The Challenge of Temptation: desire, emotion, and stability.

by

Soraya J. Gollop

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in The University of Michigan
2007

Doctoral Committee

Professor James M. Joyce, Chair
Professor Elizabeth Secor Anderson
Professor Peter A. Railton
Associate Professor Mika T. Lavaque-Manty

To Nitin, my darling husband.

Acknowledgements

This work would not have been possible without the support and generosity of my family, friends, and colleagues, to whom I tender my sincerest thanks. To my husband Nitin, for his endless support and patience which have made all the difference. To my friend Anna Gotlib whose pep talks, coffees, and endless trips to the airport have been of incalculable help. To my family, both old and new, who have had such faith in me, especially my sisters Annora and Alicia who have been there even in the smallest hours of the morning. And all of my other friends and colleagues in the department at Michigan who have made this experience so much better than it could have been. This would not be complete without acknowledging the extraordinary efforts and intellectual generosity of James M. Joyce, my primary advisor, without whose advice this would not be the work that it is.

Table of Contents

List of Figures

Abstract

Desires are usually presented as simple states whose contribution to action, choice, and deliberation are understood simply in terms of motivational strength and object. The challenge of temptation is to give an account of desires that explains why temptations should not be treated on a par with other desires in rational deliberation. Desires *qua* simple states fail this challenge as *ex hypothesi* what it is to be tempted to do some thing is for doing that thing to be your strongest motivation. I examine the nature of desire through the lens of temptation, and create a more complex picture of desires identifying and defending various properties of desires. These properties are: an account of how desires may be more or less stable with respect to reflection and new information; the emotional component of desires (desire responses), in variously forming and undermining our reflectively stable desires (considered preferences); how it is that we are psychologically disposed to value goods over time (hyperbolically), and the way in which it is rational to value goods over time (exponentially); finally an account of the difference between warranted and unwarranted changes in the strength of desires. Temptation is a consequence of a pervasive natural tendency to discount the future hyperbolically. If you discount the future hyperbolically, then the proximity of the good causes the comparative strength of your desire responses to reverse. Desire responses are produced by a perception-like system within the agent, the conative system. This reversal occurs because proximity constitutes abnormal operating conditions for the conative system, thus producing unreliable desire responses. As the visual system produces misleading visual perceptions under abnormal conditions, the conative system can produce misleading desire

responses under abnormal conditions—temptations. We can identify and compensate for these misleading desire responses by paying attention to the relative stability of our desires. Thus desires are not simple states which contribute to deliberation solely in terms of object and strength, but rather more complex states that contribute to deliberation in terms of their stability and emotional components, as well as their object and motivational strength.

CHAPTER 1

INTRODUCTION

Desires are the most neglected element of rational deliberation. In discussions of deliberation, rational choice, and action theory they are generally mentioned and dismissed as merely a brute source of motivation or guide to the agent's pleasures. What is intriguing is that such a blunt dismissal of desires is not in line with folk discussions of desire. We are temporally extended creatures, and in the course of living and acting, our desires change. We acquire new desires, we dispense with old desires, and the desires that we have change. Desire understood solely in terms of object and strength give a picture of rational deliberation which reduces it to a simple weighing of the relative strength of desires. But intuitively desires have many more attributes than simply strength and object. Desires can be enduring or fleeting, serious or frivolous, passing fancies or lifetime aims. I can desire objects, people, personal experiences, conceptual truths or abstract ideals. My desire may motivate actions, or be frustrated by my acting so as to bring them about. My desire may lie in the pleasurable contemplation of a state of affairs that I would be positively displeased if it occurred. A theory which treats all desires as given, and pays no attention to the circumstances of their inception or change, cannot adequately distinguish between one's most dearly held life aim, and an intense but fleeting temptation to act in ways that will undermine that aim. It is these considerations which generate what I will call the 'challenge of temptation'.

The idea that temptations should (at least some of the time), be resisted in order to satisfy longer term desires whose motivation is weaker at the moment of choice just seems obvious. To be tempted to do *x* is, *ex hypothesi*, for it to be the case that at that moment your strongest motivation is to do *x*. If all there is to desires is object and motivational strength, then there should be no question of resisting temptation being a desirable thing to do. After all, if what I am doing in giving in to temptation is simply acting on my strongest desire, and desires are considered only in terms of strength and object, then we have no resources left for explaining why I should resist temptation in order to satisfy my longer term desire. Thus the view that it is rational to resist temptation (in at least some cases) is at odds with the view that desires are simple entities which can be wholly understood in terms of their object and strength. The challenge of temptation is to explain why we do not consider temptations to be on a par with other desires, and any plausible theory of desire should be able to give an account of what is wrong with temptations.

I propose that what we need to do in order to solve the problem of temptation is to pay attention to the emotional component of desires, as temptations occur when our motivations are suborned by intense experiences of desiring. Ultimately I will argue that such intense experiences of desiring should not be treated on a par with other desires because they are the product of certain pervasive mistakes, mistakes that are identifiable in deliberation through the stability properties of desires.

The Challenge of Temptation

We are no strangers to temptation. Most of us have, on a daily basis, the opportunity to satisfy a short term desire which, if acted upon, will undermine the fulfillment of some more important, longer term, desire. A simple case of this would be someone, let us call her Kim, who wants to lose weight in order to fit into a particular dress for her sister's wedding. Generally, Kim's desire to maintain her diet is stronger than her desire for forbidden

foodstuffs, but she has a weakness for chocolate. When the opportunity to eat chocolate cake is to hand, she is overwhelmed by her desire for the cake. Consider a particular instance: Kim is passing her favorite bakery at 4pm on Tuesday. At that time, she wants to eat cake more than to maintain her diet. But when she is temporally separated from the opportunity to eat cake, she wants to maintain her diet more than eat cake. Over the day, the interaction between her desires for cake and dieting look like this:

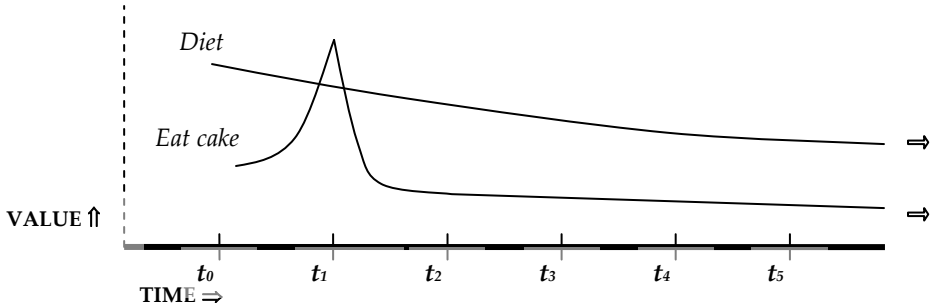

Figure 1: Kim's temptation

This reversal in the strength of Kim's desires raises the question of what she should do in this case. Should she give in to temptation and eat the cake, or resist temptation and maintain her diet? Common sense tells us that she should resist the temptation to eat the cake. Indeed, in the act of calling this a 'temptation' we are implicitly marking it as different from her other desires. And the difference so marked designates the temptation as, in some robust sense, less important than the currently weaker competing desire. However, if we understand the contribution of desires to choice only in terms of their motivational strength, then she should eat the cake.

The aim of this dissertation is to defend a theory of desire that can make sense of our intuitions about temptations, and give an account of why temptations are identifiably different from other desires. My view is that the problem of temptation will only be solved by giving an account of desires that builds in an account of why it is that temptations are flawed desires, rather than simply desires on a par with all other desires.

Before I can consider how various views of desire cope with the challenge of temptation, I need some more resources for talking about the properties of desires. In the next section I will suggest a characteristic of desires that will both add to our understanding of desires, and be helpful in considering the challenge of temptation. This is the property of stability.

The Stability of Beliefs and Desires

Now, one identifiable feature of temptations is that they tend to be *unstable*, that is, the power that they exert over the agent is not constant. For instance, the sight or scent of the chocolate cake makes it more tempting. These changes in the strength of desires are, I will argue, expressions of a distinct property of desires, their 'stability'.

The view that beliefs may be more or less stable is common in the literature on epistemology,[1] and I contend that desires may be more or less stable in analogous ways. Stability for beliefs is a function of the propensity of beliefs to change in the face of new information. The lower the propensity for change the more stable the belief; the higher the propensity for change the less stable the belief.

Stability is a property of beliefs with respect to bodies of information. This property, with respect to a particular set of information, is called *resilience*. Thus, '*A's* belief in a proposition *x* is **resilient** with respect to some piece of evidence *e* just in case learning *e* will not greatly change the credence that *A* has in *x*'. The rough, overall level of stability of a particular belief can be expressed in terms of the bodies of evidence that the belief is resilient with respect to. By considering the relevant sets of evidence, we can have some sense of the

[1] Stability concerns in the epistemology literature also appear as discussions of 'diachronic coherence'. See, for instance, (Polanyi, 1952; Rott, 2004; Skyrms, 1984; Sobel, 1990; Sorell, 1981). Loeb interestingly interprets Hume as putting forward some sort of stability constraint in his (Loeb, 1991; Loeb, 2002).

degree to which that belief is stable. The stability of a belief is a measure of its resilience compared to the resilience of another belief, making stability an essentially relative notion.

The fundamental notion of stability is the stability of a belief with respect to a body of evidence. However, there are other ways in which we can talk about stability. It is possible to compare the stability of different beliefs with respect to the same body of evidence, in order to get a sense of their comparative stability with respect to that piece of evidence. We can talk about a belief being stable with respect to a wide range of bodies of evidence, and use this to compare the general stability of beliefs. These two latter senses of stability are both derivative of the fundamental relative notion of stability, and it is the final sense that I employ in talking about the unqualified stability of a belief (or any other state).

The cases of stability which I am interested in are ones in which the instability of a belief gives us information about something that is going wrong with that belief. I want to emphasize the point that not all cases of instability are problematic. That my belief that it is going to rain has a high propensity to change with respect to information directly relevant to the weather is a sign of my belief working well. Compare this with an unstable belief which has a high propensity to change with respect to information that is evidentially irrelevant to the subject of the belief, such as cases of wishful thinking. Here the changes in my beliefs look like a sign that my beliefs are not functioning well. It is these latter cases that I will focus on.

A case of wishful thinking can be found in the familiar tale of Charlie Brown and the football. At the beginning of every football season, Charlie Brown tries to kick the football that Lucy was holding, and every time, Lucy pulls it away.[2] Let me take a little liberty in filling in Charlie Brown's psychological states here. Each time Charlie Brown is standing in front of Lucy, he has to choose whether to try and kick the football, or whether to walk away.

---

[2] This example was suggested by Elizabeth Anderson. The football gag appeared 40 times in Schulz's Charlie Brown comic strips between 1952 and 1999.

Before he is standing in front of Lucy, Charlie Brown swears to himself that he won't try to kick the football this time, because he *knows* that Lucy will pull it away. After each attempt, Charlie Brown swears never to try and kick it again. But when Charlie Brown is standing right in front of Lucy, he believes that this time he will be able to kick the football. In believing this Charlie Brown is engaging in wishful thinking. If we look at Charlie Brown's belief that he will be able to kick the football over time, it fluctuates—when he is not standing in front of Lucy he places little credence in the belief 'I will be able to kick the football when Lucy is holding it', while he is in front of her places a lot of credence in that belief. His belief that he will not be able to kick the football behaves in the opposite manner. The relative strength of Charlie Brown's beliefs over time look like this:

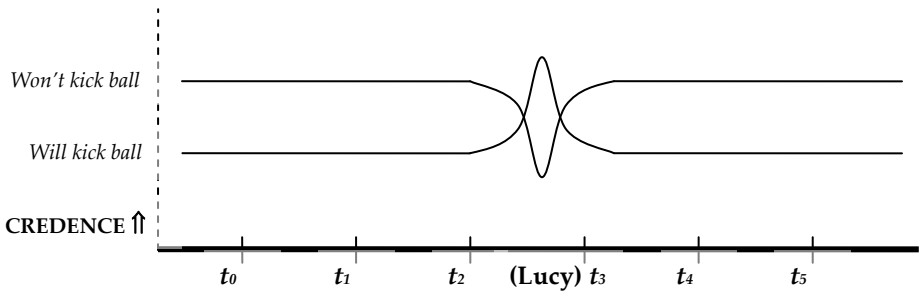

Figure 2: The case of Wishful Thinking

The credence that Charlie Brown places in the belief that he can kick the football when Lucy is holding it changes when a certain state of affairs is true of him, the state of affairs in question being standing in front of Lucy. There is something fishy about what is going with Charlie Brown's beliefs here. The change in his beliefs is caused by an event that has nothing to do with the *truth* of his beliefs. Intuitively, whether or not he is standing in front of Lucy at this particular point in time will not by itself cause it to be more or less likely that Lucy will pull the football away when he is standing in front of her. And it seems that the credence that he places in the proposition 'I will kick the football when Lucy is holding it' should only be responsive to information about the *likelihood* of him kicking the ball when Lucy is holding it,

which is in turn dependant on the likelihood of Lucy pulling the football away. Because the information that he is presently in front of Lucy is evidentially irrelevant to Charlie Brown's belief that he can kick the football when she is holding it, it should have no effect on the credence that he places in this belief, yet it clearly does. I will call such a belief change 'unwarranted'. When changes in beliefs are warranted, then, they will be caused by some change in the world which is *evidentially relevant* to the truth of the proposition.

I propose that we can generalize the type of stability story that I have given for beliefs to all propositional attitudes that have similar attributes, including desires. Schematically, the stability of a propositional attitude will be a function of the propensity of that attitude to change under conditions conducive to change in that attitude. So, we have some propositional attitude such as a belief, desire, intention, or hope. To be susceptible to a stability story the propositional attitude must have some measure of strength, and some identifiable class of events that are capable of changing the attitude, a 'change event'. The measure of strength, and particular change events, will vary with the attitude in question. Strength for beliefs is the credence which the agent has for the truth of a proposition, while strength for desires is the desire's measure of motivational force. [3] For beliefs, the primary change event is learning new evidence. Learning new evidence will also be a change event for other propositional attitudes insofar as they depend on beliefs. Reflection is another type of change event for propositional attitudes. Now, for each distinct type change event for a kind of propositional attitude, we can construct a measure of stability for that propositional attitude. That measure of stability captures the resilience of any particular state with respect to a particular change event. So, '*A's* propositional attitude toward *x* is resilient with respect to change event *e* just in case *e's* occurrence does not change *A's* propositional attitude toward *x*. We can then talk about such

[3] This is a view proposed by (Davidson, 1980; Schroeder, 2004; Smith, 1998) among others.

stability comparatively, with respect to either single change events or various sets of change events.

I propose that there are two types of change events for desires, and thus two types of stability. The first is stability with respect to new information, which is a property of desires that depend upon other propositional attitudes (such as beliefs). The second is stability with respect to reflection, which is a property of all desires. I will discuss these in turn.

We can get some sense of how many desires are dependent upon other propositions (and thus capable of being more or less stable with respect to new information), by considering the decision theoretic distinction between epistemically basic and non-basic desires. A basic desire is a desire for which there is no proposition such that were the agent to learn it to be true, her desire would change. A desire for *p* is epistemically basic just in case the agent's desire for *p* does not depend on any of her beliefs. We can understand the dependence of a desire *p* on the truth of some proposition *q* in terms of how the truth of *q* affects the desirability of *p*. If you have an **epistemically basic** desire for *p*, then (*p* & *q*) will be *just as* desirable to you as (*p* & ~*q*), both of which will be just as desirable to you as *p*.[4] In contrast, if you have an **epistemically non-basic** desire for *p* (which is dependent on *q*), then (*p* & *q*) will be *more (or less)* desirable to you than (*p* & ~*q*).[5]

Only highly specific propositions (such as propositions which describe whole possible worlds) will count as basic desires using this definition. The main property of basic desires that

---

[4] This account of the basic-ness of desires is drawn from standard decision theoretic accounts such as those given by James M. Joyce. (Joyce, 1999)

[5] A larger class of desires than basic desires is the class of practically basic desires. Practically basic desires are not actually basic because there are in principle certain things that you could learn which would cause you to change your desire. But such desires for p are practically basic because for any q that you have a realistic probability of learning, you are indifferent among the desirability of p, (p & q), and (p & ~q).

is of interest to us here is that basic desires so defined will be stable no matter what you learn.[6] New information will not be a change event for such basic desires, although it will be for all other desires. Given that it is likely that no or few such basic desires exist, it is the case that most desires will be dependent upon other propositions. These desires will then be responsive to evidence regarding the truth of those propositions on which they depend. Thus new information will be a change event for the vast majority of desires.

Common instances of epistemically non-basic desired are desires for a means to some end. For instance, I desire a hammer as a means to drive in a nail. If I learn that rubber duckies are the best tool for the job, then my desire for a hammer (given that it depends on my believing it to be the case that 'hammers are the best tools for driving in nails'), will change.[7]

Now, considering new information as a change event for desires we can construct a measure of resilience (and thus a measure of stability) for desires. Resilience, as in the belief case, will not be a property of individual desires; rather it will be a two-place relation between particular desires, and specific pieces of new information (evidence). As the stability of a belief is a function of its resilience with respect to new information, the stability of a desire will also

---

[6] If I have a basic desire, and I learn something, the fact of my learning that thing will not change my desire. It will have no evidential impact on my desire. However, there is another way in which learning some thing may change my desires, even if they are basic. This is the case in which learning some thing has a causal impact on my desire. For instance, I have a basic desire for chocolate cake, and learning that my nana has died causes me not to desire chocolate cake any more. But learning this information changes my desire through brute causal force, as it were, not through any process of reasoning. So the claim is not that learning new information can never change a basic desire, but rather that learning new information cannot change a basic desire in virtue of its evidential impact.

[7] There are several common types of information that epistemically non-basic are often responsive to. Another type of information on which desires commonly depend is information about what the world would be like if the desire were fulfilled. Desires for certain experiences will be responsive to information about what the experience will be like. If I learn that the chocolate cake that looks so delicious has been baked with salt rather than sugar, then I will no longer desire that cake, as I (correctly) infer that the experience of eating the cake will not be that experience in virtue of which I generally desire to eat chocolate cake. The second is information about particular experiences. One of the reasons I desire to engage in certain activities is because I have certain expectations about what that experience will be like. Information that I learn now about what that experience will be like for me is relevant to my desire in that in (at least) some cases, it seems that my desire should change if I learn that my expectations about the experience are mistaken.

be a function of its resilience with respect to new information. The level of stability of a particular desire can be expressed in terms of the sets of evidence that the desire is resilient with respect to. So, 'A's desire for a state of affairs *x* is resilient with respect to some set of evidence *e* just in case learning *e* will not change the desire that *A* has for *x*.' Thus the stability of a desire is a function of how likely that desire is to change with respect to learning new information. Stability for desires, like stability for beliefs, is essentially a relative notion.[8]

As desires differ in the extent to which they depend on the truth of other propositions, so they will differ in how susceptible they are to change as a response to new evidence. Using these differences, we can construct a hierarchy of desires. At the most stable end of this hierarchy will be basic desires. At the least stable end of this hierarchy are those desires which depend on beliefs which can be undermined by learning any new information at all, where that new information is likely to be learned by the agent. The majority of desires, however, will fall somewhere between these two extremes. Thus epistemically non-basic desires have the structure that, like beliefs, they may be compared in terms of how resilient they are with respect to new information.

The second type of stability that I will propose is the stability of desires with respect to reflection. Reflection is a change event for desires because we are agents with multiple concerns and aims, and our various desires do not exist in isolation from one another. In order to act in any principled way at all, we must seek to balance this plurality of ends. An agent

---

[8] As desires differ in the extent to which they depend on the truth of other propositions, so they will differ in how susceptible they are to change as a response to evidence. Using these differences, we can construct a hierarchy of desires. At the most stable end of this hierarchy is what we might call practically basic desires. Practically basic desires are not actually basic as there are in principle certain things that you could learn which would cause you to change your desire. However such desires for p are practically basic in that for any q that you have a realistic probability of learning, you are indifferent among the desirability of p, (p & q), and (p & ~q). At the least stable end of this hierarchy are those desires which depend on beliefs which can be undermined by learning a wide range of new information, where that new information is likely to be learned by the agent. Most desires will fall somewhere between these two extremes.

who frequently acts on the impulses of the moment, rather than reflecting upon her corpus of beliefs and desires and determining which of her desires are most important, will, by her own lights, do badly at satisfying those desires. An agent requires a certain amount of consistency amongst her desires if she is ever to be successful in achieving her ends at a time, let alone in achieving her ends over time. When such consistency is achieved by an agent, it is done so through reflection. The experience of desiring two contradictory things, reflecting upon them and establishing which should be pursued and which foregone is far too common an experience to deny. I may have a certain desire right now, but if I were to reflect upon the corpus of my beliefs and desires, this desire may strengthen or weaken as a result of my reflection.[9]

As with the case of the resilience of desires with respect to new information, we can construct a measure of resilience (and thus a measure of stability) for desires with respect to reflection. Each act of reflection can be defined in terms of the information which is considered in that act. The set of possible information for each act of reflection is the full corpus of the agent's beliefs and desires, and each possible combination of this information constitutes a distinct act of reflection. The level of stability of a particular desire can be expressed in terms of the acts of reflection that the desire is resilient with respect to. So, 'A's desire for a state of affairs x is resilient with respect to some act of reflection r just in case performing r will not change the strength of the desire that A has for x.' Thus the stability of a desire with respect to reflection is a function of how likely it is to change under reflection.

The type of reflection that I am considering is rational reflection. I take it that reflection is capable of producing change in desires in two ways. One of these is purely causal. If I reflect on some horrible experience, and the trauma of this reflection causes me to abandon

[9] Note that this is a distinct change event from new information, as the agent does not require any external input to change her desires through this method. Reflection may change desires even in the absence of any newly learnt or discovered evidence.

all my desires through some sort of psychological schism, then the mechanism of this change is not the content of the reflection—my reasons—but rather it is the causal power of the fact of the reflection that causes my desire to change. In contrast, the type of reflection that I am concerned with here is rational reflection, which causes desires to change in virtue of the conclusion reached in that act of reflection.

Achieving consistency among desires through reflection is a matter of retaining those desires that are consistent with one another, and discarding those that are not. However, given that there is no fact of the matter about which desire should be discarded in such a conflict and which desire retained, each agent must establish a rough hierarchy of importance amongst her desires in order to solve this question.[10] The view that such hierarchies exist stems from the idea that there is no 'view from nowhere' from which we can reflect on our desires. In the process of reflection we treat certain desires as fundamental, those which are the starting points for our deliberation. In cases of conflict amongst desires, fundamental desires are those that cause changes in competing desires, rather than being changed themselves. Using the propensity of different desires to change under reflection, we can create a hierarchy of reflective basic-ness for desires. The more susceptible a desire is to change through reflection, the less reflectively basic it is. Reflection is not a change event for a fundamental or reflectively basic desire in such a hierarchy. Thus an agent's corpus of desires at any given time can be organized into a hierarchy of reflective basic-ness. Just as in the belief case, where desires lie in this hierarchy of importance has implications for their stability. Desires at the

[10] In saying this I do not want to commit myself to the view that all desires are comparable. I take it that an agent who has pluralistic concerns that cannot be reduced to a dimension of comparison will construct multiple such hierarchies, and achieve consistency in this way. Nor do I want to imply that is seeking such consistency an agent only succeeds when she establishes a complete ordering of her desires, I take it that in many cases of reflection very rough hierarchies are sufficient.

lower end of this hierarchy will be unstable with respect to reflection, while desires at the higher end will be stable with respect to reflection.[11]

Now, paying attention to the stability features of desires not only gives us another way of talking about them, but one which can begin to make some sense of why temptations are different to other desires. Both instrumental and intrinsic desires can be more or less stable. The instability if instrumental desires can be understood in terms of changes in the cognitive structure which supports such desires. If my desire for *P*, depends upon my belief that *Q*, and my belief in *Q* is unstable, it will follow that my desire for *P* is unstable with respect to new information. Suppose that my belief in *Q* is actually an instance of wishful thinking, where the change in *Q* is triggered by the proximity of some particular thing. In this case my desire for *P* will demonstrate the same shift in intensity as Kim's desire for cake in fig. 1. But this cannot be the whole story of temptation, as mere instability with respect to new information does not make a desire a temptation (although it may be a clue that there is something fishy about that desire). Temptation usually takes place in the absence of new information about the about the object of one's desires. For instance, the mechanism of Kim's temptation seems to be a shift in the strength of her intrinsic, rather than instrumental desires.

In the case of Kim, the fluctuation of the felt intensity of her desire to eat the cake (her emotional experience of desiring) is not a response to changes in the properties in virtue of which she finds the cake desirable. The strength of Kim's desire for the cake is changing in

---

[11] Stability with respect to reflection, like stability with respect to new information, is essentially a relative notion. The stability of a desire under reflection is a measure of its resilience compared to the resilience of another desire. It tells us which of the desires presently to hand is more likely to change if the agent were to examine them in light of her full corpus of beliefs and desires. Desires with no resilience under reflection at all will be changed by the most cursory act of reflection. This is in contrast to desires with maximal resilience with respect to reflection, which will not change under reflection no matter how rigorous the reflection that the agent subjects them to. However the stability of desires comes in degrees. The vast majority of desires will lie somewhere between these two extremes.

response to the mere proximity of the cake, rather than any change in either her certainty about receiving the cake, or the relevant properties of the cake. Like the case of wishful thinking, this instability in her desire for the cake is not moving Kim toward a better grasp of the underlying goodness of the outcome for her. Instead it appears to be a shift in her intrinsic desire for the cake, a shift that is not motivated by any changes in the properties of the cake in virtue of which she finds it desirable. By taking these stability properties seriously we can treat desires as being more complex than is allowed by the standard, simplistic view of desires. Moreover this additional property gives us an analysis of what is going on in the case of temptation, which is a first step to answering the challenge of temptation.

The Structure of the Dissertation

The challenge of temptation is problematic for current theories of desire, especially those views of desire which treat desires as simply having an object and a measure of motivational strength. The tricky case, and the subject of this dissertation, is the case of temptation which is generated by the experience of feeling an urge for some thing. I propose that this is a case which can only be understood if we first understand how the emotional components of desires work. Thus a view of desire capable of answering the challenge of temptation needs to address not only how these emotional components of desire generate temptations, but why they should not be treated on a par with other desires in deliberation. The bulk of the dissertation is an attempt to establish an account of what these emotional components of desire are.

In the next two chapters I consider various views of desire, and how they accommodate both stability and temptation. In the second chapter, I consider two views of desire that treat desires as various species of cognitive, rather than conative, attitudes. The first is the view that desires are a species of judgment, which is the view of 'Judgmentalism' about desire. The second is Scanlon's argument that desires should be understood as a type of

reason. I argue that both of these views fail the challenge of temptation in the same way—they ultimately hold both that desires which the agent does not endorse (i.e. conflict with longer term, weaker although plausibly more important aims, for instance, temptations) are simply not desires. Judgmentalism cannot answer the challenge of temptation because only internally coherent desires turn out to count as desires, so they deny the existence of the phenomena. Scanlon's reasons view similarly reduces to the claim that we can only desire that which is, in a relatively robust sense, consistent with our other views. Thus these approaches only solve the challenge of temptation in an uninformative way, by eliminating temptations, and thus the challenge of temptation, by fiat. This does not respect the robustness of experiences of temptation.

In the third chapter I consider three conative views of desire; two of them familiar, and one less so. In historical and contemporary philosophical, psychological, and neuro-physiological discussions of desire, desires *qua* conative states have been presented in three distinct ways. These are the three faces of desire: motivation, pleasure, and reward. The first two of these inform various skeletal and widely used concepts of desire such as: 'mental states capable of motivating action'; or 'positive affect toward some end', which capture the Motivational Theory of Desire, and the Hedonic Theory of Desire respectively. The third view, that of desire *qua* reward, is championed by Timothy Schroeder in "The Three Faces of Desire". This is the Reward Theory of Desire.

Neither the Motivational nor Hedonic theories of desire can answer the challenge of temptation because temptations, in terms of the relevant characteristics of motivation and pleasure, look just like other desires. The reward view is plausible in many ways, but it still cannot answer the challenge of temptation. Representations of tempting things produce reward signals in the same way as other desires, so this view alone cannot distinguish temptations from other desires.

I propose that what is missing from the views of desire canvassed in chapters two and three is an explicit recognition of the fact that our desires are intimately connected to certain kinds of emotional responses. It is the nature and origin of these emotional responses of desiring, that both cause temptations, and explain why it is that they should not be treated on a par with other desires in rational deliberation. Proving this point is the aim of the fourth and fifth chapters.

In the fourth chapter I argue that not only do these emotional components of desire exist, but that they affect our desires, and can be manipulated in certain ways. I give a basic outline of what this felt emotional component of desires looks like, and defend its psychological plausibility. These emotional components of desires are the characteristic phenomenological corollaries of desires which are a type of automatic affective responses, 'desire responses'. I contrast these desire responses with what I call the 'considered preferences' of the agent. A considered preference is a full-fledged desire that is stable with respect to minimal reflection, over time, and across circumstances.[12] These requirements of minimal stability are meant to exclude desires which are degenerate in a variety of ways. As a species of automatic affective responses, desire responses are produced by some system in the agent which processes such representational and affective responses, which I call the conative system.

I then give an account of how such emotional responses work, and what constraints we can place on them by employing a loose analogy with D'Arms and Jacobson's response-dependent theory of emotion. The constraint that I will propose for the emotional response of desiring is that such responses can more or less 'fitting' in the same way that emotions can be more or less fitting according to D'Arms and Jacobson. What it is for an emotion to be fitting,

---

[12] By minimal reflection, I mean a fairly cursory consideration of the desire in light of your beliefs and other desires.

then, is that it is the *appropriate* response to the relevant property. That is, the response gets the property right. What it is for the response to get the property right is that the response tracks the norm established by the responses of normal subject under normal conditions. A desire response that does not 'fit' in this sense is mistaken in some identifiable way—it fails to reflect that which it is supposed to reflect.

Thus I argue that what it is for such a response to be mistaken is that it fails to track the 'normal' response for that agent, a norm which is established by appealing to both the 'normal' desire responses of the agent—those responses that she has in the absence of any interference—and the vanishing point desires of that agent. Considering the connection between our desires and these emotional components of desire I argue yields an account of what causes temptations, and why they should not be treated on a par with other desires in deliberation. The hypothesis that I propose about the abnormal conditions for operation for the conative system is that proximity undermines the fit of our desire responses in the same way that distance undermines our visual perceptions of relative size. The desire response that results from such a manipulation is flawed, and should be discounted in deliberation, which explains why temptations should not be treated the same as other desires in rational deliberation.

Now, the real challenge of temptation is not that we sometimes have an intense emotional experience of desiring some thing when it is proximal, even if that thing is contrary to our longer term aims. But rather that it is not obvious to us at the moment of desiring that satisfy this urge is contrary to our considered preferences. Temptation is not merely an emotional overlay of preferences that can easily be dismissed when it leads us astray, it also misleads us about our considered preferences are by undermining our ability to effectively reflect on what they actually are. All temptations have in common the effect of distorting the agent's reflections on her considered preferences. What we need is an explanation of how it is

that such proximal desires interfere with an agent's ability to reflect on what her considered preferences actually are, as well as an account of why it is that the proximate desires which count as temptations should be discounted in deliberation. This is the subject of chapter five.

In the fifth chapter I fill out the argument proposed in the fourth chapter by identifying why it is that the conative system reliably produces mistaken desire responses under the conditions of proximity, and how it is that proximity interferes with our desire responses. Temptation, I argue, is best understood as a consequence of a pervasive natural tendency to discount the future hyperbolically. I rely on George Ainslie's seminal work in establishing that the default rate at which we discount the future is hyperbolic. If you discount the future hyperbolically, then the mere proximity of the good causes the comparative strength of your desire responses to reverse. This reversal occurs, I argue, because proximity constitutes abnormal operating conditions for the conative system, thus producing unreliable desire responses. These responses are unreliable because they are the result of the inappropriate intensity of the emotional response of desiring.

Our desires are (partially) caused by the properties in virtue of which we take the objects of our desire to be desirable. These properties not only cause in us 'full-fledged' desires that we reflect upon and endorse (considered preferences), but also automatic affective responses that constitute the emotional experience of desiring that thing (desire responses). Both of these states have a measure of motivational strength, and we tend to assume that they track one another. Thus our motivations can be supplied either by our considered preferences or our desire responses. However when these two measure of motivational strength come apart, we must reconcile them. Hyperbolic discounting is just such a case when these two measures of motivational strength come apart, and it is through this mechanism that hyperbolic discounting creates temptations.

I argue that, in the case where these two measures of motivational strength come apart, we aim to achieve a loose reflective equilibrium between the strength of considered preferences and the strength of desire responses. So, there is a range of conditions under which I desire the cake, and I desire it in virtue of its desirable features. The automatic affective response that is the desire response also occurs over a range of conditions. Both considered preferences and desire responses are in part produced by a common cause—the properties in virtue of which the thing is desirable to you. However, we tend to encounter difficulties when these two conditions come apart. If one of these measures of motivation is mistaken, then it should be discounted in achieving this equilibrium. I propose that the mistaken desire responses should be discounted in achieving reflective equilibrium between the motivational strength of desire responses and the motivational strength of considered preferences, on the grounds that it results from the conative system operating under abnormal circumstances.

