COGNITIVE ARCHITECTURE AND THE FUNCTION OF HUMAN COGNITION

by

Nathan Josephe Fox

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

The Faculty of Graduate Studies

(Philosophy)

The University of British Columbia
(Vancouver)

May 2010

ABSTRACT

A number of models of cognitive architecture have been advanced with the intention of providing some sense of the psychological processes that subserve a range of behaviours. For instance, Sober & Wilson (1998), C. Daniel Batson (1988) and Robert Frank (1988 and 1990) attempt to account for contrasting (if not contradictory) behaviours – respectively, hedonistic and altruistic behaviour, self-oriented behaviour and other-oriented behaviour marked by empathetic reactions, and behaviour that reflects rational self-interest in material incentives and behaviour that tends to produce long-term benefits in social interactions. However, the approaches that I have examined encounter difficulties. One difficulty in basing psychological models on empirical data is that the mental states that precede and accompany motivations may be ambiguous or obscure. Those states may be composite states consisting of components that are inextricably linked. For instance, it is not clear whether an altruistic act has some desire for pleasure lurking in the shadows. In Sober & Wilson's approach, cognitive structure is predicted largely on the basis of *general* factors in the natural selection of cognitive devices, e.g., their availability for selection, energetic efficiency, and reliability. However, the *particular* factors that play a role in the aetiology of traits depend upon the function that those traits evolved to perform. For instance, while the *reliability* of a physical system component may certainly be an important general factor in natural selection, it may be a detriment for a device that has as a *particular* biological function the production of phenotypic flexibility. To avoid the problems that I identified in these approaches, I derived a model of cognitive architecture that is intended to predict motivations and actions that are consistent with aspects of evolutionary theory about the function of cognition. The theory upon which I depended is advanced in Peter Godfrey-Smith's book *Complexity and the Function of Mind if Nature*. He proposes that there is a single overarching adaptive function for the mind: to subserve adaptive plasticity. Accordingly, my model suggests a general pattern in the sequencing of human mental states that would tend to maximize behavioural flexibility as a means of maximizing inclusive fitness.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I am grateful to Dr. Christopher Stephens for introducing me to the primary writings upon which my thesis depends and for his many helpful, supportive and constructive comments provided throughout the process of developing and writing it. I thank Dr. Andrew Irvine for his many helpful comments. I am grateful also to my wife Susan Fox for her support and patience.

INTRODUCTION

Those who have advanced models of human cognitive architecture – describing the nature and organization of cognitive devices predicted to produce human motivation and behaviour – have depended variously on *empirical data* of one kind or another or on *theories* of one kind or another. For instance, unsystematic observations of behaviour or reflections on the nature of motivations often underlay the advancement of hedonistic motivational theories (roughly, those theories that assume that the ultimate desire is for pleasure and to avoid pain). Similarly, psychologists such as C. Daniel Batson infer motivational models from empirical data gleaned from controlled psychological experiments. In contrast, Elliott Sober and David Sloan Wilson (1998) propose models of human motivation that largely rely upon theoretical predictions about the characteristics of the cognitive devices that produce those motivations. Their predictions are based on three recognized factors in natural selection: a) the presence or the *availability for selection* of traits within a population, b) the *energetic efficiency* of competing traits, and c) the *reliability* of alternative cognitive mechanisms at producing those outputs for which there is adaptive pressure. The approach that I take in this investigation also depends upon theoretical predictions. However, this approach also emphasizes an aspect of evolutionary theory that Sober & Wilson largely disregard in their analysis, viz., theory that specifies the essential attributes of those motivations and behaviours that the cognitive architecture in question evolved to produce and that describes how it is that essential attributes are plausible. (Distinctions between these different theoretical aspects are considered repeatedly throughout this thesis.)

The approach I take avoids certain problems that may be inherent to other approaches. One difficulty in basing psychological models on empirical data is that the mental states that precede and accompany motivations may be ambiguous or obscure. Those states may be composite states consisting of component states that are inextricably linked. It is all too easy to disagree about whether an altruistic act has some desire for pleasure lurking in the shadows. It is hard to adjudicate whether pleasure usually accompanies liking or that pain is an adjunct to envy. It is difficult to be confident that empathy is a basic mental state: it might be the name given to the composite experience of pain and the desire to help others. Further, model building that mainly relies on

observed behaviour and not on theory requires assumptions about the representativeness of the data – a kind of sampling problem.

