive answer comes from the fact that what's wrong with the causal answer is not its *not* locating the realization in the later time, but the fact that it *does* locate it in the earlier time, and likewise, what's wrong with the belief answer is not its *not* locating the realization in the earlier time, but the fact that it *does* locate it in the later time. That is, the problem is not that the realization needs to be in both but that it can't be in either.)

we have left for thinking about desire. Beliefs are corrigible, which is to say that if one of my beliefs is false or unjustified, from the very fact that it is a belief I am enjoined to *do* something about this: give it up, or seek justification, or call it into question, etc. It is part of what it is to *be* a belief that that sets standards for goodness and badness among my beliefs, standards which any given belief must (normatively speaking) live up to. "Rationality" is just our name for these internal or constitutive norms. I have been arguing that there are no such norms in the case of desire.

So the old thesis that desire is directed at the apparent good, must be false. *Belief* is directed at what appears to be true, and what appears to be justified, in the sense that belief is corrigible in the light of truth and justification. That is, belief can be directed at *apppparent* truth precisely insofar as it misdirected from, and redirectable back onto, *real truth*. But desire is not corrigible in the light of goodness, or anything else. So desire is *not* directed at the apparent good. There is nothing that desire aims at or seeks which could serve to guide or redirect an errant desire—because, if I'm right, such guidance or redirection is simply inconceivable.

Someone might, hearing this argument, jump to the conclusion that the realm of norms have nothing to do with each other—desires are just causes. Let's consider the picture of desire we have on this view. A desire does not even try to tell us a way the world should or ought to be, it just makes the world be one way as opposed to another. On this view, it's not that desire *is supposed to* or *tries to* or *should* make the world one way as opposed to another. It just does—for if desires don't aim, then one thing they don't aim *at* is the satisfaction of desire.

But notice that there is an internal connection between the content we assign to a mental state, and the norms that such a state is governed by. The propositional content of a belief just is the proposition whose truth determines whether that belief lives up to the standards of belief, whether it succeeds at being a belief.

In the case of desire (on the view in question) there is no proposition or state of affairs such that, if that state of affairs is good, or if that proposition comes true, then the desire will have succeeded in doing what it was trying to do. Because *there is nothing that desires are trying to do*. Again, we repeat the mantra, desires are just causes. The point is not just that desires can't be distinguished from any of the many other kinds of mental states that play a part in the causal story of the genesis of human bodily movements (feelings, perceptions, beliefs). The point is that desires can't be distinguished from any of the other causes that give rise to non-human, or inanimate movements (impacts, electric charge, etc.) If desires are *just* causes, there is nothing distinctive about the way certain animate beings, in virtue of their mental lives, cause their own movements.

We can still say, a desire combines with a belief to cause a movement, but we are using the word combine in a way where there would be nothing wrong with saying that a desire of mine 'combines' with a belief of *yours*, or with a nonmental item such as a rock or a color or an impact. It is just a contingent fact if these combinations don't happen. There is nothing in the kind of thing that desire is that regulates what it combines with and how it combines with those things.

I think it is safe to call such a view, that is, the view that there nothing which is both *essentially* a cause of movement and *essentially* a mental state, skepticism about the existence of desire.

Thankfully, there is an alternative. We don't need to conclude from the fact that desires are incorrigible, that they can't provide normative guidance. Let's say I have a friend who is utterly insensitive to reasons. Nothing that anyone tells him ever makes any kind of impact on him—it is as though he is blind and deaf to the world. Should I do what he tells me to do, when he orders me around? Ordinarily, no. But what about the special case in which my friend is a prophet, in direct communication with God? Desires might have something to do with norms by having everything to do with norms; it may be that we don't desire the apparent good because what we desire is the *real* good. There might yet be standards that an incorrigible mental state must satisfy in order to be that mental state, it's just that the standards would have to be strict: something only qualifies as that mental state if it satisfies them.

But what about the case where you want something that *isn't* good, for instance when you want to enter the room with the lion in it? We will say what we already do say: you didn't really want that. You didn't want to turn the door handle, you didn't want to push the door open, you didn't want to enter the room, any more than you wanted to be mauled by the lion. But *something* got you into the room, if not a desire. So what was it? What made you enter the room?

I say "I see a lake up ahead", and my companion, who has traveled this desert many times, says, "no you don't, that's a mirage." I don't reply, "if there's one thing I know, it's that I *see* a lake, whether there is one or not." Likewise, you do not say, "If there's one thing I know, its that I *want* to go into the room, lion or no lion." We allow ourselves to be corrected out of illusions—and someone who says "I want p"—where, unbeknownst to him, p is bad—such a person is under an illusion.

It's an illusion that made you enter the room. How? The illusion moved you in just the way that the skeptic about the existence of desire thinks that everything moves everything. It caused you to enter the room in the way a wind blowing a leaf causes it to move. But it did not motivate you to enter the room. In order to motivate you, the illusion would have had to move you *to do something*. Only a state with satisfaction conditions, such as a desire, can do that. That's why skepticism about the existence of desire amounts to skepticism about the possibility of motivation.

It is only if the thing we want is really good that we really count as wanting it; and so it is only if the thing we want is really good that we can be motivated to get it, that is, to move ourselves with motions that in some way speak to some state of our mind. If desire is to be a distinctive kind of cause, namely, a cause of what might satisfy it, and if desires are incorrigible, then I can desire X only if X is really good.

There is something remarkable about the Socratic classification of mental states: he distinguishes knowledge from belief but he does not make a corresponding distinction on the side of desire. Aristotle does, he distinguishes what he calls boulêsis, 'rational wanting,' sometimes translated 'wish,' from mere appetite (epithumia). Aristotle thinks that just as there are kinds or levels or cognitive state, there are kinds or levels of conative state. For Socrates, all wanting is the conative analogue to knowledge—we might be able to believe something that merely *seems* true, but we cannot want something unless it really *is* good. What I have been doing, in this talk, is trying to bring out why that is not a crazy thing to think.