The final element of my argument is to propose that the stability of desires gives the agent indirect but reliable access to the strength of her considered preferences even when she is in the grip of temptation. We can appeal to the comparative stability of our motivations in order to identify flawed desire responses and discount them in deliberation, thus giving an indirect way for the agent to identify, and therefore resist, temptations. I propose that appealing to the stability of desires will give us an account of what precisely it is that is wrong with Gauthier's proximal preferences (desire responses), which in turn will generate an account of why they should not be treated the same as other desires in rational deliberation. Gauthier, however, leaves it a mystery why we have these proximal shifts in our preferences, and does not address the question of precisely what it is that is wrong with them.

Hyperbolic discounting shows why it is that the conative system yields felt emotional components of desiring which are not fitting under conditions of proximity. But more than

this, it shows how pervasive mistakes in intensity occur in the conative system in the formation of proximal desires that count as temptations. Isolating the mechanism of hyperbolic discounting thus explains what causes these proximal shifts in our desires in temptation cases, and why they are mistaken.

However, the underlying aim of the dissertation is not simply to argue for the particular answer to the challenge of temptation that I defend, but rather to demonstrate that the challenge of temptation requires us to treat desires as far more complex states than we generally do. Indeed, I take the main contribution of this dissertation to be the characteristics of desires that I identify, analyze, and defend in the course of giving this response to the challenge of temptation. Specifically: the account of how desires may be more or less stable with respect to reflection and new information; the role of the emotional component of desires, desire responses, in variously forming and undermining our reflectively stable desires (our considered preferences); the uneasy relationship between how it is that we are psychologically disposed to value goods over time (hyperbolically), and the way in which it is rational to value goods over time (exponentially); and the account of the difference between warranted and unwarranted changes in the strength of desires.

CHAPTER 2

THE METAPHYSICS OF DESIRE

This chapter is a preliminary investigation what kinds of theories of desire can both answer the challenge of temptation, and accommodate the stability properties of desires. Its aim is to disprove a tempting general thesis about desires, the view that desires can be reduced to some other mental phenomena. In this chapter I will consider two such reductive theses. The first is what I will term 'Judgmentalism' about desire, which is the view that desires are a type of judgment. The second is the view that Scanlon proposes in 'What We Owe to Each Other', which identifies desires with a type of reason. Both of these views are in the cognitivist tradition, in that they are aiming to explain desires in terms of a cognitive, rather than conative, attitude. Considering such views is necessary for my overall project because if a view of this cognitivist sort is correct, then how it is that desires may be more or less stable is not an independently interesting question, rather the stability properties of desires will be a matter of the stability properties of these other states. There are various possible views about desire which explain desires in terms of other mental states such as beliefs, judgments, or reasons. In this chapter I will argue against two of the most prominent examples of such views. Ultimately I will argue that neither is a satisfactory alternative to a broadly Humean (conative) conception of desires, in part because these views cannot address the problem of

temptation. Indeed, I will argue that such views eliminate the challenge of temptation through the somewhat implausible mechanism of excluding temptations from the class of desires.[13]

Judgmentalism

The first possibility that I will address is the view that desires are judgments, which I will call Judgmentalism. To be a judgmentalist about desire is to hold that all desires are judgments, rather than some simpler state. There is more than one way of understanding judgments about desires, and thus more than one version of Judgmentalism. I will consider two types of Judgmentalism, each of which depends upon a different way of understanding judgments about desires. To begin with I need to explicate the nature of judgments, as this is necessary to fully grasp what it means to say that desires just are judgments.

Within the generic category of 'judgments', there are two much discussed subsidiary forms, theoretical judgments and practical judgments. Broadly speaking theoretical judgments are taken to be judgments about matters of fact, while practical judgments are taken to be judgments of an agent regarding how they ought/will/should act. For those who hold that practical judgments are a distinct type of judgment, the mark of a practical, as opposed to theoretical, judgment, is that it should somehow cause (or at least be capable of causing), actions. Theoretical judgments, on the other hand, are taken to be motivationally inert by those whose accept the existence of practical judgments.[14] The relationship between these two

---

[13] Note that a more traditional cognitivist view of desires, like the rationalist view attributed to Kant, is not contrary to my project as it still requires an account of intrinsic desire that can make sense of temptation. This view "…sees reason as, in addition to its instrumental function, capable of in some sense determining the agent's ends independently of desire and in some cases of motivating the agent to act independently of, and potentially contrary to, the ends given by the "lower faculty" of desire." (Hurley, 1989) There are, then, two different sources of or grounds for practical reason on this account, one in desire and the other in reason itself." Temptations are ends that arise in the 'lower faculty' which produces intrinsic desires, and the Kantian does not want to treat all intrinsic desires as they do temptations. Thus the Kantian still requires an account of how it is that temptations differ from other intrinsic desires.

[14] "Practical reasoning in this more or less technical sense leads to (or modifies) intentions, plans, and decisions. Theoretical reasoning in the corresponding technical sense leads to (or modifies) beliefs and

forms of judgment is controversial. What is readily agreed by all is that there is such a thing as epistemological or theoretical judgment, so I shall take epistemological judgment as the model for judgments.

Svavarsdottir gives an account of the standard epistemological usage of the term 'judgment' that I will use as a framework for discussion:

> "[The term] 'Judgment' is standardly used to designate a mental event (a cognitive act) closely related to the cognitive state of believing something: the belief that such and such is a state that grounds the disposition to judge such and such. When judging such and such, an agent is affirming, in thought or language, that such and such is the case."[15]

This standard conception of judgment is composed of two elements. One is a belief, while the other is a commitment to, or endorsement of, the good epistemological standing of that belief. In the interests of simplicity I will analyze judgments by postulating the fewest possible elements. Given the general understanding of mental states that we are working with—propositional attitudes— I will in the first instance treat all of the elements of judgments as propositional attitudes. For notational purposes, I will express attitudes in all capital letters (ATTITUDE), and indicate the propositional content of the attitude with angular brackets (<that *p*>). Now, it is clear that one of the elements of an epistemological judgment is a belief. What sets epistemic judgments apart from other beliefs is an explicit element of endorsement of the epistemic status of that belief. The question is how we should understand this second endorsement element of the judgment.

The candidates for this element of endorsement are those mental states which are taken to be elements of reasoning broadly construed, beliefs and desires. The distinction between beliefs and desires is standardly considered to be one between cognitive and conative attitudes. Cognitive attitudes are those whose epistemic status depends on such and such being

---

expectations." Here Harman differentiates between the two types of reasoning (and hence the two types of judgment), on the basis of the content and function of that reasoning. (Harman, 2004, p. 45.)

[15] Svavarsdóttir, 2006.

the case. Conative attitudes are those desires, hopes, wishes etc. that express the way that the agent wants the world to be. The endorsement element of a judgment is something in the territory of an affirmation of the content of the judgment. The question is whether this affirmation takes a cognitive or conative form. Conative states reflect the way that the agent wants the world to be, rather than the way that they take the world, or their beliefs, to be. Cognitive states reflect the way that the agent takes the world, and her beliefs, to be. If we take endorsements to be conative, then the endorsement element of the epistemological judgment will look something like: WANT/HOPE/WISH/DESIRE <to BELIEVE <that *p*>>. This clearly does not capture the endorsement element of epistemic judgments, as it expresses something about the way that the agent wants the world to be, rather than the way that she takes the world to be. And a prerequisite for endorsing some thing the way that it is, is that you take the way that it is to be a part of the way that the world is. Thus conative states are not plausible candidates for the endorsement role. This leaves cognitive states to fulfill the role of endorsement in judgments.

I propose that we can capture this endorsement element through a second belief, thus modeling epistemic judgments as a pair of beliefs. This second belief is a belief about the question of whether or not your evidence is sufficient to support your belief. Specifically, it is a belief that the credence that you place in *p* fits your evidence that *p*. Note that the picture of judgment that I am proposing here is an explicitly internalist one. The agent judges only in those cases where she has access to the evidence for her belief that *p*, and believes that her credence in *p* fits this evidence. What makes this an internalist picture is that I am proposing that the agent must have access to the evidence for her belief in *p* in order to judge that *p*. This picture satisfies the ideal of ontological minimalism in that it does not posit any entities other than those widely accepted.

There are a variety of mental phenomena in the immediate locale of judgments that must be kept separate. These are: the act of judging; the mental state of judgment; and the process of reasoning that leads to judging. The mental state of judgment is what we get after an act of judgment. The act of judgment bridges the gap between reasoning, and having a (mental state of) judgment. Reasoning leads to judgment, although it is not the same as judgment. In the clearest cases, all three of these elements work together, while remaining conceptually separate. I will focus on the relationship between reasoning and judging as it highlights an important characteristic of judgment.

So, you have an epistemological judgment that *p* when you believe that *p*, and you believe that the credence that you place in *p* fits your evidence that *p*. One thing that you do in judging that p is that you make and an implicit connection, between believing that *p* and endorsing the BELIEF <that *p*>, explicit. This explicit endorsement is generated by reasoning your way to a judgment. Now, as a rational believer, in believing that *p* to degree .7, I implicitly commit myself to believing that this is the degree of belief which is supported by my evidence as if I believed that a .7 credence in *p* was *not* supported by my evidence, then I would not believe *p* to degree .7. In the case of garden variety beliefs such an endorsement is part and parcel of having a belief.

In the case of mere belief, it is possible (if incoherent) for an agent to hold a belief without an endorsement, however an endorsement is necessary for a judgment. So what distinguishes judgment from mere belief is that the existence of this endorsing belief is guaranteed. Note that I am not claiming that this sort of endorsement is only present in cases of judgment, rather that this endorsement is *explicit* in cases of judgment.

Thus we have a model of judgment where what it is to judge that *p*, is to: BELIEVE <that *p*>, and BELIEVE <that the credence that you place in *p* fits your evidence that *p*.> Less formally we can understand this endorsement element of a judgment as an evaluative

proposition of the form: this is the right belief to hold in light of the evidence. Now that we have some sense of the nature of judgments in the clearest case of epistemic judgments, we can consider what judgments capable of explaining desires might look like.

Desires as Judgments

The aim of this part of the discussion is to show that a theory of desire which takes desires to be a species of judgment cannot answer the challenge of temptation. I will argue that such a view must hold that temptations are not desires at all, as they are not the kind of states that tend to be endorsed by the agent. This, I propose, is more a case of ignoring a problem than a plausible explanation of why it is not a problem at all. A plausible view of desires must encompass an account of temptations as a type of flawed desires, rather than merely deeming them not to be desires by fiat. So my agenda is to argue against all forms of Judgmentalism.

There is two ways of understanding judgments about desires, and thus two types of Judgmentalism. One form of Judgmentalism is phenomena driven, in that it moves from a natural way to understand judgments about desires, to the claim that desires just are this type of judgment. The second form of Judgmentalism is theory driven in that it begins with the aim of giving a certain type of account of desires, which entails that judgments about desires be understood in a certain way. I will consider these in turn.

Humean Judgmentalism

Here is a picture about judgments about desires that I think is correct. This is the picture there are desires broadly understood as conative states that have the ability to motivate, whose object can also be the object of beliefs such as the belief that this thing is desirable. Call desires construed as propositional attitudes capable of motivation *Humean desires* to distinguish them from various other models of desire. Given the presence of such states, it is pretty straightforward to give an account of how judgment works in the case of

such desires that is similar to the way in which judgment works in the case of belief. In the belief case, we move from a belief to an epistemic judgment by adding a further belief which is an affirmation of the good epistemic standing of the initial belief. In the case of Humean desires is it easy to begin with such a desire, and move to an analogous state of conative judgment by adding a belief which is an affirmation of the desire. An endorsement of a desire of the kind required by judging cannot merely be the affirmation that you possess the desire. Rather, it needs to say something about the 'good standing' of this desire. I take it that what it is for a desire to be in good standing is for its object to be desirable in light of your reasons, or some similar formulation.

Let us call judgments whose contents are such attitudes 'conative judgments'. The content of the conative judgment would be the object of the associated desire. The endorsement element will be an affirmation of the 'correctness' of this desire in the form of a belief about the desirability of that thing: BELIEF <*p* is desirable>. What it is for something to be desirable in this sense is something like it being the right thing for you to want in light of your reasons. It is, in a sense, natural to understand judgments about desires as conative judgments, as conative judgments have a natural source of motivation. This is, in a nutshell, the advantage of the Humean approach.

I propose that Humeans, very broadly construed, can and should accept conative judgments as a model of how *judgments* about desires work. The view that we can (and clearly do), make judgments about our desires in very basic cases such as deciding that a particular desires is the right one to act on, is a natural extension of any view of desire. A nice feature of such conative judgments for my view is that the endorsement element of the desires makes explicit the agent's reasons for favoring that desire over other unendorsed desires. This awareness of the grounds of a desire is going to tend to make desires that are the objects of conative judgments more stable than simple desires.

Now, insofar as this is simply an account of how judgments about desires work I have no objections to it. What is contentious is the further, judgmentalist, claim that desires *just are* in some necessary sense, such conative judgments. Having admitted the existence of standard desires, the only way that this could function as a theory of all desires is if it makes the further claim that the basic propositional attitude of desire is *always* accompanied by the endorsement. Now, for the rational agent, although beliefs and epistemic endorsements are not the same thing they tend to go together. Even in the case of garden variety beliefs, epistemic endorsements are implicit. However there is no such natural tendency for desires that *p* to be so closely identified with beliefs that *p* is actually desirable.

The claim that desiring and beliefs about what it is correct to desire always go together is plausible only if we hold that you should only desire that which you believe to be desirable. It seems plausible to say that this is true of some of our desires. I, for instance, take it to be true that I desire not to torture kittens because given my reasons torturing kittens is undesirable. But in many other cases of desire, there is no such harmony between my desires, and my beliefs about what the right thing to desire given my reasons is. I desire to drink red wine, but given my plans for tomorrow morning, my beliefs about disrupted sleep, and my prediction of the wine causing a headache I do not think that this is the right thing to desire in light of my reasons. My desire for wine is incontinent in the sense that it contradicts my reasons, but it seems to be a desire nonetheless. The one proposing that all desires are conative judgments must hold that *all* desires are of this former type – because in making a conative judgment, the agent must believe that the object of her desire is that which is the right thing to desire in light of her reasons. The implication of this view is that it is impossible for an agent to desire that which she does not take to be fully rational. This claim of constant companionship between desiring and endorsing is not necessary if we simply hold that conative judgments are one of the places that desires figure in our mental economy, which is

an eminently plausible position. Thus although the view that conative judgments are desires is at best a psychologically implausible view of *all* desires, it is a plausible view of the subset of *rational* desires.

Cognitive Judgmentalism

The second version of Judgmentalism can informatively be called cognitivist Judgmentalism. As in the case of Humean Judgmentalism, this second version depends on a particular account of judgments about desires, in this case, a cognitivist account of such judgments. Call such judgments desire-judgments. The motivation for this view is the project of giving a cognitive account of desires. A thought that could motivate such an account is that judgments, in that they have a reflective element, are a more likely candidate for a cognitive account of desire that beliefs, as beliefs have very little chance of accounting for the motivational capacity of desires'.

Now, as this is a cognitive account of desires, the desire-judgments cannot include conative elements, so the judgments in question are not conative judgments. To satisfy the cognitive constraint both elements of the judgment—the content and the endorsement—must be cognitive states. An intuitive place to start is with beliefs about the good. After all, it is not completely counter-intuitive to say that a way of understanding what it is to desire some outcome is to believe it to be good. If we use this as the content of a judgment, then we end up with a species of epistemic judgment, judgments about the good. The content of such a judgment would look like: BELIEF <*p* is good>. In the case of epistemic judgment, the endorsement element follows the form: BELIEF <that the credence that you place in *p* fits your evidence that *p*.> If we take desire-judgments to be a straightforward species of epistemic judgment then the endorsement element would be essentially the same. Thus we can

understand desire-judgments as: BELIEF <*p* is good> and BELIEF <that the credence that you place in the goodness of *p* fits your evidence for *p* being good.>

The problem with such a formulation is that it does not capture one of the two important characteristics of desires. It is generally assumed that desires fulfill the functional role of motivating and justifying action in our conception of practical reason. Thus a belief that plausibly fulfills the desire role would need to have some pretension to motivational efficacy. It would also need to be able to explain or justify action. To fail to fulfill either of these desiderata would make it a very implausible theory of desire. Beliefs about the good satisfy this second criterion, as beliefs about the good could justify action in the following way: If I am going to a movie and you ask me why I am going to that particular movie, if I respond that I believe it to be good, I have in some plausible sense given a *sufficient explanation* of my action. However beliefs about the good have difficulty with the criterion of motivational efficacy. That I believe something to be good simply does not speak to my motivation to do that thing, as my motivation to act in accordance with such belief is arguably external to the belief. I believe that peonies, yogis and chocolate are good, but none of these beliefs necessarily motivate me to act in any way. That I believe that *p* is good tells me nothing about what role this information will play in my deliberative processes. In the case of justification, a belief about the good can play the appropriate role because its connection to action has already been established because the action that needs justifying has already been performed. Justification comes after action, and is linked to that particular action by the professed or imputed states of the agent. Motivation, however, comes before action, and so the connection between a given belief and motivation needs to be internal to that belief. If it is the case that beliefs about the good always carry motivational force, then such a connection is plausible, but this is simply not the case. Thus beliefs about the good alone are not persuasive

candidates for the role of the content of desire-judgments because they do not have a clear connection to motivational force.

What else is there in the cognitivist's armory that might be capable of capturing this missing element of motivation? I take it that the most plausible candidate amongst beliefs is going to be beliefs about reasons. In contrast to the belief that swimming is good, the belief that I have *reason* to swim has a much closer connection to how I should act in light of this information. Thus the most plausible view of the content of desire-judgments will include some recognition of reasons. This recognition could occur in either the content or the endorsement element of the judgment. The content of the desire-judgment could then look something like: BELIEF <I have reason to *p*>.[16] The endorsement element of a desire-judgment would be an analogue of the evaluative proposition that constitutes the endorsement element of the epistemic judgment. In the case of epistemic judgment this is the belief that: and BELIEF <that the credence that you place in *p* fits your evidence that *p*>. The claim that a belief 'fits' evidence is an appeal to the idea that given your evidence, you are justified in placing that level of credence in that belief. There are two parts to your belief, its content, and the credence that you place in it. Such a belief 'fits' the evidence only if it gets both of these elements correct. In the case of desire-judgment, the content is also a belief, so, the endorsement element should be pretty similar. In the case of epistemic judgments, the content of the belief which fits this evidence will be the belief that, given the evidence that you have, is most likely to be true. Thus the endorsement element of a desire-judgment will have the form: BELIEF < that the credence that you place in having reason to *p* fits your evidence for having reason to *p*.>

[16] At this point I am putting aside concerns about just where these reasons will come from if we eliminate the conative view of desires.

Putting these two elements together, we get the picture that a desire-judgment is: BELIEF <I have reason to *p*> and BELIEF < that the credence that I place in having reason to *p* fits my evidence for having reason to *p*.>. So, a desire-judgment is a belief that I have reason to *p*, combined with a belief that this reason to take *p* is correct. The view of cognitivist Judgmentalism is that *all* desires are desire-judgments so understood.[17]

Objections

One immediate disadvantage of the cognitivist judgmentalist view of desires is that, as a cognitive view of desire, it is going to have difficulty satisfying our conative intuitions about desires. Specifically, such a view is going to have difficulty accounting for the motivational aspect of desires. My recognition, however sincere, that I have a reason (even an objectively good reason) to do some thing does not necessitate that I am motivated to do that thing. In proposing an account of desire that appeals only to beliefs about reasons, the view faces the problem that the recognition of reasons does not entail that an agent accepts, or is motivated to comply with, those reasons. Thus a significant cost of understanding desires in terms of beliefs about reasons is that it is not compatible with the view that having a desire to *p* means being motivated to *p*. Of course, a proponent of the view may be willing to bite this bullet; however, such a move is only attractive if there are other benefits to the theory. I propose that these benefits are too elusive to justify this move.

Another problematic aspect of this view is the reliance of the endorsement element of the desire-judgment on beliefs about what it is correct to do given your reasons. I, like many others of my acquaintance, desire things that I firmly believe are *not* the right thing to desire given my reasons. The list of my desires that are contrary to reason is quite extensive, and I don't think that I am alone in this. The apparent existence of such desires creates a dilemma

[17] The outline of a view such as this is proposed, although not defended, in Hurley, (1989).

for the judgmentalist: On one hand she can deny that any 'desire' for an object that the agent does not take herself to have good reason to do is a desire, thus excluding a large group of apparent desires from her view. On the other hand, she can accept these as desires, which requires us to believe such things to be the right thing to do in light of the reasons. This can only be the case if we do not have access to our beliefs about what it is that we have reason to do, despite holding an internalist view of judgment. In either case the cognitivist judgmentalist picture of desire falters.

So we have a step from desires to conative judgments, which seems quite plausible. What the cognitivist anti-Humean tries to do is take these conative judgments and turn them into some species of epistemic judgment. It is this second step which is not possible. You cannot take this last step without making the motivational ability of these judgments very mysterious. Thus the view that judgments about desires are epistemic judgments can only be had at the price of making the motivational properties of desires deeply mysterious.

The Plausibility of the Judgmentalist Program

There are several objections that can be made to the judgmentalist program, in addition to those objections which apply to each form of Judgmentalism. If Judgmentalism is correct, then any desire-like state which is not endorsed by the agent will turn out not to count as a desire. This result is analogous to what happens when it is proposed that what it is to *act*, is to *act rationally*. On such an account irrational 'actions' don't turn out to be actions at all—they are relegated to the realm of mere behavior.[18] Thus Judgmentalism about desires gives us too pristine an account of agency. It simply denies that irrational desires (in the sense of internally incoherent desires – those that the agent herself does not endorse), are desires at all. If we want to preserve space in our moral psychology for desires that are fleeting,

[18] This is the view of action that is generated by a literal reading of Kant's account of agency.

ungrounded, knee-jerk, or generally unsatisfactory in a wide variety of ways, then we must reject judgmentalist views of desire.

I also have a concern with the general approach of Judgmentalism. The strategy is to argue that an entire class of mental states should be understood in terms of judgments about that state. Judgments have two elements, endorsement and content. In the sense that the endorsement is an endorsement *of* the content, it is clear that the endorsement is always one step beyond the content of the state. Specifically, it is the result of the agent reflecting on the content of the state. Such an act of reflection requires the agent to entertain the proposition that she is reflecting on. But what it means to be able to reflect upon the content of a judgment is for the agent to entertain that proposition without endorsing it. Thus, it seems as if the content of the state can, and must, exist in the absence of the endorsement element. Judgmentalism, particularly Humean Judgmentalism, presupposes the existence of desires which are not accompanied by an endorsement. So, the worry is, why can the content itself not count as the state, given that it exists independently of the state? If the desire exists as the content of the judgment, then positing the necessity of the endorsement element is a step too far. All versions of Judgmentalism are vulnerable to this objection.

In conclusion, the judgmentalist view of desire, which takes all desires to be species of judgment, is unsatisfactory because it requires a major conceptual adjustment. In accepting the view of cognitivist Judgmentalism that all desires are in fact desire-judgments we must give up on the idea that desires are necessarily motivational, which is a radical revision of the concept of desire. Even if we are willing to bite this bullet, there is a further toll extracted by Judgmentalism instantiated through both desire-judgments and conative judgments. Both views of judgment require that it is a prerequisite of desiring that a desire meets some standard of 'correctness' for desires. That is, both views hold that we desire things only if we take them to be the 'right' thing to desire in light of the reasons. Thus it is a price of

Judgmentalism that 'irrational' desires cannot count as desires. I propose that this is too high a price to pay, making Judgmentalism about desires an implausible view.

Desires and Reasons

In this section the aim is to consider the second of the two cognitivist theories of desire, the view proposed by Scanlon in 'What We Owe to Each Other'[19] that desires are best understood as reasons. A central claim of Scanlon's argument is that we should reverse the normal order of explanation between desires and reasons, and explain the most distinctively action related aspects of desires in terms of reasons. In proposing this reversal Scanlon is in part motivated by the same view as I, the view that the term 'desire' requires greater explanation. However, the conclusion that Scanlon draws from this observation is radically different from mine. What Scanlon does is to take a standard normative understanding of reasons—justificatory reasons—to be primitive, and explains two functions that philosophers commonly attribute to desires, motivation and justification, in terms of such reasons. Thus instead of explicating the nature of desires *qua* motivational states, he eliminates them as distinct philosophical entities. The sense of desire that is left standing—the directed-attention theory of desire that Scanlon takes to capture the folk meaning of desire—motivates and justifies only by appealing to reasons external to the desire. In short, Scanlon holds that desires, as we understand them in talk of action and deliberation, *just are reasons*.

Scanlon presents the view that desires are reasons as an alternative to a certain Humean conception of desires. Both Scanlon and the Humean view he objects to hold that there must be some state which fulfills the functional role of motivating and justifying action in our conception of practical reason. The difference between the two views is what is taken to realize this functional role description. Standardly, the realizer of this role is taken to be

---

[19] Scanlon, 1998.

Humean desires, which is the essence of the Humean view. Scanlon, in contrast, argues that only *reasons* are capable of filling this role. His strategy is to argue that in order for desires to fulfill this functional role the Humean requires a state that satisfies: (1) the constraints generated by this functional role description (in order to fit accounts of practical reasons); and (2) the constraints generated by the folk conception of desire (in order to count as desires). He concludes that there is no desire-like state that fulfills all these constraints. Rather, he argues that the folk sense of desire is captured by the 'directed-attention' theory of desires. This is the view that what it is to have an occurrent desire for some thing is for your attention to be repeatedly directed toward that thing. However, desires in the directed-attention sense are not independently capable of either motivating or justifying action. It is reasons that play these motivational and justificatory roles in the directed-attention theory of desire. Thus, Scanlon argues, it is reasons that realize the desire role in practical reason. Scanlon's ultimate claim is that because the desire role is fulfilled by reasons, what philosophers are really referring to when they engage in desire talk (particularly in the context of practical reason) is reasons. As he puts it: "…the notion of a desire…needs to be understood in terms of the idea of taking something to be a reason."[20] Thus desires as conceived by philosophers, he proposes, just are reasons.

The Argument

The argument that Scanlon gives is very complex, so I will begin with a schematic of his argument, and then focus on the key points. In broad terms, Scanlon's argument is as follows:

> Philosophers claim that when they talk of desires (the philosopher's sense of desire, desireΦ), they are talking about the same psychological states that are picked out by ordinary usage of the term 'desire', the folk theory of desire, desireF.

---

[20] Scanlon, 1998, p. 7-8.

The fundamental characteristics of desiresΦ are that they motivate and justify action.[21]

DesiresΦ are generally assumed to be instantiated by states such as pro-attitudes.

Pro-attitudes cannot be desiresΦ because they do not satisfy the Humean's own constraints on the desire-role of practical reason.

The only remaining possibility for desiresΦ is basic urges.

Basic urges cannot be desiresF because they do not satisfy the folk view that to desire something is to perceive that thing in the guise of the good.

So, desireF and desireΦ are not the same phenomena.

DesiresF are desires in the directed-attention sense.

DesiresF possess the characteristics of motivation and justification only derivatively through the cognitive process of 'taking something to be a reason'.

The roles of motivation and justification that are central to desiresΦ are played by reasons in desiresF.

Therefore, talk of desire in the philosophical sense should be replaced with reasons talk.

Indeed, Scanlon's conclusion could be expressed in stronger terms. As he puts it "…the philosophical use of 'desire' is not a harmless choice of technical terminology but a seriously misleading one."[22] So, Scanlon seeks to reconcile the apparently motivational aspects of action with its rational aspects through replacing the concept of desireΦ with reasons.

One of the keys to Scanlon's argument is a distinction between the philosophical sense of 'desire', and the way that the folk understand 'desire'. Scanlon proposes that we can distinguish between the folk sense of desire, and the philosophical sense of desire, in the

---

[21] "Desires are commonly understood in philosophical discussion to be psychological states which play two fundamental roles. One the one hand, they are supposed to be motivationally efficacious: desires are usually, or perhaps always, what moves us to act. One the other hand, they are supposed to be normatively significant: when someone has a reason (in the standard normative sense) to do something this is generally, perhaps even always, true *because* doing this would promote the fulfillment of some desire which the agent has"(Scanlon, 1998, p. 37).

[22] Scanlon, 1998, p. 55.

following way: The folk sense of desire is just what people mean when they invoke desires in non-technical conversation. The philosophical sense of desire is 'whatever fills the desire-role that is stipulated in practical reason'. We can call the 'folk' or ordinary view of desire (desireF),[23] and the philosophical view of desire (desireΦ). A part of the subtext of Scanlon's distinction between the two senses of desire is that he doesn't want it to be the case that desiresΦ just are desiresF, as it seems that he cannot easily deny the existence of desires in the folk sense, which is his intention with respect to desiresΦ.

The Constraints on DesiresΦ

Four of the five constraints on desiresΦ that Scanlon appeals to are furnished by what he calls the 'standard Humean' view of desire. This is presented as an uncontroversial, if simplistic, Humean view of desire. Scanlon does not present the view at a single point, but rather refers to it in a variety of ways in a variety of places. The basic outline of Scanlon's standard Humean view of desire is:

> Desires are commonly understood in philosophical discussion to be psychological states which play two fundamental roles. On the one hand, they are supposed to be motivationally efficacious: desires are usually, or perhaps always, what move us to act. On the other hand, they are supposed to be normatively significant: when someone has a reason (in the standard normative sense) to do something this is generally, perhaps even always, true *because* doing this would promote the fulfillment of some desire which the agent has.[24]

There are two distinct roles that desires must fulfill: (1) desires must be motivationally efficacious; and (2) it is the fulfillment of desires which grounds our reasons. Scanlon presents the standard Humean position as capturing the extreme version of both of these claims, by adding a universal rider. So, the first claim of Scanlon's standard Humeanism is: (1) desires are the *sole* source of motivation for action, and thus always occur in a correct explanation of

---

[23] Scanlon refers to this variously as the ordinary or common view of desire.
[24] Scanlon, 1998, p. 37.

action. Call this thesis 'motivational Humeanism'. The second claim is: (2) *all* of our justifying reasons depend on our desires. Call this thesis 'justificatory Humeanism'.[25]

In addition to these two relatively uncontroversial theses about Humean desires, Scanlon also attributes I will call 'independence requirements' to the standard view. These references are scattered throughout his discussion, so I will gather them here to suggest the tone of these independence requirements. Scanlon says that a substantial thesis of the standard Humean view is that it is "…claiming a special role for desires in moving us to act…"[26] At a later point, he adds:

> According to this familiar model, *desires are not conclusions of practical reasoning but starting points for it*. They are states which simply occur or not, and when they do occur they provide the agent with reason to do what will promote their fulfillment.[27]

He variously describes desires in his standard Humean sense as being "*sources* of motivation", "…original sources of reason…"[28], as well as "…independent sources of reasons…"[29]. These various glosses on the 'special role' of desires in practical reason amount to Scanlon placing two more constraints on the desire-role in practical reason. Both of these are 'independence requirements' in that they are claims about desires being *original* and *independent sources* of motivation and justification. Thus, Scanlon's standard Humean view has two more theses: (3) all desires$_{\Phi}$ (non-instrumental) desires are reason-independent sources of motivation for action; and (4) all basic desires$_{\Phi}$ are reason-independent sources of justification. The third thesis can be called the "motivational independence requirement", and the fourth the "justificatory independence requirement".[30] [31] These theses are grounded in Scanlon's

---

[25] These first two constraints, and the position names are drawn from Arkonovich, (2001, p. 499).
[26] Scanlon, 1998, p. 37.
[27] Scanlon, 1998, p. 43. Emphasis original.
[28] Scanlon, 1998, p. 45.
[29] Scanlon, 1998, p. 46.
[30] Constraints one and three, and the position names are drawn from Arkonovich, (2001, p. 499).
[31] "I have argued in this section that when we consider the various states that might be identified as desires we find none that can play the role in justification [that is] commonly assigned to desires—that

presentation of the functional role description of desiresΦ in practical reason. The source for all of these constraints is ostensibly the claim of pure Humeanism that desires are the sole source of motivation and justification in practical reason.

So, in Scanlon's terms the standard Humean view is meant to capture the philosophical sense of 'desire', and it will do this by satisfying these four theses about desires: (1) desires are the sole source of motivation for action, and thus always occur in a correct explanation of action; (2) all of our justifying reasons depend on our desires; (3) all desires are independent sources of motivation for action; and (4) all desires are independent sources of justification. What is necessary for a state to count as a desire on Scanlon's standard Humean view is that it possesses all four of these properties.

Scanlon proposes a fifth and final property of desiresΦ which is generated by the claim that desires in the philosophical sense are the same as desires as conceived by the folk. Scanlon appeals to the case of Warren Quinn's radio man to pump the intuition that, barring external theoretical agendas, we have a robust intuition that a part of what it is to desire some thing is to perceive it in the guise of the good. The case is of a man who experiences an urge to turn on every radio he sees. *Ex hypothesi*, Quinn claims that the radio man does not see the turning on of radios as good. The radio man does not want to hear the music, he does not appreciate the tactile qualities of the knob, rather he just has an urge to turn the radio on.[32] Such an urge,

---

of states which are independent of our practical reasoning and which, when they occur, provide reason for doing what will promote their fulfillment." (Scanlon, 1998, p. 49.)