A problem inherent in Sober & Wilson's approach in which cognitive structure is predicted largely on the basis of theory related to the natural selection of cognitive devices[1] is that the resulting model does not need to be able to explain behaviours that we may reasonably expect it to explain. Their approach is like trying to infer the inner workings of a particular man-made device in a process that depends largely on general preferred characteristics of man-made devices and not so much on the particular reasons for making the device in the first place. To illustrate (crudely), let us say that aliens abduct a hard-rock guitarist together with his guitar, guitar amplifier and portable power supply. Say, that they want to model the workings of the amplifier but are not permitted to disassemble it. They know only that it is man-made and that the device is capable of producing sounds of varying frequencies. Clearly, they would not do well by attempting to infer the nature of the amplifier's components mainly on the basis of a theory about the qualities that humans are prone to building into their devices. They would not be able to predict too much about the inner workings of the guitar amplifier from the theory that humans are prone to making their devices out of resources that are *available*, and that are *energy efficient* and *reliable* (the factors that Sober & Wilson consider). However, their prospects would improve were the aliens to attempt to model the workings of the amplifier by relying more on a theory of the *essential attributes of that which the device was made to produce* and how it produces it. A good theory for the aliens would be that the device was made for the purpose of producing sound waves that humans are able to detect and that these waves can be produced by cyclically compressing and decompressing air. They may further find that humans are able to detect such cyclical compressions and decompressions at frequencies between 20 and 20,000 cycles per second. On the basis of such theories, the aliens could predict that behind the fabric front of the amplifier there is a membrane that is capable of vibrating within that frequency range and that some other devices included in the box energizes that process.

My goal is to derive a model of human cognitive devices that would be able to produce motivations and actions with the essential attributes specified by aspects of evolutionary theory; and to do so in

[1] In addition to depending on factors in natural selection, Sober & Wilson depend on the assumption that human psychological devices evolved to cause parents to attend to the wellbeing of their offspring. However, cognitively unsophisticated organisms also attend to the wellbeing of their offspring. So, their assumption about human psychological devices seems to disregard theory about the adaptive pressure that produced sophisticated human cognitive devices.

a way that is consistent with how that theory suggests that those attributes are plausible. The theory that I will use to make those specifications is elaborated in Peter Godfrey-Smith's book *Complexity and the Function of Mind if Nature*. He proposes that there is a single overarching adaptive function for the mind: to subserve adaptive plasticity. According to his Environmental Complexity Thesis, "[t]he function of cognition is to enable the agent to deal with environmental complexity" (p. 3). This may be viewed as an ultimate function toward which many particular instrumental functions of the mind are oriented. The Environmental Complexity Thesis suggests that the adaptive pressures that are presumed to have been contributory causes of an evolutionary history of the mind relate to phenotypic plasticity that enhances inclusive fitness. Thus, motivations are predicted to be flexible while at the same time enhancing inclusive fitness. However, increasing flexibility can easily diminish inclusive fitness. This potential problem is one of the central issues in Chapter 2.

In Chapter 1, I will introduce cognitive models advanced by Sober & Wilson, C. D. Batson and Robert Frank. My emphasis will be on presenting counterexamples that suggest weaknesses in the fundamental structure of those models. (I will then examine the three models in greater detail in chapters 2 and 3 for illustrative purposes.) In Chapter 2, I present predictions about human cognitive architecture and human motivation that follow from adaptive plasticity theory, signal detection theory, and other evolutionary theories associated with the Environmental Complexity Thesis. Finally, in Chapter 3, I will propose a cognitive model based on the predictions and inferences presented in Chapter 2. That modelling of cognitive devices presents an alternative to the approaches that are taken by Sober & Wilson, Batson and Frank. It is a model that does not have the weaknesses that I find in those other models.

1 THREE COGNITIVE MODELS THAT PERTAIN TO THE PRODUCTION OF MOTIVATION

Sober & Wilson (1998), C. Daniel Batson (1988) and Robert Frank (1988 and 1990) each advance models of cognitive architecture that are intended to provide some sense of the psychological processes that subserve a range of behaviours. They are particularly interested in accounting for contrasting (if not contradictory) behaviours – respectively, hedonistic and altruistic behaviour, self-oriented behaviour and other-oriented behaviour marked by empathetic reactions, and behaviour that reflects rational self-interest in material incentives and behaviour that tends to produce long-term benefits in social interactions.

