

A PUZZLE ABOUT RATIONAL DESIRE

Chase B. Wrenn

Department of Philosophy

University of Alabama

ABSTRACT: The following four assumptions plausibly describe the ideal rational agent. (1) She desires to believe only truths. (2) She knows what her beliefs are. (3) She does not both desire that P and desire that $\sim P$, for any P. (4) Whenever she desires that $P \rightarrow Q$ and knows that P, she desires that Q. Although the assumptions are plausible, they have an implausible consequence. They imply that the ideal rational agent never believes and desires contradictory propositions. She never desires the world to be any different than she thinks it is. Preserving our intuitions about desire, without embracing the implausible conclusion, is what I call “the Wishful Thinking Puzzle.” In this paper, I examine how this puzzle arises and I argue that it is surprisingly difficult to solve. Even the decision theoretic conception of desire is not immune to the puzzle. As I argue, the contrastive conception of desire avoids the puzzle, but that view is not worked out in enough detail to win our full confidence.

1. Introduction

Let S be an ideal rational agent. Each of the following principles is plausibly true of her:

- (1) *Awareness.* If S believes that P, then S knows that S believes that P.
- (2) *Epistemic Responsibility.* For any proposition P that S considers, S desires that she believes that P only if P is true.
- (3) *Conative Consequence.* If S desires a conditional, $P \rightarrow Q$, and she knows that P, then she desires that Q.
- (4) *Consistency.* S does not desire that P and also desire that $\sim P$.

We should construe “desire” in (1)–(4) as full desire, all things considered. That is the attitude of really wanting something to be the case, wholeheartedly, and upon reflection over all the relevant considerations. Desiring something fully, all things considered, is more than just having a pro attitude toward it or a slight preference that it be the case rather than not. Principles (1)–(4) may not be plausible when construed in terms of anything less than S’s full desires, all things considered.

S is “ideal” in two ways. First, she does not suffer cognitive limitations, such as those that prevent real agents from satisfying *Awareness*. Second, she constitutes a *normative* ideal. When we fall short of the standard set by (1)–(4), we often see that as a failure in need of correction. We aim to emulate S as closely as possible. At least, it is plausible that we do.

There is a serious problem with the model of rationality (1)–(4) provide. Although each principle is plausible, together they have a very implausible consequence. They imply that S never desires that $\sim P$ while believing that P. So, if (1)–(4) correctly describe our ideal of rationality, it is irrational to believe and desire contradictory propositions. That consequence is unacceptable. Thus we have a puzzle: How can we reconcile the intuitions behind (1)–(4) with further intuition that it is not irrational to desire $\sim P$ while believing P? I call this the Wishful Thinking Puzzle, and it proves surprisingly difficult to solve.

For example, a likely move in response to the puzzle is to give up the idea of full desire all things considered. Perhaps we could replace it with a conception of desire on which desiring is a matter of more and less, rather than all or nothing. The trouble, as I will argue below, is that this move does not work. The most familiar, graded conception

of desire comes from decision theory, but decision theory also faces the Wishful Thinking Puzzle.

Here is the outline of this paper. First, I show how (1)–(4) lead to the unacceptable conclusion that it is irrational to desire and believe contradictory propositions. Then I consider the plausibility of (1)–(4) themselves, for perhaps we could evade the puzzle by rejecting one of those principles. None of the most likely objections to the principles is a clear success, though, so I consider a few other approaches to the puzzle. These include appealing to the notion of “direction of fit,” construing (1)–(4) as “pro tanto” norms rather than absolute principles, and replacing the notion of full desire all things considered with the decision theoretic conception of desire. None of those approaches work.

Finally, I consider the prospects for a “contrastive” conception of desire. On that view, to desire a proposition is to prefer the way things would be if it were true to the way they would be if it were false. That conception of desire has the virtue of solving the Wishful Thinking Puzzle without being *ad hoc*, as it arise for both full desire all things considered and decision theory. However, the contrastive conception of desire is too inchoate to be declared the definitive solution to the puzzle. Ultimately, the right conclusion to draw is just that the Wishful Thinking Puzzle is a genuine and difficult puzzle, and the contrastive conception of desire is a promising response to it.

2. The Wishful Thinking Puzzle

The Wishful Thinking Argument, below, is a *reductio ad absurdum* of the assumption that S, who satisfies (1)–(4), believes that P while desiring that $\sim P$. I use

'B(ϕ)' to symbolize 'S believes that ϕ ', 'K(ϕ)' for 'S knows that ϕ ', and 'D(ϕ)' for 'S desires that ϕ ':

- | | | |
|------|------------------------------|----------------------------------|
| (5) | B(P) | Assumption |
| (6) | D(\sim P) | Assumption |
| (7) | D(B(P) \rightarrow P) | <i>Epistemic Responsibility</i> |
| (8) | K(B(P)) | 1, <i>Awareness</i> |
| (9) | D(P) | 7, 8 <i>Conative Consequence</i> |
| (10) | D(P) & D(\sim P) | 6, 9 <i>Conjunction</i> |
| (11) | \sim (D(P) & D(\sim P)) | <i>Consistency</i> |

Because (10) and (11) conflict, we must either reject the assumption that S believes that P but desires \sim P.

Imagine Sarah, an ideally rational Yankees fan, and suppose for reductio that she thinks her team is losing and desires that they not be losing. By *Epistemic Responsibility*, she desires that she believe they are losing only if they are losing. By *Awareness*, she knows she believes they are losing. So, lest her belief that they are losing be false, she desires that the Yankees be losing. But then Sarah would violate *Consistency*, for she would desire that the Yankees be losing and also desire that they not be losing. Thus Sarah does not both believe they are losing and desire that they not be losing.