[32] Quinn, in his presentation of the case simply stipulates this example, whilst acknowledging that it is bizarre. He says: "Suppose I am in a strange functional state that disposes me to turn on radios that I see to be turned off. Given the perception that a radio in my vicinity is off, I try, all other things being equal, to get it turned on. Does this state rationalize my choices? Told nothing more than this, one may certainly doubt that it does. But in the case I am imagining, this is all there is to the state. I do not turn the radios on in order to hear music or get news. It is not that I have an inordinate appetite for entertainment or information. Indeed, I do not turn them on in order to hear anything." His target in this example is what we might call the functionalist interpretation of Humean desires, where desire is presented as bare dispositions to act. The conclusion that Quinn draws from this case is that it is deeply mysterious how such functional states could possibly rationalize action. He goes on to say: "I cannot see

Scanlon argues, is a purely functional state that "…lacks the power to rationalize actions".[33] Because it lacks this power it cannot, according to Quinn and Scanlon, count as a desireΦ. Thus Scanlon uses the Quinn example to propose that desiresΦ include an 'evaluate element'. The evaluative element of desire that Scanlon derives from the radio man case is that desiring involves "…the judgment that there is something good—pleasant, advantageous, or otherwise worthwhile—about performing the action."[34] Ultimately, this evaluative requirement comes from Scanlon' account of the folk view of desire, desireF, as 'desire in the directed-attention' sense. I will come back to this part of Scanlon's argument. This, then, yields the fifth constraint on desiresΦ, the 'evaluative requirement' that a necessary condition of desiring something is perceiving that thing *sub specii boni*, in 'the guise of the good'.

So, Scanlon's view is that a state must possess all five of these properties to count as desires in the philosophical sense.[35] Of these five constraints, only two of them play crucial roles in Scanlon's argument, these are the motivational independence requirement and the evaluative requirement. He uses the motivational independence requirement to reject the wider class pro-attitudes as candidates for Humean desires. The evaluative requirement plays double duty: in the first instance he uses it to reject basic urges as Humean desires, then he uses it to argue that the folk theory of desire—desire as directed-attention—locates motivation in takings to be reasons. It is this latter use of the evaluative requirement which is the most implausible element of Scanlon's view, so it is this part of the view that I will focus on.[36]

---

how this bizarre functional state in itself gives me even a prima facie reason to turn on radios, even those I can see to be available for cost-free on-turning." (Quinn, 1995, pp. 189-90.)

[33] Scanlon, 1998, p. 38.

[34] Scanlon, 1998, p. 43.

[35] "According to Scanlon, this model attributes to desires the following three features: 1) "desires are not conclusions of practical reasoning but starting points for it" (p. 43; 2) "desires are states which simply occur or not" (p. 43; 3) "when they do occur they provide the agent with reason to do what will promote their fulfillment" (Arkonovich, 2001, p. 43).

[36] One of the puzzles about how Scanlon manufactures these constraints on desire is that it by focusing so tightly on the requirements of motivation it does not clearly capture other types of desires. There are

The Directed-Attention Theory of Desire and Motivation as Reasons

Scanlon's aim in proposing the directed-attention theory of desire is to show that the necessary attributes of desireΦ, justification and motivation, can only be provided by reasons, and from this conclude that desiresΦ are not desires in the folk sense, but are in fact reasons. This is, in a nutshell, Scanlon's position. So we can conceive of the debate that Scanlon is having as between two extreme positions: At one end is the Humean position that desires are those psychological states that are uniquely capable of the subjective motivation and justification required by rational action. At the other end is Scanlon's view, which is that reasons are the states that are uniquely capable of the subjective motivation and justification required by rational action, and desiresF have these attributes only derivatively. The aim of this section is to examine his argument for the latter proposition.

Scanlon derives the directed-attention theory of desire from two sources: the phenomenology of occurrent desiring, and the evaluative requirement generated by Quinn's case of the radio man. The 'directed-attention' theory of desire is the view that what it is to desire something is for your attention to be repeatedly favorably directed toward that thing.

> A person has a desire in the directed-attention sense that P if the thought of P keeps occurring to him or her in a favorable light, that is to say, if the person's attention is

---

various types of desire that do not seem to entail motivation at all. I desire that my husband surprises me with flowers, but my doing anything to bring this end about (such as mentioning it), robs the gesture of the element of spontaneity which is part of what I desire in it. My desire requires no *motivation*, although it clearly seeks *satisfaction*. Other examples of such desires are desires for outcomes that you have no control over, or desires such as Freudian wishes, that find their satisfaction in fantasy. The satisfaction of a desire only requires motivation in the case of desires for one's own future actions. But, as is highlighted by these examples, desires for action do not exhaust the category of all desires. The claim that desires should motivate cannot be understood as the claim that motivating is a necessary attribute of desires. The strongest claim about motivation that is compatible with the existence of such desires is that a necessary attribution of a desire is that it has the *ability* to motivate. As Arkonovich puts it, "…while all desires might necessarily seek *satisfaction*, we need not think that all satisfaction involves *motivation*. Here I am thinking of the Freudian account of the wish. The wish is clearly a desire-like state as such states are usually defined in contemporary philosophical psychology. Yet a Freudian wish is precisely a state which finds satisfaction in fantasy as opposed to action, and consequently does not motivate the agent to do anything." (Arkonovich, 2001, p. 43).

> directed insistently toward considerations that present themselves as counting in favor of *P*.[37]

Scanlon attributes this view of desire to reflecting on the following two cases. The first case is of having a positive evaluation of some action such as "…seeing something good about drinking a glass of foul-tasting medicine…"[38], but failing to desire to drink the medicine. The second case is seeing some action as pleasant, but having no desire to do it. What stops the first case from being a case of desire is that you lack the directed-attention element of desire. Your attention is not directed toward the prospect of drinking the medicine in the right way. What stops the second case from being a case of desire is that although you take the action to be pleasant, you fail to count this as a consideration in favor of performing that action.

Scanlon claims that the idea of 'desire in the directed-attention sense' thus

> …capture[s] an essential element in the intuitive notion of (occurrent) desire. Desires for food…and sexual desires are marked by just this character of directed-attention. And this character is generally missing in cases in which we say that a person who does something for a reason nonetheless "has no desire to do it…"[39]

There are three factors here which Scanlon takes to be particularly apposite to the folk notion of desire. The first is that desires in the directed-attention sense are occurrent psychological states with identifiable and typical qualia. The second is that such desires "…capture…the familiar idea that desires are unreflective elements in our practical thinking—that they "assail us" unbidden and that they can conflict with our considered judgment of what we have reason to do."[40] The last is that it satisfies the evaluative requirement, as a central part of desiring something in the directed-attention sense is perceiving it in the guise of the good (or at least the guise of the pleasant).

---

[37] Scanlon, 1998, p. 39.
[38] Scanlon, 1998, p. 39.
[39] Scanlon, 1998, p. 39.
[40] Scanlon, 1998, p. 39.

This view of desires is not particularly complex as Scanlon is explicit that it does not and cannot capture the motivational and justificatory properties of the desire-role in practical reason. How it is that desires in the directed-attention sense are connected to motivation and justification is embedded in Scanlon's understanding of the evaluative requirement.

Central to the idea that what it is to desire something is to have your attention directed toward it, is that the thought of *P* repeatedly occurs to you *in a favorable light*. Scanlon further glosses this element of 'directed-attention' as "…the person's attention [being] directed insistently toward considerations that present themselves as counting in favor of P.[41] Scanlon proposes that what it is to count certain considerations as being in favor of *P*, is to *take these considerations to be reasons for P*. It is this characteristic of the directed-attention theory of desire that makes it clear why it is that Scanlon holds that desires should be ultimately understood in terms of normative reasons.

So, Scanlon's account of how motivation operates in the case of desires in the directed-attention sense contains a shift from "having your attention directed toward considerations that appear to count in favor of", to "taking these considerations to be reasons". When the shift occurs in his presentation of the view, Scanlon treats the change as merely terminological.[42] But this is not at all clear to me. It seems that taking something to be a reason is a much more committed state than simply having your attention repeatedly directed toward considerations that count in favor of something. I think that Scanlon equates these two things because he has a particularly strong interpretation of the evaluative requirement in mind.

Insofar as desiring something is to be attracted to it, then it is at least plausible (if not universally accepted) to say that when you desire some thing you perceive it, in some very

---

[41] Scanlon, 1998, p. 39.
[42] Copp and Sobel make this point extremely clear. (Copp & Sobel, 2002, p. 256-7).

minimal sense, to be good. Call this weak intentionalism. But Scanlon is not interested in such a weak sense of evaluation. If he was satisfied by this interpretation, then he would not make the crucial move in his formulation of the directed-attention theory from 'seeing some thing in a favorable light' to 'takings to be reasons', as surely weak intentionalism is satisfied by seeing some thing in a favorable light. The only justification for this shift is if he is implicitly interpreting the evaluative requirement as stronger requirement, such as 'judging yourself to have a reason'. Call this strong intentionalism. I think that there is evidence for Scanlon's being a strong intentionalist in how his notion of 'taking' something to be a reason turns out to be motivationally efficacious.

Scanlon argues that the element of desires in the directed-attention sense that does all the work to make them look like desiresΦ is in taking considerations in favor of *P* to be reasons for *P*. It is then these reasons that fulfill both the motivational and justificatory roles that are central to desireΦ. As he puts it: "…when a person *does* have a desire in the directed-attention sense and acts accordingly, what supplies the motive for this action is the agent's perception of some consideration as a reason, not some additional element of "desire"."[43] Thus desires in the directed-attention sense motivate only in virtue of distinct reasons. The notion of 'taking' or 'perceiving' something to be a reason that he uses is a complex one. He does not appear to mean that in every case the agent must 'judge' or 'believe' there to be a reason. At one point, Scanlon states that having a desire for *P* in the directed-attention sense "…involves a tendency to judge that [you] have [a] reason…"[44] to get *P*. At another he presents the idea as it at least being the case that it 'seems to' the agent that there is a reason. However in still other places he treats the idea of seeming to be a reason as equivalent to believing that there is

[43] Scanlon, 1998, p. 40-1.
[44] Scanlon, 1998, p. 43.

a reason. In the latter case, Scanlon explicitly commits himself to giving some account of how beliefs can motivate. It is in this account that we see his commitment to strong intentionalism.

Scanlon's argument for the view that beliefs can motivate is a form of a 'one thought too many' argument. In the case of beliefs, he argues that:

> A rational person who judges that there to be sufficient grounds for believing that *P* normally has that belief, and this judgment is normally sufficient explanation for so believing. There is no need to appeal to some further source of motivation such as "wanting to believe."[45]

The same, he proposes, is true of intentions. There is no reason to suppose some specifically motivational entity like desireΦ in addition to a judgment of how to act and the reasons that such a judgment recognizes in order to account for how motivation works in the practical case. When a rational person judges that she has 'compelling' reason to *P*, under normal circumstances she will form the intention to *P*. When you have a desire in the directed-attention sense, you have a tendency to make such judgments. According to Scanlon, "…this judgment is sufficient explanation of that intention and of the agents acting on it."[46] So, what is capable of motivating an agent to act are intentions of the form: 'I judge that I have reason to *P*'. Thus it is something that looks like a judgment about reasons that does the motivational work for desires in the directed-attention sense. This is just a retelling of the thesis of strong intentionalism. It is the claim that what it is to be motivated when you have a desire in the directed-attention sense is for you to form an intention by judging yourself to have a reason to act in accordance with your desire. So it is strong intentionalism about desires that yields Scanlon's final conclusion that what it is to be motivated by having a desire for *P* in the directed-attention sense is in essence to be motivated by judging yourself to have a reason for *P*.

---

[45] Scanlon, 1998, p. 33.
[46] Scanlon, 1998, p. 33-4.

So, how plausible is strong intentionalism? The first issue is that this is a cognitively demanding interpretation of both desires and action. In order to desire, you have to have the concept of a reason. Scanlon does not make this constraint as explicit as I have, but he clearly takes having reasons to be necessary for being motivated when you have a desire in the directed-attention sense. He makes this clear in his discussion of the possibility of akratic actions. He states:

> Even when desire in the directed-attention sense runs contrary to our reason (that is to say, our judgment) in [the sense of seeing something as a reason that I judge not to be one], however, it remains true that *the motivational force of these states lies in a tendency to see some consideration as a reason.*[47]

Insofar as desire in the directed-attention sense involves this tendency to see something as a reason, then he makes possessing the concept of a reason central to desiring in the folk sense. One case of desire that this excludes is the desires of creatures who lack the concept of a reason, such as babies or toddlers seeking their parents for comfort, eating, drinking, or dancing to music. Desires in the directed-attention sense clearly motivate in these cases, but their motivation must, according to Scanlon, proceed via conceptualization as a reason. I would hesitate to say that babies and infants take themselves to have reasons, but take it that denying that they have desires that motivate them to be very far from the folk understanding. Another case is that of what can be called spontaneous action, such as scratching an itch without really thinking about it, or hugging your spouse in passing just because he is there. In these cases, where a desire pops out of nowhere, and you act on it without thinking about it, to posit that you judge yourself to have a reason seems to be one thought too many. Maybe in some cases you do, but in those where you don't, I presume that the folk theory would still want to count them as desires, and fairly central cases of desire at that. Given that what Scanlon claims of the directed-attention theory of desire is that it captures the folk meaning of

---

[47] Scanlon, 1998, p. 40. Emphasis added.

desire, rather than being a stipulative view of what desires are, this is a serious objection. If he goes beyond the folk view in a way that excludes various intuitively clear cases of desire from counting as desires, then it is so much the worse for his theory.

A further objection to strong intentionalism is that in order for your behavior to count as action you must in essence judge yourself to have reason to so act (rather than it merely seeming to you that you have a reason to so act). This suffers from the same objections regarding the possibility of imperfect action as the view of Judgmentalism about desires discussed in the first half of this chapter.

So, what could lead Scanlon to having such an implausibly strong interpretation of the evaluation requirement? One possibility is if he is assuming that desiring some thing involves seeing that thing as desirable, where what it is to see some thing as desirable is to make a judgment about its desirability. This is only going to be plausible if we equate 'desiring *P*' with 'judging *P* to be desirable'. This is a very specific reading of the evaluative requirement, which is not really supported by why we might turn to the evaluative requirement in the first instance. One of the attractions of the evaluative requirement is that it gives an intentional gloss to desiring that is absent from an unnaturally bare functional state such as that of Quinn's radio man. As Copp and Sobel put it: "We do not merely act under the force of desire, but in typical cases we intend to act in light of our desires."[48] But it is really plausible to say that desiring something involves judging that thing to be desirable? I think not, and for the same reasons that it is implausible to say that all desires are conative judgments. Interestingly, it seems as if Scanlon's view that desires just are reasons actually ends up relying on the claim of Humean Judgmentalism, that all desires are conative judgments, and is the worse for it.

---

[48] Copp & Sobel, 2002, p. 276.

Thus Scanlon's argument is that the proper understanding of desiresF is the 'directed-attention sense of desire. However desires in the directed-attention sense only motivate tangentially, by causing the agent to judge that she has a reason to so act. Thus what distinguishes desiresF from desiresΦ, and prevents them from fulfilling the desire role in practical reason, is that desiresF are only capable of justifying or motivating action derivatively, through appealing to independently existing reasons. However, as we have seen Scanlon's reliance on an implausibly strong interpretation of the evaluative requirement undermines his claim that desires in the directed-attention sense capture the folk theory of desire.

Conclusion

Scanlon's argument is between two extremes: At one end is the Humean position that desires are those psychological states that are uniquely capable of the subjective motivation and justification required by rational action. At the other end is Scanlon's view which is that reasons are the states that are uniquely capable of the subjective motivation and justification required by rational action, and desiresF have these attributes only derivatively. The structure of the argument that he gives is to reject the possibility of Humeanism, thus leaving his reasons view as the sole explanation. However, there are many moderate Humean positions which do not subscribe to both motivational and justificatory Humeanism, let alone the independence requirements. Smith, for instance, defends a version of Humeanism that endorses a version of motivational Humeanism, but denies the justificatory Humean view that desires are uniquely capable of justifying action.[49] Another version of moderate Humeanism which is perfectly plausible endorses both motivational and justificatory Humeanism, but is

[49] Smith, 1995.

moderate in that eliminates the claims that desires are the *unique* sources of justification and motivation. Thus the dichotomy that underpins Scanlon's argument is false.

The larger difficulty with Scanlon's view is the implausibility of his interpretation of the folk theory of desire, the directed-attention theory of desire. Scanlon's argument that we should understand desires in the philosophical sense as (1) distinct from desires in the folk sense; and (2) in terms of reasons, fails in a number of ways. Not only is his take on the standard Humean account of desire implausible, his account of the folk theory of desire also fails. But there is value in the views that are driving Scanlon's arguments. The final reason that Scanlon gives for preferring reason talk to desire talk is because he takes the standard Humean view of desire to commit us to a specific, and problematic, picture of practical reasoning:

> A desire is naturally understood as have a two-part structure: it has an object and a weight. It is a desire *for* something, typically taken to be some state of affairs, and it counts in favor of that thing with a certain degree of strength, on this view, when our desires come into conflict, rational decision is a matter of balancing the strengths of competing desires. If we take desires, along with beliefs, as basic elements of practical thinking, then this idea of balancing competing desires will seem to be the general form of rational decision-making.[50]

I agree with Scanlon that insofar as we treat Humean desires a having only two contributions to make to rational deliberation, its object and its strength, we get a troubling picture of rational deliberation. In fact it is this precise view of desires and rational deliberation that I described in the first chapter as so comprehensively failing the challenge of temptation. However I propose that we can avoid this picture of deliberation, while retaining the benefits of the Humean view of desires, by considering the relative *stability* of desires in deliberation. It is this view that I will be defending in later chapters. So, I take Scanlon's motivations in proposing the reasons view of desires to be sound, but think that his arguments do not deliver the required conclusions.

---

[50] Scanlon, 1998, p. 50.

I take it that a reasonable conclusion to draw from this discussion is that these cognitivist theories of desire bring us no closer to solving the challenge of temptation, except through eliminating generally cognitivist views of desire on the grounds that they implausibly attempt to eliminate the problem by fiat, rather than explaining what it is that is flawed about temptations. Having narrowed the field to broadly Humean views of desire, in the next chapter I will consider the various permutations of the Humean view to explore in greater detail how it is that we should conceive of desires, in order to reach a view that is capable of answering the challenge of temptation.

CHAPTER 3

THE NATURE OF DESIRE

The aim of this chapter is to examine precisely what type of conative state desires might be, and take one step closer to an understanding of desire that can answer the challenge of temptation. In the previous chapter I considered the broad question of whether desires should be understood in terms of other mental states, such as reasons or judgments, and rejected this possibility. In this chapter, the aim is to explore what resources I can appeal to in order to discover some limits on what can be said about desires *qua* conative entities.

Timothy Schroeder, in "The Three Faces of Desire"[51], points out that in both historical and contemporary philosophical, psychological, and neuro-physiological discussions of desire, desires have been presented in three distinct ways. These are the eponymous three faces of desire: motivation, pleasure, and reward. The first two of these inform various skeletal and widely used concepts of desire such as: 'mental states capable of motivating action'; or 'positive affect toward some end'. The third view, that of desire *qua*, reward is championed by Schroeder. Schroeder's aim is somewhat different to mine in that he is defending a thesis within the philosophy of mind about how desires are physiologically instantiated, while I am defending a view about how desires should be conceived (and thus treated) in rational deliberation and practical reason. The instructiveness of Schroeder's view lies in its

[51] Schroeder, 2004.

engagement with what recent neuroscience tells us about how desires are physiologically instantiated, and I take a desideratum of any concept of desire to be that it is consistent with neuroscience. The bulk of the chapter will be spent on Schroeder's argument that the most fundamental face of desire in creatures like us is reward. I will be arguing that Schroeder succeeds in demonstrating that reward is the neglected face of desire. What Schroeder fails to do is to prove that desires just are rewards, although I think his failure is instructive, and suggests the account of the emotional component of desire offered in chapter four.

Each of Schroeder's three faces of desire can be represented by a distinct and independent theory of desire, all of which could count as 'the' theory of desire. The strongest candidate, and currently prevalent understanding of desire, is captured in the Motivational Theory of Desire (MTD), which takes the motivational face of desire to be central.[52] The weakest candidate is the Hedonic Theory of Desire (HTD), which is grounded in a long tradition which identifies desires with pleasure and displeasure. The final and least familiar candidate, the Reward Theory of Desire (RTD), is the view of desire championed by Schroeder. All of these views are amenable to being understood in terms of propositional attitudes, so I will treat them as such in discussing their differences.

Motivational Theory of Desire

The Motivational Theory of Desire roughly satisfies the functionalist intuition that mental items are defined by what they characteristically do, and what desires characteristically do is motivate us to act.[53] One source of this view is a version of an Aristotelian 'ergon' argument. According to the ergon argument, 'the' characteristic function of a thing is that

---

[52] Schroeder calls this view the Standard Theory of Desire. I will be referring to it as the Motivational Theory of Desire for ease of identification.

[53] "Desires are distinguished by what they do, and what they do is move us to act." (Schroeder, 2004, p. 11.)

which distinguishes it from other things. The Motivational Theory of Desire takes motivation to be the characteristic function of desire because it is this action-guidingness which distinguishes desires from beliefs and other mental entities.[54] Thus what it is to desire some thing is to be motivated to get that thing.

For some mental state (such as a desire) to have motivational capacity is, in its most basic form, for it to be capable of moving the agent to act. But such a simplistic account is not sufficient in the face of deviant causal chains. Donald Davidson influentially concluded that for a mental state to be counted as motivational in the sense required for *action* as opposed to *mere behavior*, we not only care that mental states have the capacity to move an agent to act, but that they cause action in the right way. A part of what it is to cause action in the right way is that the mental state in question moves the agent to act in virtue of its content. Another concern is that having the capacity to motivate for a mental state cannot be the same as actually causing action. Rather, it is the dispositional claim that the mental state in question would cause the agent to act in the absence of some stronger, motivational state. Thus the motivational face of desire must be framed (roughly), as the principle that:

> **Motivational Theory of Desire (MTD)**: To desire that *P* just is to be motivated so as to bring it about that *P*, under normal conditions.

On this view, what it is that distinguishes desires from other pro-attitudes is that they have the capacity to motivate.[55] Indeed, by definition, in the MTD it is the case that mental states which motivate are just are desires. This seems to accord nicely with our intuitions that

---

[54] This is then allied with a propositional conception of desire. On this view we do not have desires simpliciter. It is never the case that I desire watermelon, rather, my desire is 'that I eat watermelon right now'. In this view desires can always be, and are properly, expressed in terms of a 'that' clause. The motivational view presents desires in terms of a lean model of two central characteristics – propositional content and motivational force.

[55] This cannot plausibly be the stronger claim that desires are motivationally efficacious—that is, that they will always result in action—as it is clear that there are cases where we have competing desires and only the stronger one will in fact produce action. This does not mean that the weaker desire in this struggle was incapable of producing action; it had the *capacity* to do so, it just, as a matter of contingent fact, did not do so in this instance. Thus it is the capacity to produce action which is important here.

to want something (to desire it) is intimately connected with trying to get that thing. Arguably any theory of desire which said that desiring had nothing to do with trying to get, would be such a perversion of the common usage as to constitute a theory of desire*, rather than desire. Of course, it does not follow from this simple point that motivation is necessarily *all* there is to desire.

Hedonic Theory of Desire

The Hedonic Theory of Desire, in contrast, is grounded in the intuition that desires are all about affective states such as pleasure and pain. On this view, all desires are, at base, conative attitudes directed toward propositions about the state of the world which dispose the agent to experience pleasure when fulfilled and pain when denied.

> **Hedonic Theory of Desire (HTD):** To desire that *P* is to be disposed to feel pleasure if *P*, and displeasure if not *P*.[56]

The HTD and its close cousins was the dominant view of desire in eighteenth and nineteenth century British psychology, as exemplified in the writings of Bentham and Hume. Although it has fallen out of favor, it is not without substantial influence. Our hedonic

---

[56] There is a point that needs to be made here about what the object of the desire is, and how this differs from the conditions that must hold for the agent to be aware that her desire has been satisfied. Schroeder conflates these two points by arguing that the only plausible way of understanding this view is in terms of what it *seems* to the agent, as the agent's pleasure and pain does not track the way that the world is, but rather the way that the world seems to the agent. In Schroeder's view the 'seems that' locution is included here to account for cases in which the agent has mistaken perceptions about the world. If someone falsely believes that a deeply held desire of theirs has been fulfilled, intuitively they will experience just as much pleasure as someone who truly believes that some similar desire has been fulfilled.(Schroeder, 2004, p. 27.) However, it cannot be the case that the object of my desire that *P* is really 'that it seems to me that *P*'. If this were true, then I would be indifferent between my desire that my husband is happy being satisfied by my husband being happy, or it seeming to me that he is happy in the experience machine. It is clear that the object of my desire is the former, not the latter, thus the 'seems to' locution is out of place in talking about the *content* of such a desire. Thus there are two ways that we could understand the Hedonic Theory of Desire. One is that the object of my desire that *P* is directed toward the pleasure that I will be disposed to feel if *P*. The other is that I desire that *P* as a means to the pleasure that I am disposed to feel if it is the case that *P*. In this latter case the object of my desire is *P*, whereas in the former case the object of my desire is the pleasure that will result from *P* obtaining. I take the former to be the most plausible view.

experiences are at least important markers of the existence of certain desires. Pleasure and pain provide us, in many cases, with important information about whether that which we desire is what we want. Also, it seems that for at least certain desires that *P*, their strength should track the amount of pleasure (or pain) that results from *P* obtaining.[57] Thus hedonic tone is importantly connected to our understanding of desire, although the theory that desires just are such hedonic states is not, in itself, taken to be that plausible.

Reward Theory of Desire

There is a third player in this debate which has not had the exposure of the motivational view or the hedonic view. This is the Reward Theory of Desire proposed by Schroeder.[58] I will be discussing this in much greater detail because of its novelty and interest.

Before outlining Schroeder's Reward Theory of Desire I should make it clear that Schroeder's project is in the scientific tradition of discovering what instantiates natural kinds, in this case the natural psychological kind of desire. What Schroeder identifies as this natural psychological kind is the subset of intrinsic pro-attitudes that are called 'desires', 'wants', and 'wishes' in the common parlance. His aim is not to eliminate desire talk, rather to understand the nature of desire by discovering the source of its surface features—motivation and pleasure. In the way that discovering that H_20 is water teaches us something about the essential nature of water, Schroeder proposes that discovering the underlying nature of desire will allow us to

---

[57] This claim is meant to be understood in light of the caveats expressed in the previous footnote.
[58] An early version of this theory was first expressed in the philosophical literature by Fred Dretske: "The third face of desire, stressed by Fred Dretske in his *Explaining Behavior* (Dretske, 1996) but largely neglected by both popular thinking and philosophical research, is that there is a link between desire and reward: desires determine what counts as a reward and what counts as a punishment for an organism. A rat, according to this view, can only be rewarded with food when it wants food; when it does not want food, food is no reward, and indeed force-feeding associated rat would be constitute a punishment. In the motivational theory, this can be accommodated as another accidental, but evolutionary sensible, feature of desires. Just as it makes sense for pleasure and displeasure to be linked to desire, so it makes sense for desires to determine what will serve to operantly condition an organism." (Schroeder, 2004, p. 15.)

go beyond these surface features in our understanding of intrinsic desires. Now, there are two distinct claims that we can see Schroeder as defending. The first is the 'physical' view that he is simply identifying the fundamental source of desire-like phenomena in creatures like us as being instantiated is the area of the brain which is the seat of reward. The second is the 'conceptual' view that he is arguing that our *concept* of desire should be understood in terms of reward, which entails that the other two surface features of desire, motivation and pleasure, should be understood as being *caused by* desires, rather than as elements of desires.

I will begin with a quick outline of the view. Schroeder's Reward Theory of Desire is the view that an intrinsic desire is a mental representation which has been constituted by the agent as a reward. A reward, on Schroeder's view, is a neuroscientific conception of a certain type of learning signal released in a particular subsystem of the hypothalamus.[59] This subsystem is, according to Schroeder, the mechanism by which rewards are realized in the brain. When I get the desired object/experience, this triggers a certain type of learning signal in this system, which reinforces the representation of that thing as a reward. What it is for an agent to constitute some representation as a reward is for that representation to drive the production of the appropriate kind of learning signal in the agent's reward system.

The biggest challenge that Schroeder faces in defending a theory of desire as reward is to present and motivate a view of reward that will make sense of the claim that what it is to desire some thing is to constitute that thing as a reward. So, in order to understand what is going on in Schroeder's view, we need understand in detail what he means by 'reward', and how it is that our representations can be 'constituted as' rewards.

---

[59] "There are two output structures of the biological reward system in organisms like us, twin structures immediately adjacent to one another deep in the brain: the *ventral tegmental area*, or *VTA*, and the *pars compacta* of the *substantia nigra*, or *SNpc*." (Schroeder, 2004, p. 49.)

Schroeder's theory of reward

There are two central and not completely coincident concepts of what constitutes a reward in the way that the term 'reward' is commonly used: (1) The agent-relative view that what is a reward for an individual is giving them what they want, whether or not the giver intends the thing as a reward; and (2) the view that whether or not something counts as a reward is related not only to the agent's response to that thing (as in (1)), but also to the intention that was expressed by the conferring party in so conferring that thing. It is important to note that these common sense understandings of reward are **not** what Schroeder is linking to intrinsic desires. His use of the term 'reward' is a technical one, where reward is taken to be a specific type of learning. The formal statement of Schroeder's theory of reward is:

> Contingency-based Learning Theory of Reward (CLTR): For an event to be a reward for an organism is for representations of that event to tend to contribute to the production of a reinforcement signal in the organism in the sense made clear by computational theories of what is called 'reinforcement learning'.

Note that Schroeder's aim here is not to give an account of what constitutes a reward in terms of necessary and sufficient conditions. Thus the fact that CLTR does not easily incorporate several elements of the folk theory of reward such as the intentions of the conferrer (if there is an agent involved in this role) or the efforts of the receiver, does not, Schroeder claims, undermine his view. Rather, his aim is to "…give a theory of the psychological aspect of reward: what must be the case inside an organism in order for it to make things into rewards."[60] [61]

---

[60] Schroeder, 2004, p. 67.

[61] "When an organism like us is rewarded by, say, being given a bicycle, the first thing that happens in its brain is that it represents being given a bicycle. This representation causes activity elsewhere in the brain, which categorizes the represented event as a reward. Meanwhile, other brain structures have been attempting to predict the rewards and punishments the organism was going to receive at this moment. The combination of current reward information and predicted reward information is used by the brain to calculate the difference between the rewards that had previously been predicted and the

According to CLTR there are three main elements of reward: representation, reinforcement and learning. The most straightforward, and least contentious, of these elements is the claim that rewards are representational states. The representations in question are those of propositional content, as well as the less cognitive representations of things such as thirst found in the hypothalamus. [62,63] The idea that we have such representational states is uncontroversial, as is the view that such states can participate in our various psychological processes. That rewards have representational content in this wide sense puts Schroeder's view on a par with many other views of reward, as it is essentially the claim that such states have objects. It is the latter two elements—reinforcement and learning—that require explanation and defense. Both of these elements appear in the CLTR because Schroeder's theory of reward understands the learning element in terms of Sutton and Barto's theory of Reinforcement Learning.

Reinforcement Learning

The theory of Reinforcement Learning builds on a long discourse about 'reinforcement' as understood in terms of operant conditioning within psychology.[64]

---

rewards that have actually materialized. The result is released to the rest of the brain in the form of a very specific signal, one causing a very specific form of learning. This signal has effects upon the short-term operation of the brain and upon its long-term dispositions, effects that, in organisms like us, affect our feelings and modify dispositions to act, think, and experience, all in ways that tend to increase the acquisition of rewards and the avoidance of punishment." (Schroeder, 2004, p. 49.)

[62] "All that is required for something to be counted as a perceptual or cognitive representation, so far as RTD is concerned, is that it be a content-bearing thing, making up some perceptual or cognitive attitude, localized in or distributed through some perceptual or cognitive center of the brain, capable of passing output to the reward system." (Schroeder, 2004, p. 134.)

[63] "A desire that justice be served is an entity that involves, as a part, the capacity to perceptually or cognitively represent that justice is being served, and it is from the content of this representation that the desire acquires its content." (Schroeder, 2004, p. 133.)

[64] The seminal definition of reinforcement is expressed in Thorndike's Law of Effect, 1911: "Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur The greater the satisfaction or

Reinforcement Learning is a computational approach to explaining how it is that we learn by interacting with our environment.[65] The main focus of the theory is "goal-directed learning from interaction"[66]. The general idea motivating this view is that we learn many things through interacting with our environment. A child discovers the pleasures of sweet, the aversiveness of hot things, and the attraction of bright colors by touching, tasting, and seeing the world around her. Not only does interacting with the environment teach us things about the environment and our reaction to it, it also poses certain types of learning problems such as: 'How do I get the chocolate?'; 'Where should I go in the maze to get the cheese?'; and other more complex questions. The focus of this theory is learning problems where the agent has a specified goal, and the problem is to achieve the goal through interacting with the environment. What the agent is aiming for in achieving this goal, according to the theory, is to maximize a measurable reward signal which is released by the achievement of the goal. Such a reward signal is generally taken to be quantifiable, and thus able to be expressed numerically.

The two main elements of Reinforcement Learning are the processes of search and memory. The 'search' element is captured by allowing the agent to try many responses to the

---

discomfort, the greater the strengthening or weakening of the bond." (Thorndike, 1911, p. 244) The Law of Effect is somewhat controversial, but is generally taken to capture in broad strokes, a general intuition in psychology about how certain behaviors may become more or less likely. This general intuition is at the root of the procedural understanding of 'reinforcement' in psychology. It "…refers to the presentation of a rewarding event following a response, which results in the phenomenon of an increased probability of that action re-occurring."(Evans, 2001) In turn, that which reinforces behavior in this way is taken to be a reward. The relevance of the Law of Effect for Sutton and Barto is that it captures two central elements of reinforcement learning: search and memory.