1.1 Elliott Sober's and David Wilson's Support for Psychological Altruism

Sober & Wilson (1998) argue in favour of psychological altruism – with a focus on altruistic motivations associated with the welfare of one's children. They argue that evolutionary theory predicts that some psychological desires, particularly desire for the wellbeing of one's offspring, are truly oriented toward the interests of others; i.e., they do not trace back to egoistic desires. Thus they oppose psychological egoism, roughly the view that *all* desires are ultimately oriented toward one's own interests, which prominently include pleasure and the avoidance of pain. Further, they suggest that behaviour intended to enhance the wellbeing of one's offspring is even more reliably produced if their psychological altruism works in concert with psychological egoism in an approach that they call "Type-Two Pluralism": You "[p]erform an action if and only if you believe that it will maximize pleasure and minimize pain *or* that it will do the best job of improving the welfare of your children" (p. 319). Type-Two Pluralism allows that some altruistic behaviour toward your offspring may be subserved by cognitive mechanisms that are self-oriented (for example, via beliefs that helping them will be pleasurable or lead to an avoidance of pain) and some such altruistic behaviour is subserved by altruistic ultimate desires.

As I suggested earlier, Sober & Wilson's motivational model is largely based upon inferences drawn from recognized factors in natural selection:[2] Biological devices that are energy efficient, available and reliable tend to be favoured. They find their principal support for altruistic ultimate desire and Type-Two Pluralism in the third factor, reliability. However, I worry about the validity of basing evolutionary arguments that pertain to motivational systems so squarely on reliability. While the reliability of a physical system component may certainly be an important general factor in natural selection, it may be a detriment for a device that has as a *particular* biological function the production of phenotypic flexibility. Reliability may entail inflexibility – doing X under varying conditions or under any conditions. However, it may be that under some of those conditions, it is better to do Y. For example, there may be some conditions under which the inclusive fitness of an individual is not enhanced by helping its offspring – such as under severe resource scarcity, the illness of the child, or conditions of war. For this reason, even if altruistic ultimate desire devices are inherently more reliable in the sense of producing the same response under varying environmental conditions, they may nonetheless be less beneficial. (I examine generally how inclusive fitness may be taken into account in the choices we make throughout Chapters 2 and 3.)

Let us consider another aspect of their position. They suggest that their model affords a multiplicity of high priority desires (besides the ultimate desire to attend to the wellbeing of one's child): It "leaves open the possibility that people may acquire *other* motives with the net result that they change their *behaviour*" (p. 303). For instance, Sober & Wilson recognize cases in which parents sacrifice their children; and they reference a Daly and Wilson (1988) survey in which infanticide is normal under certain conditions. These competing motives can function alongside the ultimate desire for the wellbeing of one's children: "the most one can conclude is that [individuals who killed their offspring] cared about something else *more*" (p. 302). Further, they report the finding that parents experience regret when they kill their children and "regret implies the existence of stronger desire that trumps a weaker one" (p. 302). However, Sober & Wilson do not seem to specify how an individual makes choices when they are subject to conflicting motivations. Thus, they seem to allow the possibility of conditions under which individuals could care about some people more than they care about their own offspring. If one can care more for others under some conditions, then it would be strange if one could not also care more for one's self than one's offspring under some conditions.

[2] Although, they are not apparently opposed to empirical investigations of the sort that Batson (1988) undertakes. However, they do not find that investigation to this point have been sufficient to decide the matter.

This accommodation for competing desires adds flexibility to their model. However, making that accommodation without providing a mechanism for prioritizing these competing desires leaves them with a model that does not regulate behaviour for the sake of inclusive fitness. They do not seem to identify adaptive pressure for a device that would prioritize competing desires so that individuals will tend to enhance their inclusive fitness. (Some high priority desires might include survival of self, survival of other relatives, and the possession of resources necessary for life.) I will argue in Chapters 2 and 3 that it is plausible that there is a mechanism to ensure that inclusive fitness is taken into account in assigning such priorities, and this mechanism is part of a particular kind of hedonistic psychological system.