How Sarah manages this is anybody's guess. She might be a Buddhist free from desire or a Pyrrhonist free from belief. She might be a Stoic whose beliefs guide her desires, or a wishful thinker whose desires guide her beliefs. Or maybe she sometimes follows one of those courses and sometimes follows another. However she does it, Sarah never believes and desires contradictory propositions, if she satisfies (1)–(4).

If (1)–(4) characterize the rational ideal, then it is irrational to believe something while desiring its denial. (1)–(4) seem to characterize the rational ideal, but it is not irrational to believe something while desiring its denial. The Wishful Thinking Puzzle is the puzzle of resolving the intuitions behind (1)–(4) with the rationality of desiring what you believe to be false. It is not an easy puzzle.

3. The Plausibility of the Principles

The most natural response to the Wishful Thinking Puzzle is to object to one of the principles driving the Wishful Thinking Argument. If any of (1)–(4) can be shown to be false, on independent grounds, then we can dismiss the argument as invalid and declare the puzzle solved. Each principle, though, is plausible in its own rights, and so it is hard to find any independent justification for rejecting them. Let us consider the principles one by one, starting with *Awareness*.

Awareness, it must be granted, is false of actual agents. If the ideal rational agent has unconscious beliefs, then it is false of her as well. Dropping *Awareness* would not solve the Wishful Thinking Puzzle, for the Wishful Thinking Argument does not require the principle in its full generality. We could run another version of the argument, without *Awareness*, that shows S never desires that $\sim P$ while *knowingly* believing that P. That conclusion is no more welcome than the conclusion of the original Wishful Thinking Argument.

The principle of *Consistency* is plausible because we are considering full desire, all things considered. We can argue for it in the following way. Suppose S has a certain pro attitude, *A*, toward both the proposition P and its denial, $\sim P$, and suppose S maintains that

attitude toward both propositions upon careful reflection over all the relevant considerations. If S is rational, it is hard to believe that A is full desire, all things considered. It might be some other form of desire, but her equal endorsement of P and of \sim P seems to stand in the way of its being *full* desire or a desire *all* things considered.

Conative Consequence encapsulates the idea that desiring a conditional is, in part, being disposed to desire its consequent when one knows the antecedent is satisfied. It has some precedent in the literature on conditional desires (e.g, Goldstein 1992; Bradley 1999), and it follow from other plausible assumptions. In particular, it is plausible that the ideal rational agent would desire whatever is obviously necessary and sufficient to satisfy one of her desires. If she desires $P \rightarrow Q$ and knows that P, then Q is obviously necessary and sufficient to satisfy her desire. So, just as *Conative Consequence* says, she would desire that Q.

Not everyone accepts *Conative Consequence*. F. C. T. Moore (1994) has argued against the principle. According to Moore, the principle is false because the belief (or knowledge) that P and the desire that $P \rightarrow Q$ can directly motivate Q-promoting action, without the mediation of an additional desire that Q. This argument is inconclusive. It draws a conclusion about the rational commitments of one who believes that P and desires that $P \rightarrow Q$ from premises about the capacity of the beliefs that P and the desire that $P \rightarrow Q$ to cause action. Even if the belief that P and the desire that $P \rightarrow Q$ can motivate action directly, it does not follow that one who believes that P and desires that $P \rightarrow Q$ is not thereby committed to desiring that Q. Moore's objection does not reveal *Conative Consequence* to be implausible, and so it does not provide a clear solution to the Wishful Thinking Puzzle.

We are left with *Epistemic Responsibility*. That principle is also plausible, for it is natural to think of truth (or knowledge, which entails truth) as the aim of belief. Bernard Williams famously argued that an attitude not aimed at truth simply could not be the attitude of belief (Williams 1970). A rational person wants her beliefs to be true. I assume ideal rationality requires that to be a full desire, all things considered. Anything less would be either a limitation on one's rationality or a limitation on one's commitment to being rational. The ideal rational agent suffers no such limitations. She is fully rational, and she is fully committed to being rational.

The philosophical precedent for *Epistemic Responsibility* is broad. In addition to Williams's claim belief is necessarily aimed at truth, there is Timothy Williamson's suggestion that knowledge is the norm of belief (2000, pp. 255–6). According to Williamson, a person ought to believe only what she knows. Knowledge entails truth, so, on Williamson's view, one ought to believe only what is true. The ideal rational agent would care about the norm of belief, and she would desire to conform to it. So, if Williamson is right, the ideal rational agent would satisfy *Epistemic Responsibility*.

In the second edition of *Theory of Knowledge*, Roderick Chisholm claims it is our basic intellectual obligation to do our best to ensure that, for any proposition we consider, we believe it if and only if it is true (Chisholm 1977, p. 16). The ideal rational agent desires to fulfill her basic intellectual obligation. So, if Chisholm is right, she satisfies *Epistemic Responsibility*.

Several philosophers have recently turned their attention directly to the value of truth (e.g., Lynch 2004; Blackburn 2005; Williams 2002; Kornblith 1993). Often, they argue that truth has a special role in a rational person's cognitive life. That rational person

adopts epistemic practices aimed at leading her to believe what is true and not what is false. If this widely held view is correct, the rational person deeply desires that her beliefs be true. She satisfies *Epistemic Responsibility*.

When I have shown the Wishful Thinking Puzzle to others, they have often thought *Epistemic Responsibility* must be to blame for it. It is thus worth considering the objections they have raised to the principle.