[65] "The idea that we learn by interacting with our environment is probably the first to occur to us when we think about the nature of learning. When an infant plays, waves its arms, or looks about, it has no explicit teacher, but it does have a direct sensorimotor connection to its environment. Exercising this connection produces a wealth of information about cause and effect, about the consequences of actions, and about what to do in order to achieve goals. Throughout our lives, such interactions are undoubtedly a major source of knowledge about our environment and us. Whether we are learning to drive a car or to hold a conversation, we are acutely aware of how our environment responds to what we do, and we seek to influence what happens through our behavior. Learning from interaction is a foundational idea underlying nearly all theories of learning and intelligence." (Sutton & Barto, 1998, p. 3.)

[66] Sutton & Barto, 1998, p. 3.

situation, which will be selectively reinforced or undermined. The 'memory' element is where an association is formed (or strengthened) between a response to the situation and an experience of satisfaction, allowing the agent to remember which responses were most effective.[67]

To take a computational approach is to adopt the perspective of someone designing an artificial intelligence system, and construct a computational process that is capable of solving the problem at hand. There are two parts to such an approach, the framing of the situation, and construction of the process. The problem is framed through the specification (or mapping) of the environment faced by the agent (the situation). The process is constructed by considering various ways that the agent may interact with the environment (actions), each of which constitutes a policy action. The aim in every case is for the agent to pursue that policy for action that maximizes a numerical reward signal. Thus "[r]einforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal."[68]

There are four main parts to Sutton and Barto's theory of Reinforcement Learning: a policy for action; a reward function; a value function; and an environment.[69] The *policy* determines what the agent will do in the face of different states of the environment, which "…[corresponds] to what in psychology would be called a set of stimulus-response rules or associations."[70] A policy for action can be understood as a disjunctive description of the moves that an agent may make in an extensive form game. The *reward function* determines what the goal of the agent is in a given situation, by "…map[ping] each perceived state (or state-action pair) of the environment to a single number, a *reward*, indicating the intrinsic desirability of

---

[67] Sutton & Barto, 1998, p. 18.
[68] Sutton & Barto, 1998, p. 3.
[69] Sutton & Barto, 1998, pp. 7-8.
[70] Sutton & Barto, 1998, pp. 7.

that state."[71] The reward function captures the intrinsic desirability of being in a particular state at a particular time. In contrast, *value function* of a state-action pair is the sum of rewards that the agent achieves over time if they begin with that state-action pair. So while reward measures the immediate positive impact of a state on an agent, value measures the sum total of desirable outcomes resulting over time from that starting point. The final element of the view, the environment, is the state of the world within which the agent seeks to maximize her rewards.

How it is that this mapping maximizes a numerical reward signal is cashed out in the 'Temporal Difference' (TD) learning algorithm. The TD algorithm is based on the idea that through making temporal 'errors' in valuation—by estimating the reward of some state differently at different times—the agent learns something about the actual value of that state to her.[72] In the Reinforcement Learning model the quantity estimated is that of *value*, understood as the *cumulative reward* of the state. The TD algorithm is based on two assumptions: the assumption "…that the computational goal of learning is to use the sensory cues to predict a discounted sum of all future rewards…within a learning trial…"; and the Markovian assumption that future rewards depend only on current perceptions—they have no connection with past perceptions.[73] The Markovian assumption captures how it is that the agent can better predict the actual value of a reward by improving her successive estimations of the reward during a trial. This is important because during a trial the agent only has access to *estimations* of the sum of all future rewards, as the sum itself can only be known at the end

---

[71] Sutton & Barto, 1998, pp. 8.
[72] Sutton & Barto, 1998, p. 21.
[73] Let *V(t)* be the sum of all future rewards; r*(t)* be the reward at time *t*; *E*[·] be the expected value of the sum of all future rewards in the trial; let the rate at which the agent discounts future be some γ such that $0 \leq \gamma \leq 1$. Then the future value of all rewards to the end of a particular learning trial can be predicted with the equation
$V(t) = E\left[\gamma^0 r(t) + \gamma^1 r(t+1) = \gamma^2 r(t+2) + ...\right]$
(Schultz, Dayan, & Montague, 1997,p. 1595)

of the trial, yet her aim is to maximize her future rewards.[74] The attractiveness of this view is that learning about the actual value of states will allow the agent to better maximize her reward signal, as she will learn which state-action pair will lead to the most rewards over time, thus allowing her to maximize the satisfaction of her intrinsic desires.

Thus what it is for something to be a reward, according to Schroeder's Contingency-based Learning Theory of reward, is for it to drive a learning signal that predicts an error between actual and expected rewards. For this to count as a claim about the physiological basis for reward Schroeder needs to give an account of how our representations tend to produce a physiological reinforcing signal that matches the form required by the Temporal Difference algorithm central to Sutton and Barto's theory of reinforcement learning.[75] At the heart of Schroeder's theory of reward is the view that it is demonstrable that such a reinforcing signal exists in the reward center of the brain, which is comprised of a pair of neural structures immediately adjacent to one another: the *ventral tegmental area*, or *VTA*, and the *pars compacta* of the *substantia nigra*, or *SNpc*, the VTA/SNpc for short. Schroeder is in good company in making this claim. As Dayan et al put it:

> The idea that dopamine cells in the vertebrate midbrain report errors in the prediction of reward has been a powerful (although not undisputed) organizing force for a wealth of

---

[74] Let *V(t)* represent the sum of all future rewards, and $\hat{V}(t)$ represent an estimation of that sum. If the agent in fact acts on V(t) during the trial, an estimate $\hat{V}(t)$ of the sum of all future rewards must exist. Sutton and Barto argue that the possibility of such an estimate is present within *V(t)*, because the changes of *V(t)* are consistent through time. Through this consistency in *V(t)*, $\hat{V}(t)$ can be estimated at successive time steps. Errors in successive estimations $\hat{V}(t)$ are the temporal difference errors at the center of the reinforcement theory. These temporal differences amongst instances of $(\hat{V}(t))$ act as a proxy prediction error for the value of all rewards available to the agent in the trial, which may be utilized during the trial in order to improve the estimates $\hat{V}(t)$ of *V(t)*. Improving the estimates of *V(t)* during a learning trial is the primary mechanism by which, according to Sutton and Barto, the agent aims to maximize her rewards.

[75] To construct and use an error signal similar to the TD error above, a neural system would need to possess four basic features: "(i) access to a measure of reward value r(t); (ii) a signal measuring the temporal derivative of the ongoing prediction of reward

$$\gamma\hat{V}(t+1) - \hat{\dot{V}}(t)$$

; (iii) a site where these signals could be summed; and (vi) delivery of the error signal to areas constructing the prediction in such a way that it can control plasticity." (Schultz et al., 1997, p. 1595)

> experimental data. This theory derives from reinforcement learning [as framed by Sutton and Barto], which shows how a particular form of the error (called the temporal difference error) can be used to learn predictions of reward delivery and also how the predictions can be used to learn to choose an adaptive course of action in terms of maximizing reward and minimizing punishment.[76]

So, Schroeder's view of reward is that firings of the VTA/SNpc function as a contingency-based learning system (based on the TD algorithm) because they fire in the pattern necessary for predicting differences between expected and actual rewards. As Schroeder puts it:

> …VTA/SNpc neurons fire in a pattern that carries information about the difference, at time *t*, between the rewards received and expected at *t* versus those rewards the organism was predicting (at *t*-1) it would receive or expect at *t*…Such a signal has been shown to be exactly the sort of signal required for reward-based reinforcement learning. That is, such a signal is exactly what is most computationally useful if a system is going to modify itself adaptively on the basis of rewards received.[77]

Schroeder's essential claim here is that the firings of the VTA/SNpc function as a contingency-based learning system because they fire in the pattern necessary for predicting differences between expected and actual rewards. The conclusion that Schroeder draws from this is that there is empirical evidence that we have a structure in our brains capable of fulfilling the desiderata of reinforcement learning. Representations of events that produce such learning signals make the represented events rewards for Schroeder.

From this Schroeder concludes that the VTA/SNpc provides the biological basis of what it is for something to count as a reward. If it is in fact plausible to conclude that this is just what reward is, then this is persuasive evidence in favor of his Contingency-based Learning Theory of Reward. And vindicating this theory of reward is an important step for Schroeder in proving his Reward Theory of Desire. So, should we be persuaded by Schroeder's theory of reward?

---

[76] Dayan & Balleine, 2002, p. 285.
[77] Schroeder, 2004, p. 50.

Schroeder's argument for the Contingency-based Learning Theory of Reward (CLTR) is given against a background of three possible theories of reward—CLTR, the Behavioral Theory of Reward, and the Hedonic Theory of Reward. The Behavioral Theory of Reward is the view that reward is a stimulus which tends to increase any behavior which it follows. The classic example would be 'rewarding' a rat with sugar water. If giving the rat sugar water every time it presses a lever tends to increase the rat's lever pressing behavior, then the sugar water is a reward. The Hedonic Theory of Reward is the view that a reward is anything which gives pleasure. Schroeder argues for CLTR by identifying "…straightforward theoretical benefits to thinking of contingency-based learning as the nature of reward and punishment."[78] The theoretical benefit that Schroeder has in mind is that of explanatory power.

Schroeder argues that both the behavioral and hedonic theories of reward are vulnerable to the following objection: If behavior (or pleasure), just is reward, then reward *cannot cause* behavior (or pleasure). That reward causes both behavior and pleasure is an important element of the common-sense understanding of reward. But if we *identify* reward with behavior (or pleasure), then, he claims, any such causal story we may tell is trivial. CLTR, in contrast, allows us to maintain this common-sense causal intuition. Reward, if understood in terms of contingency-based learning, can be easily and reasonably understood as a cause of both behavior and pleasure. Schroeder argues that any theory of reward which does not suffer from this objection gains an advantage due to the robustness of this causal intuition, and the CRLT is just such a theory. [79]

A second advantage of CLTR is along the same lines—it captures a variety of other common sense intuitions about reward:

---

[78] Schroeder, 2004, p. 62.
[79] Schroeder, 2004, p. 62.

> …rewards **cause** emotional changes, they directly motivate via deliberation and inculcate unconscious effective behavioral dispositions, they inculcate intrinsic desires, they modify intellectual dispositions, and they can, perhaps, modify sensory capabilities.[80]

Most of these effects have, to some degree, been demonstrated to be mediated in certain cases through VTA/SNpc teaching signals. That is, the effects have been demonstrated as being caused by temporal difference errors in reward prediction. Schroeder is very optimistic that those effects which have not yet been demonstrated as resulting from this system in certain cases (the inculcation of intrinsic desires and the unconscious modification of intellectual dispositions)[81], will be so when the appropriate experiments are performed. Thus CLTR gives an account of how these aspects of reward work.

The final theoretical benefit of the CLTR identified by Schroeder is that the theory captures everything learnt about reward through behaviorism (although not necessarily in precisely the same terms.) That is, those things considered to be rewards and punishments in the behaviorist tradition continue to count as rewards, and rewards are still capable of modifying immediate behavior, behavioral dispositions, and intellectual dispositions.[82] Moreover the theory is better than a purely behaviorist theory of reward as we still get to explain how reward *causes* these things, rather than merely identifying reward with these behaviors. The robustness of this argument for the CRLT depends upon the plausibility of the claim that we take rewards to cause pleasure and motivation. I am inclined to accept Schroeder's argument for this point as this does seem to be a *prima facie* plausible interpretation of the folk theory of reward.

So, as an account of reward, Schroeder's view looks better than the alternatives, the hedonic and behavioral views. It is both physiologically and psychologically more plausible, and has greater explanatory power in a number of dimensions than its competition. Thus, the

---

[80] Schroeder, 2004, p.63. Emphasis added.
[81] Schroeder, 2004, p. 62.
[82] Schroeder, 2004, p. 64.

claim of CRLT that for an event to be a reward for an organism is for representations of that event to contribute to the production of a reinforcement signal in the organism in the sense made clear by the theory of reinforcement learning is the most persuasive theory of reward. Having established his view of reward, we are now in a position to make clear what the reward face of desire is, and how it is that this could be the fundamental face of desire.

The formal statement of Schroeder's Reward Theory of Desire is:

> Reward Theory of Desire (RTD): To have an intrinsic (positive) desire that *P* is to use the capacity to perceptually or cognitively represent that *P* to constitute *P* as a reward. To be averse to it being the case that *P* is to use the capacity to perceptually or cognitively represent that P to constitute *P* as a punishment.[83]

The first thing to note about this theory is that it is a theory of intrinsic desires, a theory about what it is to desire to that your spouse be happy, that you have a good life, that you not smell the skunk, that you see Balanchine, etc.. The next point is that it distinguishes between desires and aversions, the positive and negative aspects of desire. This distinction is, for Schroeder, grounded in intuition and justified by physiology. He claims that the brain distinguishes between these two cases in the working of the reward center. This particular element of his argument is beyond the scope of this discussion and will have no bearing on the outcome, so I will put it to the side.

The final element of Schroeder's view that needs to be explicated is what it is to 'constitute' something as a reward. The account he gives of this constitution relation is functionalist: "…to constitute something as a reward or punishment is to use a representation of it to drive the production of a reward or punishment signal."[84] So a representation which is

---

[83] Schroeder, 2004, p. 131.

[84] "Reward and punishment signals, in turn, are to be understood in terms of learning theory. A reward signal is an event that causes a characteristic, mathematically describable form of learning, and a punishment signal is an event that causing as opposing form of learning…[T]his is learning in a very specific sense: it is a change in the connectivities of units that are themselves describable at an appropriately abstract level…Hence, if something is a causal system that is mathematically describable

constitutive of a reward is a representation that produces a learning signal of the kind specified by the CLTR. In sum, to have an intrinsic (positive) desire that *P* is to use the capacity to perceptually or cognitively represent that *P* to produce a reinforcement signal, thus constituting *P* as a reward.[85]

I take it that by giving a physiologically plausible account of reward, and demonstrating how it is linked to desires, Schroeder has succeeded in showing that reward is a neglected face of desire. This leaves his further conclusions to consider. The first is the issue of how the reward face of desire stands in relation to the other two faces of desire—pleasure and motivation—and which of them is the most fundamental (both physiologically and conceptually), and thus the best candidate for instantiating desires. The second is Schroeder's main conclusion, the claim that intrinsic desires just are rewards.

Desire: Reward, Motivation, or Pleasure?

Schroeder makes the argument that the reward face of desire is more fundamental than the motivational or hedonic faces of desire in two parts, one physical, and one conceptual. Schroeder's central, and persuasive, argument for the physical claim that the reward face of desire is more fundamental than the other two faces of desire is a physiological one. His

---

as instantiating contingency-based learning, then it is the site of such learning." (Schroeder, 2004, p. 134-5.)

[85] An issue that may give one pause about the RTD is its reliance on representations, and question of whether or not you have a desire in the absence of the relevant representation. Schroeder carefully points out that the RTD does not commit us to an episodic account of desire in the following extract: "The fact that desires include perceptual and cognitive representations as proper parts is likely to strike many as confusing. Does this mean that every time one perceives a pie one also desires it? Does it mean that one cannot desire the well-being of a child without thinking of that child's well-being? No. According to the theory, desires need not involve tokened representations, need not involve actual episodes of representing pies or well-being. Rather, to desire is to be so organized that tokened representations of pie or well-being, if they occur, will contribute to the production of reward signals…This is why RTD requires a link between representational *capacities* and reward signals, rather than a link between occurrent representations and reward signals." (Schroeder, 2004, p. 134.)

conceptual argument is an argument about the relative explanatory power of the three views of desire.

The first, physiological, argument depends upon the claim that the activity that Schroeder identifies in the VTA/SNpc (reward center) as the reward signal is plausibly the cause of activity in the other centers of desire in many simple cases of intrinsic desiring. He begins by identifying the locations in the brain which are primarily responsible for reward, pleasure, and motivation. It is uncontroversial that the reward center is connected to those structures in the brain most implicated in hedonic tone and the motivation of action. However Schroeder goes one step further in making a claim about the specific nature of this connection. He contends that the physiological structures that instantiate motivation and pleasure are less fundamental in the brain than the reward structure as identified by Schulz et al. As Schroeder puts it:

> Desire's best-known face, motivation, seems to stem from the brain's reward system. Desire's other well-known face, pleasure, seems to represent the activity of the reward system. And desire's neglected face, reward, is constituted by the activity of the reward system.[86]

So, Schroeder argues that the reward is the physical instantiation of desire because it is the most fundamental physical cause of desires. This is the physical claim that desires are rewards. His evidence for this view is that the reward structure (the VTA/SNpc), is **physiologically prior** to the structures which instantiate motivation and pleasure. In this discussion I will refer to the VTA/SNpc as the **reward center**. The *perigenual anterior cingulated cortex* (PGAC) is identified as the neural home of hedonic experiences; I will call this the **pleasure center**.[87] I shall use the term **motor center** to refer to the motor striatum, which is primarily implicated in voluntary action as it is the system that selects among the possible actions that are formed

[86] Schroeder, 2004, p. 131.
[87] Schroeder, 2004, pp. 78-83.

in the motor cortex, which is the neural seat of what Davidson refers to as 'primitive acts'.[88] Thus the motor center receives input from the motor cortex, the *supplementary motor area* (SMA), and the motor region of the *anterior cingulate cortex* (AC), which are generally held to be the important seats of motivation.

Vindication of this claim requires that there is evidence that the activities of the reward center *cause* activities in both the pleasure center and the motor center, and that there is *no direct causal connection* between the pleasure center and the motor center. Thus there are three claims about causal relationships amongst these centers that Schroeder must defend: (1) <reward center *directly causes activity in* pleasure center>; and the (2) <reward center *directly causes activity in* motor center>; while (3) <motor center *does not directly cause activity in* pleasure center> and < pleasure center *does not directly cause activity in* motor center>

The first causal connection, the claim that activity in the reward center causes activity in the pleasure center is supported by evidence of the neural pathways between the reward center and pleasure center, and the operation of dopamine on these two structures. These structures are such that, as Schroeder puts it: "…the activity of the reward system is a normal *cause* of pleasure."[89] However note that these two systems are independent as although the operation of the reward center is a *sufficient* cause of activity in the pleasure center, it is not a *necessary* cause. Berridge et al demonstrated that rats with extensive lesions to the reward center were still able to experience pleasure and displeasure, as indicated by facial expression.[90]

---

[88] Primitive actions are those actions which are performed directly, without performing any other actions to bring them about. Lifting a glass is a primitive action, while eating chocolate is not. (Davidson, 1980, ch 3.)

[89] Schroeder, 2004, p. 36; pp. 76-83.

[90] Berridge and Robinson (1998) cited by Schroeder, (p. 81). In addition Schroeder argues: "There are at least two tremendously important facts about the PGAC. The first is that it is wholly distinct from the VTA/SNpc: that the neural basis of pleasure, therefore, is not identical to the neural basis of reward,

The second causal connection, the claim that activity in the reward center causes activity in the motor center is supported by studies which demonstrate that destruction of the reward center causes the complete absence of voluntary movement[91] as evidence that reward is causally prior to motivation.

The lack of the third connections—that pleasure does not cause motivation, and motivation does not cause pleasure—is shown by the lack of neuroscientific evidence for a direct connection between the motor center and the pleasure center.[92] Experiments have shown that the connections between these two areas are indirect. Individuals who have significant damage to the pleasure center (and thus lose hedonic tone) are still capable of goal-directed action. However due to the lack of hedonic tone they fail to get distressed or excited about the prospects of the success or failure of goal-directed action.[93] It is also possible to manipulate motivation in organisms without changing their hedonic responses.[94]

So both the hedonic and motivational faces of desire seem to be the effects of a common cause, this common cause being the reward face of desire. Thus, the reward system is the seat of a physiologically more fundamental face of desire than either motivation or

---

and so pleasure is not identical to a reward signal. The second is that the PGAC is far from the only source of excitatory input to the VTA/SNpc: that pleasure, therefore, is not the only thing that is rewarding." (Schroeder, 2004, p. 36.)

[91] Berridge & Robinson, 1998; Langston & Palfreman, 1995.

[92] "Work done in localizing the neural basis of motivation has centered around the *motor cortex*, the main region of the brain sending signal directly to the spinal neurons controlling the voluntary muscles. However, the selection of an actual action from the range of plausible actions is not performed by the motor cortex, but by another structure deep in the brain, known as the motor center of the *basal ganglia*. The basal ganglia guide both conscious, pre-planned action and spontaneous action. It may come as a surprise to learn that the connections between the neural basis of pleasure, in the PGAC, and the control of the voluntary muscles appears to be fairly modest. Instead of pleasure dominating motivation, motivation appears much more influenced by the neural basis of reward in the VTA/SNpc. Reward signals from the VTA/SNpc appear to have a very important influence upon the basal ganglia, both in the short term (influencing immediate motivation) and in the long term (guiding the formation of behavioral tendencies). (Schroeder, 2004, pp. 36-7.)

[93] Foltz & White, 1962.

[94] Berridge & Robinson, 1998.

pleasure. Schroeder's argument for the physiological primacy of the reward center is well supported by current neuroscience, so this claim is plausible.

Schroeder argues for the conceptual priority of the reward face of desire by appealing to the relative explanatory power of the three views. That activity in the reward center cause activity in both the pleasure center and motor center allows for appeals to reward in order to *explain* what happens in both the pleasure center and the motor center. It is a part of the folk theory of desire that our desire *cause* our actions, and that having desires satisfied *cause* pleasure. Having a view of desire that allows for these causal connections is thus, he claims, doubly satisfactory. In the first place it accords with our common sense intuitions about the causal connections amongst these states, which is always a desirable characteristic in an otherwise counter-intuitive theory. Maintaining these causal relations is also explanatorily beneficial. Separating the essence of desire from its 'public faces', as it were, increases the explanatory power of the view because such a theory of desire gets to say informative things about why desires motivate, and why having our desires satisfied is pleasurable. In these respects the RTD is explanatorily more powerful than its rivals. This argument rests on both the strength of our intuitions about the precise nature of the causal connections amongst these states, and whether or not we are willing to bite a bullet on the loss of this explanatory power. It is certainly correct that we take there to be close connections amongst reward, motivation, and pleasure. However I dispute Schroeder's claim that the folk understanding of these connections is so clearly causal. The folk theory, but its very nature, is not that specific about the types of connections that exist amongst these states, and I think that Schroeder is going beyond the available evidence in drawing this conceptual conclusion.

This then leaves the issue of Schroeder's final conclusion, the claim that the Reward Theory of Desire is the correct theory of intrinsic desire. Even if it is true that the reward face of desire is physiologically more fundamental than its two rivals, it is still an open question

whether or not it is a plausible theory of intrinsic desire, or simply a more fundamental feature of some deeper account of desire. I will argue that the latter is true, on the grounds that Schroeder's theory of reward, the CRLT that is at the center of his Reward Theory of Desire, appeals to an antecedent (and thus more fundamental) notion of intrinsic desire.

The Reward Theory of Desire as 'the' Theory of Desire.

The aspect of Schroeder's view that is problematic for his final conclusion is his view of what it is for a representation to be *constituted* as a reward or punishment. The account he gives of what it is to constitute some representation as a reward or a punishment is functionalist. It is that "…to constitute something as a reward or punishment is to use a representation of it to drive the production of a reward or punishment signal."[95] Thus a representation which is constitutive of a reward is whatever it is that produces a reward signal.

Now, there must be space in this view for representations that do, and do not, constitute rewards. Throughout the course of any day we all have multiple representations that should not turn out to be rewards. So, what is the difference between a representation that produces a reward signal, and a representation that does not? However, Schroeder's account does not have an answer to this question. This is a very suggestive gap in his view. I think that Schroeder is obscuring a crucial assumption in his view by giving such a bare functionalist account of this constitution relation. This assumption is buried in his reliance on the reinforcement theory of reward.

---

[95] "Reward and punishment signals, in turn, are to be understood in terms of learning theory. A reward signal is an event that causes a characteristic, mathematically describable form of learning, and a punishment signal is an event that causes an opposing form of learning…[T]his is learning in a very specific sense: it is a change in the connectivities of units that are themselves describable at an appropriately abstract level…Hence, if something is a causal system that is mathematically describable as instantiating contingency-based learning, then it is the site of such learning." (Schroeder, 2004, p. 134-5.)

Let us take a closer look at the actual cases that Schroeder relies upon in defending his Contingency-based Learning Theory of Reward. The question of how it is that psychological concepts such as reward, pleasure, and motivation are instantiated in the brain is the territory of a group of relatively recent scientific disciplines which operate at the intersection of neuroscience and psychology, such as biopsychology, neuropsychology, affective neuroscience, behavioral neuroscience, etc. As interdisciplinary enterprises, these endeavors take their concepts from both sides of the fence. Neuroscience contributes information about the grey squishy bits, while organizing concepts such as pleasure, motivation, and reward come from psychology. These psychological concepts organize neurological research by identifying the experiences and activities that are characteristic of such concepts, thus establishing what it is that is happening in the brain when the agent experiences certain things, or behaves in certain ways. However this is not a one way street. If research begins with the view that liking something and wanting to get it are identical, yet through brain manipulations it discovers that liking and wanting can be disconnected, then we learn that these are in fact two distinct concepts. This process of exploration results in a type of reflective equilibrium among psychological concepts and neural instantiations. So, in order to judge how plausible Schroeder's position is, we need to be clear about the experiments which he is basing his conclusions on, specifically the concepts, and the characteristic experiences and behaviors that are taken to fall under them.

The theory of Schulz et al which is so central to Schroeder's claim relies on just such a process of reflective equilibrium. The idea of reward is taken from psychology. The instantiation in the brain is discovered through the neuroscientific technique of monitoring the firing of particular neurons during some event which will count as reward for the creature in psychological terms. The principle investigations which are taken to support their theory entail the tracking of dopamine receptors in the VTA/SNpc of monkeys who are being 'rewarded' by tiny amounts of fruit juice. Schulz's hypothesis was that understanding reward

processing at a neuronal level would help us to understand voluntary, goal-directed behavior. There were three trials, each with a red square trigger stimulus. In the first condition the monkey would be rewarded for moving (pressing a small lever under where their right hand was resting); in the second condition the monkey would be rewarded for not moving its hand; in the final condition the monkey was not rewarded for moving its hand. In order to receive the reward of 1.5ml of apple juice, the monkey had to perform the required (non)action 3.5 seconds after the trigger stimulus. The reward was delivered 1.5 seconds after the action was performed.[96] The structure here is of operant conditioning achieved through association between the conditioned stimulus and the unconditioned. The conditioned stimulus was the red square, and the unconditioned stimulus was the apple juice.

The experiment recorded the activation of individual neurons within the target zones while the monkeys were performing computer controlled behavioral tasks. The fruit juice functions as a reward because it is intuitively an unconditioned stimulus for the monkeys.[97] So, Schroeder's claim that reward is a certain type of learning signal *qua* neuroscientific, is in its fullest expression the claim that: reward *understood as that response which is elicited in monkeys by fruit juice* is a certain type of learning signal. I think a key question that must be asked of this study is why it is that fruit juice elicits the reward signal that Schroeder identifies as 'reward'.[98] [99]

---

[96] Schultz et al., 2000.
[97] Schultz, Tremblay, & Hollerman, 2000.
[98] Schulz et al define reward as "…an operational concept for describing the positive value that a creature ascribes to an object, a behavioral act, or an internal physical state. The function of reward can be described according to the behavior elicited. For example, appetitive or rewarding stimuli induce approach behavior that permits as animal to consume. Rewards may also play the role of positive reinforcers where they increase the frequency of behavioral reactions during learning and maintain well-established appetitive behaviors after learning." So, Schroeder is proposing a theory of desire as reward, where the notion of reward at issue is operational—essentially, that which is capable of conditioning the subject. (Schultz et al., 1997, p. 1593.)
[99] Not unsurprisingly, talk of a 'reward-based' analysis of desire in this sense causes the specter of behaviorism and Pavlovian conditioning to loom, an association that Schroeder is eager (but perhaps cannot) do away with. He draws a hard distinction between what he calls the 'doctrinaire' behaviorist,

Recall that Sutton and Barto take their reinforcement theory of learning to capture the Rescorla-Wagner model[100] of what is going on in classical and operant conditioning—the idea that learning is a matter of predicting the unconditioned stimulus by using the conditioned stimulus as evidence for its imminent occurrence. In such conditioning the organism has certain automatic responses to certain stimuli. Conditioning occurs when a behavior is elicited from some subject by repeatedly pairing the conditioned stimulus with some unconditioned stimulus which the subject naturally finds attractive. It is the aim of TD algorithm to capture what is going on in such a model of conditioning. In such cases it is the unconditioned stimulus that drives the subject's behavior, in that it is the unconditioned stimulus that generates the reward function by determining the intrinsic desirability of each state, and thus the relative values (the sum of rewards available for the outcomes of a given plan of action) of possible policies for action. According to the TD algorithm, the subject learns what the value of some state-action pair is through changes in successive estimations of the value of implementing particular policies of action. This then gives the agent a way of identifying that policy of action which has the greatest value, so she can act so as to maximize her numerical

---

who holds the unflinchingly bullet-biting position that rewards and punishments just are their effects on behavior, and the neuroscientific conception. Then he points out—correctly, if somewhat simplistically—that neuroscientists are engaged in finding the underlying structure through which reward is expressed in the brain, and thus they are not behaviorists. By looking at reward in neuroscientific terms, which gives us a way of understanding reward independently from the behaviorist's focus on behavior and pleasure, Schroeder claims to avoid this implication. If, however, we take the in principle objection to behaviorism to be that they fail to take seriously the complex interior life which accompanies these phenomena in beings like us, then it is not clear how much better current neuroscience is doing. But given that his neuroscientific understanding is so heavily rooted in this tradition, does his disavowal really work? This is an interesting question, although not one which is strictly relevant to the present discussion, so I will leave it to a later time.

[100] There is some question about what happens in both classical and operant conditioning to the organism when 'learns' some behavior. The Rescorla-Wagner model is an account of how such learning proceeds. According to this model, animals learn to *predict* the imminent advent of the unconditioned stimulus by the presence of the conditioned stimulus from repeated pairings of the unconditioned stimulus with the conditioned stimulus. This prediction hypothesis can be understood as the conditioned stimulus behaving as *evidence*. If we apply the Rescorla-Wagner model to operant conditioning cases, where a conditioned stimulus cues a behavior which is followed by an unconditioned stimulus, then the conditioned stimulus will be evidence for the organism of a forthcoming unconditioned stimulus, if the organism performs the CR.

reward signal. So, if we ask the question of what determines the reward function and thus the value of the outcomes to the agent, the answer will be given in terms of the intrinsic desirability of the unconditioned stimulus to the subject.

Thus Schroeder's reliance on the theory of Reinforcement Learning leaves him vulnerable to the following argument:

> What it is for something to be an intrinsic desire is that it is constituted by the agent as a reward. (Reward Theory of Desire)
>
> A representational state is constituted by the agent as a reward if it produces a reward signal. (Contingency-based Learning Theory of Reward)
>
> A reward signal measures the difference between actual and expected value. (Theory of Reinforcement Learning)
>
> The value of a state of affairs is determined by a reward function. (Theory of Reinforcement Learning)
>
> A reward function attaches positive value to a representation if and only if the object of the representation is already intrinsically desirable to the agent. (Theory of Reinforcement Learning)
>
> Thus a representation produces a reward signal (is constituted as a reward in Schroeder's terms) iff the representation is of something that the agent antecedently takes to be intrinsically desirable.
>
> ---
>
> Therefore, the essential nature of an intrinsic desire cannot be that it is constituted by the agent as a reward.

In short, representations are not intrinsic desires *because* they are constituted by the agent as rewards, but rather that they are constituted by the agent as rewards *because they are intrinsic desires.* The model of reinforcement learning relied upon so heavily by Schroeder presupposes a fully fledged view of intrinsic desires. Thus Schroeder's reward view of desire is not a plausible view of what intrinsic desires are. There is a more fundamental notion of intrinsic desire in play that explains why some representations contribute to a reward signal while others do not.

How is it that Schroeder missed such an important step in his explanation of reward in terms of contingency-based learning? A part of the clue is to be had in his somewhat abrupt

dismissal of, and thus lack of exposition about, the types of computational learning theory that do much of the work in his view of reward.

> Reinforcement learning is driven by…information about the difference between actual and expected contingencies…But these details are just that: details…So I propose to gloss over the particular mathematical details, leaving them to those interested in computational theories of learning and neural modeling.[101]

Unfortunately for Schroeder, the devil, as is so often the case, is in the detail. On close examination it turns out that Schroeder's view of reward, and thus his view of desire, relies on the presupposition of desiring built into the theory of Reinforcement Learning that grounds his CLTR. His view yields an account of intrinsic desires because it presupposes them. The argument that he gives has the form of a 'desire in—desire out' argument, so he cannot, by relying on such an argument, discover the fundamental nature of desires. Now, this argument does not completely invalidate Schroeder's view, just his claim that intrinsic desires are rewards. What, then, is the import of Schroeder's Reward Theory of Desire? An in depth discussion of this issue is beyond the scope of this chapter, but I will give a sketch of a response.

If we take Schroeder's central claim that reward is best understood as a *process of cognitive learning* seriously, then the reward face of desire can be understood as a theory of desire change. There are two points in favor of this interpretation. The first is that this interpretation of the RTD goes some distance toward explaining why Schroeder may have taken it to be a theory of nature of intrinsic desires. What Schroeder is in essence claiming in the Reward Theory of Desire is that desires are a certain type of learning function.[102] This is suspicious because it seems to be conflating what a desire is at a given time, with a process through which desires change. Once we take seriously the idea that desires are dynamic states

---

[101] Schroeder, 2004, p. 66.
[102] I am unsure whether I should be talking about processes, structures, or functions here, but I cannot see anything that hangs on the difference between these cases at this point.

that persist over time, and assume that they can change in response to certain experiences, then it makes more sense to understand Schroeder's RTD as a theory of *desire change*. Understanding Schroeder's account of desire as an account of desire change also helps make sense of why he would argue for it in the first place. It is relatively common that someone who takes himself to be giving a theory of *p* is in fact giving a theory of change in *p*, and it is this error which I am attributing to Schroeder. He takes himself to be giving an account of intrinsic desires, while he is in fact giving an account of how it is that such intrinsic desires change.