Sober & Wilson suggest four distinct reasons that the reliability of an altruistic desire device associated with the wellbeing of one's child ought to be greater than that of purely hedonistic motivations for one's offspring's wellbeing. I will introduce one of those arguments now, the Tricky Engineering Argument.[3]

> Hedonism asserts that whenever the organism believes that its children are well off, it tends to experience pleasure; whenever the organism believes that its children are doing badly, it tends to feel pain. What is needed is not just that *some* pleasure and *some* pain accompany these two beliefs. The amount of pleasure that comes from seeing one's children do well must exceed the amount that comes from eating chocolate ice cream and from having one's temples gently massaged to the sound of murmuring voices. This may require some tricky engineering [and] these causal connections [are] anything but trivial (p. 315).

Sober & Wilson support their tricky engineering argument by appealing to the apparently poor correlation between levels of physical pain and degree of corporeal injury. We can appreciate how this appeal supports their argument. Physical trauma to finger tips can be exceedingly painful while a brain tumour may be completely painless. Perhaps at the time that these traits evolved it was on balance useful to have sensitive finger tips (for manipulating objects) and detrimental to have sensitive brain tissue (which could not in any case be tended to in the event of trauma). Further, it is plausible that there would not have been any adaptive pressure to reduce finger tip pain because

[3] The argument addressed here is examined in some detail in §2.5 and the remaining three arguments are examined in §3.3.3 and §3.5.1.

individuals can carry on even if their finger tips are sore. So, tricky engineering issues may arise under conflicting adaptive pressures. *This* tricky engineering problem involves making finger tips sensitive and brain injuries painless while making it so that pain correlates with the biological significance of injuries.

However, this analogy involving the correlation between *physical* pain and corporeal injury may have been made incorrectly. The pain and pleasure that is relevant to making choices about child care is *psychological* – it does not originate in the stimulation of pain receptors or in nerve damage. It does not involve tissue damage or homeostatic causes and is not like hunger, fatigue or nausea. Thus, the production of psychological pain may not be subject to the same sort of conflicting adaptive pressures. The psychological pleasure produced by the wellbeing of one's child would seem only to need to represent the value of that event or state from the standpoint of inclusive fitness. So the need for *tricky* engineering may simply not arise.

1.2 Daniel Batson's Empathy-Altruism Hypothesis

Daniel Batson and a number of colleagues (1988) conducted a series of studies to gather data that would either support two associated hedonistic motivational models or support an "empathy-altruism hypothesis." The two hedonistic models together predict that prosocial or altruistic behaviour is directed toward either garnering self or social rewards (such as praise, honour and pride) or avoiding self or social punishments (such as censure, guilt and shame). In contrast, the empathy-altruism hypothesis predicts that "prosocial motivation associated with feeling empathy for a person in need is directed toward the ultimate goal of benefiting that person, not toward some subtle form of self-benefit" (p. 52). Thus, according to the empathy-altruism hypothesis and under standard conditions, altruistic behaviour is produced by the interaction of three devices: an *altruistic ultimate desire*, a *belief* that a person is in need, and *feelings of empathy*. The inclusion of empathy in their model follows from a consistent empirical finding that "when empathy is low, helping drops dramatically if escape is easy [and individuals can avoid helping]. When empathy is high, however, helping remains high even if the empathically aroused individuals can easily reduce their arousal by escaping exposure to the suffering victim" (p. 52).

Their studies involve establishing conditions that are intended to arouse empathy in their subjects. The subjects face fellow students, confederates, who appear to be dealing with painful circumstances. In studies 2 and 5, a fellow student has apparently lost both parents and a sibling in a car accident. In studies 1, 3 and 4, subjects listened to confederates express their fear and anxiety about having to endure a series of uncomfortable electric shocks. (I examine Study 4 in some detail in §2.4.3.2.) Subjects are then offered the opportunity to help the confederates under varying conditions.

The conditions that subjects then encounter are intended to cause them to hold the view that they would not receive subsequent self or social *rewards* for helping or plausibly not receive subsequent self or social *punishments* for not helping. For instance, in Studies 3 and 4, the experimental design is intended to cause subjects to believe that they are *justified* in not helping. In Study 3, subjects are told that most other subjects in their position choose not to help the confederate. In Study 4, subjects are lead to believe that it is difficult to meet the qualifications for being permitted to help. Batson assumes that such justifications ought to diminish subjects' anticipated self or social rewards for helping and anticipated self or social punishments for not helping. Even though subjects encounter these conditions, the studies seem to indicate that test subjects who empathize with their confederates are willing to help them. Batson concludes that their tests fail to support the hedonistic hypotheses and succeed in supporting the altruistic hypothesis. I will point out two worrisome aspects in his analysis.