According to one objection, *Epistemic Responsibility* is implausible as formulated, although the general idea behind it is correct. That objection runs as follows:

Rational people do desire, in general, to believe only truths, but that is different from desiring-true every proposition one believes. For example, I believe that there is now injustice in the world, but I do not desire it true that there now be injustice in the world. *Epistemic Responsibility* is incorrect because it says we desire-true every proposition we believe, when it should say we desire to believe only truths.

The objection fails for two reasons.

First, it presupposes that believing something, such as that there is injustice in the world, in no way commits one to desiring that thing. That is not exactly correct. I believe there is injustice in the world, and I want to have only true beliefs. I could reason as follows: If there were no injustice in the world, I would have a false belief. But I want not to have false beliefs. So, I desire that there be injustice in the world.

Of course, I also have very powerful reasons to desire that there be *no* injustice. That is why the Wishful Thinking Puzzle is puzzling. Rationality seems to pull one in

incompatible directions—toward desiring that there be injustice (on pain of having a false belief) and also toward desiring that there be no injustice (because injustice is bad).

The second reason the objection fails might be more persuasive. *Epistemic Responsibility* does *not* say that S desires-true every proposition S believes. It says S desires-true a bundle of propositions with the form ‘ $B(P) \rightarrow P$ ’, where P is a proposition S considers. But $D(B(P) \rightarrow P)$ does not imply $B(P) \rightarrow D(P)$. To get from the former to the latter, we need to apply *Awareness* and *Conative Consequence*. So, the objection does not show that *Epistemic Responsibility* is wrong. If it shows anything, it shows that the conjunction of (1), (2) and (3) is implausible. We knew that already; that conjunction implies that S violates *Consistency* if she desires $\sim P$ while believing P.

This objection, then, amounts to claiming that we should drop *Epistemic Responsibility* because it is implicated in the Wishful Thinking Argument. That would be an ad hoc maneuver, not a solution to the puzzle. A solution would show that we have *independent* reason for rejecting one or more of the principles.

A better objection to *Epistemic Responsibility* accuses it of incorporating a quantifier confusion. The principle says that, for every proposition, P, that S considers, S desires to believe that P only if P is true. We might symbolize it like this (where ‘ \forall ’ is a universal quantifier ranging over propositions S considers and ‘P’ is a propositional variable):

$$(12) \quad (\forall P)D(B(P) \rightarrow P)$$

Now, if our intuition is that S has a general desire to believe only truths, we might think (12) expresses it wrongly. To express it correctly, we would need to put the quantifier *inside* the scope of the desire operator:

$$(13) \quad D(\forall P)(B(P) \rightarrow P)$$

So, one might object, the Wishful Thinking Argument depends on (12) when (13) is the right expression of our intuitions.

This is a nice move, but it is not entirely successful. Given one more plausible assumption, (13) implies (12). In that case, though, expressing *Epistemic Responsibility* in the form of (13) does not block the Wishful Thinking Argument after all.

Along with any of the following principles, (13) implies (12):

- (14) S's full desires all things considered are closed under implication.
- (15) S's full desires all things considered are closed under universal instantiation.
- (16) S's full desires all things considered are closed under the instantiation of universally quantified conditionals.

Though each principle has some plausibility, I will focus on (16), which is the weakest principle and so the most modest.

To see why (16) is plausible, suppose I have a certain pro attitude *A* toward the proposition that all my students do well, but I do not have that attitude toward the proposition that Mr. Ornery, whom I despise, does well if he is one of my students. Then, it seems, *A* is something less than *full* desire, *all* things considered. If it were full desire, all things considered, then I should not only have it toward the proposition that all my students do well, but toward the proposition that Mr. Ornery does well if he is one of my students. So, it seems, desiring the universally quantified conditional ('All my students do well', in this case) commits me to desiring its instance ('Mr. Ornery does well if he is one of my students', in this case).

To get around the Wishful Thinking Argument by replacing *Epistemic Responsibility* with (13), we must also reject (16), but (16) is also plausible. If we were to reject it, we would need to do so on grounds other than the Wishful Thinking Puzzle, on pain of ad hoc-ery.

Epistemic Responsibility and similar principles have also been attacked by philosophers who think we should not aim to believe truths. On their view, our aim should be more modest. We should aim what it is reasonable for us to believe or what would be justified. This view also has precedent. In the third edition of *Theory of Knowledge*, Chisholm revises his view. He says there that our basic intellectual obligation is to endeavor to believe only what is reasonable (1989, p. 1). Richard Rorty (1998), Larry Laudan (1984) and Bas van Fraassen (1980) have all argued that true belief, over and above reasonable or justified belief, is not a rational goal.

It would be too lengthy a digression to address all the issues bearing on whether we ought to aim for truth over and above reasonability in our beliefs.¹ I will mention just two general points here. First, the view that we ought to aim for reasonability rather than truth is a controversial and unpopular position, as even its advocates acknowledge. So, even if it is debatable whether *Epistemic Responsibility* is true, the principle is still *plausible*, and that is all that is necessary for the Wishful Thinking Puzzle to be puzzling. Second, there are some good reasons to reject the view that we should aim for reasonability rather than truth. One is just that the concept of “reasonable” or “justified” belief is already bound up with the idea of probable truth. Laurence Bonjour and Alvin Goldman, whose views are otherwise very different, agree on this much: whatever accounts for the justification of a

¹ For discussions, see Blackburn (2005), Lynch (2004), Williams (2002) and Wrenn (2005), and almost any issue of *Philosophy of Science*.

belief must make it likely that the belief is true, and that is a conceptual fact about the justification of belief. So, even if we should aim for reasonability, it hard to see how we could do that except by aiming for truth.