I claim that Schroeder's picture is suggestive of, in its most plausible form, this process. The RTD, taken as a theory of desire change, captures a mechanism through which our desires can change as a response to information about the actual and estimated value of an outcome to the agent. If we follow a desire through time, then this describes a process where this desire is either *reinforced* or *undermined* by experiences. Intrinsic desires enter into the reward system, and are altered through the cognitive and conative processes that are instantiated in, and driven by, the reward system. Simplistically we can say that the learning element of the theory describes a process such that if the rewarding-ness of a state, experience or outcome exceeds the way in which it is predicted to be it will be reinforced. If the rewarding-ness of the state does not live up to expectations of its value it will be undermined. The details of how this learning occurs are the details of reinforcement learning. I consider that Schroeder's view, taken as an account of desire change, is an important contribution to a more realistic understanding of desires.

In sum, Schroeder argues persuasively that reward is a physiologically more fundamental face of desire than pleasure or motivation. Although I take Schroeder's Contingency-based Learning Account of Reward to be the correct account of how our desires function as rewards, and connect to both the pleasure and motivational centers; it is not a

plausible theory of the essential nature of our concept of intrinsic desires. I do not want to take a strong stand of which of these three faces of desire is 'the' face of desire, in part because I take the most plausible view of desire to be a cluster concept which includes all three of these elements. Now, none the three theories presented here can answer the challenge of temptation as temptations do not differ from other desires in terms of their hedonic aspects, motivational structure, or rewarding-ness. However, there is a suggestive gap in Schroeder's Reward Theory of Desire which indicates where we should look next for an answer to the challenge of temptation. The gap is the question of why it is that some representations are constituted as rewards, while others are not.

CHAPTER 4

THE EMOTIONAL COMPONENT OF DESIRES

I propose that what is missing from the views of desire canvassed in the previous chapters is an explicit acceptance that our desires are intimately connected to certain kinds of emotional responses that we have to things. I will argue that not only do these emotional components of desire exist, but that they affect our desires, and they can be manipulated in certain ways. Considering the connection between our desires and these emotional components of desire will yield an account of what causes temptations, and why they should not be treated on a par with other desires in deliberation.

The aim of this chapter is to argue that what is needed in our understanding of desires is an acknowledgment of the form and influence of the emotional component of desires. My aim is to defend the idea that there are characteristic phenomenological corollaries of desires—what I will call the 'emotional responses of desiring'—which are intimately connected to our desires.

The view that there is a felt emotional component of desires is driven by two distinct concerns. The first, and most influential, is the evidence of the phenomenology of certain common cases of desiring. This is the phenomenology of simply wanting some thing, in the absence of any robust explanation of why this should be the case. Such emotional experiences of desiring behave, in many ways, like a response to that thing on a par with various

perceptual responses. Like the scent of a rose, the wanting has distinctive and identifiable qualia. The second concern is more theoretical. This is the gap in Schroeder's Reward Theory of Desire. What was lacking in Schroeder's Reward Theory of Desire was an answer to the question of why it is that some representations drive reward signals, while others do not. By paying attention to the felt emotional aspect of desires, I propose that this gap can be filled.

These emotional responses of desiring are the phenomenological states characteristic of experiencing an urge, yearning, craving, desire response, attraction, want or desire for some thing. Such felt emotional components of desires are not the same as desires, but they are intimately connected with them. If I look at a painting and have an emotional response of desiring, then this can count as evidence that I have a desire for the painting. It is not usually evidence for me that I have a desire, but if I express this kind of emotional response through comments, expression, gestures, etc., then from the third personal point of view, this counts as evidence that I desire the painting.[103] If I desire to pursue some career, and represent it to myself in great detail yet have no emotional response of desiring to the prospect, then this looks like evidence that I don't 'really' desire that career. However such responses are neither necessary nor sufficient conditions for desires, as both of the connections I have described are defeasible. For instance, someone who is close to his brother, and looks at his sister-in-law and feels such an emotional response of desiring, may well squelch this response rather than forming the attendant desire for his brother's wife. Conversely, someone who desires to grade papers, and represents to herself the prospect of grading those papers may well lack the emotional response of desiring to this representation, but this is not likely to call into question

[103] In most cases, the agent herself will not use this as evidence that she has the desire, as she has access to the desire and is thus in no need of evidence one way or the other. However, if her access to her desire is being disturbed in some way, then she can take a third personal perspective in order to 'discover' what her desires are, and in this case the emotional response can count as evidence of her desires.

her desire to grade the papers. Thus the felt emotional component often goes along with desires, but they are neither necessary nor sufficient for having a desire.

In the normal case, when we have such a felt emotional component of desire, we have the desire too. So the felt emotional component of desires is an emotional correlate of certain desires. Although there is a close connection between full fledged desires and the felt emotional component of desires, I am not taking a stand on what this relationship is, as the precise relationship between such felt emotional components of desire and desires proper is not clear. However these emotional responses of desiring are important as they are often natural accompaniments to full-fledged desires, and their existence provides certain types of evidence about the accompanying desires.

My presentation of how this felt emotional component of desires works will be broken up into several sections. In the first instance, I will be concerned with motivating the existence of responses of the type required for the felt emotional component of desires. The second section will explore how such a response would behave by drawing an analogy to the types of responses relied upon in various theories of response dependence. The third section will again rely on an analogy to a response-dependence theory. This time I will use a looser analogy to the response-dependent theory of emotion proposed by D'Arms and Jacobson, in order to argue that the felt emotional component of desires can be more or less fitting. Such an account of fitness requires explicit constraints on responses in order to generate an account of when such a response is appropriate in the sense that it 'fits' the stimulus. This then leads to the next and final chapter, where I will identify various instances of failure in the conative system, failures that produce characteristic cases of temptation through manipulation of the felt emotional component of desires. It is this account of fitness that will ultimately show how paying attention to this emotional component of desires can give an answer the challenge of temptation.

Automatic Affective Responses

The felt emotional component of desire is meant to capture the experience of considering something and wanting it. Although note that by 'wanting' I do not mean to include the idea that such wanting is necessarily motivational. I desire that my husband surprises me with flowers, but my doing anything to bring this end about (such as mentioning it), robs the gesture of the element of spontaneity which is part of what I desire in it. My desire requires no *motivation*, although it clearly favors *satisfaction*. What I am interested in here is the experience of having a particular emotional response to some outcome, a response which is characteristic of the felt experience of desiring.

I propose that the felt emotional component of desires is a type of **automatic affective response** to some object, call this the 'desire response'. Such responses occur when an agent has a particular characteristic response to external stimuli. A response is automatic if it is outside of the conscious control of the agent, and it is affective if it encodes an attitude (affect) of the agent. Such a response encodes several pieces of information. It has a valence, which determines whether the object of the response should be approached or avoided. It also has a felt intensity, which acts as a proxy (although not a necessarily reliable one), for a variety of information about the agent's attitude toward the thing. The fact of such responses can provide evidence for how much the agent is motivated to get that thing (in cases where the desire is a desire for some thing), or evidence for how strong the agent's desire for that thing is, or how appealing that thing is to the agent, etc.

The view that our representations can have a valence and felt intensity which is supplied by some automatic affective response is not a novel one. Baumeister et al argue that emotions cause behavior through a feedback mechanism which relies on a dual process view of emotions. According to the dual process view, emotions are sorted into two distinct processes, automatic and consciously controlled, both of which participate in the feedback mechanism.

Such dual process views are becoming widely accepted in talk of cognition.[104] In emotions, the controlled process yields those things that count as what we tend to take as 'full-blown' emotions, such as fully worked out desires for specific careers, or particular activities. These are consciously experienced complex states. In contrast, the automatic process yields far simpler states that may not even be consciously detected by the agent. These are something like the 'unconscious emotions' proposed by Wienkelman and Berridge,[105] although I have no commitment to the idea that these states be unconscious.

Automatic affective responses are composed of a representation, and at least one association.[106] The association is with other mental items such as patterns or memories. The mechanism through which such responses are formed is similar to the way in which representations and associations comprise the recognition of objects. The felt emotional component of desires constituted by such automatic affective responses need no inferences, [107] they simply occur to the agent. The idea that this felt experience of desiring is automatic—that is, is outside of the conscious control of the agent—reflects the idea that the emotional aspect of desiring, like other emotions, is not the sort of thing that we can *choose* to experience.

A feature of such desire responses is that they attach to other representations. Desire responses are additional to the basic sensory representation of an object, state of affairs, possibility, idea, etc. If this were not the case, then representing something would be sufficient for having a desire response to that thing, but this can't be right, as is shown by the gap in Schroeder's Reward Theory of Desire. I take the claim that there is such a felt emotional component of desires to be continuous with Schroeder's Reward Theory of Desire. What it

---

[104] Chaiken & Trope; Wilson.
[105] (Winkielman & Berridge, 2004) Their work provides persuasive evidence that "…people have automatic affective reactions (such as liking and disliking something) that are simple and rapid and may well guide online behavior and quick reactions."(Baumeister, Vohs, DeWall, & Zhang, p. 168.)
[106] Such automatic affective states are proposed by (Kunst-Wilson & Zajonc; Murphy, Monahan, & Zajonc; and Winkielman & Berridge, 2004).
[107] Baumeister et al., p. 168.

does is supply the missing element of Schroeder's view, by providing a way to differentiate between representations that drive a reward signal, and representations that do not. What plausibly provides such emotional responses a connection with many of the familiar clothes of desire is that it drives a reward signal in the agent, and thus engages her motivational and hedonic systems. Now, I do not mean to say that this is the only way in which a reward signal can be driven in Schroeder's account. Rather that in his trickiest case, the case of spontaneous intrinsic desires that have no cognitive elements, such automatic affective responses can differentiate between representations that drive a reward signal and representations that do not. Another way in which such a reward signal could be differentiated in cases such as the 'cold' intrinsic desire, that with no clear emotional affect, is through representations being representation of the good, the moral, the noble, etc. Thus the desire response can play the role of a missing element in Schroeder's view, how it is that reward signals are produced in cases of spontaneous intrinsic desires. [108]

These felt emotional components of desires are the kinds of things that are often at the end of the explanatory road in Humean accounts of action. When the agent gets to the point that she cannot provide any other explanation for her desire than the response "I just like it" or "I just want it", what she is talking about is an affective response that is out of her conscious control—she is simply wired in such a way that she responds to representations of that thing positively. By saying this I do not mean to say that there is no explanation of why

---

[108] It is this aspect of automatic affective responses that Baumeister identifies as making them adaptive learning mechanisms for creatures like us. "Imagine an early human encountering a dangerous predator. For conscious emotion to mediate the flight, a sequence something like this would be necessary. The person must recognize the animal and cognitively appraise the danger. This gives rise to physiological arousal, which spreads through the person's body. The bodily response then triggers a further cognitive process involving the brain, which recognizes the bodily state as fear and on that basis initiates a motor response, and the person flees. This sequence is plausible, but it would take some time (at least seconds, more likely minutes), during which the person is continually exposed to danger. Humans or animals whose responses depended on such a sequence might therefore make relatively easy meals for quick-acting predators.
In contrast, automatic affect would arise in perhaps a tenth of a second, almost as soon as the predator is recognized" (Baumeister et al., p. 169.)

she has this response to the thing, but rather than the explanation is going to be psycho-physical—in the way that explaining why I perceive stop signs as red is psycho-physical—rather than cognitive or rational. In other words, such desire responses are the product of some sort of automatic system within the agent which I will call the 'conative system'. The idea of the conative system is similar to that of the perceptual system, although it is meant to capture the agent's psychological dispositions to value goods. As product of such an automatic system—a system that is outside of the direct conscious control of the agent—the appropriateness or otherwise of desire responses will be explained in terms of how well the conative system is working. It is in this sense that explanations of desire responses will by psycho-physical rather than cognitive. Note that I am not committing myself to the view that having such responses is either necessary or sufficient for desiring, rather, I am proposing that such responses are corollaries to desire in many common cases.

What is interesting about these desire responses is that they are the type of affective element of a representation that (1) arises as a response to the relation between the subject and the object; (2) are exactly what is needed to distinguish representations capable of producing reward signals from representations that are not; and (3) are unconsciously caused in the agent, so are a sufficient explanation of any action that they may cause.

In sum, the type of response that is at the center of the felt emotional component of desires that I am proposing is an automatic affective response. Such a response is the product of a representation of the thing, and various associations that the agent has with that thing. These states are both philosophically and psychologically plausible in that there is a place for such states in our folk and philosophical theories of desire. Scanlon, in his discussion of appetitive unmotivated desires, references such states in enumerating the elements of a desire for a drink. In Schroeder's discussion such states appear in the guise of unconditioned stimuli. In the folk view such states appear in the guise of appetites and urges. The view that there is

such a felt emotional component to our desires is deeply entrenched in the literature, yet an in-depth discussion of the import and characteristics of this element of desire is conspicuous in its absence.

Now, in order to make the view that there are such desire responses plausible, I need to do two things. The first is to flesh out the way in which such emotional responses work. The second is to show how such felt emotional component of desires can answer the challenge of temptation. I propose that both of these aims can be achieved through considering a loose analogy with D'Arms and Jacobson's response-dependent account of emotions. The first aspect of this view that I will consider is what it is to understand certain mental states as essentially relational, that is, in terms of certain characteristic responses to perceived or imagined properties. This will give us the sense of a response that is appropriate for understanding desire responses. The second is an account of when it is that such a response is 'fitting', that is, when it is the appropriate response to have. I will use this account of fitness to argue that desire responses can be similarly more and less fitting, and it is this account of fit that will ultimately solve the challenge of temptation.

Desires as Responses

The kind of responses that I am proposing as the felt emotional component of desires is an expression of a relation between the desirer and the object of the desire. One context in which such relational aspects of mental states have been discussed is in terms of the distinction between primary and secondary properties. Locke, in his *Essay Concerning Human Understanding* outlines the distinction between primary and secondary properties (qualities in his terminology) in the following way: Primary qualities are 'utterly inseparable from… [a] body', no matter how small (II. viii. 10), while secondary qualities 'are nothing in…objects

themselves, but powers to produce various sensations in us'.[109] The classic examples are shape for primary properties, and color for secondary properties. The intuition behind this distinction is that primary properties are somehow objective aspects of the world, while secondary properties exist in the relationship between the mind and the world. According to Locke, a secondary quality is not a part of the object itself, but rather a power of the object to produce certain sensations in us. Response-dependent properties are a way to capture these secondary qualities, while the theory of response-dependence is a way of cashing out just where these properties exist within this relation. My interest in this view is not in the properties themselves, but with the idea of such a response capturing information about the relation between the agent and the world—the kind of relation that is at work in desire responses.

The heart of a response-dependent analysis of a property is an identity claim of the form:

> **Response Dependent Property (RDP)**: *x* has *P* = *x* elicits response *R* in [normal] subjects *S* under [normal] conditions *C*.[110]

One classic use of the idea of response-dependence is in order to understand what properties such as *NAUSEATING*.[111] If we apply the RDP schema to *NAUSEATING*, then we get the result that "*NAUSEATING* is the [property] of [being] disposed to produce nausea in normal people (perhaps in other animals as well) under normal conditions."[112] The class of nauseating things will be whatever falls under the concept NAUSEATING, while the property of *NAUSEATING* will be the property by virtue of which any particular thing is a member of this class.

---

[109] Locke & Nidditch, 1975, II. viii. 10.
[110] LeBar, 2005, p. 181.
[111] Capitalisation of a term will indicate talk of a concept, while italicized caps will represent talk of a property.
[112] LeBar, 2005, p. 176.

Standardly, a response-dependent property is taken to possess a robust element of *invariance*. The RDP schema builds in a pair of normalcy conditions, one on the subject, and the other on the conditions. So, what it is for *x* to have *P* is for x to elicit *P*-responses in normal subjects under normal conditions. Each of these normalcy constraints guarantees a certain type of invariance in the possession of the property. The former excludes odd tastes, strange desires, and abnormal psychologies. The latter excludes improbable, miraculous, and simply odd conditions. Deployed together, these two constraints preserve what we might call the "anthropocentric objectivity" of the property. This is the idea that the thing in question is the same for all normal human beings in normal circumstances. Response-dependent properties are generally not taken to vary in virtue of the different responses of different agents. This is the type of objectivity that we are familiar with from the paradigmatic Lockean secondary property of color.

Now, these two constraints—normal subjects and normal conditions—play two roles in response-dependent accounts. In the first instance, they ensure that the properties have the type of anthropocentric objectivity required by allowing only responses from those subjects and conditions that track the normal type of subjects and conditions to determine the possession of the property. Once such a norm has been discovered[113], it functions as a second constraint in that the response of any particular subject in any particular condition is taken as appropriate if it tracks this norm. It is this latter constraint from norm to appropriateness of responses that generates an account of what it is for a response to be appropriate, and thus functions a constraint on the responses themselves. In order to get a sense of how these constraints work, we need to have some idea of how the norm that defines what counts as an appropriate response is determined. D'Arms and Jacobson cash out this aspect of how

[113] This could be understood in terms of the norm being established through the discovery of the response-dependent property.

invariance can be interpreted as a constraint on the appropriateness of responses in their account of the 'fit' of emotional responses.

Emotions, according to D'Arms and Jacobson, are complex psychological states which involve characteristic qualia and evaluative presentations. Emotions involve these *evaluative presentations* because: "…they purport to be perceptions of such properties as the funny, the shameful, the fearsome, the pitiable, et al."[114] D'Arms and Jacobson argue that there are properties, like fearsomeness, that should be understood in response-dependent terms. What it is for an object to have the property of fearsomeness, on this account, is for it to elicit the response of fear in normal subjects under normal conditions. Such a property is relational in that it is an artifact of the relation between the perceiver and the object.

Central to what it is to have an emotion, on this view, is to have an evaluative presentation of, for instance, a person as pitiable, an action as shameful, or a state of affairs as fearsome.[115] But what, precisely, are these evaluative presentations presentations of? D'Arms

---

[114] (D'Arms & Jacobson, 2000a, p. 66.) In making this claim Jacobson and D'Arms point towards the work of (De Sousa, 1987; Greenspan, 1988; Roberts, 1988; and Solomon, 1976) on the emotions.

[115]There are two ways in which such evaluative presentations could be construed. On one hand the view could be that emotions are constituted by *judgments* about objects/states qua the bearers of these evaluative properties. On the other hand, emotions could be taken as states that involve these evaluative presentations, which do not necessarily involve evaluative judgments. The former position is what we can call a *judgmentalist* view of emotion. The latter position is a *non-judgmentalist* view of emotion. To be a judgmentalist about emotions is to hold that emotions constitute *judgments* about the evaluative properties of the object. To be a non-judgmentalist is to deny that this is the case. The aim of Jacobson and D'Arms is to "…deny that [emotions] constitute judgments, while attempting to capture what is right about judgmentalism." (D'Arms & Jacobson, 2000a, p. 67.) Judgmentalism seems right insofar as having an emotion seems intimately connected to various kinds of evaluative judgments. But the claim that having an emotion just is making such a judgment is to deny that we can ever feel an emotion without making the associated judgment—a claim which D'Arms and Jacobson oppose on the grounds that the nature of judgments is that they presuppose some form of endorsement. When I judge that the dragon is fearsome, I am in effect saying that it is appropriate for me to be afraid of the dragon. In the case of dragons, this seems harmless enough. But what about cases where the agent simultaneously feels some emotion and repudiates it. Some (although not all) cases of fear of flying fit this mold. We can present the person who is deathly afraid of flying with all the statistics, moreover she may well agree that it is irrational of her to be more afraid of flying than driving, but this will not alter her fear. The agent who fears, and yet acknowledges at the same time that her fear is irrational, is clearly not endorsing this fear. Yet the judgmentalist must ascribe to this understandable individual the state that she both does, and does not, judge that flying is fearsome. It is in cases such as this that we see the limits of Judgmentalism about emotion. What the judgmentalist gets right, in the view of D'Arms and

and Jacobson take them to be presentations of certain properties, such as the property of fearsomeness, pitiful-ness, and so on, which are best understood as *response-dependent* properties. In the case of emotions, these are response-dependent *evaluative* properties. In essence, the appeal is to a response-dependent account of value. If fearsomeness is a response-dependent property, to say that something is fearsome is to say that it is so partly in virtue of the response (fear) that we have to that thing.

Now, at the heart of D'Arms and Jacobson's response-dependent approach to the emotions, is the claim that what it is to fear something is to perceive the property of fearsomeness in that thing. Thus D'Arms and Jacobson, in their understanding of emotions, are talking about emotional responses as determining response-dependent properties in the objects.[116] Although they restrict their account to a small number of basic emotions like fear, a

---

Jacobson, is that the agent who both fears flying, and judges that flying is not fearsome, is in an unstable state. "Such conditions put psychological and rational—that is, causal and normative—pressure on us to alter our feelings or our judgments in order to bring them into harmony."(D'Arms & Jacobson, 2000a, p. 67.) What the judgmentalist gets wrong, then, is that she holds that there cannot be an emotion without a judgment, by substitution, the claim is that an agent cannot have an emotion without endorsing it in this way. In the face of the clear multitude of plausibly irrational fears, envies, and jealousies, this latter claim just seems wrong. Judgmentalism about emotions is not a tenable position to hold, as it cannot account for this large class of emotions, however respecting the elements of emotion that Judgmentalism gets right—that feeling an emotion and not endorsing it is an unstable state for creatures like us—is also important. Thus D'Arms and Jacobson hold a non-judgmentalist view of emotions—the view that emotions *involve* evaluative presentations of objects/states, but they do not take the further step of holding that emotions constitute *judgments* about the evaluative properties of the object/state.

[116] I am talking about response-dependent *properties* here, which may strike you as odd. Historically, response-dependence has been taken to be a way of understanding *concepts*, rather than properties. However, giving a response-dependent account of properties has certain advantages. The strength of the property view, and the weakness of the concept view, is that concepts need to be grasped a priori, while we can discover what properties are. Clearly, a prioricity is plausible in cases such as nausea, but not so in other cases of widely promoted response-dependent concepts such as colour. Mastery of the concept of nauseating does seem to entail knowing that it will produce nausea in normal subjects under normal conditions—without knowing this, it seems, you simply do not understand what it is for something to be nauseating. Mastery of the concept of green, on the other hand, seems to be possible without it being the case that the individual has any complex story to say about green being the kind of colour response produced by normal subjects gazing upon some thing in normal conditions. That little Johnny can accurately distinguish and report on a wide selection of green things seems to be a pretty good grasp of GREEN. As LeBar puts it: "I propose instead to think about evaluative properties, such as *good* and *bad* (or *evil*), and to take the position that we can best understand such properties as response-dependent. The advantage of this move is that the nature of properties, unlike concepts, is not tied tightly to the

danger of such an approach is what we can call the problem of metaphysical profligacy. It is very easy to have an emotional response to some thing, and if each emotional response succeeds in attributing the emotion-eliciting property to the thing, then such properties are going to abound in a fairly suspicious manner. What such a view needs in order to be plausible is some constraints on when the response *succeeds* in discovering the property.

D'Arms and Jacobson argue that such a constraint can be supplied by a theory of when an emotional response is 'fitting'. On this view, a response only succeeds in attributing the property to the thing if the response is *appropriate*. D'Arms and Jacobson[117] maintain that the fitness of an emotion tracks how accurately the emotion "…presents its object as having certain evaluative features."[118] The notion of fittingness proposed here is a relation between the presentation of the object by the emotion and the world that is analogous to the relation between a belief and the world. The analogue is that they are both veridical notions—as the accuracy of the belief is related to how well it reflects the world, so the fit of the emotion is related to how well it reflects the world.

One distinction that it is important to keep in mind when considering D'Arms and Jacobson's view is that between the 'correctness' of an emotion (whether or not that emotion accurately presents its object), and whether or not the agent is correct in feeling that emotion (in other words, is the agent justified in feeling that emotion in the circumstances). In talking of the fit of an emotion, D'Arms and Jacobson are exclusively addressing this first issue, the issue of whether or not the emotion accurately presents its object. This is the analogue of the

---

way we represent them. We are at liberty to discover, through investigation and theory, that properties are unlike how they naively appear to us, and this is just the sort of case I want to make for the response-dependence of value." (LeBar, 2005, p. 181.)

[117] D'Arms and Jacobson have coauthored a series of papers that work on this concept including (D'Arms & Jacobson, 2000a; D'Arms & Jacobson, 2000b)

[118] D'Arms & Jacobson, 2000a, p. 65.

success condition of accuracy in the belief case. It is the appropriateness of these evaluative presentations that is the concern of this account of fitness for the emotions.

D'Arms and Jacobson present their account of appropriate emotional responses in the following extract:

> What it is for something to be fearsome it is for fear of it to be appropriate; a joke is funny if and only if it is appropriate to be amused by it; and, in general, for an evaluative property φ to be instantiated by X is for an associated response F to be appropriately held toward X.[119]

The property is defined by the appropriate response that normal subjects have under normal conditions. On this view, what it is for Jack to experience fear is for him to perceive some object or state of affairs as fearsome. The object is presented to him by the emotion as having the property of *FEARSOMENESS*. The property that figures in this explanation, the property of *FEARSOMENESS*, is a response-dependent property. The normal subject/normal conditions constraint means that not every fear response is a response to the property of *FEARSOMENESS*. A fear response is only a response to that property if it fits, that is, if it is an appropriate response for the agent to have to that thing.

D'Arms and Jacobson propose that there are two dimensions of fit for emotions, intensity[120] and shape. Roughly speaking (and D'Arms and Jacobson note that this analysis, for principled reasons, can only be given roughly) the shape of an emotion is determined by the evaluative features that it presents an object as having. To the extent that the object has those features, the emotion is fitting in terms of its shape. Consider 'fearsomeness' to be a higher level role property which, in some specific case such as snakes, is instantiated in a

---

[119] D'Arms & Jacobson, 2000b, p. 732.
[120] D'Arms and Jacobson use the term 'size' to denote this dimension of fit, however I take the term 'intensity' to be a more appropriate term for the type of emotional response at issue.

group of lower-level realizer properties.[121] Snakes are fearsome in a way which is plausibly biologically hardwired, and this fearsomeness is constituted by their dangerousness. This dangerousness is itself a higher level role property which is instantiated in a particular snake by still other lower-level realizer properties. In the case of water moccasins, a member of the pit viper family, their dangerousness is realized in their venomousness and their aggression. The emotional response is fitting in terms of its shape if it correctly attributes these properties to the thing. If I fear a long shape on the ground in the dark because I take it to be dangerous (I erroneously perceive it as a snake), then my emotional response is not fitting in terms of its shape—it gets the properties of the object wrong. The intensity of the emotion is concerned with whether or not the intensity of the reaction correctly tracks the magnitude of the evaluative features.[122] In the case of the snakes, fearing a mildly venomous and non-aggressive snake with the same intensity as a water moccasin is getting the intensity of the emotion wrong. I am, in other words, taking the snake to be more fearsome than it actually is. Thus there are two dimensions of fit for emotions, intensity and shape. An emotion which is not appropriate—it does not 'fit' the stimulus—is mistaken in one of these quantifiable ways.

What it is for an emotion to be fitting, then, is that it is the *appropriate* response to the relevant property. That is, the response gets the property right. What it is for the response to get the property right is that the response tracks the norm established by the responses of normal subject under normal conditions. Note that this property, as a response-dependent property, is a higher-level role property which exists in the relation between the agent and the object, which is instantiated in the object by various lower-level realizer properties. With a property such as *FEARSOMENESS*, there are elements of both fact and value. The facts reflect the lower-level realizer properties which instantiate the value element, which is the

---

[121] The distinction between higher level role properties and lower level realizer properties is due to Frank Jackson and Phillip Pettit in (Jackson & Pettit, 2002, p. 106.
[122] D'Arms & Jacobson, 2000a, pp. 73-4.

higher-level role property. In the case of response-dependent accounts of *properties* it is easy to understand this sort accuracy talk because although response-dependent properties are defined and discovered through our responses, they are properties of the object itself. A fear response is fitting, then, only if it accurately attributes the property of *FEARSOMENESS* to the object/state of affairs. Thus, D'Arms and Jacobson's theory of the fittingness of emotions is veridical, in that it purports to track, in the sense of appropriateness, the 'right' response. However, it is not necessary to appeal to such a property in establishing a norm for such responses. All we require is that a response is invariant in some way—that is, reliably elicited under identifiable conditions. So, in order to understand how such an account of fitness could apply to emotional responses such as desire responses, I need a detailed account of how invariance operates within the response-dependent account.

There are two normality constraints cited in the analysis of response-dependent properties, and thus two dimensions of invariance—subjects and conditions. The 'normal conditions' constraint ensures that the agent's perceptual apparatus is working properly, and thus a certain type of consistency among the agent's own experiences, while the 'normal subject' constraint ensures consistency amongst different agents. Moreover, it is these two aspects of invariance—invariance among subjects; and invariance across conditions—which allows the 'normal' response to be established, and thus functions as the norm which responses must track is order to count as appropriate. But the felt emotional component of desires does not meet this second condition, by virtue of being a 'desire' response. If a theory claims that desires are (or should be) invariant with respect to subjects, then this looks to be an implausible theory of any element of desiring, as desires are essentially subjective.

I will begin by considering how such appeals to invariance amongst responses functions in a clear instance of a response-dependent property, that of *RED*, and then

extrapolate this analysis to the case of desire responses.[123] What do the two dimensions of invariance—normal conditions and normal subjects—do in this case? In the case of the 'normal subjects' constraint the aim is to respect the idea that properties such as color should be invariant amongst subjects. The color of the rose does not change when it is viewed by different agents, although under normal conditions someone who is color blind may perceive it as a different color than someone who is not color blind. An alteration in the perceptual system such as color blindness is taken to be a flaw because it occurs in a relatively minor portion of the population. In a certain sense, what we take to be red is determined by a democracy of perceptual systems. It is 'group-determined', if you will. So, if most perceptual systems take a color to be one way, then the color will enter the language in that way and this will become the norm. This process is initiated by the intuition that the property itself is (or should be) invariant amongst people. Color is not a property that is plausibly individually-determined in the way that certain more subjective properties may be. I propose that in the absence of the view that a particular response *should* be anthropocentrically objective, there is no need to impose a 'normal-subject' constraint on a response-dependent account—a circumstance which is true of the felt emotional component of desires. In this there should be no requirement of consistency amongst agents, as any plausible account of desiring must do more than lip service to the platitude '*de gustibus non disputandum*', by allowing desires to be, in large part, determined by the agent herself. So, the type of response that I am concerned with need only contain a 'normal conditions' constraint.

What, then of the 'normal conditions' constraint? If I look at a red object in blue light, it will generate a 'purple' response in me. If I look at the same object in very low light, it will appear almost black. If I am wearing yellow tinted glasses, it will appear orange to me. We easily recognize that the color of the object is not changing with these altered conditions;

---

[123] Clear, that is, for anyone who accepts the existence of response-dependent properties.

rather, what alters the appearance of the object is that my perceptual system is being affected by the conditions. The 'true' appearance of redness is taken to be the way that it appears in the most normal conditions for seeing, that is, reasonably bright daylight. These are the conditions under which we 'calibrate' our color vision, and so it is these conditions that count as normal for viewing colors. In one way, we can see the normal conditions constraint as providing paradigm working conditions for our perceptual apparatus. My response of 'redness', and thus the perceptual apparatus that generates this response, is only reliable under these normal conditions.

Such an account of the 'normal conditions' constraint on responses relies upon the view that the responses in question *should* be invariant. In order to show that our desire responses should be invariant, I will begin by motivating the idea that full-fledged desires have such characteristics of invariance. Then I will argue through the connection between desires and the emotional response of desiring, that our desire responses are invariant in similar ways. I propose that the folk theory of desire builds two identifiable types of invariance into the definition of desires.

The first type of invariance is invariance across circumstances. A part of what it is to desire some thing is to desire that thing over a wide range of circumstances. If desire chocolate cake, it seems that there are lots of different circumstances that I may be in, and shifting among these circumstances should not change my desire for the chocolate cake. Intuitively, it should make no difference to my desire whether I had skate or cod for dinner, if I am dining beside the river, or on the balcony, if I am wearing red or black and so on. Moreover, if you only desire some thing under very particular circumstances, it seems that those circumstances should enter into the description of the desire. For instance, if I only desire fruit cake on Christmas day, then it seems that my desire is not for 'fruit cake', but rather 'fruit cake on Christmas day', or possibly even 'Christmas cake'. But in most desires we omit such

conditions, and that omission is telling, because it expresses the presumption that desires do not, in general, have these conditions built in. An additional indication that we generally assume desires to be invariant over some wide range of circumstances is that when we attribute a desire for some thing to someone only in very specific conditions, we are usually implying that they really don't desire that thing at all. If I say that President George W. Bush only desires the freedom of the people once in a blue moon, I am not implying that he desires the freedom of the people on the second full moon in a calendar month; rather, I am implying that he does not really desire this at all. Now, in saying this I am not proposing that desires should be invariant with respect to *all* circumstances. Whether or not I have a mouthful of sardines surely affects my desire for chocolate cake at that moment, as does whether or not I am running for my life, or have just eaten a large amount of chocolate cake and so on. Thus, the presumption of desiring some thing is that you desire for that thing is **invariant across circumstances** to some reasonable degree—in other words, our desires are *stable* across circumstances.