First, Batson assumes that the only mental states that count as hedonistic motivations in their studies are the anticipated self or social rewards or punishments – including praise, honour, pride, censure, guilt and shame. However, it is plausible that the empathy that seems necessary for the altruistic behaviour may play a role in hedonistic motivation. It is plausible that the subjects in Batson's studies were sensitive to the suffering and anxiety of their classmates, just as it is plausible that people generally have sensitivity to the suffering and anxiety of others in their communities. We are prone to paying attention to such matters. Further, it may be that the suffering and anxiety of their classmates was unpleasant for the test subjects – plausibly worrying or distressing. Batson does not consider the possibility that the prosocial behaviour of their test subjects depended upon the unpleasantness that plausibly accompanied their feelings of empathy. Perhaps subjects helped in order to reduce this unpleasantness. Alternatively, those feelings that accompanied the empathy

may have acted as a rational prompt to help: Perhaps others would find out about the helping behaviour and subsequently hold the subjects in higher esteem, in which case, they would have been rewarded. Thus, the unpleasantness associated with the empathy may have led to the prosocial behaviour Batson found and screened off the motivational effects that might have been produced by subsequent shame, pride or censure.

Second, even if the altruistic subjects in Batson's tests were not influenced by psychological pain that accompanied empathy, the data that supports their empathy-altruism hypothesis may not be adequately representative. In particular, their hypothesis does not account for cases in which subjects are empathetic but fail to exercise prosocial behaviour when given the opportunity. To reiterate their assumption on this: "When empathy is high . . . helping remains high even if the empathically aroused individuals can easily reduce their arousal by escaping exposure to the suffering victim" (p. 52). However, it seems as though helping may not remain high if that helping conflicts with higher priority behaviours – like helping others who need that help more or helping oneself even if one does not need the help more. For instance, say that John feels empathy for a neighbour and that prior to being able to help that neighbour John comes to believe that his family would be endangered by doing so. John may feel empathy but decide not to help because of the conflict. It may be counter-argued that learning about the endangerment to his family caused John's empathy to wane. However, it is not clear that this would in fact occur. Even if it did occur, it would still suggest that empathically aroused helping is more opportunistic than would seem from their model.

1.3 Robert Frank's Pluralistic Cognitive Model

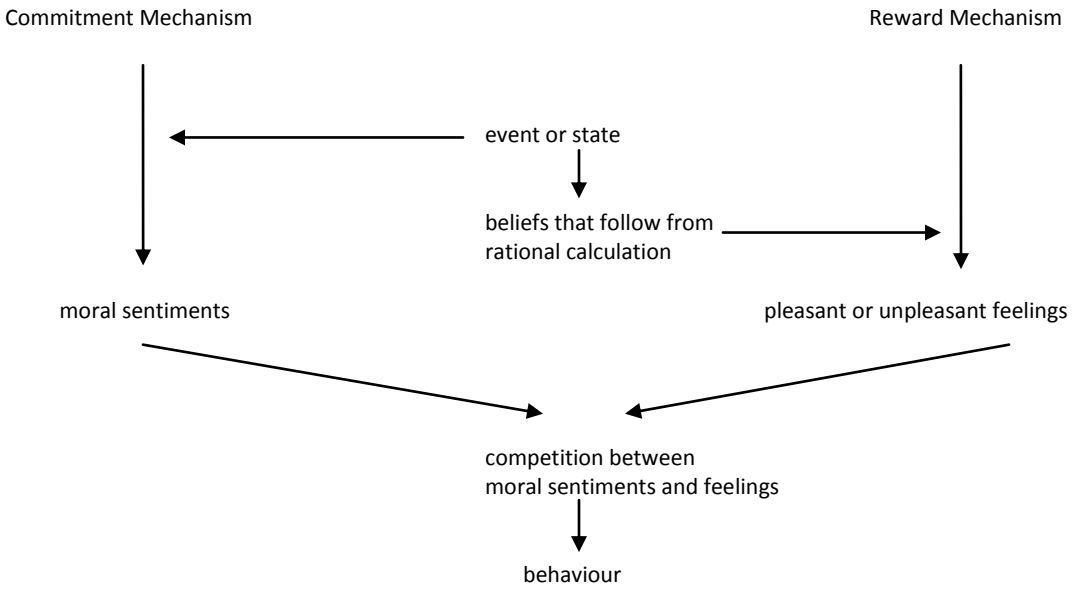
Robert Frank (1988 and 1990) advanced a pluralistic cognitive model to account for human motivation and behaviour. It consists of a "self-interest model" and a "commitment model." Crudely, the former deals with the pursuit for predicted material benefits while the latter deals with social behaviours that are not predicted by agents to produce such benefits but that may nonetheless tend to be beneficial over the long run.