A straight solution to the Wishful Thinking Puzzle would show, independently of the puzzle itself, that (1), (2), (3) or (4) is false. The principles are all plausible, though, and they survive the most likely objections to them. That is part of why the puzzle is puzzling; if there is a straight solution, it is not clear what that solution is. So, it is worth considering other approaches.

One other approach is to give up the idea of full desire all things considered in favor of a notion of desire that admits of degrees. I consider that move in Section 6. First, though, I consider a couple of moves that preserve the idea of full desire all things considered: the appeal to “direction of fit” and the “pro tanto” interpretation of the principles.

4. Direction of Fit

Taking a cue from Elizabeth Anscombe (1957), some philosophers apply a metaphor of “fitting” to our beliefs and desires.² Our beliefs, they say, have “mind to world direction of fit;” we do (or should) adjust our beliefs to reflect how things stand in the world. Our desires, in contrast, have “world to mind direction of fit;” we do (or should) adjust the world to reflect our desires. This difference in direction of fit is supposed to account for the difference between beliefs and desires with the same contents.

One might think this metaphor holds the key to the Wishful Thinking Puzzle. After all, the Wishful Thinking Argument seems to contradict the principle that beliefs and

² See, for example, Smith (1994).

desires have opposition directions of fit. Its conclusion is that one's beliefs and desires should fit *one another*. According to the Wishful Thinking Argument, one who desires $\sim P$ while believing P should either change what she believes or change what she desires. To change what she believes would be to indulge in wishful thinking. To change what she desires would be to impose an incorrect, belief-like direction of fit on her desires, obliterating the distinction between belief and desire. Either way, the Wishful Thinking Argument seems to require us to impose the wrong directions of fit on our attitudes.

Even if this tells us *that* the Wishful Thinking Argument goes wrong, it does not tell us *where*. The principles on which the Wishful Thinking Argument depends are all compatible with the idea that beliefs and desires have opposite directions of fit. *Epistemic Responsibility*, the only principle that actually ascribes a belief or desire to S , says that S wants her beliefs to fit the world. It seems to codify the mind to world direction of fit for belief, not to contradict it. Step (9) is the only place in the argument where it is inferred that S has a certain desire. *Conative Consequence* licenses that inference, but *Conative Consequence* does not apply the wrong direction of fit to desires. Rather, it just specifies part of what it means to desire a conditional fully, all things considered. Similarly, *Consistency* just expresses a formal constraint on full desires, all things considered. It does not apply the wrong direction of fit either. Thus, the Wishful Thinking Argument has an interesting feature. Although its conclusion violates the principle that belief and desire have opposite directions of fit, no step in the argument depends on misconstruing the direction of fit of either attitude.

The direction of fit metaphor might be a useful, pre-theoretic way of characterizing the difference between belief and desire.³ It does not solve the Wishful Thinking Puzzle, though. To solve the puzzle, the metaphor would have to give us a way to block the Wishful Thinking Argument, but it does not. Instead of telling us how to avoid the unacceptable conclusion of the Wishful Thinking Argument, the metaphor only tells us something about why the conclusion is unacceptable. The “pro tanto” strategy, in contrast, does involve an effort to block the argument.

5. The Pro Tanto Response

One might think of (1)–(4) not as absolute requirements of rationality, but merely “pro tanto” requirements. Such requirements hold in normal circumstances, other things being equal, but they also allow for exceptions in abnormal circumstances, when other things are not equal. It is not objection to a set of pro tanto requirements that they sometimes conflict, so long as the conflicts arise only in exceptional circumstances.

This suggests a way of solving the Wishful Thinking Puzzle, which I call the “pro tanto response.” According to this response, at least one of (1)–(4) is a pro tanto requirement that does not apply in the circumstances the Wishful Thinking Argument presupposes. So, the Wishful Thinking Argument is invalid, and the Wishful Thinking Puzzle does not arise. The argument goes wrong at whatever step invokes a principle that does not apply in the envisioned circumstances.

The pro tanto response does not work. To solve the Wishful Thinking Puzzle, it is not enough to proclaim that some of the principles (1)–(4) are pro tanto norms. We also need an explanation of *precisely what* is abnormal about the situation assumed in the

³ See Sobel & Copp (2001) for an excellent discussion of the metaphor’s limitations.

Wishful Thinking Argument, such that the ordinary rules of rationality do not apply. We need an explanation of what “other thing” are not “equal” in that situation. We also need to know exactly which principle is to be suspended under those circumstances. In the absence of such explanations, the pro tanto strategy is less a solution to the puzzle than an ad hoc refusal to admit that it arises.

The pro tanto strategy runs into trouble because there is *nothing* unusual about the situation the Wishful Thinking Argument envisions. The argument assumes only that S believes P and desires \sim P. That is not abnormal, exceptional or unusual. We frequently desire things we believe to be false. Often, we desire them *because* we think they are false. According to the pro tanto strategy, though, the case in which one believes P while desiring \sim P is not only unusual, but *so* unusual that the ordinary rules of rationality do not apply. That claim seems no less bizarre than the Wishful Thinking Argument’s conclusion itself. A solution to the puzzle will have to come from a different direction.

6. Decision Theory

A natural response to the Wishful Thinking Puzzle is to blame it on the fiction of full desire, all things considered. Maybe there is no such attitude. Most of our desires are matters of more and less, not all or nothing affairs. Even if there is such an attitude as full desire all things considered, maybe we are wrong to formulate *Epistemic Responsibility* in terms of it. Once we allow for degrees of desire, (2)–(4) look much less plausible, and so the puzzle might seem to go evaporate. One could even see the puzzle as a reduction ad absurdum of the idea that there are full desires all things considered.