The second type of invariance is the stability of desires with respect to the passage of time. Another part of the folk understanding of what it is to desire some thing is that desire should, to some extent, be temporally invariant—that is, stable with respect to the mere passage of time. If you desire to be an opera impresario for 1/30th of a second, and the thought never crosses your mind again, then this really should not register as a desire at all. If you desire for 10 minutes Tuesday next, then this still does not look like a desire.[124] The term for such a state is a 'whim'. What will count as a sufficient duration for something to count as a real desire will be a function of the magnitude of the aim expressed in that desire. It seems

[124] To be explicit: If you have the desire to be an opera impresario for only 10 minutes Tuesday next, then it doesn't seem as if you actually have the desire to be an opera impresario. However, if you stably desire to be an opera impresario for only 10 minutes Tuesday next, this looks like a desire, if a slightly odd one. What is at issue is not the duration of the object/experience/thing desired, but rather the duration of the desire itself.

that 10 seconds is too short to count as a desire to be an opera impresario, but may well be sufficient for a desire to scratch an itch. I am not concerned to give an account of what this duration may be, only to suggest that such a notion is built in to our understanding of desires. I think that it will turn out that most of our desires are quite temporally invariant. What it is to 'really' desire some thing is to be disposed to desire that thing over some significant period of time. That is, we assume that a part of what it is to desire some thing is that your desire is stable with respect to time—our desires are **invariant over time**. Thus there are two types of invariance for desires, invariant across circumstances, and invariance over time. And these senses of invariance are best understood in terms of the notions of stability for desires that I proposed in the first chapter.

There are two reasons that such invariance should extend to the felt emotional response of desiring. I propose that, to a significant extent, our desires and our desire responses should track one another. If this is true that our desires have these stability properties—invariance across circumstances, and invariance over time—our desire responses should also have these stability properties. Now, I need to introduce a distinction in order to motivate the idea that our desires and our desire responses track one another. The distinction that I am proposing is between what I will call 'considered preferences', and desire responses.

A considered preference is a full-fledged desire that is stable with respect to minimal reflection, over time, and across circumstances. By minimal reflection, I mean a fairly cursory consideration of the desire in light of your beliefs and other desires. I take it that a desire which is not stable with respect to such a minimal act of reflection is a degenerate desire, if it counts as a desire at all. In short, a desire which does not survive such an act of reflection looks more like a whim than a desire. The minimal standards of stability across circumstances and over time are in a similar spirit. If a desire changes as a result of the mere passage of small periods of time—that is, if it is very unstable with respect to time—then it is a degenerate

desire. These requirements of minimal stability are meant to exclude desires which are degenerate in a variety of ways. As was discussed earlier, a certain amount of stability with respect to a time is built into our concept of desire. Similarly, the idea that a desire is stable with respect to a fairly wide range of normal circumstances is built in to the notion of a desire.[125] Thus a considered preference is reasonably stable with respect to time, circumstance, and reflection.

Such considered preferences are distinct from desire responses, in that a desire response is the felt experience of desiring something—the automatic affective response of desiring. Desire responses, in virtue of their nature as felt experiences, are occurrent states. As automatic affective responses triggered by representations, their existence is closely tied to the period that the agent is consciously aware of that representation. Desire responses, as occurrent states, cannot be said to be stable in the same way that considered preferences are, but there is still a presumption of invariance. What produces the effect of invariance in desire responses is going to be the reliability of the mechanism that produces them—in this case the conative system. It seems that if I have an emotional response of desiring some thing, and neither I, nor the properties of the thing in virtue of which I desired it change, then my response to that thing should be roughly the same.

Considered preferences and desire responses are not necessarily connected for the same reason that desires and the felt experience of desiring are not necessarily connected. One way to think of them is as separate effects of the same cause, where that cause is the properties of the thing in virtue of which it is desirable to the agent. So, although considered preferences

[125] Note that this is true even of desires that hold only in specific circumstances, when those circumstances are built in to the content of the desire. So, if I have a desire to drink egg nog at Christmas, then it seems that I should have this desire whether or not it is raining, if I am wearing read or black, if I am in New Zealand or America, and so on.

and desire responses are distinct, they tend to track one another as effects of a common cause are wont to be.

The second reason to think that such invariance is true of our desire responses is a function of such responses being automatic affective responses that are produced by some system within the agent—the conative system. As we saw from the color case, a constraint of 'normal condition' is generally meant to ensure the accuracy of such automatic psycho-physical systems. This is the condition in which we 'calibrate' our perceptual system, and is meant to capture the paradigm working conditions for that system. I propose that this is also true of our conative system. There are identifiable conditions under which the system generates the wrong response. What will count as abnormal conditions for the conative system are those conditions in which it reliably makes mistakes. What counts as a mistake for the conative system will be violations of the two types of invariance I have just presented, and thus will be responses that do not 'fit' in D'Arms and Jacobson's terms. I propose that just as misperceptions of color are subject to a type of epistemic criticism (if you perceive green as red, there is something wrong with either the conditions of your perception, or your perceptual apparatus, and the burden is on you to compensate for this), inappropriate desire responses are also subject to a type of criticism. In order to motivate this claim, I need to say more about the notion of fit with respect to desire responses.

Before I can give an analogous account of fit for desire responses, I need to stipulate two caveats on the account. The first is that I can only give an account of the fit of desire responses in the case of changes in the intensity of the desire response. In the case of desire responses, the initial acquisition of a desire response pegs it into your motivational/evaluative system at a certain strength. Once a desire response has been acquired, and accorded its 'level' as it were, then it seems that any changes in the strength of that desire response should be explicable. However I am making no claims about how the initial strength of the desire

response is, or should be, established. Note that this is explicitly a claim about *changes* in desire responses, not a claim about the initial acquisition of a desire response. The second caveat is that I take the fit of desire responses in the interesting cases to be mainly a matter of the intensity of these responses, rather than their shape. I will not have much to say the question of the 'shape' of these desire responses—how the properties of the desirable thing actually are connected to the existence or initial intensity of the affective response—but what I will be focusing on to illuminate the temptation cases is the way in which the intensity of such desire responses *changes*.

How such changes in intensity of an emotional response may be more or less mistaken lies in the view that desire responses may be inappropriate under abnormal conditions. First let me describe the case of mistakes in desiring that I take to be fairly straightforward, and not terribly surprising. According to D'Arms and Jacobson, our emotions can fail to fit on two dimensions, shape and intensity. When an agent makes a mistake about the shape of the thing—takes it to have properties that it does not—this can clearly lead to desire responses that are flawed in some sense. If I desire a piece of that cake because I falsely believe that it was made with sugar, rather than salt, then I am making a mistake about the shape of the thing in that I am taking it to have the property of sweetness when it does not. A mistake about the shape of the thing is going to be a mistake in my beliefs, rather than a mistake in my desires, and thus not of particular interest here. This is at bottom a cognitive mistake, a mistake about the way the world is, rather than strictly a conative mistake in desiring. It is the other case that I am primarily interested in, the case of disproportionate desire responses. An inappropriately intense desire response can, I believe, truly be thought of as a *conative* mistake—a mistake in the system that generates the desire response to the thing.

How do such conative mistakes work? Let us say that I desire a caramel in virtue of its sweetness. Sweetness is a higher level role property which is realized in many different ways –

sugar, agave nectar, golden syrup, etc. And it seems that sweetness is a response-dependence property. What it is for something to be sweet is a matter of the particular response that we have to the chemical make-up. Moreover, sweetness is inherently motivating. All mammals like sweet, and absent confounding factors, mammals, including infants, rats and monkeys will seek out sweetness. Consider a case of temptation involving a sweet thing such as caramel. A standard case of temptation looks to be a mistake in intensity. If I am strongly tempted to eat the caramel when it is right in front of me, but only mildly attracted to the prospect of eating the caramel at some distance, it is not as if I believe the caramel has different *properties* when it is right in front of me. What changes between the two cases is the *intensity* of my desiring response. In both cases I am correctly attributing the property of sweetness to the caramel, but I am intuitively putting too much emphasis on sweetness in my emotional response when it is right in front of me. It is this shift in *intensity*, rather than *shape*, which changes my desire from one perspective to the other. So, in a case of temptation such as this, my *cognitive* states remain the same, but there is this shift in my *conative* states—it is the *felt emotional component of my desire*, my desire response, which changes.

So, what makes a subjective response such as a desire response a more or less appropriate response? I contend that the appropriateness of a response has something to do with what the 'normal' response for that agent is. Now, there is more than one way that we can understand such a norm, and many of the ways of understanding these norms will give coincident determinations of what is, and what is not, an appropriate response. I do not want to argue for a particular interpretation of such a norm, just for the idea that such norms exist. One way of determining such a norm, which looks like the kind of account that would be given for the appropriateness of finding a joke funny, is in terms of how frequent a response is in a population. If the majority of people find a joke funny under normal circumstances, then amusement is an appropriate response to that joke. What determines the normal response (the

norm) in this case is the frequency of responses amongst the majority of people, and any particular response is appropriate if it tracks the norm. Another sense of norm which is in play with respect to such emotional responses is the case in which the response has some biological function such as fear. It seems that a fear response is appropriate when it fulfills the biological function of avoiding danger. The norm in this case is a biological one, and the appropriateness of a particular response is determined by how well it fulfills the biological function of the emotion. Another sense of normal that can determine a sense in which an emotional response is appropriate is the idea of the 'normal observer'. This is what the response of the normal observer—the agent who is free of prejudices, biases, imperfect perceptions etc.—would be under normal circumstances. In this idea of the norm, what determines the appropriate response is the response which is free of any of a number of confounding or distorting influences.[126]

The analogy to the case of the desire responses is that there is a normal response which is determined by the responses that the agent has to that thing over time. The norm is some combination of the types of norms outlined above. I propose that what determines a 'normal' desire response for any given agent is the response that the agent has when she is free from confounding influences. Then what it is for the agent to have an appropriate response is that it tracks this norm. Why is this plausible? Because these emotional responses are connected to the properties of objects, and this connection is not accidental. The properties in virtue of which the thing is desirable cause me to desire that thing in some way. Moreover, these responses should track the properties in virtue of which you desire the thing in the sense that the intensity of your desire response should only change if there is either a change in the

[126] Note that these norms do not prevent the agent from reasonably and legitimately changing the way that they evaluate outcomes, and thus the intensity of their desire responses to things. In such a case, it is not true that both the agent, and the properties of the thing in virtue of which it is desirable to the agent, have remained the same. Rather, there is a change in the agent, which justifies the change in the agents desiring response.

properties of the thing, or in the desiring agent. Moreover, this norm captures the presumption of invariance in our considered preferences and desire responses. Thus what it is to have a mistaken desire response is for it to be inappropriately intense in that it does not conform to the agent's normal responses to that thing.

How plausible are such mistakes in the conative system? I take it that a conative mistake in the intensity of a desire response is on a par with other well known perceptual mistakes. Consider the case of perspective in visual representation. When I am standing looking down an avenue of oak trees, the trees appear smaller to me as they dwindle into the distance. If I attribute the property of size to the trees solely on the basis of their relative size in my perception, then I will believe the trees that are further away from me are smaller than the ones that are close to me. Of course, I am not in fact going to believe this. In the case of visual perception, we are clear about what the apparatus is that produces the apparent effect of similar sized objects appearing smaller as they get further away, and we are also pretty adept at compensating for this mistake. It is only the very young (or significantly impaired) who do not automatically account for these variable appearances in their attributions of relative size. I think that there are similar flaws in our conative system that can produce systematic mistakes in our perceptions of the 'intensity' of the attractive properties that are reflected in inappropriately intense desire responses. However the conative system differs from the visual system in that we are not particularly adept at compensating for these flaws. Thus our conative systems have abnormal conditions for operation, which can be identified by the pattern of mistaken desire responses that they produce. This claim is central to my response to the challenge of temptation as I will argue that temptations are instances of these mistaken desire responses. What happens in a case of temptation is that the agent's desire response is out of proportion to her considered preference, and what causes the challenge of temptation is that the same thing that causes the mistaken desire response interferes with the agent's ability to reason about her considered preferences. Thus it is easy for her to mistake the inflated

desire response for her considered preference, and treat it as such in her deliberations. So, in order to motivate my answer to the challenge of temptation I need an account of what constitutes the abnormal conditions of operation for the conative system.

Gauthier on proximal and vanishing point preferences

I propose that the normal and abnormal conditions for the operation of our conative systems can be understood in terms of Gauthier's distinction between vanishing point and proximal desires. My hypothesis is that the conative system tends to produce inappropriate emotional responses of desiring, that is, desire responses whose intensity is out of proportion with the desirability of the object for the agent, under conditions of proximity. Thus it is proximity which constitutes abnormal conditions of operation for the conative system.

As temporally extended beings, we both form, and consider, our desires from more than one temporal perspective. We form desires now that can only be satisfied much later. We form desires that will be satisfied now, or not at all. We desire both in the present moment and the far future and at all points in between. Thus our desires can conflict not only at a time, but also across time. David Gauthier takes the intertemporal nature of this conflict seriously when he distinguishes between 'proximal' and 'vanishing point' desires.[127] Gauthier notes that for certain types of choices, agents have a pervasive tendency to desire one thing when a choice is imminent, and another thing when the choice is not imminent. The former is the agent's **proximate desire**, while the latter is the agent's (temporal) **vanishing-point desire**. Gauthier argues:

> Now, at any given time, although a person may want to act on his now proximate [desires], he does not want to act at other times on what would be his then proximate [desires], where these are in conflict with the vanishing-point [desires] that he now holds. Recognizing this, he is able to understand that if, given proximate [desires], he chooses the

[127]Gauthier terms these preferences. I am calling them desires for the sake of the consistency of the paper. In the context at hand there is very little difference between these two things. (Gauthier, 1997)

> action that best realizes his immediate concerns, he is deliberating in a way that may not lead him to the best realization of his overall concerns, as viewed at that or any other time.[128]

Vanishing point desires, according to Gauthier, are the desires that an agent has when a choice is not imminent. This is contrasted with the agent's proximal desires, which are those held by the agent when a choice is imminent. In many cases, an agent's vanishing point and proximal desires will be the same, but in various instances such as cases of temptation, hyperbolic discounting, or weakness of the will, the agent's proximal and vanishing point desires come apart. This distinction between vanishing point and proximal desires is coincident, to a great extent, with the distinction that I have proposed between considered preferences and desire responses.

The existence of these two perspectives on one's desire can be illustrated in cases where they come apart. Now, certain people who loath going to the dentist but desire good teeth will exemplify just such a split between their proximal and vanishing point desires. In the period immediately prior to a dental appointment, such an agent will have a proximal desire not to go to the dentist. At any other time, her vanishing point desire will be to do what is necessary to maintain her teeth, including attending dental appointments. That these two types of desire are capable of separating in such cases indicates that they are two distinct perspectives. Gauthier talks of such cases as instances of temptation.

Using such examples, Gauthier motivates the intuition that there is a difference between vanishing point and proximal desires by highlighting the strangeness of proximal desires. Then, using an appeal to the agent's overall good, he argues for a rule of choice which favors vanishing point over proximate desires whenever the benefit of having a proximate desire satisfied falls beneath a certain threshold. This is, in many ways, an attempt to answer the challenge of temptation. However, by arguing in terms of a brute appeal to the overall

---

[128] Gauthier, 1997, p. 20.

good of the agent, Gauthier leaves it a mystery why we have these proximal shifts in our desires, and does not address the question of precisely what it is that is wrong with them. This is particularly worrying for his view as not all proximal desires look bad in the way that he describes, as it is only proximal desires that are in conflict with vanishing point desires that are problematic. However on Gauthier's view all proximal desires are tarred with the same brush. Thus he does not give an account of what is peculiarly wrong with temptations, although he identifies properties of desires that are relevant to this question.

Proximity, in the case of desire responses, has the following effect: When you have a change in the strength of your desire response when a good is proximal, it is not that you take the thing to have different properties at that point; the difference is in the response generated by the conative system. In the proximal case you are drawn more to the sensory features of the thing that support your desire, without in fact taking those properties to be any different to how you understand them in the vanishing point case. It is that these features engage your desiring response (and thus your motivations) to a greater extent in the proximal case—so your desire response is disproportionate in that it is anomalous. If you were to contemplate precisely those same features from many other, less proximal, perspectives then your desire response would be less intense. Thus the case where we have the difference between proximal responses and considered preferences in play is one in which your beliefs about the properties of the thing remains the same, although the response of desiring is more intense in the proximal case than it is in the vanishing point case.

In the case of desire responses, I propose that the abnormal conditions of operation for the conative system—that system (whatever it is) that produce the automatic affective responses that are the emotional components of desires—are present in the proximal case. As creatures we are pretty reliable in what we do and do not have a desire response to. If some object regularly evokes a certain desiring response under certain conditions, you would expect

it to continue to do so. This is a reflection of the characteristics of invariance in desire responses that I introduced earlier. What differentiates an appropriate desire response from a mistaken one is that appropriate desire responses track, in some sense, the norms for such responses introduced earlier. The norm of desire responses is the desire response that the agent has to some thing when she is free of distorting influences. This is the idea of the norm being determined by her response *qua* 'normal observer'. In the proximal case, the agent departs from these norms.

Of course, as there is no 'normal subject' constraint in this account, desirable for you may well not be the same as desirable for me which seems to be exactly the right result. Thus what it means for conditions to count as normal for the operation of the agent's conative system will be linked to the vanishing point perspective in the sense just outlined. Through this connection to the vanishing point perspective, there is also a connection between the normal responses of the agent and the stability of certain desires and desire responses over time, across circumstances, and with respect to reflection.

My hypothesis about the abnormal conditions for operation for the conative system is that proximity undermines the fit of our desire responses in the same way that distance undermines our visual perceptions of relative size. On this picture the vanishing point for an agent represents a perspective on her desires which is free from distorting influences. It is freedom from distorting influences which determines what the 'normal' response for the agent is, and thus what the appropriate desiring response is. So, the importance of the vanishing point is that it gives us evidence of what the appropriate desire response would be. Thus what it is to have an appropriate desire response will be for it to track the norm determined by responses in the absence of distorting influences, and the vanishing point desire give us evidence about what that normal response is. Differences among proximal responses and vanishing point desires are evidence that there is some distorting influence at work in

producing the proximal desire response. Defending this hypothesis is the aim of the fifth chapter.

I began with two claims about the felt emotional component of desires. I said that it could distinguish representations that are capable of producing a reward signal from representations that are not, and that it could answer the challenge of temptation. Even with only this basic sketch of the view in hand we can see how it fulfils the first claim. According to the view that there are such felt emotional components of desires, what distinguishes a mere representation from a desire is that the desire includes a desiring response to the object of the representation produced by the agent's conative system. Jack has a representation of a piece of chocolate cake which drives a reward signal, thus engaging his motivations, and a representation of a piece of yellow cake, which does not drive a reward signal, and thus does not engage his motivations. I propose that Jack's representation of the chocolate cake causes a certain response— a craving (desiring response) which is a felt emotional component of desire —which his representation of the yellow cake lacks. Thus a representation that will produce a reward signal is one which possesses a felt emotional component of desiring in virtue of being a particular kind of conative response, a desire response.

The second claim was that I could answer the challenge of temptation by paying attention to such felt emotional components of desires. So far I have laid the groundwork for this solution, which will be presented in full in the next chapter. There I will defend this analysis of abnormal conditions of operation for the conative system by identifying systematic mistakes in intensity in emotional responses of desiring that occur in the condition of proximity, and the mechanism through which they are produced. Ultimately I will argue that it is these mistakes that generate the phenomenon of temptation, and make it the case that temptations should be treated differently from other desires in rational deliberation.

CHAPTER 5

THE CHALLENGE OF TEMPTATION

The aim of this chapter is to answer the challenge of temptation. I will identify a systemic mistake in the conative system that is (1) the product of the conative system working under the abnormal conditions of proximity; and (2) that systematically produces the flawed desire responses that constitute a significant type of temptation. In addition I will argue that paying attention to the relative stability of our desires will yield a method for identifying such flawed desire responses in the course of our deliberations, thus allowing us to ameliorate their influence in determining what it is that we should do.

We are agents that exist and act over time. One aspect of being temporally extended beings is that we face intertemporal conflicts amongst our desires. We are often called upon to weigh a desire that is right in front of us against a desire for some future good. Temptations are a class of such intertemporal conflicts. What it is to experience temptation is to have difficulty delaying gratification by resisting the smaller, more immediate satisfaction of the temptation in favor of achieving the later, larger satisfaction of the conflicting desire. Thus temptation is a product of being agents who act not only at a time but also over time, and is thus endemic to the human condition.

Now, the insidiousness of temptation is not that we sometimes have an intense emotional experience of desiring some thing when it is proximal, even if that thing is contrary

to our longer term aims. Rather, it is that it is not obvious to us at the moment of desiring that satisfy this urge is contrary to our considered preferences, and thus irrational. Temptation is not merely an emotional overlay of desires that can easily be dismissed when it leads us astray; it misleads us about what our considered preferences are by undermining our ability to effectively reflect on our considered preferences. I am only considering a specific, if fairly common, class of temptation here. This class of temptations is those temptations which are generated by some sort of temporary shift in the apparent motivational strength of the agent's desires, which is exemplified by the case of Kim and the chocolate cake that I introduced in the first chapter. However, I take it that all types of temptations have in common this effect of undermining the agent's ability to reflect on her considered preferences, which is what makes temptations such pervasive deliberational difficulties.

In chapter four I proposed that we can begin to understand temptations in terms of a conflict between vanishing point and proximal desires. However, the mere distinction between proximal and vanishing point desires cannot tell us what is wrong with temptations, as not all proximate desires are temptations. A proximate desire which does not conflict with any vanishing point desires looks to be just as rational a candidate for acting upon as many vanishing point desires. There are two characteristics which make a proximal desire a temptation: (1) it conflicts with some more important, longer term, goal, (2) there is some effect at work that interferes with the agent's ability to reflect on her considered preferences. It is through these mechanisms that temptations pose deliberational difficulties. In order to flesh out how temptations can be understood in terms of this distinction, I need an explanation of how it is that such proximal desires interfere with an agent's ability to reflect on what her considered preferences actually are, as well as an account of why it is that the proximate desires which count as temptations should be discounted in deliberation.

I propose that temptations should be discounted in deliberation because they are, in some robust sense, mistaken. As was outlined in the previous chapter, the kind of mistake that I will be looking for is cases of mistaken strength in the desire responses that constitute the felt emotional component of desires. I will argue that in the case of temptation you are being misled. Specifically, you are being misled about the motivational strength of your considered preference by the motivational strength of your desire response. What goes on in the temptation case is that the strength of the felt emotional component of desire is out of proportion with the strength of your considered preference. By overestimating the motivational strength of your considered preference because of the increased strength of your desire response, you end up treating a desire that is all-things-considered weaker, as stronger than some other, competing, desire which is not affected by these proximity effects.

My case relies on several elements. The first issue that I need to address is the question of how the strength of considered preferences and desire responses is determined. Then I will explore how it is that the strength of these states can go wrong, by examining the phenomena of hyperbolic discounting. The next step is to use these mechanisms to analyze the case of temptation, and show why it is that temptations occur, how it is that they are mistaken, and thus why it is that they should not be treated on a par with other desires in deliberation. Then I will suggest an explanation of why it is that our ability to reflect upon our considered preferences is impaired in cases of temptation. Finally, I will argue that we can use the relative stability of desires to determine which desires (or desire responses) should be discounted in deliberation, thus answering the challenge of temptation.

Considered preference, desire responses, and motivational strength.

In the previous chapter I began motivating the idea that the considered preferences of an agent are not the same as her desire responses. In this section I want to explore these states in greater detail, paying particular attention to what determines the strength of these states,

and what it is for these state to track one another. I propose that there should be a loose reflective equilibrium between our considered preferences, and our desire responses such that a mismatch between these two is evidence that something is wrong somewhere.

Recall that a considered preference is a full-fledged desire that is minimally stable with respect to reflection, over time, and across circumstances. By considered preference I do not mean a desire that has been considered in light of all possible information, as I am not positing that considered preferences are an 'ideal observer' phenomena. In this way, the position that I am proposing is different from the view put forward by Michael Smith in defending a rationalist form of instrumentalism. [129] Smith argues that the desire which it is rational for an agent to pursue in means-end reasoning is the desire that the 'ideal' version of her would have. That is, it is what it is that she would desire if she had *all* the relevant information about herself and the world. In my view, a **considered preference** is what the agent would want if she took the information *that she possesses now* into account. Smith's ideal version will, on many occasions, rely on information that the agent does not, at this point in time, possess. It may even rely on information that the agent cannot in principle possess at any time. My view of a considered preference is that this is the desire that the agent would have were she to take the information that she has at her disposal now, including information about recent changes in her desires, and the causes of those changes, and come to some type of equilibrium state through reflection. Such considered preferences are distinct from our desire responses.

Considered preferences and desire responses are not necessarily connected for the same reason that desires and the felt experience of desiring are not necessarily connected. One way to think of them is as separate effects of a common cause, where that cause is the properties in virtue of which the thing is desirable to the agent. Under normal circumstances,

---

[129] Smith, 1995.

we have the default expectation that desire responses and considered preferences will track one another in certain ways. The main respect in which we expect there to be a connection between our desire response and considered preferences is in terms of their motivational strength. [130]

Both desire responses and considered preferences have the capacity to move the agent to action—that is, they can both be motivationally efficacious. A characteristic of any motivationally efficacious state is that its motivational strength can be measured. It is easy to understand what the motivational strength of a desire response is, as it is captured in the felt intensity of the desire response. So the motivational strength of a desire response is a function of the intensity of the emotional response of desiring.

The motivational strength of a considered preference is a little more complex. I will borrow from the decision theoretic presentation of desire and represent the strength of desires in terms of what the agent is willing to pay for a state of affairs in which the desire is satisfied. There are various problems with this way of measuring the strengths of desires, but it is close enough for these purposes. Given that a requirement for considered preferences is that they sustain a certain degree of stability in these respects, I take it that this is also a requirement of the motivational strength of such states. This measure of motivational strength, like the considered preferences that it is a characteristic of, can be more or less stable with respect to reflection, across circumstances, and over time. Now, we can discover the motivational strength of a considered preference at any particular time or in any particular circumstance through the price that the agent is willing to pay for the satisfaction of that preference at

---

[130] There is another dimension in which we can think of considered preferences and desire responses tracking one another. This is the question of the states that they are stable with respect to. Both considered preferences and desire responses can be more or less stable with respect to reflection, time, or circumstance. So another dimension on which they can be more or less the same is in terms of the acts of reflection, times, and circumstances that they are stable with respect to. This is not a dimension of matching which figures in this analysis, so I will not dwell on it here.

different times and in different circumstances. Through this we can identify the approximate motivational strength of the considered preference that is stable with respect to those same circumstances, acts of reflection, and periods of time that the preference itself is. Thus in light of the stability constraints on considered preferences, I propose that the motivational strength of a considered preference is this stable motivational power.

I propose that a case in which the strength of the considered preference does not match the strength of the desire response generates grounds for questioning the legitimacy of the motivational strength of both states. After all, in a rational agent there is an underlying drive for consistency amongst our motivations, because as a matter of metaphysical fact we can only act on one in a set of competing motivations, and it seems to be a matter of rationality that we act on the most important or valuable of any such competing options. I take it that this most important motivation is the one which is 'rationally' strongest, where what I mean by 'rational' strength is that the strength of the state which is relatively stable (over time, across circumstances, and with respect to reflection). It is this idea that drives my view that we aim at a loose reflective equilibrium between our considered preferences and desire responses to some state of affairs, such that a mismatch between the strength of these states is evidence that something is wrong somewhere, and that something needs to change.

How, then, does such a reflective equilibrium work? There is a range of conditions under which I desire the cake, and I desire it in virtue of its desirable features. This is true of both my considered preferences, and my desire responses. Thus these states, when directed toward the same object, are in part produced by a common cause, which is the property in virtue of which the thing is desirable to you. However, we tend to encounter difficulties when the strength of these two states does not match.[131] For instance: If I have a strong considered preference to play the Cello and I reflect on that considered preference by generally

[131] By 'match' have the fairly loose notion of approximately the same motivational strength in mind.

representing to myself in detail what it would be like to have it satisfied, yet I lack a matching (roughly equally strong) desire response, then I have reason to question that considered preference. Conversely, if I have an intense desire response to the idea of playing the Cello and I represent to myself what it would be like to have that desire satisfied, and this fails to generate in me a matching considered preference to do that thing, then I have reason to question that desire response. In the case where all is well with our desires, it seems that the strength of our considered preferences matches the strength of the related desire responses. However, when this is not the case, one of these states must either change, or be discounted in deliberation, as the agent's actual motivation to obtain a particular state of affairs must have only one measure of motivational strength.

So, in those cases where the strength of the considered preference and the desire response state do not match, the agent can either reject the motivational strength of one or the other, or change one in order to match the other. However, there is no one right way for an agent to reconcile such inconsistencies. Either of these cases can be made consistent by rejecting either the desire response or the considered preference as flawed, and thus in need of being brought into line with the other form of motivational strength. I take it that the question of which state is rejected in any particular case is going to be determined on the basis of a variety of relevant reasons. For instance, if it is a case where your considered preference is to do your duty, and the strength of your desire does not match, then it seems likely that your desire response will be rejected, as gaining emotional satisfaction is not a part of the aim of doing your duty. Conversely, there are other cases where it seems that the aim of desiring just is to gain such emotional satisfaction, as in the case of choosing a career in order to live a satisfying life. A mismatch here seems to provide a reason for rejecting the considered preference in light of a lack of a matching desire response. There are also other reasons that might come into play in making such a choice. The type of case that I am interested in is one in which one of the states is plausibly mistaken in the level of motivational strength that it is

displaying, and thus should be rejected in reaching equilibrium amongst motivations on these grounds. The specific case is one in which our conative system reliably produces desire responses which demonstrate an inappropriate strength. It is this, I will argue, which is going on in the case of temptation.

Because the motivational strength of a desire response is determined by the intensity of that emotion, the motivational strength of that desire response is inappropriate if and only if the intensity of the desire response is inappropriate. The intensity of a desire response is inappropriate when it does not track the 'normal' response that the agent has to that thing when she is free from distorting influences, as was discussed in the previous chapter. I propose that what explains a violation of this norm, such as is evidenced in cases of temptation, is the conative system operating under abnormal conditions, conditions that interfere with her conative system and thus distort her desire responses.

In the next section I will identify the conditions in which our conative system reliably produces desire responses whose strength tracks neither the normal strength for that desire response, nor the strength of the associated considered preference. In the last chapter I suggested that it is proximity which counts as abnormal conditions for the operation of the conative system. In the next section I will offer an analysis of why this is plausibly the case, by presenting a systematic way in which the strength of our desire responses increases as a result of the mere proximity of the good.

Discounting the Future

I propose that the mechanism that generates mistakes in our conative system under the condition of proximity is hyperbolic discounting, which is a particular way of discounting the future. Now, there is nothing troubling about discounting the future *per se*. It is natural that in some sense our evaluation of goods in the future is mitigated by how distant we are

from those goods. Experiencing a delicious cup of coffee is worth more to me at this moment than the same experience occurring two years, two weeks, two days, or two hours in the future. So, the intuition goes, the agent's perception of how desirable or valuable some thing or event is to her, should, in some way, take into account the agent's temporal distance from that thing. There are three closely connected questions that occur when we talk about agent's discounting the future: (1) why do agents discount the future; (2) how agents should discount the future; and (3) how agents do in fact discount the future. It is through answering these questions that I can demonstrate that hyperbolic discounting is a mechanism through which our conative systems reliably yields mistaken felt emotional components of desire—desire responses whose intensity is not appropriate—under the condition of proximity.

The first answers to the 'why' question were economic explanations of discounting the future produced in the thirteenth century debate over the moral permissibility of charging interest on loans. In hindsight, we can see that core of the issue was that people were willing to take money now on the condition that they repay a greater amount in the future. There was something about these people's subjective attitudes toward money that made money possessed now worth more to them than the same amount of money possessed in the future. Conrad Summenhart, a late scholastic theologian at the University of Tübingen, in his 'Treatise on Contracts'[132] argued from the recognition of such time preferences to the conclusion that the fair price of a newly created debt was below the amount lent.[133] This insight led to the argument that if the *present value* of a sum of money is greater now than it is in the future, then interest is simply making up this difference. The revolution behind this view was moving from valuing money in terms of its face value (objectively), to valuing money in terms of how much it is worth to the person (subjectively).

---

[132] Summenhart, 1499.
[133] Rothbard, 1987.

When economists talk about discounting the future, part of what they have in mind is 'present value' calculations. Consider an agent receiving $100. If the agent receives the sum today, according to her present value calculation, it will be worth $100. But if she receives $100 one year from now, her present value for it might be $95. Why might the present value of a good consumed now differ from its value if it is consumed in the future? There are three reasons.

The first is that the good might be what is called a 'dated commodity'. A 'dated commodity' is a commodity whose time of delivery causes fluctuations in its value to the agent. For instance, a bottle of wine that improves with age will gain value as time passes. In this case, the properties in virtue of which the wine is desired improve as time passes—the wine gets better and thus intuitively more valuable to the agent. Another sense in which a good can be a dated commodity is if there are opportunities that are gained or forgone by possessing the good at a particular time. For instance, if I have $100 today, I can invest it for the year and earn $5, so the $100 is worth $5 more to me now than it will be in one year's time. This $5 difference is often referred to as an 'opportunity cost', and this represents a different type of change in the properties in virtue of which the good is valued. In this case, an earlier time of arrival of the $100 is an identifiable benefit to the agent, which should be included in her subjective valuation of the good.