The *self-interest model* includes a *reward mechanism* and *pleasant and unpleasant feelings* that reward and punish agents respectively. The reward mechanism produces the pleasant and unpleasant feelings in response to either physical stimuli (pleasurable or painful) or material incentives. Also, the self-interest model includes *rational calculation* that is employed to predict costs and benefits and anticipate future pleasant or unpleasant feelings. Frank explains:

> Suppose, for example a hungry person calculates that being fat is not in his interests, and for this reason refrains from eating. His rational calculation has clearly played a role, but it is an indirect one. It is still the reward mechanism that directly governs his behaviour [in causing pleasant / unpleasant feelings]. The rational calculation informs the reward mechanism that eating will have adverse consequences. This prospect then triggers unpleasant feelings . . . Rational calculations, understood in this way, are an input into the reward mechanism (1990, p. 75).

According to Frank, the self-interest model does not account for an important range of social behaviours that are accounted for in the *commitment model*. These behaviours are presumably not mediated by the anticipation of pleasure or avoidance of pain. Rather, they are mediated by moral sentiments – including guilt, anger, envy, gratitude, liking to retaliate, disliking an unfair bargain, and feeling bad when cheating. As opposed to producing foreseeable rewards, they may merely tend to produce benefits over the long run. For example, retaliations may contribute to a reputation that deters future aggressions. (This potential benefit may or may not outweigh predicted costs of such aggressions.) Frank illustrates his point using the rather extreme case of the feud between the American McCoy family and the Hatfields. At one point the Hatfields set fire to the McCoy's farmhouse and shoot family members as they escape the blaze. The commitment model applies to such behaviours as abstaining from cheating which may contribute to a reputation for being honest and so invite potentially beneficial mutual cooperation. Similarly, a promise of future support, made credible by the appearance of being in love, invites partners for long-term commitments. Self sacrifice may make agents more attractive to others in their community and permit future beneficial reciprocal transactions. Thus, it may be that altruistic behaviour, moral behaviour, the maintenance of values, acting according to conscience, and exercising self control tend over the long term to be beneficial in a social context. When the moral sentiments that mediate such behaviour conflict with anticipated pleasure or pain produced by the self-interest model, those sentiments compete with and often overcome the pleasure or displeasure produced by the reward mechanism (see Figure 1).

Figure 1 – Robert Frank's Pluralistic Model



Frank's model is pluralistic. However, that does not entail that the reward and commitment mechanisms must be activated simultaneously. The pleasant and unpleasant feelings that are produced by the reward mechanism may arise either in a competition with moral sentiments or in the absence of competing moral sentiments. Moral sentiments may arise either in competition with pleasant or unpleasant feelings or in the absence of pleasant or unpleasant feelings. These possibilities follow from sentiments and feelings having causal antecedents that are independent in the right way.

One initial worry that I have about Frank's model is that it seems as though pleasant and unpleasant feelings in the reward model are associated with the moral sentiments in the commitment model in a way that contradicts the separation between the models that Frank suggests exists. It seems as though the sentiments, envy, anger, pride, shame, guilt, liking and disliking are usually preceded or accompanied by either a pleasant or unpleasant feeling.[4] This connection is more apparent when those sentiments are more intense – when the states in question are rage, love or hate as opposed

[4] My suggestion here may seem to contradict the worry I express about the role of empathy in Batson's model. Empathy may play a role in hedonistic motivation. However, further on I will advance a monistic model in which both self-oriented and other-oriented behavior follow from similar sequences of mental states.

to anger, liking or disliking. This unaccounted connection between pleasant and unpleasant feelings and those sentiments are further supported by Batson's assumptions that pride is a reward and that guilt and shame are both punishments.