I will confine my discussion here to the most familiar notions of graded desire, which come from decision theory. If we adopt a decision theoretic conception of desire, we do not escape the Wishful Thinking Puzzle. There is a decision theoretic version of the puzzle.

Decision theory presupposes that the rational agent has preferences among possible gambles, and those preferences satisfy constraints that make it possible to define both a *value function* over the possible outcomes of the gambles and a *credence function* over propositions. I will denote the value function ‘ $v(\)$ ’, and I will use ‘ $C(\)$ ’ for the credence function. The credence function is a probability function, and it is taken to measure the subjective probabilities or degrees of belief the agent assigns to each proposition. Because it is a probability function, we can define $C(A | B)$, the agent’s conditional credence for A given B, as the ratio $C(A \& B)/C(B)$.

We can also define two quantities, the *expected value* and the *expected utility* of a prospective action, A. The expected value of A is the sum of the values of the possible outcomes, weighted by the agent’s conditional credence that each outcome will come about given that she takes A. Where the S_i ’s are the possible outcomes, the expected value of A is:

$$(17) \quad \sum_i v(S_i) C(S_i | A)$$

The expected *utility* of A is similar. Where the K_j ’s are hypotheses about the causal relationships between earlier states of the world and the possible outcomes, the expected utility of A is given by:

$$(18) \quad \sum_j \sum_i C(K_j) v(S_i) C(S_i | A \text{ and } K_j).$$

According to traditional, “evidential” decision theory, rational agents act so as to maximize expected value. According to “causal” decision theory, they act so as to maximize expected utility. I will not adjudicate between evidentialism and causalism here.⁴

The *desirability* of a proposition can be measured by either its expected value or its expected utility. We can define two desirability functions, des_e and des_c , as follows:⁵

$$(19) \quad des_e(P) = \sum_i v(S_i) C(S_i | P)$$

$$(20) \quad des_c(P) = \sum_j \sum_i C(K_j) v(S_i) C(S_i | P \text{ and } K_j)$$

We can think of $des_e(P)$ as the “news value” of P; it measures how good the agent would find the news that P. We can think of $des_c(P)$ as the “instrumental value” of P. It measures how much, in the agent’s view, P’s truth would promote good outcomes.

One could view $des_c(P)$ or $des_e(P)$ as measuring directly the extent to which the decision theoretic agent desires that P, but it is better to identify the strength of one’s desire that P with $D_e(P)$ or $D_c(P)$:

$$(21) \quad D_e(P) = des_e(P) - des_e(\sim P)$$

$$(22) \quad D_c(P) = des_c(P) - des_c(\sim P).⁶$$

The decision theoretic Wishful Thinking Puzzle arises because $des_e(\sim P)$, $des_c(\sim P)$, $D_e(\sim P)$, and $D_c(\sim P)$ are all *undefined* for an agent who fully believes that P. To fully believe that P is to assign it a credence of 1, so $C(\sim P) = 0$. But when $C(\sim P) = 0$, the conditional probabilities in (19) and (20) are fractions whose denominators are 0. So, if

⁴ I take this way of distinguishing causal and evidential decision theory from Lewis (1981). The use of ‘expected value’ and ‘expected utility’ as labels for the different functions is due to him.

⁵ Jeffrey’s (1965) definition of desirability is (19); (20) is its causalist analog.

⁶ Here is why it is better. Suppose we identify $des_e(P)$ with the degree to which one desires that P. If $des_e(P) > 0$ and $des_e(P) = des_e(\sim P)$, then the agent qualifies as desiring that P and also qualifies as being indifferent to P (because she neither prefers P to $\sim P$ nor prefers $\sim P$ to P). The attitude of desiring that P, however, should be incompatible with the attitude of being indifferent whether P.

the decision theoretic agent fully believes that P , *there is no degree to which she desires that $\sim P$.*

This feature of decision theory is not entirely unknown. The usual response is to ignore it. In a decision problem, the desirability of a proposition with 0 subjective probability is never relevant. David Lewis (1981) dismisses the problem as a curiosity one should never allow to arise, for he thinks it is “rash” to assign a contingent proposition a credence of 0 or 1.

Lewis’s response is common, but it is far too cavalier. To give a proposition a credence less than 1 is to believe it less than fully. It is to entertain some doubt about the proposition, however small. Maybe it is true that we should fully believe *very few* propositions, but some propositions do merit our full belief. Consider the proposition that I am now suffering agonizing pain. There are certain clear cases in which it is rational, not rash, for me to believe fully that I am suffering agonizing pain. They are cases in which I neither have nor ought to have any doubt that I am then suffering agonizing pain, and they are also cases in which it is perfectly reasonable for me to desire that I *not* then be suffering agonizing pain.

On the standard, decision theoretic account of desire, the Wishful Thinking Puzzle might arise only rarely, but it does arise and it is a problem. That account makes it impossible for a rational person to desire not to be suffering agony unless she also entertains some doubt as to whether she is actually suffering it. The move to decision theory does not solve our problem. To the contrary, it reintroduces the problem in a new form.