The second and third ways in which the present value of a good may be influenced are both variations on the theme of the risks or certainties related to receiving a good now vs. receiving the same good after a delay. One case is uncertainty about the delivery of the good itself. If I am certain that I will receive a good if I take delivery of it in the very near future, but I am to some degree uncertain that the good will be delivered if I contract to have it delivered at a future point, then there is an identifiable difference between what the good is worth to me now and what the promise of the good in the future is worth to me now. Let us

say that a piece of cake bought and consumed right now is worth $3 to me. However, if I purchase a contract to have an identical piece of cake delivered to me in three months time, there will be certain risks that something will go wrong in the interim. Perhaps I will fall under a bus; it is conceivable that the cake supplier may go out of business, and so on. All of these risks contribute to a measure of uncertainty about the outcome of receiving the cake in three months, which make the promise of a future good now, worth less than possessing the good itself now.

The other case is that I may be uncertain about my tastes for cake in the future. The value of a good to an agent comprises not only of the good itself, but also what we might call the 'uptake' of the good. The goodness of the cake to me depends not only on the objective properties of the cake, but also my tastes for the cake. It is clear that having a cake when I greatly desire it will be worth more to me than having a cake when I am already sated, or have resolved to eat no more cake, or have developed an allergy to the cake, and so on. As agents it is both plausible and possible that our tastes will undergo such endogenous changes. These are causes for uncertainty, entailing that the promise of a cake in three months will be worth less to me than cake right now.

These three cases all highlight some difference between *what it would be worth to me now* to have the good now, and *what it would be worth to me now* to have the promise of the same good later. The contrast is between my *present* view of *present* goods, and my *present* view of *future* goods. Thus the difference between the subjective valuation of money in the present and the subjective value of a good in the future is generally explained in terms of three factors: (1) changes in the properties in virtue of which the good is valued and opportunity costs; (2) uncertainty about the future delivery of the good, contrasted with the relative certainty about delivery in the very near future, and (3) uncertainty about my future taste for the good, contrasted with the relative certainty about my tastes in the very near future. A

present value calculation is the agent's subjective valuation of a present or future good at the present time which takes into account these three types of change in the value of the good, which captures why it is that discounting the future is a natural, and in some form rational, aspect of our evaluative process.

Exponential and Hyperbolic Discounting

The second question is how it is that we do in fact discount the future. The default rate at which we discount the future is a feature of the way that we are psychologically disposed to value goods. In other words, this is an element of the conative system that produces the felt emotional response of desiring, our desire responses. The way in which any agent discounts the future is captured by the shape of her discount curve. There are two natural possibilities for the shape of this discount curve: hyperbolic and exponential. The shape of the agent's discount curve is dictated by the equation which captures the way in which she discounts the future, and represents the agent's present estimation of how desirable the promise of that good will be at various times. To borrow some terminology from Gideon Yaffe: Let G = the actual value of consuming the good to the agent; Y = the present value of the good with an actual value of G; C = the rate at which she discounts the future; and D = the delay between the present time, and the time at which the good will be consumed. Exponential discounting is expressed in the equation $Y = G \times CD$; while hyperbolic discounting is captured in the equation $Y = G (1 + D)$.[134] An agent discounts the future exponentially when she discounts the future at some *constant* rate. If an agent discounts the future *hyperbolically*, when she is furthest from the good she discounts it to some large degree, and the rate at which she discounts reduces rapidly as the good becomes temporally proximal.

---

[134] Yaffe, 2001, 195-6.

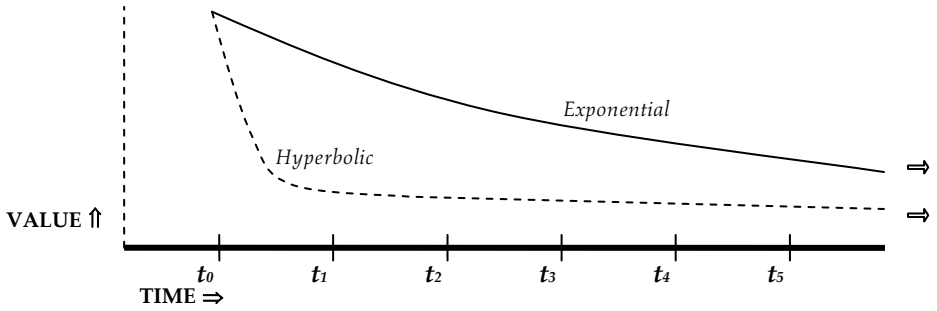

Figure 3: Discount Curves: Hyperbolic and Exponential

In this diagram, both curves represent an agent's views at time t_0 with respect to the delivery of a good at t_1, t_2, t_2,…, t_n. The **exponential** curve demonstrates a constant rate of discount over time. The **hyperbolic** curve demonstrates a rapid decrease in the rate of discounting in the temporal proximity of the good, and a very large rate of discount when it is not in the temporal proximity of the good. Thus the difference between hyperbolic and exponential discounting is not a question of the value of the good in the present or the value of the good in the far future—both curves converge at these points—rather it is a question of how it is that the agent values the good in between.[135]

Consider what I will call a case of 'simple' goods. I will define a 'simple good' as a good which is not a dated commodity. In the garden variety of cases that I am going to consider, all simple goods will be subject to roughly the same uncertainties, and these uncertainties will be essentially constant over time. It might help to think of these uncertainties as the base line of general existential uncertainty that we all live with on a day to day basis—we may fear that we will have the flu or a headache tomorrow, or extra work of

[135] The goods that I am talking about are a type of 'Savage outcome'. In his seminal work of expected utility theory, the 'Foundation of Statistics', L. J. Savage defined an outcome as a 'state of affairs in the world' where that state of affairs includes not only a description of the world, but also a description of the way that the agent is interacting with the world. It is important to include the state of the agent in the outcome, because goods are valued by agents not merely in virtue of their properties, but also in virtue of how those properties affect the agent. No matter how laudable spinach may be on its own merits, the value of the outcome of having spinach also depends on the agent's taste for spinach. (Savage, 1972).

some kind, a crisis in the family (minor or major) and so on, but these are pretty constant fears. So, for any pair of simple goods, it seems that the mere passage of time should not change the agent's desires for the goods.

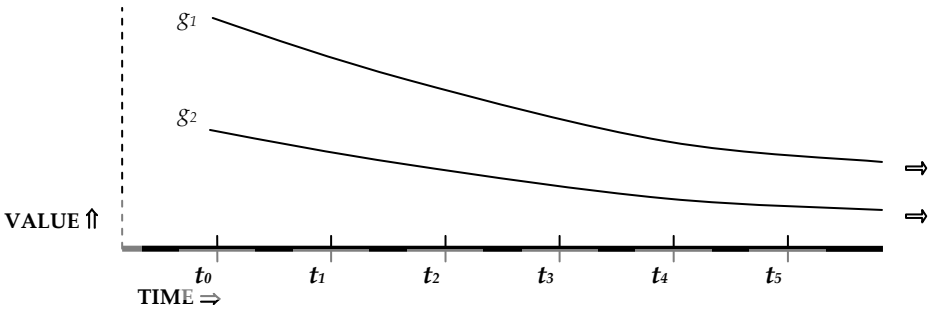

Figure 4: Exponential Discounting Curve

The intuition is that if there is no change in the properties of the good in virtue of which it is valued, nor an endogenous change in the agent that would make the good more or less valuable to her, then there should be no change in the comparative desirability of the good to her. A way of putting this thought that we explored in the previous chapter is that there is an *invariant* character to our desires: your present value at time t for some simple good g consumed at t will be the same for every t which is the present. This invariant character does not imply that the agent should always value g to the same degree in another sense—it allows that when you view two instances of the same good, which are promised at different times, from the *same* temporal location, the good may be valued differently. It is consistent with this view about the invariant character of desires that your present value at t_0 for some simple good g consumed at $t_0 + \varepsilon$ is different to your present value at t_0 for g consumed at t_1. The invariant characteristic of such non-dated goods is that the passage of time does not make them intrinsically better or worse, so they are not intrinsically better or worse for the agent to possess at any particular time. Thus in all cases of simple goods it is plausible to say that for each possible *t qua* the present, this graph should represent the agent's relative desires for two simple goods, g_1 and g_2.

Despite these plausible intuitions in favor of exponential discounting in the case of simple goods, as a matter of psychological fact, we do not naturally discount the future exponentially. In *Picoeconomics* and *The Breakdown of the Will*, George Ainslie presents a persuasive case that the default rate at which we discount the future is **hyperbolic**.[136] In hyperbolic discounting, the rate at which the agent discounts future goods decreases as the temporal proximity of the good increases. This results in the agent having a steeply bowed discount curve.

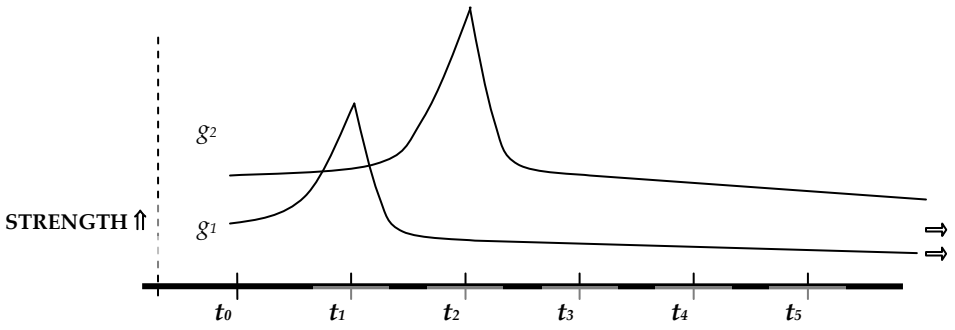

Figure 5: Hyperbolic Discounting Curves

g_1 is a hyperbolically discounted simple good delivered at t_1; g_2 is a hyperbolically discounted simple good delivered at t_2. If an agent discounts the future hyperbolically, the rate of discounting will be large when she is some temporal distance away from the good, but will rapidly decrease when the good is proximal. The problematic consequence of hyperbolic discounting is that it allows the possibility of earlier, smaller, goods being desired more than later, larger, goods when they are proximal. The period in figure 3 where the peak of the line of desirability of the smaller good g_1 protrudes above the line of desirability for the larger good g_2 illustrates how such a reversal in comparative desirability occurs.

Consider the case of choosing a desert at a restaurant. Jim has been to this restaurant before, and knows that the molten chocolate cake is by far his favorite desert, although there is

[136] Ainslie & Haslam, 1992; Ainslie, 2001.

a key lime pie which comes a distant second. Now, Jim has gone to this restaurant with the express intention of ordering the chocolate cake, and he knows that if he had a choice between the key lime pie and the chocolate cake being delivered at the same time, he would, at all points prefer the chocolate cake to the pie. Moreover, he knows that if he does not have the chocolate cake he will regret it. Usually, when he is at this restaurant, the waiter takes orders for the molten chocolate cake before the entrée, as it takes 30 minutes to prepare. On this particular day, the restaurant is trying to get rid of its key lime pie, so the waiter doesn't request orders for the chocolate cake before the entrée. Instead, the waiter comes to the table at the end of the meal carrying a piece of pie, and offers Jim the following choice: Either have the key lime pie now or the molten chocolate cake in 30 minutes. Let us assume that Jim is in no hurry, the ambience of the restaurant is lovely, and he and his companion are enjoying their conversation. There are no bad effects in the mere fact of waiting 30 minutes for desert (indeed, he may even enjoy it more after his dinner has settled, and he will certainly enjoy the conversation). Let t_1 in the above diagram be the point at which Jim is offered this choice, t_2 be the point at which he could have the chocolate cake, g_1 the key lime pie, and g_2 the chocolate cake. Because of his hyperbolic discounting, at the time that he is offered the pie his desire for the pie is stronger. So, it seems that as a result of hyperbolic discounting and the passage of time, Jim will choose the pie over the cake, even though in another perfectly reasonable sense, he 'always' prefers the cake to the pie.[137] This example illustrates how it is that discounting the future hyperbolically can lead to reversals in preferences, and thus intuitively perverse choices as a result of the mere passage of time. Thus if an agent discounts the future hyperbolically, a shift in time alone will affect the comparative desirability of future goods for her. In short, the strength of her desire response will alter when the good is proximal.

[137] This example is due to Jim Joyce.

Temptation

The default hyperbolic shape of our discount curves is exemplified in many common cases of temptation. For instance, the difficulty of: keeping a diet, not procrastinating, refraining from scratching an itch, going to bed the optimal time, cleaning house, etc. These are not terribly exotic cases—they occur all the time. Such cases fall into the category of 'temptations', as they have the structure of a hyperbolically discounted earlier good appearing to the agent to be more attractive than some longer term goal when the earlier good is temporally proximal.

Such common cases of temptation work in the following way: Our considered preferences and desire responses are both effects of a common cause, which is the property(s) in virtue of which the thing is desirable to you. Now, these properties are nothing as general as a property of desirability. What happens in these temptation cases is that we have a property, such as tastiness (that may be a response-dependent property), where from both the proximal and the vanishing point my attribution of that property to the object remains the same. So, to use the terms of D'Arms and Jacobson about the fit of emotions, the object has the same *shape* for me from both perspectives. What changes between the vanishing point and the proximal cases is the *strength* of my desire response.

We are no strangers to such temptations. Most of us have, on a daily basis, the opportunity to satisfy a short term desire which, if acted upon, will undermine the fulfillment of some presently weaker yet ultimately stronger, longer term, desire. A simple case of this would that of Kim which I introduced in the first chapter. Now, Kim's desire to maintain her diet is stronger than her desire for forbidden foodstuffs, but she has a weakness for chocolate. Most of the time Kim succeeds in maintaining her diet, but when the opportunity to eat chocolate cake is to hand, she is overwhelmed by her desire for the cake. Consider a particular instance: Kim is passing her favorite bakery at 4pm on Tuesday. At that time, she is more

motivated to eat cake more than to maintain her diet. But when she is temporally separated from the opportunity to eat cake, she is more motivated to maintain her diet more than to eat cake. If we look at the strength of her desires over the course of Tuesday, she will begin the day desiring to maintain her diet more than to eat cake. But as she approaches the bakery (and thus the opportunity to eat cake) at 4pm, she is tempted to eat cake. Over the day, the interaction between her desires for cake and dieting look like this:

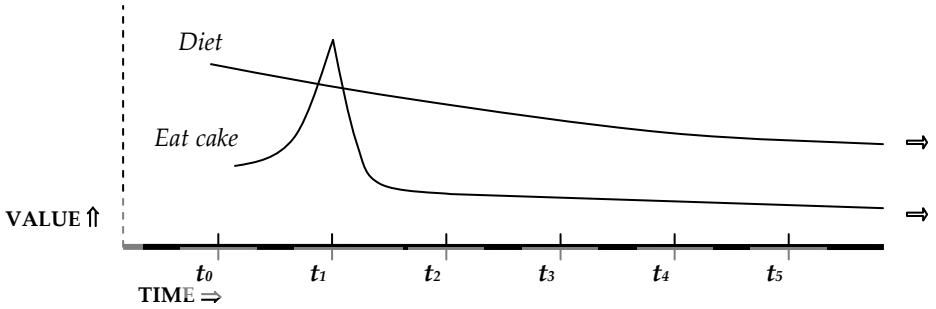

Figure 6: Kim's temptation

So Kim has a stronger motivation to eat cake than to maintain her diet around 4pm. Because the felt intensity of the urge to eat cake spikes when the opportunity to eat cake is proximal, she is discounting the eating of cake hyperbolically. However, dieting, as the kind of good that has no specific date of delivery, is going to behave more like an exponentially discounted good—it has no spikes in strength as there is no specific time of delivery where proximity effects can manifest. In sum, Kim's problem is generated by the fact that her hyperbolic discounting of the cake causes a reversal of her comparative motivational strength of her desires for diet and cake when the possibility of the cake is proximal.

Warranted Change in Desire

Let me consider then, the general question: In which cases would such changes in the motivational strength of a desire, and thus when such a reversal in the comparative strength of two desires, be warranted? (Note that in this section I will use the generic term desire to apply

to both considered preferences and desire responses, as the same points apply to both.)

Consider a case:

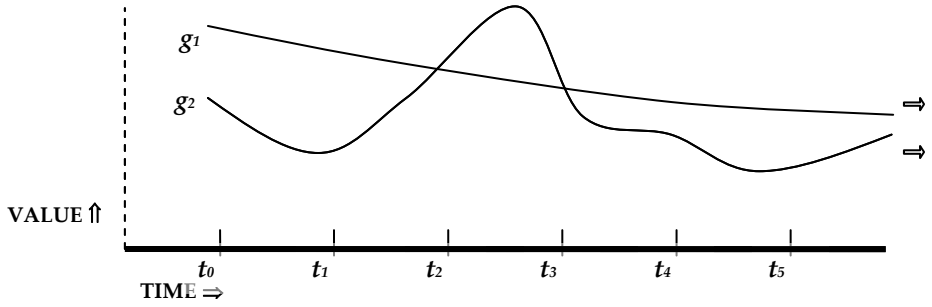

Figure 7: Changeable Discount Curves

In this example g_1 and g_2 have different discount curves. The shape of the discount curve represents the agent's present estimation of how desirable the delivery of that good is at various times. Here the agent's discount curve at time t_0 reflects her present values at t_0 for a good g_n promised at times $t_1 \ldots t_n$.

There are, I propose, (at least) three ways in which the strength of a desire can change over time, and yet be rational. One case in which changes in the strength of a desire over time looks rationally irreproachable is the case of dated goods where the properties in virtue of which the thing is desired change as a result of the passage of time. Take the case of a wine which gets better as it matures. If my desire for the wine is grounded in the sensory pleasure of consuming the wine, then the properties in virtue of which I desire the wine are those very properties which will improve as it ages. It seems correct that the amount that I am willing to pay for the wine should track these properties. As the wine improves (or declines) so my price should increase (or decrease). I will call this the case of 'property change'. This is a case where the underlying properties of g_2 may change in a way that the underlying properties of g_1 do not. Let us take g_2 be a nice bottle of Chablis. Unlike most white wines, this fictional Chablis is good when it is young, declines a little with the loss of the freshness of youth, improves into

middle age, and then severely declines in old age. In this case changes in the agent's discount curve are explained by changes in the properties in virtue of which the good is desired.

The second and third cases in which it seems perfectly reasonable that the strength of my desires fluctuate over time, are both generated by my uncertainties about the future. For instance, it seems reasonable to be willing to pay less for a piece of cake to be eaten in three weeks time than a piece of cake to be eaten now. In paying in advance for a piece of cake I am running two risks that I do not run if I eat the cake immediately. I run the risk that the cake will not be delivered at the later time (uncertainty about the good). In general, the closer I am to some good the more certain I become of receiving that good. Such certainty (uncertainty) about the promise of a good in the future being fulfilled underwrites the second type of reasonable change in desires over time, an uncertainty about the external world that I will call 'delivery doubt'.

The third case is a type of internal uncertainty. The value of a good to me resides not only in the properties of the thing itself, but my uptake of those properties as an agent. If I contract to receive a good in the future I run the risk that I will not be in the relevantly same state at the time of the delivery of the good. For instance, in the case of the cake I run the risk that my hunger for the cake will not be the same in three weeks time. I may simply not feel like eating chocolate cake at the appointed time, and chocolate cake eaten when you don't feel like it is clearly less valuable than chocolate cake eaten when you really want it. Call this 'internal state uncertainty'. Another case of this which is pertinent in the case of goods promised in the far future is the concern of mortality.

Thus warranted changes in the strength of desires happen in three ways: (1) there is some change in the properties of the thing in virtue of which it is valued (property change); (2) uncertainty about the likelihood of getting the good at the future time (delivery doubt); and (3) uncertainty about what your future attitude toward the thing might be (internal state

uncertainty). Note that these are derived from the three reasons that we canvassed earlier to explain why it is that we discount the future at all.

The fact that such an account of warranted change in the strength of desires is available tells us that when desires are working properly, changes in desires—both in our considered preferences and desire responses—are *explicable*. We can trace them back to changes in ourselves or changes in our information about the desired thing, which are causally and reliably linked to the desirable properties of the thing. What is so strange about the change in desire which results from hyperbolic discounting is that this change is not a result of property change, delivery doubt, or internal property change. There is no alteration in Kim's capacity to enjoy the cake, nor the cake itself, such that having the cake at that point would be a better outcome for her. The fluctuation of the felt intensity of her desire to eat the cake (her desire response) is not a response to changes in the properties in virtue of which she finds the cake desirable. The strength of Kim's desire for the cake is changing in response to the mere proximity of the cake, rather than any change in either her certainty about receiving the cake, or the relevant properties of the cake. The claim that certainty is not causing this effect is the most delicate, so let me make this clear by considering what would happen if Kim was dining at the probabilistic restaurant. At the probabilistic restaurant, you do not simply order chocolate cake, but you order the gamble that you will receive a piece of chocolate cake with a probability of .6, and nothing with a probability of .4. The waiter will come to your table not only with the glorious piece of cake that smells deliciously chocolaty, but also the list of gambles that you can purchase.[138] In this case, any shifts in Kim's desire when the *possibility* of the cake is proximal are clearly not a matter of mere certainty about the outcome of having the cake, as the delivery of the cake is not certain. Yet if Kim discounts the possibility of having the cake hyperbolically, then she will have precisely the same reversal of preferences as

[138] This example is due to James M. Joyce.

she does in the case where the delivery of the cake is certain. The reversal in Kim's preferences that occurs through hyperbolic discounting cannot solely be caused by the effects of certainty and delivery doubt. Thus the change in Kim's desire for the cake is not warranted, it is generated by some mistake that she is making about the characteristics of the desire, a mistake which is generated by hyperbolically discounting the future. It is this mistake which makes Kim's case a case of temptation, rather than simply a case of changing desires. What, then, is this mistake?

Hyperbolic discounting and temptation

In the case of temptation generated by hyperbolic discounting, the time at which the good is delivered influences the felt intensity of the desire response *in the absence of any changes in the properties in virtue of which the outcome is desired*. So, what is it rational for Kim to do, in this case? Should she give in to temptation and eat the cake, or resist temptation and maintain her diet? The challenge of temptation is that if we deliberate just by weighing the relative motivational strengths of desires (and desire responses) then arguably she should give in to temptation, as *ex hypothesi*, her strongest motivation is that which she is tempted to do. However, this does not seem like the right response, as a part of what it means for something to be a temptation in the folk sense is that it is somewhat irrational compared to the other desires of the agent. I propose that through understanding how hyperbolic discounting works in such a case we can understand why it is that temptations should not be taken at their face value in terms of motivational strength, and thus should be rejected in deliberating about what it is rational to do in such cases.

The case of hyperbolic discounting is one in which the difference between the strength of the desire from the proximal and the vanishing point perspective is a difference in the evaluative impact of the properties in virtue of which you desire that thing. What happens is

that the sensory features of the tempting alternative—its immediate properties—loom larger in your desire responses in the proximal case than they do in the vanishing point case. The properties of the objects of our desires can be (very roughly) divided into the categories of the attractive and the unattractive. It is seldom the case that all of the properties of some possible object of desire are univocally attractive or unattractive, and indeed, in these cases there is somewhat less room for conflicted responses to that thing. In most cases, objects of desire have a mix of such properties, which a desirous response tracking a balance of attractive over unattractive properties, and an undesirous response tracking the opposite balance. What happens in the case of hyperbolic discounting is that your desiring response emphasizes the immediate sensory properties of the thing, whether positive or negative, while minimizing its distal consequences, when the good is proximal. This is generally called a 'salience' or 'focalizing' mechanism. In this way it is not that your response is getting the properties of the thing wrong—they are all still there—but rather that the intensity of the response to the properties is altered. This salience mechanism interferes with your conative system, which then produces a desire response that does not track your norm for such responses. I propose that Kim's desire response (her desire for cake) is mistaken because it fails to track her normal response to the cake because of proximity effects, and in doing so comes apart from the strength of her considered preference. Thus hyperbolic discounting gives an explanation of why it is that temptations are mistaken desire responses, rather than desires on a par with all other desires.

So, in cases of hyperbolic discounting, the earlier, smaller, good looms larger in the agent's perceptions than does the later, larger good. This is effectively a spilt between her desire responses and her considered preferences. Insofar as the felt intensity of the agent's desire tracks the goodness of the outcome for the agent, the desire responses and considered preferences of the agent will not conflict. But in cases of hyperbolic discounting, they come apart, because of the agent's inappropriately intense desire response.

Recall how the reflective equilibrium between the strength of considered preferences and the strength of desire responses is supposed to work. There is a range of conditions under which I desire the cake, and I desire it in virtue of its desirable features. The automatic affective response that is the desire response also occurs over a range of conditions. Both considered preferences and desire responses are in part produced by a common cause—the properties in virtue of which the thing is desirable to you. However, we tend to encounter difficulties when these two conditions come apart, and thus require, for the purposes of action, to establish a rough equilibrium between them. There is no set way in which this should be done *except for the case in which one of the states is mistaken*. If one of these measures of motivation is mistaken, then *it* is the one that should be discounted in achieving this equilibrium. This is precisely what is going on in the case of temptation, and thus conflicts between the stronger motivational strength of temptations generated through hyperbolic discounting, and the weaker motivational strength of the associated considered preference, should be resolved by adopting the motivational strength of the considered preference.

The trick, however, is to be able to identify when your motivations are being supplied by such a flawed desire response rather than a considered preference when you are in the grip of temptation. One of the characteristics of temptations is that it is not necessarily obvious to us at the moment of desiring how contrary this urge is to our considered preferences. I contend that temptation is not merely an emotional overlay of preferences that can easily be dismissed when it leads us astray; it also misleads us about what our considered preferences are in this case by undermining our ability to effectively reflect on what our considered preferences actually are. All temptations have in common the effect of distorting the agent's reflections on her considered preferences. Due to the same focal and salience effects that generate the mistaken desire response, it is difficult to engage in clear-headed reflection about just what it is that your considered preferences are in such a situation, and thus to act rationally through resisting temptation. Specifically, given that we have more immediate

access to the felt intensity of the emotional response than to the strength of the desire, especially in cases of temptation when our evaluations of the properties in virtue of which we desire things have been distorted (thus distorting our reflections on our considered preferences in that moment), it is easy to confuse the strength of the desire response with the strength of the considered preference.

Now, I began with the claim that there are two characteristics which make a proximal desire a temptation: (1) it conflicts with some more important, longer term, goal, (2) there is some effect at work that interferes with the agent's ability to reflect on her considered preferences. I have identified the cause, nature, and structure of the conflict between the motivationally stronger yet ultimately weaker desire response, and the ultimately stronger yet presently weaker competing desire. What I now need to address is precisely what the effect at work that interferes with the agent's ability to reflect on her considered preferences is, and how this is connected to hyperbolic discounting and inappropriately strong desire responses.

How temptation undermines reflection

That hyperbolic discounting is connected to the problematic aspect of temptation of interfering with the agent's ability to clearly reflect on what her considered preferences are, can be discovered through considering *why* it is that our default rate of discounting the future is hyperbolic. I will describe three representative studies that support the conclusion that the default rate of discounting the future is hyperbolic, and also suggest why hyperbolic discounting tends to go along with the agent's ability to reflect on her considered preferences being impaired.

In all of these studies, the degree to which an agent will choose a smaller, sooner (SS) reward over a larger, later (LL) reward is taken as a measure of his impulsivity. The degree to which an agent resists this temptation is taken to be a measure of his self-control. The

experiments all display a preference reversal in preferences for LL over SS rewards when the delay before both rewards increases the same amount. The greater the tendency the agent has to choose SS over LL rewards as the delay increases, the more impulsive the agent. The discounting function which makes sense of such choices is hyperbolic, hence the findings of pervasive impulsivity in these studies is evidence for the existence of hyperbolic discounting.

The first of these experiments is a series of amount versus delay trials using aversive sounds.[139] The subjects were exposed to an uncomfortable level of white noise, with the reward being the cessation of that noise for some period. The delay in the experiment was the time before the noise would cease, and the amount was the duration of time for which the noise ceased. As the delays before the LL reward increased, the participants chose SS rewards in significant numbers, even though the LL reward was four times the size of the SS reward. Solnick and Navarick express this in terms of a demonstration of the impulsivity of the participants.[140] These types of choices also demonstrate a robust tendency to hyperbolically discount the future.

The second experiment also uses amount versus delay trials, with a period of access to a video game as a reward.[141] This differs from the previous experiment in that the reward is a positive reinforcement, although it is similar in that it is a reward which is immediately consumed. The subjects of this experiment were asked to choose among a variety of different combinations of access to the game, and waiting time, with an initial series of trials establishing that the subjects, on balance, preferred playing the game to waiting. In trials where the two options had an equal amount of delay, and unequal amount of reward, the subjects preferred the greater playing time to the lesser. In trials where the amount of delay

---

[139] This description is taken from two such studies, (Navarick, 1982/8; Solnick, Kannenberg, Eckerman, & Waller, 1980/2).
[140] Navarick, 1982/8; Solnick et al., 1980/2.
[141] Millar & Navarick, 1984/5.

was unequal and the amount of reward equal for the two options, the subjects preferred the lesser to the greater delay. As in the previous case, the trials that supported the hyperbolic discounting hypothesis were the ones where one option embodied an SS reward, and the other a LL reward. If the delays were short, the agents preferred the LL to the SS reward. However, in a series of trials where the delay before both options was increased equally, a significant number of subjects reversed their preferences, choosing the SS over the LL reward. This is another situation that is naturally explained by the hypothesis that the default rate of discounting the future is hyperbolic.

The third is a series of experiments testing the self-control of food-deprived adults using food as a reward.[142] In this study, equal numbers of experimentally naïve men and women from 18-47 years were given controlled access to their favorite juice in an amount versus delay trial after refraining from eating or drinking for six hours. This experiment worked with a positive reinforcement that was immediately consumed, although juice for sustenance-deprived subjects differs in type and accessibility to the video games of the previous study. As in the previous two studies, a significant number of subjects displayed choices consistent with preferring LL to SS rewards where delays were short, and in trial where delays for both options were extended by the same amount, coming to prefer SS to LL rewards.

All of these studies either appeal to agents who are not capable of exposing their choices to a rigorous cognitive screening process, or to choices amongst rewards that are not amenable to such cognitive oversight such as those that are consumed on delivery, or those that confound easy quantification. The relevance of such characteristics of the experimental condition is that they all have the effect of interfering with the subject's ability to reflect upon the comparative goodness of the outcomes. All of the experiments in which hyperbolic

---

[142] Forzano & Logue, 1992/8.

discounting appears to be prevalent create situations in which the ability of the agent to reflect is compromised in some way. In cases where these conditions which compromise reflection do not exist, such as gambles using money, most people appear to discount the future exponentially. What the condition of proximity does to us is to introduce various salience and focalizing effects that tend to compromise our ability to reflect by emphasizing the good in the desired alternative, and minimizing the bad. It is these effects that block what I have been calling 'clear-headed reflection' about what our considered preferences are in the condition of proximity.[143]

The structure of amount vs. delay experiments reveal an agent's 'unreflective' preferences better than experiments which employ easily quantifiable rewards such as money. Moreover, these 'unreflective' preferences are precisely the type of felt emotional components of desires that are generated by the conative system. The amount vs. delay experiments are a fruitful ground for such behaviors because it is evident that although we care about the delay before our reward, we are just not terribly good at quantifying and comparing amounts and delays. In essence, such experiments reveal our felt emotional components of desires before they are modified by reflection. It is no surprise, then, that the kinds of desire responses which are subject to hyperbolic discounting and tend to constitute temptations, are directed toward objects or outcomes that are not easily quantifiable, and have a tendency to elicit emotional responses of desiring, that is, desire responses.

One reason that hyperbolic discounting may be counter-intuitive, which is a fairly common response, is because we tend to *learn* to discount exponentially. This in turn implies that we recognize the value of exponential discounting, and tend to take exponential discounting as a norm for measuring choices. Note that this is entirely consistent with it being

---

[143] Note that my claim is not that we do not have *any* access to our considered preferences in cases of temptation, but rather that our access is blocked in such a way that we have an imperfect grasp of the stable motivational strength of our considered preferences.

the case that our default rate of discounting the future is hyperbolic, however it highlights the question of why such a default rate might evolve.

I am not particularly committed to a specific account of the origin of hyperbolic discounting, as I am more concerned with the fact of its existence. It is plausible that an evolutionary story could be told about its origin, although this, like many other evolutionary tales, is a 'just so' story with all of the attendant difficulties. We can readily imagine a time in the past where hyperbolic discounting might confer significant evolutionary advantage, and it is only relatively recently that the disadvantages of hyperbolic discounting have become relevant. In a society of hunter gatherers whose time is spent in the service of merely surviving against a background of scarce resources, it is plausible that 'over-valuing' proximal resources by systematically magnifying the strength of desire responses to things such as that piece of food, that element of shelter, or that sexual partner, would be an excellent survival strategy. For a human who is starving, eating what is available now could be a critical difference in whether or not she survives. However, as I said earlier, this is simply a sketch of an evolutionary story which is merely meant to hint at the origins of hyperbolic discounting, without making any firm claims.