A second worry is that Frank's model does not accommodate the *possibility* that some behaviours accounted for in the commitment model (retaliation, abstaining from cheating, making credible promises of future support, or sacrificing oneself) either a) have both moral sentiments and rational calculation as positive causal factors, or b) are the effect primarily of rational calculation. I am suggesting that it seems plausible that an agent *reason* about long-term social benefits of behaviour and even that he or she do so while not under the powerful influence of moral sentiments. It seems plausible that on occasion an individual could reason that potential long-term gains may outweigh short-term incentives. Sometimes people are strategic and relatively unemotional. Perhaps career diplomats, strategists and negotiators are particularly adept in this regard. Thus, Frank's model implies a limitation on the learning of strategic responses that may not exist.

Last, Frank does not provide an account of how sentiments and feelings compete. It seems problematic in that sentiments and feelings are incommensurable. Analogizing to perception, we may judge one flavour to be stronger than another or one taste to be stronger than another. However, comparisons between a taste and a smell are harder. Similarly, it seems difficult to judge say that one's *disliking* an unfair bargain is stronger than the unpleasant feeling associated with losing the material incentive that would follow from the transaction. This competition is the model's only mechanism for balancing motivations produced by the reward mechanism and those produced by the commitment mechanism. So, if in fact sentiments and feelings are incommensurable, then we are left without an account of this important function.

1.4 Summary

I have presented three pluralistic models that describe sequences of mental states that lead to motivation or action. In all three cases, the pluralism is part of a solution to the problem of accounting for human behaviour that sometimes seems to be motivated by self-interest and at other times seems not to be so

motivated. *One* difficulty that all three models encounter involves cases in which individuals run into simultaneous motivations that conflict. They do not provide a means by which inclusive fitness may plausibly be taken into account in the resolution of those conflicts. The Sober & Wilson account does not address the question of how individuals prioritize competing desires. Batson does not deal with cases in which empathically aroused individuals deal with powerfully conflicting motivations that may undermine helping. Frank proposes that internal conflicts are resolved in a competition between seemingly incommensurable mental states. The psychological model that I will develop does not encounter any of these difficulties. Also, I questioned in this chapter some of the reasons that have been offered for rejecting hedonistic accounts of other-oriented behaviour. Later, I will show that these reasons (and others that these same writers propose) are not well grounded.

2 CHARACTERISTICS OF HUMAN COGNITIVE ARCHITECTURE PREDICTED BY ADAPTIVE PLASTICITY THEORY

2.1 Introduction

Models of adaptive plasticity show how an organism can do better employing a multiplicity of behaviours or strategies than by adopting a generalized strategy under a broad range of environmental conditions. Notionally, specialized strategies potentially produce a higher payoff in terms of inclusive fitness. They may be more finely adapted to the conditions in which individuals find themselves.[5] Further, adaptive plasticity theory predicts that organisms are sensitive and responsive to signals that are elements or aspects of their proximal environment and that are correlated with distal environmental conditions. These signals can be used by organisms to choose that (relatively specialized) strategy available to it that tends to produce the best payoff (always in terms of inclusive fitness) relative to its other available strategies and given the likelihood of alternative distal conditions suggested by those signals. From the organism's point of view, the probabilities of these distal conditions depend on the strength of the correlation between signal and distal condition. If there were perfect correlations between them, then organisms would have the luxury of choosing the available strategy that tended to produce the highest payoff for the future world state predicted by a signal and of ignoring alternative world states. However, in the absence of perfect correlations, adaptive plasticity theory predicts that organisms will adopt that strategy that produces the highest "expected payoff" – a theoretical return based on probabilities.

I will illustrate how expected payoff is theoretically determined first for inflexible organisms that do not respond to signals and adopt a single generalized strategy. (The expected payoff of strategies under these circumstances is not a calculation that an organism needs to make because it will always adopt the same strategy regardless of extant conditions. It will nonetheless be helpful to consider this theoretical calculation.) Let W1, W2 and W3 be the only three possible distal world

[5] An important factor in the cost of plasticity is the degree of specialization of the devices that produce that plasticity. If specialized behaviour X is mediated by a sensing device and physiological reaction that have no other function, then their cost may be fully attributed to behaviour X. If on the other hand, several behaviours are subserved by a particular cognitive mechanism, then the cost of that mechanism can be allocated to those various behaviours. Hence, it would be difficult to attribute the energy and resource costs associated with general purpose cognitive devices.

states an organism may encounter at time t. Let Pr(W1), Pr(W2) and Pr(W3) be the probabilities of W1, W2 and W3 arising at t respectively. Because W1, W2 and W3 are the only possible world states at t, the sum of Pr(W1), Pr(W2) and Pr(W3) is equal to one. Let PW1, PW2 and PW3 be the payoffs that this organism will get in adopting the sole strategy it has available to it given W1, W2 or W3 arising at t respectively. The expected payoff associated with the single generalized strategy is equal to PW1 X Pr(W1) + PW2 X Pr(W2) + PW3 X Pr(W3). This organism will encounter adaptive pressure for an alternative generalized strategy that produces a higher expected payoff.