7. Counterfactual Decision Theory and the Contrastive Conception of Desire

The contrastive conception of desire provides solutions to both versions of the Wishful Thinking Puzzle. It is based on an insight that is already apparent in (21) and (22)'s definitions of desire. The desire that P is indistinguishable from the preference that P rather than \sim P. According to standard decision theory, to prefer that P rather than \sim P is to find the way the world is if P more valuable (in terms of either expected value or expected utility) than the way it is if \sim P. That makes it impossible to desire that \sim P while fully believing that P. The contrastive conception of desire involves a different account of what it means to prefer P rather than \sim P.

Suppose I know I am now being tortured, and I desire that I not now be tortured. This is not because I think the world *is* a better place if I am not being tortured. It is because I think the world *would be* better if I were not being tortured. To a first approximation, contrastivism is the claim that desiring P is preferring the way the world would be if P were true to the way it would be if P were false.

'The way the world would be' is a notoriously context-sensitive expression. In some contexts, possibilities are relevant that are irrelevant in other contexts. Consequently, "the way the world would be" can vary from one context to another. Sometimes, the variation is striking.

Consider the way the world would be if kangaroos lacked tails. What would kangaroos be like? In the context of taxidermy, the answer might well be that tailless kangaroos would be shaped nearly the same as actual kangaroos. They would just be missing their tails and a little less expensive to stuff. In the context of evolutionary biology, it is far less clear what tail-less kangaroos would look like. They would not be

shaped much like actual kangaroos, for such a critter would have untenable body mechanics. An evolutionary trajectory resulting in tail-less kangaroos might well result in kangaroos that look more like actual koalas (which are nearly tail-less marsupials) than like actual kangaroos. There is no single, context-invariant answer to the question, “What would kangaroos look like if they did not have tails?”

Which possibilities are relevant, and so how the world would be if various things were the case, is a contextual parameter that varies from conversation to conversation. It can also shift in the span of a single conversation (Lewis 1979; 1996). It is a consequence of contrastivism that one’s desires can vary, across and within contexts, as a function of which possibilities are contextually relevant. A taxidermist’s desire that kangaroos lack tails might well have a different content from a biologist’s desire. They are attitudes toward different contrasts.

Suppose Bill has looked over the dessert menu and decided that he wants the Black Forest torte for dessert. The relevant possibilities are that he have no dessert, that he have the torte, and that he have something else. The torte possibility is his favorite. After Bill announces his decision, Amanda asks if anyone wants to go next door for their excellent cheesecake. Her question makes a possibility relevant that had been irrelevant before. Bill must now decide between having the torte here and having the cheesecake next door. His previous desire for the torte does not commit him to desiring it now. He might well find that he no longer desires the torte, because he finds the cheesecake possibility preferable to it.

There are thus two important components to the contrastive conceptions of desire. First, it treats desire as a form of counterfactual preference. To desire that P is to prefer

the way things would be if P to the way they would be if $\sim P$. Second, it acknowledges that what a person counts as desiring can vary as a function of which possibilities are contextually relevant. Here is a more precise formulation of contrastivism:

- (23) ‘S desires that P’ is true at context C if and only if S prefers the P-possibilities relevant in C to the $\sim P$ -possibilities relevant in C.

This formulation surely needs refinement. It fails to specify what makes a possibility relevant in a given context, what it means to prefer one set of possibilities to another, and whether the truth of a desire-attribution depends on the desirer’s context or the attributor’s. These are defects I will mostly ignore. The formulation is enough to let us solve both versions of the Wishful Thinking Puzzle.

First consider the decision theoretic version of the puzzle. According to contrastivism, to desire that P is not to prefer how things *are* if P rather than $\sim P$, but to prefer how things *would be* if P rather than $\sim P$. That means we should alter the definitions of expected value and expected utility to apply a *counterfactual* notion of conditional probability, rather than the more familiar notion applied in (19) and (20). Robert Stalnaker’s (1970) account of counterfactual conditional probability will work nicely here.

Let $C(!)$ be what Stalnaker calls an “extended probability function.” It is a function that satisfies the following requirements:

- (24) (a) $C(A ! B) \geq 0$
 (b) $C(A ! A) = 1$
 (c) If $C(\sim A ! A) \neq 1$, then $C(\sim A ! C) = 1 - C(A ! C)$
 (d) If $C(A ! B) = C(B ! A) = 1$, then $C(C ! A) = C(C ! B)$

$$(e) \quad C(A \& B \mid C) = C(B \& A \mid C)$$

$$(f) \quad C(A \& B \mid C) = C(A \mid C) \cdot C(B \mid A \& C)$$

Where t is an arbitrary tautology, we can define the unconditional credence $C(A)$ as $C(A \mid t)$. Three features of $C(\mid)$ are important for our purposes. First, $C(A \mid B) = C(A \mid B)$ whenever $C(B) > 0$. Second, $C(A \mid B)$ has a value even when $C(B) = 0$. Third, as Stalnaker shows, $C(A \mid B)$ can be understood as an agent's counterfactual conditional credence in A , given B . It measures the agent's estimation of what her degree of belief in A would be *if B were known*, even in cases where $\sim B$ is known.⁷

Let "counterfactual decision theory" be the result of replacing standard decision theory's definitions of expected value and expected utility with (25) and (26):

$$(25) \quad \text{des}_{e+}(P) = \sum_i v(S_i) C(S_i \mid P)$$

$$(26) \quad \text{des}_{c+}(P) = \sum_j \sum_i C(K_j) v(S_i) C(S_i \mid P \text{ and } K_j)$$

Counterfactual decision theory identifies the expected value of P with how good the agent thinks the world would be if P , and it identifies the expected utility of P with how good the agent thinks P would make the world. Counterfactual decision theory defines degrees of desire in the obvious ways:

$$(27) \quad D_{e+}(P) = \text{des}_{e+}(P) - \text{des}_{e+}(\sim P)$$

$$(28) \quad D_{c+}(P) = \text{des}_{c+}(P) - \text{des}_{c+}(\sim P).$$

When $C(P) > 0$, (25)–(28) agree with their counterparts in standard decision theory.