The next question is what evidence is there that the hyperbolic curve is fundamental. I suggest that it is demonstrated by its persistence in the face of the clear advantages of exponential discounting. Consider: Not only can the person with an exponential curve exploit a person with a hyperbolic curve; Ms. Hyperbolic is doubly at a disadvantage as she will consistently undermine her long term goals by giving in to short term temptations. So, it seems clearly in Ms. Hyperbolic's interests to discount the future exponentially. Moreover, Ms. Hyperbolic seems to recognize this in that most people, in cases of goods that are not immediately consumed, or when diabolical psychologists are not confounding their computations about outcomes, act as if they discount the future exponentially. Why does the

hyperbolic curve persist? Because it seems as if the rate in which we discount the future is not in our direct control. If it were, then we would all be manipulating our utilities in order to make things worth more to us. If I know that I am receiving a particular present *g* on my birthday, and my discounted valuation of *g* is six utiles, one consequence of my discount curve being in my direct control would be that I could change it so that *g* is worth 10, 12 or 20 utiles to me. We could coin our own rewards, and would, insofar as more reward is better, have a significant incentive to do so. Yet we do not do this, which suggests that this control is not something that we possess. This is not to deny that we can discount the future exponentially—we clearly do in many cases—but to claim that this is not our default rate of discounting the future. Our default rate is not a matter of choice, but rather a fact about our psychological dispositions to value goods. As Ainslie puts it: "The banker-like curve seems to represent an added accomplishment, not a fundamental change."[144] What makes this an added accomplishment is that it looks like exponential discount curves are a function of the agent's reflection on, and deliberation about, her desires. Given that hyperbolic discounting is the default rate of discounting—that is, that hyperbolic discounting is a feature of the way that we are psychologically disposed to value goods—then exponential discounting represents a triumph of reason over our natural impulses in evaluation. This accomplishment is demonstrated in our considered preferences, which tend to act as if their objects have been exponentially discounted, but is lacking from our desire responses, because our desire responses are the product of the same conative system that represents our psychological dispositions to value goods. Thus it is not that hyperbolic discounting causes the proximity effects which undermine the agent's ability to clearly reflect on her considered preferences. Rather, such proximity effects appear to be a precondition for hyperbolic discounting to manifest itself.

---

[144] Ainslie, 2001, p. 37.

The hypothesis that temptations generated by hyperbolic discounting are connected to conditions which interfere with the agent's ability to clearly reflect on what her considered preferences are is supported by the fact that all of the experiments in which hyperbolic discounting appear to be prevalent have one thing in common, they are situations in which the ability of the agent to reflect is compromised in some way. One of the experiments used cognitively impaired adults as subjects; other experiments use goods that do not lend themselves to quantification (and thus comparison), and so on. Thus hyperbolic discounting is plausibly the result of desire responses similar to the perceptual responses characteristic of various visual illusions. However what makes hyperbolic discounting so pernicious is that we are not good at correcting for this warp in our conative system in cases where the objects of our desires are not easily quantifiable, in the way that we do in the case of visual illusions. In cases where these conditions which compromise reflection do not exist, such as gambles using money, most people appear to discount the future exponentially. What the condition of proximity does to us is to introduce various salience and focal effects that tend to compromise our ability to reflect by emphasizing the immediate properties in a desired state, and minimizing the distal consequences of that state. These effects undermine the agent's ability to fully reflect on what her considered preferences are in the condition of proximity.[145]

---

[145] One issue that I should touch on is the relationship between my project and Ainslie's views. Ainslie's own project is descriptive, rather than normative. He is aiming to reconcile his empirical discovery of the prevalence of hyperbolic discounting with the utility model: "Hyperbolic discounting is a shock for utility theory. Suddenly the pavement moves beneath our feet, and we have to take the simple concept of maximizing expected reward not as a description of basic human nature but just as a norm that we try to implement." (Ainslie, 2001, p. 38.) His goal is to give an account of how it is, on the assumption that we discount the future hyperbolically, that utility theory can describe the phenomena of self-control, impulse, and addiction. Ainslie is aiming to form a hypothesis about the nature of the will that makes sense of akrasia and other paradoxes of motivation which "…does not violate the conventions of science as we know it." (Ainslie, 2001, p. 12.) By modeling the mind as a population of temporally divergent and competing interests, shaped by hyperbolic discounting, Ainslie finds this theory. Insofar as Ainslie's project is to describe an extension of utility theory, his aims are orthogonal to mine. Thus the uses to which he puts hyperbolic discounting are suggestive of, but neither coincident with, nor contradictory to the view that I am defending.

Thus hyperbolic discounting is an identifiable mistake in the conative system because it systematically produces desire responses whose motivational strength does not match the strength of the correlated considered preferences. Such desire responses cause difficulties with deliberation because the conditions that produce these temptations also interfere with the agent's ability to reflect upon her considered preferences. In such cases it is easy to treat the strength of the desire response as the strength of the considered preference, especially as in many cases where the conative system is operating normally, the felt intensity of such desire responses is correlated with the motivational strength of the considered preference.

Temptation and deliberation

Now, this explanation of what it is that is wrong with temptations, and how it is that they are produced by the conative system operating under the abnormal conditions of proximity, still leaves a final issue to be addressed. This view is not particularly helpful unless I can solve the deliberational problem of how it is that we identify such flawed desires in the absence of a complete accounting of their origin, which is unlikely to be forthcoming. The analysis that I have given so far about flawed desire responses only allows these responses to be identified if the agent has access to the mechanism by which the desire was created. But this is not a useful method for identifying such problematic desires in the midst of deliberation, and such identification is necessary to answer the challenge of temptation.

I propose that we can use the comparative stability of our desires as a proxy for information about the motivational strength of our considered preferences, and thus as a proxy for identifying mistaken desire responses. In short, the concept of stability for desires can be used to give an account of why it is that in a case like Kim's her considered preferences should be privileged over her desire response to the cake. Consider: Cases of unstable desires that result from hyperbolic discounting are analogous to cases of unstable beliefs generated by wishful thinking. In the case of Charlie Brown and the football introduced in the first chapter,

the change in his beliefs is caused by an event that has nothing to do with the truth of his beliefs. The changes in Charlie Brown's beliefs, and thus this instability, are caused by information which is *not evidentially relevant* to his opinions. Charlie Brown is simply mistaken if he treats his opinions which are generated by wishful thinking on a par with all of his other opinions, as they are not sound. Essentially his belief change (and the belief that results from this change) is unwarranted, and should be discounted in his deliberations.

In the case of Kim, the fluctuation of the felt intensity of her desire to eat the cake (her desire response) is not a response to changes in the properties in virtue of which she finds the cake desirable. Like the case of wishful thinking, this instability in her desire response is not moving Kim toward a better grasp of that which the state reflects, the underlying goodness of the outcome for her. The change in the strength of Kim's desire for cake is not warranted because, like the change in Charlie Brown's credence in the case of the football, it is not caused in the right way. Thus her desire change (and the desire that results from this change), is unwarranted and should not be treated on a par with her other desires in her deliberations.

In the cases of Charlie Brown and Kim, what makes the unwarranted change in attitude clearly problematic is that the attitude which is strengthened comes to dominate some other, better grounded, attitude. In Charlie Brown's case the attitude is his correct belief that he will not be able to kick the ball. In Kim's case it is her considered preferences to maintain her diet. Thus paying attention to types of instability gives us an account of what it is that is wrong with proximal preferences (desire responses). Specifically, the instability in the felt intensity of the hyperbolically discounted desire response is not moving the agent toward a better grasp of the underlying goodness of the outcome for her, where that underlying goodness is expressed through either the normal desire response that she has to that thing, or the motivational strength of her considered preference. Thus, we can understand temptation in

terms of this identifiable type of instability that is displayed by temptations, instability which is not caused in the right way.

How, then, does paying attention to the stability of these states solve the deliberational problem of how it is that we identify such flawed desires? Consider again the case of Kim: her mistake is that she is, in some sense, *overestimating* the goodness of the outcome of eating cake at the moment when the option of eating cake is proximal. The strength of her desire response, and thus the strength of her motivation for the cake, is abnormally strong for her. The epistemic problem is that her ability to directly evaluate precisely how good she really takes the outcome to be is distorted by the proximity effects that produce the temptation. Had she, at the time of choosing, had a different perspective to take on her desires—a perspective 'out of the moment' as it were—she would be able to see that the desire response for the cake is not her considered preference for the cake. The felt intensity of her desire for cake was weaker 10 minutes ago, and will be weaker 10 minutes from now, which is a fact about her desire responses that we have access to from this temporally extended perspective. However, Kim herself cannot simply take this third personal perspective in, so it is not clear that she is making a *cognitive* mistake in believing, *at that point in time*, that eating the cake is her considered preference. What is happening is that there is some subset of her conative machinery—that part of the mind dedicated to the processing of the felt intensity of desire responses—which 'takes over'. Kim has better access to her desire responses at the point of temptation through the lens of felt intensity, than she has to her considered preferences. From the outside, observing this shift in desires, we can see that it is not warranted, so the desire response for the cake is not warranted. But this mistake is only clear from the *third personal* perspective.

As it stands, my view apparently suffers from an epistemic problem. How is it that *Kim* is able to tell that her desire is mistaken? Especially as given the distorting effects of

temptation on the agent's capacity to reflect on her considered preferences, there is no direct way for the agent to do this. I have two responses to this issue.

The first is that there is an application of the view that need not solve the epistemic problem. In arguing that inexplicably unstable desire responses are flawed, I am not suggesting that it is *necessarily* the case that the *agent* must have access to these mistakes. I want to appeal here to Nomy Arpaly's distinction between an 'account of rationality' (a third personal theory of rationality that gives an account of when it is that an agent is acting rationally), and a 'rational agent's manual' (a first personal theory for how you, as an agent, should deliberate in order to be rational).[146] Interpreted as a third personal evaluation of rationality, my argument that temptations are flawed desire responses whose cause also interferes with the agent's ability to reflect on her considered preferences need give no account of how the agent herself has access to this information, particularly given the subtle nature of the agent's mistake.

I do think, however, that a more can be said, using the concept of stability, about how this could be understood as a part of a rational agent's manual. In other words, the agent may have access to information about her mistake, through taking the 'third personal' perspective on her competing motivations outlined earlier. The significance of the stability of beliefs in deliberation is that a measure of stability can act as a proxy for the justification of the belief. A deliberatively responsible agent can rely on stable beliefs (that is, beliefs that are *known* to be stable), in deliberation, even if they can no longer access the justification for that belief. Consider the case of auto-epistemic beliefs. An auto-epistemic belief is a belief that has the property that the fact that you have the belief at all, and that this belief is stable, is reason to think that it is true. That I have the belief that my mother's middle name is Brenda is a reason to think that my mother's middle name is Brenda. This is the case even if I have no other

[146] Arpaly, 2003.

evidence or justification for the belief. In cases of such deeply ingrained information, the *presence* of a particular belief is reason to think that that belief is justified. Moreover, it is not just the fact that I have the belief right now which is reason to think it true, but the fact that I have had this belief over a period of time. In these circumstances, the fact that the belief is a stable one is evidence that the belief is a justified one. Here the stability of the belief acts as a proxy for the justification of that belief. The result is that the agent need not revisit, nor even know (at this point in time), the justification of such a stable belief in order to utilize the belief in deliberation.

My hypothesis is that the stability of desires may play a role in deliberation similar to that of the stability of beliefs. It can act as a proxy for information relevant to deliberation — information relevant to distinguishing mistaken desire responses from considered preferences. Kim's desire for cake is a desire response, while her desire to maintain her diet is a considered preference. One difference between desire responses and considered preferences is the scope of the current information which enters into them. In understanding the strength of her considered preferences, Kim reflects on all of her reasonably accessible information about the world and herself, including her information about her past and future desires. In contrast, the strength of her desire response just is the felt intensity of her desire at a time. One thing to note about desire responses is that the intensity of such an automatic affective response is always tied to the moment of emotional experience, as they are occurrent states. If Kim takes these desire responses to be authoritative of the strength of her considered preferences, then the evidence that she is relying upon is very restricted. This would not be a problem if the information excluded from the calculation is irrelevant to the deliberation at hand, but it seems clear that information about the past and future state of the motivational strength of a

desire *is* relevant to this choice.[147] It is both possible and coherent for Kim to feel a desire for cake more intensely than her desire to maintain her diet at a particular time, while being aware that in the past she has preferred maintaining her diet over eating cake. Stability acts as a proxy for this information by flagging abrupt and inexplicable changes in desire. Considering the relative stability of her motivations with respect to the passage of time will demonstrate which is more likely to be due to a desire response, and which due to a considered preference, and thus whether one of them should be discounted in deliberation.

So, there are three pieces of indirect information about her desire responses and considered preferences that Kim has access to. The first is the stability of her motivation to diet, which is stable, thus indicating that it is likely to be a function of a considered preference. The second is the past strength of her desire to eat cake, which has been stably weaker than her desire to diet in all cases except those in which cake is proximal. The third is the sudden spike in the felt intensity of her desire to eat cake in the presence of the cake, which in its instability looks like a flawed desire. I propose that the instability of the sudden increase in the motivational strength of her desire, in the absence of any evidence of a warranted change, is evidence that this is not due to a considered preferences or normal response, but rather a desire response with inappropriate motivational strength. Thus the instability of her desire to

---

[147] Let me make a delicate point here. Many of our deliberations about what to do take place automatically. We employ various heuristic devices to limit the scope of our deliberations, and to establish what we really care about in any given case, because a full reassessment of our entire corpus of beliefs and desires every time we are deciding whether or not to eat the chocolate cake is simply not feasible. I propose that felt intensity tends to act as proxy for our considered preferences in deliberation, which is just fine so long as our considered preferences and felt motivations do not come apart. The resilience of various motivational accounts of desire is evidence that this approach is pretty good. However, in cases of temptation, the trick is not to act on temptation, but to resist it. Given the prevalence of hyperbolic discounting, and such attendant splits between the agent's desire responses and considered preferences, strict instrumentalism dictates that people should give in to temptation far more than they actually do or ought. Thus, as Ainslie might put it, the main challenge to instrumentalism is not to explain why we give in to temptation when we do, but to explain how it is that we manage to resist temptation so successfully. Moreover, it needs to say why it is that we take giving in to temptation to be some sort of failure. If strict instrumentalism were correct, then not only would we be giving in to temptation on a regular basis, we would not be perceiving this as any type of rational failure.

eat cake, when compared with the stability of her considered preference to maintain her diet, indicates that maintaining her diet is, in this instance, her strongest *stable* motivation. Thus it is the motivation she should act on if she is only considering the strength of her considered preferences. Stability, then, can act as a proxy for information relevant to deliberation, information about what is, and is not, an appropriate desire response. The instability of desire responses *that cannot be explained* is a symptom of flaws in the proximal emotional responses of desiring, thus such inexplicable instability can be used as evidence that a desire has been produced by the conative system under abnormal conditions—in other words, inexplicable instability can function as a marker of flawed desires, desires that should be discounted in rational deliberation.

Conclusion

The challenge of temptation is that a plausible theory of desire must be able to differentiate temptations from other desires, and in doing so, show why it is that they should be treated differently in deliberation. By taking the emotional component of desire—the desire response—seriously, I have argued that what differentiates temptations from other desires is that they are inappropriately intense desire responses. The mechanism that produces such temptations is hyperbolic discounting, which is the default rate at which we discount the future. That it is the default rate means that hyperbolic discounting tends to be caused in conditions where the agent's ability to reason about her considered preferences has been compromised in the manner discussed earlier. Thus Kim's desire to eat cake is irrational because the change in the strength of her desire is not warranted—it is a change caused by hyperbolic discounting and proximity effects—thus it is not stable. My view is not that, as a general matter, stability or resilience is evidence of a desire that is going to stand up to rational scrutiny. Rather it is that in this very specific case of temptation generated by hyperbolic discounting, stability can act as a proxy in deliberation for direct information about what is,

and is not, a considered preference. Thus it is evidence about which motivations should be discounted in deliberation. So, although instability is not in itself a problematic characteristic of desires, instability of the kind that generates unwarranted changes in desires is a marker of the type of conative flaw that generates temptations.

Appealing to stability in this way also gives us an account of what precisely it is that is wrong with Gauthier's proximal preferences (desire responses), which in turn generates an account of why they should not be treated the same as other desires in rational deliberation. Hyperbolic discounting shows why it is that the conative system yields felt emotional components of desiring which are not fitting under conditions of proximity. But more than this, it shows how pervasive mistakes in intensity occur in the conative system in the formation of proximal desires that count as temptations. Isolating the mechanism of hyperbolic discounting thus explains what causes these proximal shifts in our preferences in temptation cases. This is an improvement on Gauthier's view, as he leaves it a mystery why we have these proximal shifts in our preferences, and does not address the question of precisely what it is that is wrong with them.

Recall how the reflective equilibrium between the strength of considered preferences and the strength of desire responses works. There is a range of conditions under which I desire the cake, and I desire it in virtue of its desirable features. The automatic affective response that is the desire response also occurs over a range of conditions. Both considered preferences and desire responses are in part produced by a common cause—the properties in virtue of which the thing is desirable to you. However, we tend to encounter difficulties when these two conditions come apart. If one of these measures of motivation is mistaken, then it should be discounted in achieving this equilibrium. I have argued that the flawed desire responses should be discounted in achieving reflective equilibrium between the motivational strength of desire responses and the motivational strength of considered preferences, on the grounds that it

results from the conative system operating under abnormal circumstances. However, I think there are also independent reasons for reaching a reflective equilibrium in this way.

Another, and different, species of argument for favoring considered preferences over flawed desire responses in such cases is that hyperbolic discounting opens us up to various types of exploitation. One such form of exploitation is extremely common and widely successful. Every time that you go to a restaurant, and the server brings you a dessert tray to look at after you have eaten your main, your tendency to overestimate the attractiveness of dessert when it is right in front of you is being exploited. There are endless articles in marketing magazines advising various businesses how to exploit the kinds of temptations that are generated by hyperbolic discounting.

A different face of this same vulnerability is that such proximal preferences, in the sense that they short-circuit our considered decisions about what is valuable to us, undermine our autonomy in an identifiable way. They interfere with our ability to clearly reflect on what our considered preferences are in predictable ways. This is what is being exploited in the desert tray case, and it can't be good. Thus in a very straightforward sense, privileging these flawed desire responses over considered preferences makes us worse off. However, this is not to say that we can never rationally act on whims, inclinations, or other species of spontaneous desire, rather, the claim is that we cannot act on them when they are contrary to those considered preferences that we wish to maintain.

In sum, these automatic affective responses are connected to desires because they capture the relationship of desiring between agents and the objects of their desires. They solve the gap in Schroeder's Reward Theory of Desire, because such automatic affective responses provide a principled explanation for why it is that some representations are constituted by the agent as rewards while others are not. Finally paying attention to this emotional component of desires answers the challenge of temptation by highlighting the role that the conative system

plays in accurately or inaccurately generating desire responses, and shows that classic cases of temptation are generated by the conative system operating under abnormal conditions. Thus temptations should not be treated on a par with other desires in rational deliberation, because they are irrational states in the sense that they are inappropriately strong desire responses.

CHAPTER 6

CONCLUSION

There are three concluding points that I want to make. The first is to acknowledge those areas of the dissertation where I have not fully discussed the relevant issues. The second is to suggest those areas of future research that arise from the dissertation. The final and most important point that I want to make is what I take to be the main contribution of the dissertation to be. I will address these in turn.

There are several issues which I have excluded from the scope of the dissertation that should be addressed in a fully realized version of the view put forward here. I note them here both as a recognition of these gaps, and because these gaps suggest interesting areas for further exploration.

The first of these issues is that I suggested that Schroeder's Reward Theory of Desire is best understood as a theory of desire change, rather than a theory of desire. However, I did not address precisely how this might work. I take it that a full discussion of this possibility would be fruitful area of further inquiry. A particularly interesting question here is how this, as an account of desire change, would interact with the account of warranted and unwarranted changes in desire that I gave in chapter five.

Another issue is precisely what the connection between full-fledged desires (considered preferences), and the automatic affective responses of desiring (desire responses)

is. I have treated desire responses as connected to, and a type of evidence about, considered preferences. But I am sure that this is not the full story. I suspect that a comprehensive analysis of these states will yield a view of desire-like states that exists on some interesting continuum between these two extremes, with some unclear line between what should, and what should not, be treated as a pure case of desire. This is not only a gap in the argument I have given here, but also an area for further research.

The last of the elements of the dissertation which require greater attention than I have given here is the analysis of the temptation cases. I have considered only a restricted, albeit common, case of temptation—that of temporal conflicts amongst motivations resulting from hyperbolic discounting. A stronger version of the case that I have presented would present both a taxonomy of temptation cases, and a discussion of the rationality or otherwise of acting on those temptations.

Not only are these areas for discussion which are highlighted by the gaps in the discussion I have given, there are a variety of fertile possibilities for further research generated by the positive contribution of the dissertation—the various properties of desires that I have defended—the accounts of stability, warranted and unwarranted change in the strength of desires, and the existence of the emotional component of desires.

One possible such extension of the work here is in applying the idea of the stability of desires to the question of how to determine the right strategy for choice when we are choosing diachronically. Diachronic choice is what happens when we make choices that are implemented over time, rather than at a time. It differs from synchronic choice—choice at a time—in that it faces the problem of future action, which is the question of how it is that my commitment at the present time to some action in the future (in the form of a plan, course of action, or future-directed intention), can (or should) affect my actions at the relevant future point. This problem is particularly salient in cases of desire change, a subset of which I have been

exploring here. By their very nature, intentions are temporally located – they are formed at a time, in light of a particular set of desires and beliefs. If we are acting on an intention at the approximate point in time at which it is formed, then there is no difficultly with this. As I form the intention to sip my coffee and immediately reach out to do so, the beliefs and desires that lie behind this intention are, in the vast majority of cases, going to be the same throughout the intention-action sequence. But in cases like that of Ulysses and the Sirens, the agent desires one thing when he plans what to do on hearing the Sirens, and quite another when it comes time to implement his initial plan. In dynamic choice theory such a change in desire goes under the name of 'dynamic inconsistency'.

What makes dynamic inconsistency so problematic is that when I form an intention about actions that will take place in the future; I am forming a plan to act at that future time. But this causes a problem for explaining action over time, lucidly explained here by Michael Bratman:

> Future directed intentions and plans are, after all, revocable: they do not control one's future conduct by way of some mysterious action at a distance; and many times, in the face of new and relevant information, we recognize that it would be folly to stick rigidly with our prior intention period so in exactly what sense am I now committed to later action when I settle now on a plan so to act then?[148]

The mysteriousness of the motive force of such resolutions lies in the question of why my future self should adhere to some resolution that my past self made. Especially as my past self resolved to so act when she was not experiencing the desires that my future self is experiencing at the time the choice is to be acted upon.

There are two modes of choice that are often presented as solutions for this problem, resolute choice and sophisticated choice. Resolute choice is the view that we can form an intention at a particular time, and then succeed in implementing that intention no matter how our desires

---

[148] Bratman, 1999, p. 2.

change, simply by *resolving* to act on that intention. Sophisticated choice is the view that it is rational to act on that desire which is strongest at the moment of choice, so if you want to prevent your future self from acting what you know will be her strongest desire—in some cases that which she is tempted to do—you must ensure that she is not in a position to have that desire. In other words, the strategy of sophisticated choice is to avoid the mere possibility of temptation. What motivates this view is that it is always rational to act on your strongest desire, whether or not your present self thinks that your future self's desire is in fact a temptation. This view promotes the primacy of the present self. On this view, that which the agent is right to do is that which she judges to be right *at the moment of acting*. Past decisions are taken to have no deliberative weight. In essence, the agent is assumed to begin deliberating about what to do at each moment of choice with a 'blank slate', consulting only the beliefs and desires that she has at that point in time. This is in contrast to the approach of resolute choice, which holds that the agent's prior intentions/desires *should* be taken into account when making choices. Now, neither of these accounts of how it is rational to choose in the face of dynamic inconsistency is clearly more plausible than the other. Indeed, there are examples in which it looks obviously rational to privilege your present choices over your future self's views; and other cases where it seems clearly rational to privilege your future self's views over your present choices. The resolution solution to the problem of dynamic inconsistency is controversial because it doesn't seem to be the intuitively right thing to do in some cases (although it clearly is in others). I suspect that a plausible mixed strategy of rational choice in some cases and sophisticated choice in others could be defended by paying attention to the relative stability of the desires involved in the preference reversals.

Another area for future research that looks promising is in terms of using the properties of desire that I have proposed to generate norms of rational criticism for desires, thus challenging the strict Humean view that desires are not the kind of states that are

amenable to rational criticism. I think that such norms may be found in both the account of warranted and unwarranted desire change, and in the account of how it is that desire may be more or less stable with respect to reflection. The existence of norms of rational criticism for desires would have many interesting implications in decision theory, rational choice, and action theory.

In conclusion, throughout the dissertation I have been examining the nature of desire through the lens of the challenge of temptation. I began by pointing out that the challenge of temptation cannot be answered by the currently favored simplistic accounts of desire that treat them as states with just the two basic characteristics of motivational strength and object. In the intervening four chapters I have discussed accounts of desire that are inadequate in a variety of ways, and defended the view that desires have certain attributes, attributes which I appealed to in giving an answer to the challenge of temptation. I think that the reason that the challenge of temptation should be of general interest is that it tells us something about the inadequacy of most common theories of desire. Indeed, the theories which I have canvassed generally failed to meet the challenge of temptation for reasons connected to their simplicity. What the challenge of temptation does is to illustrate the necessity for a more complex account of desires. Thus the underlying aim of this dissertation is to give a more psychologically realistic picture of desires than is the norm in philosophical discussions of rationality, action, and deliberation.

Similarly, I take the main contribution of this dissertation not to be the specific response that I offer to the challenge of temptation, but rather the various properties of desires that I have identified and defended in giving this response. Specifically: the account of how desires may be more or less stable with respect to reflection and new information; the role of the emotional component of desires, desire responses, in variously forming and undermining our reflectively stable desires (our considered preferences); the uneasy relationship between

how it is that we are psychologically disposed to value goods over time (hyperbolically), and the way in which it is rational to value goods over time (exponentially); and the account of the difference between warranted and unwarranted changes in the strength of desires. However this is more of a picaresque exploration of some of the properties of desires other than motivational strength and object, than a fully realized theory of desire. I hope that it makes the point that desires are far more complex than they are generally accepted to be.

BIBLIOGRAPHY

Ainslie, G. (2001). *Breakdown of will*. Cambridge; New York: Cambridge University Press.

Ainslie, G., & Haslam, N. (1992). Hyperbolic discounting. Loewenstein, G., & Elster, J. eds. *Choice Over Time.* (pp. 57-92). New York: Russell Sage Foundation.

Anscombe, G. E. M. (1957). *Intention*. Oxford: Blackwell.

Arkonovich, S. (2001). Defending desire: Scanlon's anti-humeanism. *Philosophy and Phenomenological Research, 63*(3), 499-519.

Arpaly, N. (2003). *Unprincipled virtue: An inquiry into moral agency*. Oxford: Oxford University Press.

Baumeister, R. F., Vohs, K. D., DeWall, C. N., & Zhang, L. (2007). How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review, 11*(2), 167-203.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews, 28*(3), 309-369.

Bratman, M. E. (1999). *Faces of intention: Selected essays on intention and agency*. New York: Cambridge University Press.

Chaiken, S., & Trope, Y. (1999). *Dual-process theories in social psychology*. New York: Guilford Press.

Copp, D., & Sobel, D. (2002). Desires, motives, and reasons: Scanlon's rationalistic moral psychology. *Social Theory and Practice: An International and Interdisciplinary Journal of Social Philosophy, 28*(2), 243-276.

D'Arms, J., & Jacobson, D. (2000a). The moralistic fallacy: On the 'appropriateness' of emotions. *Philosophy and Phenomenological Research, 61*(1), 65-90.

D'Arms, J., & Jacobson, D. (2000b). Sentiment and value. *Ethics: An International Journal of Social, 110*(4, pp. 722-748), July 2000.

Davidson, D. (1980). *Essays on actions and events*. Oxford; New York: Clarendon Press.

Dayan, P., & Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron, 36*(2), 285-298.

De Sousa, R. (1987). *The rationality of emotion*. Cambridge, Massachusetts: MIT Press.

Dretske, F. (1996). How reasons explain behaviour: Reply to Melnyk and Noordhof. *Mind and Language, 11*(2), 223-229.

Evans, I. M. (2001). Reinforcement, principle of. In Neil J. Smelser and Paul B. Baltes (Ed.), *International encyclopedia of the social &Behavioral sciences* (pp. 12999-13002). Oxford: Pergamon.

Foltz, E., & White, L. (1962). Pain 'relief' by frontal cingulotomy. *Journal of Neurosurgery, 19*, 89-100.

Forzano, L. B., & Logue, A. W. (1992/8). Predictors of adult humans' self-control and impulsiveness for food reinforcers. *Appetite, 19*(1), 33-47.

Gauthier, D. (1997). Resolute choice and rational deliberation: A critique and a defense. *Noûs, 31*(1), 1-25.

Goldman, A. I. (1970). *A theory of human action*. Englewood Cliffs; New York: Prentice Hall.

Greenspan, P. S. (1988). *Emotions and reasons: An inquiry into emotional justification*. New York: Routledge.

Harman, G. (2004). *Practical aspects of theoretical reasoning*. Oxford: Oxford University Press.

Hume, D. (1900). *An enquiry concerning human understanding*. Chicago: Open Court.

Hurley, P. (1989). Where the traditional accounts of practical reason go wrong. *Logos: Philosophic Issues in Christian Perspective, 10*, 157-166.

Jackson, F., & Pettit, P. (2002). Response-dependence without tears. *Nous, 36*, 97-117.

Joyce, J. M. (1999). *The foundations of causal decision theory*. Oxford; New York: Cambridge University Press.

Kunst-Wilson, W. R., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science, 207*(4430), 557-558.

Langston, J. W., & Palfreman, J. (1995). *The case of the frozen addicts* (1st ed.). New York: Pantheon Books.

LeBar, M. (2005). Three dogmas of response-dependence. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 123*(3), 175-211.

Locke, J., & Nidditch, P. H. (1975). *An essay concerning human understanding*. Oxford: Clarendon Press.

Loeb, L. E. (1991). Stability, justification, and hume's propensity to ascribe identity to related objects. *Philosophical Topics, 19*(1), 237-270.

Loeb, L. E. (2002). *Stability and justification in Hume's treatise*. Oxford: Oxford University Press.

Millar, A., & Navarick, D. J. (1984/5). Self-control and choice in humans: Effects of video game playing as a positive reinforcer. *Learning and Motivation, 15*(2), 203-218.

Murphy, S. T., Monahan, J. L., & Zajonc, R. B. (1995). Additivity of nonconscious affect: Combined effects of priming and exposure. *Journal of personality and social psychology, 69*(4), 589-602.

Nagel, T. (1970). *The possibility of altruism*. Oxford: Clarendon Press.

Navarick, D. J. (1982/8). Negative reinforcement and choice in humans. *Learning and Motivation, 13*(3), 361-377.

Polanyi, M. (1952). The stability of beliefs. *British Journal for the Philosophy of Science, 3*, 217-232.

Quinn, W. (1995). *Putting rationality in its place*. New York: Clarendon Press.

Roberts, R. C. (1988). What an emotion is: A sketch. *Philosophical Review, 97*, 183-209.

Rothbard, M. M. (1987). Time preference. In J. Eatwell, M. Millgate & P. Newman (Eds.), *Utility and probability* (pp. 270-275). New York: Macmillan Press.

Rott, H. (2004). Stability, strength and sensitivity: Converting belief into knowledge. *Erkenntnis: An International Journal of Analytic Philosophy, 61*(2-3), 469-493.

Savage, L. J. (1972). *The foundations of statistics* (2d rev. ed.). New York: Dover Publications.

Scanlon, T. (1998). *What we owe to each other*. Cambridge, Massachusetts: Belknap Press of Harvard University Press.

Schroeder, T. (2004). *Three faces of desire*. Oxford; New York: Oxford University Press.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593-1599.

Schultz, W., Tremblay, L., & Hollerman, J. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex, March* (10), 272-283.

Sedaris, D. (1997). *Naked* (1st ed.). Boston: Little, Brown and Co.

Skyrms, B. (1984). *Pragmatics and empiricism*. New Haven: Yale University Press.

Smith, M. (1995). *The moral problem*. Oxford, United Kingdom; Cambridge, Massachusetts: Blackwell.

Smith, M. (1998). The possibility of philosophy of action. In Deliberation and Causation, Bransen, J. ed, *Human action* (pp. 17-41). Dordrecht: Kluwer.

Sobel, J. H. (1990). Maximization, stability of decision, and actions in accordance with reason. *Philosophy of Science, 57*(1), 60-77.

Solnick, J. V., Kannenberg, C. H., Eckerman, D. A., & Waller, M. B. (1980/2). An experimental analysis of impulsivity and impulse control in humans. *Learning and Motivation, 11*(1), 61-77.

Solomon, R. C. (1976). *The passions*. Garden City, New York: Anchor Books.

Sorell, T. (1981). Harman's paradox. *Mind: A Quarterly Review of Philosophy, 90*, 557-575.

Stampe, D. W. (1987). The authority of desire. *Philosophical Review, 96*, 335-381.

Summenhart, C. (1499). *Treatise on contracts*. Tübingen: University of Tübingen.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning : An introduction*. Cambridge, Massachusetts: MIT Press.

Svavarsdóttir, S. (2006). *How do moral judgments motivate?* Malden, Massachusetts: Blackwell Publishing.

Thorndike, E. L. (1911). *Animal intelligence. experimental studies*. Oxford, England: Macmillan.

Velleman, J. D. (1992). The guise of the good. *Nous, 26*(1), 3-26.

Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, Massachusetts: Belknap Press/Harvard University Press.

Winkielman, P., & Berridge, K. C. (2004). Unconscious emotion. *Current Directions in Psychological Science, 13*, 120-123.

Yaffe, G. (2001). Recent work on addiction and responsible agency. *Philosophy and Public Affairs, 30*(2), 178-221.