Because their counterparts are undefined when $C(P) = 0$, this means that counterfactual decision theory gives the same verdict as standard decision theory, whenever standard

⁷ Stalnaker does this by showing that $C(A \mid B)/(1 - C(A \mid B))$ represents the odds at which the agent would have accepted a bet that A , had B been known.

decision theory delivers a verdict at all. In that sense, counterfactual decision theory is a conservative extension of standard decision theory.

The Wishful Thinking Puzzle does not arise for counterfactual decision theory. $D_{e+}(\sim P)$ and $D_{c+}(\sim P)$ are well defined even when $C(\sim P) = 0$. Because counterfactual decision theory is a conservative extension of standard decision theory, it preserves whatever correct intuitions standard decision theory expresses. So, the move to counterfactual decision theory, which follows naturally from the contrastive conception of desire, solves the decision theoretic Wishful Thinking Puzzle. It shows how to avoid the unacceptable conclusion without sacrificing what is good in standard decision theory.

Contrastivism also solves the Wishful Thinking Puzzle for full desire all things considered. Given contrastivism, principles (2), (3) and (4) need to be reformulated to take the context-sensitivity of desire into account. The required adjustments undermine the Wishful Thinking Argument.

The easiest principle to adjust is *Consistency*. Contrastivism allows for the possibility that one might desire that P relative to one context and desire that $\sim P$ relative to another. What rationality does not allow, though, is desiring P and desiring $\sim P$ relative to the *same* context. The revised principle is thus:

- (29) *Contextual Consistency*: For any context C, ‘S desires that P’ and ‘S desires that $\sim P$ ’ are not both true at C.⁸

To revise *Conative Consequence*, we need to take two things into account. First, we should not require a person who knows P and desires $P \rightarrow Q$ in one context to desire Q in

⁸ If rational preferences are transitive and irreflexive, this actually follows from contrastivism. Suppose there were a context, C, relative to which S desires P and desires $\sim P$. By contrastivism and the transitivity of preference, S would have to prefer P to P. But rational preferences are irreflexive, so there can be no such context.

another context. Our revised principle should apply only within contexts, not across them. Second, we need to take into account that one might desire $P \rightarrow Q$ only because one very strongly prefers $\sim P$ to P and is actually indifferent to Q . Such a person might know that P and desire that $P \rightarrow Q$ without desiring that Q , relative to a context in which there are relevant $\sim P$ -possibilities. A suitably restricted version of the principle is this:

- (30) *Contextual Conative Consequence*: For any C such that there are no relevant $\sim P$ -possibilities, if S knows that P and ‘ S desires that $P \rightarrow Q$ ’ is true at C , then ‘ S desires that Q ’ is true at C .⁹

Epistemic Responsibility requires special care. Not every context is one in which it is possible to desire $B(P) \rightarrow P$, and not every context is one in which it is rationally obligatory. The idea behind the principle is that one wants not to have beliefs that are actually false. Without contrastivism, this idea is expressed by the claim that S desires $B(P) \rightarrow P$, for each P that S actually considers. With contrastivism, we can say something more precise. Let a context C be *fixed with respect to P* if every relevant possibility in C is a P -possibility. We can reformulate *Epistemic Responsibility* as follows:

- (31) *Contextual Epistemic Responsibility*: For any proposition P that S considers and context C that is fixed with respect to P but not fixed with respect to $B(\sim P)$ or $\sim B(\sim P)$, ‘ S desires $\sim B(\sim P)$ ’ is true at C .

The idea here is the S desires to have only true beliefs in the sense that she desires not to have false beliefs. *Contextual Epistemic Responsibility* captures the idea that one should desire to make one’s beliefs fit the world. If we hold it constant that P , there is no

⁹ The requirement that S know that P is unnecessary. If no $\sim P$ -possibilities are relevant, then (given contrastivism) there is no difference between desiring $P \rightarrow Q$ and desiring Q . Part of the intuitive force of *Conative Consequence* probably comes from the idea that that knowledge that P automatically renders the possibility that $\sim P$ irrelevant.

question of wanting the world to adjust itself to suit your beliefs. Rather, what is variable is what you believe. According to (31), if C is fixed with respect to P, and one's belief in $\sim P$ is variable, S prefers believing P and withholding judgment to falsely believing $\sim P$.

There may be contexts that are fixed with respect to P but where the costs of not believing $\sim P$ would be enormous. Let P be the proposition that you are a handless brain in a vat, artificially stimulated to have illusory experiences, and suppose that C is a context that is fixed with respect to P. Maybe you would be much better off believing that $\sim P$, even if it is fixed that P in all the relevant possibilities, and maybe we can know that a priori. If so, then it seems wrong to think that rationality requires you to desire not to believe $\sim P$. Some further restrictions on (31) are thus needed for it to be plausible, but I will ignore them for now.

With (2)–(4) replaced by (29)–(31), we can ask whether the Wishful Thinking Argument still goes through. It does not. Suppose that S desires $\sim P$ relative to some context C, and S believes that P. S thinks the relevant $\sim P$ -possibilities are better than the relevant P-possibilities. Now, by *Contextual Epistemic Responsibility*, if C is fixed with respect to P, S desires not to believe $\sim P$, and if C is fixed with respect to $\sim P$, S desires not to believe that P. Either way, we will not be able to use *Contextual Conative Consequence* to infer that S desires P. In fact, we cannot even use *Contextual Epistemic Responsibility* to infer that S desires not to believe P or not to believe $\sim P$. We could do that only if the context were fixed with respect to P or $\sim P$, but it is not. By assumption, S desires that $\sim P$, which means there are relevant P-possibilities as well as relevant $\sim P$ -possibilities. The Wishful Thinking Argument simply does not work on the contrastive conception of desire.

8. Conclusion

Contrastivism solves the decision theoretic Wishful Thinking Puzzle by embracing counterfactual decision theory. It solves the puzzle for full desire all things considered by exploiting the context sensitivity of desire. Both moves arise from a single insight: desiring that P is a matter of preferring the way things would be if P to the way they would be if \sim P.

It is also worth noting that counterfactual decision theory can accommodate the context-sensitivity of desire. $C(A \mid B)$ is meant as a measure of how likely the agent thinks A would be if B were the case. That depends entirely on which B-possibilities are relevant, and so we can expect it to vary from one context to another. Moreover, although traditional decision theory tends to assume that a single agent's values and credences are unchanged from one decision problem to the next, this assumption is not essential to solving any particular decision problems. In practice, values and credences are determined on the basis of one's preferences among outcomes in particular problems, and one's preferences among outcomes are determined by one's preferences among various possible gambles. Ultimately, then, one's desires depend on what one's attitudes are to the possible gambles. They depend on what alternatives are contextually relevant.

It is easy to see counterfactual decision theory as an elaboration or formalization of the contrastive conception of desire. In particular, the decision theoretic apparatus allows us to answer the question of what it means to prefer "the relevant P-possibilities" to the "the relevant \sim P-possibilities." The desirability of the relevant P-possibilities is either

$\text{des}_{e+}(P)$ or $\text{des}_{c+}(P)$, and to prefer them to the relevant $\sim P$ -possibilities is just for $D_{e+}(P)$ or $D_{c+}(P)$ to be positive.

Before the contrastive conception of desire is fully satisfactory, more issues need to be settled. An especially pressing question is whether the alternatives that matter to the truth value of ‘S desires that P’ are those that are relevant to S or those that are relevant to the person who makes the desire attribution. Also, even though the contrastive conception of desire has independent motivation (from the truism that one desires P if and only if one prefers the way things would be if P to the way they would be if $\sim P$), I know of no direct *argument* for the view. The contrastive conception avoids the Wishful Thinking Puzzle, and it does so without being ad hoc, but it would be a mistake to think that is the last word on the puzzle. The contrastive conception of desire might well be wrong.

My conclusion, then, is that the Wishful Thinking Puzzle is a real puzzle about rational desire. Some of the more obvious strategies for solving it do not work. At least one approach does work, but it is too underdeveloped to win our full confidence. I want the puzzle to have a clear and obvious solution, but I do not think it does. I hope that is not irrational.

References

- Anscombe, G. E. M. 1957. *Intention*. Oxford, UK: Blackwell.
- Blackburn, S. 2005. *Truth: a guide*. Oxford, UK: Oxford University Press.
- Bradley, R. 1999. Conditional desirability. *Theory and decision*. Vol. 47. pp. 23–55.
- Chisholm, R. 1977. *Theory of knowledge*. 2nd ed. Upper Saddle River, NJ: Prentice-Hall.
- , 1989. *Theory of knowledge*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

- Goldstein, L. 1992. A Buridanian discussion of desire, murder and democracy.
Australasian journal of philosophy. Vol. 70, No. 4. (Dec. 1992). pp. 405–414.
- Jeffrey, R. 1965. *The logic of decision*. New York: McGraw-Hill.
- Kornblith, H. 1993. Epistemic normativity. *Synthese* vol. 94, n. 3 (March 1993): pp. 357-376.
- Laudan, L. 1984. *Science and values*. Berkeley, CA: University of California Press.
- Lewis, D. 1981. Causal decision theory. *Australasian journal of philosophy*. Vol. 59, No. 1 (March 1981). pp. 5–30.
- . 1979. Scorekeeping in a language game. *Journal of philosophical logic*, vol. 8 (Aug. 1979): pp. 339–359.
- . 1996. Elusive knowledge. *Australasian journal of philosophy*, vol. 74, no. 4 (Dec. 1996): pp. 549–567.
- Lynch, M. P. 2004. *True to life: Why truth matters*. Cambridge, MA: MIT Press.
- Moore, F. C. T. 1994. Goldstein on the road to Rome. *Australasian journal of philosophy*. Vol. 72, No. 2. (June 1994). pp. 229–232.
- Rorty, R. 1998. *Truth and progress*. Cambridge, UK: Cambridge University Press.
- Smith, M. 1994. *The moral problem*. Oxford, UK: Blackwell.
- Sobel, D. & Copp, D. 2001. Against direction of fit accounts of belief and desire.
Analysis. Vol. 61, No. 1 (Jan 2001). pp. 44–53.
- Stalnaker, R. 1970. Probability and conditionals. *Philosophy of science*. Vol. 37, No. 1 (March 1970). pp. 64–80.
- Van Fraassen, B. C. 1980. *The scientific image*. New York, NY: Oxford University Press.

Williams, B. 2002. Deciding to believe. In: B. Williams, *Problems of the self*.

Cambridge, UK: Cambridge University Press.

----- *Truth and truthfulness: An essay in genealogy*. Princeton, NJ: Princeton University

Press.

Williamson, T. 2000. *Knowledge and its limits*. Oxford, UK: Oxford University Press.