Let us now assume that an organism is able to adopt just one of two mutually exclusive strategies, S1 and S2. Each strategy has a different payoff schedule for each of the possible world states. Let us further assume that organisms are able to pick up an environmental cue that is perfectly correlated with a distal condition for which one of those strategies produces a higher payoff than the other. For example, if an organism is able to pick up a cue that suggests that W2 will arise and that W1 and W3 will not, then it would be able to choose between S1 and S2 on the basis of which produced the highest payoff for W2.

The paradigm organisms that adaptive plasticity theory generally uses have limited cognitive abilities. Godfrey-Smith (1998) uses bryozoans, a marine invertebrate, as an example. Bryozoans choose the strategy that they employ – either a “normal” or “spined” morphology – by detecting a signal in the form of a water-borne chemical. That chemical signal is correlated with what is, for bryozoans, an important distal world condition, viz., the presence of sea-slugs in their vicinity. Theories that use such paradigms may seem as though they are not able to contribute to an understanding of human behaviour. However, I maintain that the principles of adaptive plasticity are so fundamental as to be applicable to cognitively sophisticated animals.

2.2 Engagement and Response

Flexibility is not necessarily better for organisms than inflexibility. Prominent factors that determine the extent to which flexibility is better include a) the degree of environmental variability or heterogeneity that organisms encounter, b) the stability of the correlation between signals and

relevant distal conditions, c) the cost in energy and resources of the structural mechanisms required to produce flexibility, and d) the availability of specialized strategies that produce higher payoffs under particular world conditions than the best available generalized strategy that may be adopted under a multiplicity of conditions. Further to the last point: Let us say that specialized strategy S1 produces an excellent payoff when the world is in state W1 and a poor payoff in W2; and specialized strategy S2 oppositely produces a poor payoff when the world is in state W1 and an excellent payoff in W2. Then *ceteris paribus*, organisms that are able to pick up cues that are highly correlated with W1 and W2 and appropriately employ either S1 or S2 will be better off than inflexible organisms that only adopt a general strategy S3 that delivers only adequate payoffs under both W1 and W2. Therefore, adaptive plasticity theory predicts adaptive pressure on organisms to connect environmental signals with such specialized strategies. Godfrey-Smith refers to these connections as conditionals because they may be stated in the form, 'if the world is in state A, then adopt strategy B'. Such conditionals are seen as automatic responses.[6]

These signal-strategy conditionals could be single iteration and disposable devices if, for example, the strategy that they produce is an irreversible change in morphology, which I believe is the case for bryozoans. However, conditionals that involve repeatable behavioural strategies may be deployed repeatedly – or at least until their use is extinguished, if an organism possesses a mechanism for that. Thus, the adaptive value of a given conditional that is deployed repeatedly is a function of the total of the benefits delivered over the course of the organism's reproductive life. Within these parameters, adaptive plasticity theory does not impose any sort of temporal schedule with respect to the payoffs delivered by strategies under possible distal world conditions. *Ceteris paribus* these payoffs may come all at once, gradually, after a lengthy delay, or in accordance with any other sort of payoff schedule. The theory only predicts that organisms inherit or acquire and then live with those signal-strategy conditionals that will but *tend* to enhance inclusive fitness *over a period*. (This situation may be compared with Frank's contention about a number of social behaviours having a *tendency* to produce benefits *over time* and not necessarily to produce *certain or immediate* benefits, or avoidance of harm, that might be rationally calculated.)

[6] However, highly flexible organisms may encounter conflicting conditionals and must possess a means by which to resolve those conflicts, or to prioritize their interests. This matter is addressed at length in Chapter 3.

Godfrey-Smith further distinguishes between the ability of flexible organisms merely to follow these conditionals from the ability to *learn new* conditionals.[7] He refers to the former as first-order plasticity and the latter as second-order plasticity. New conditionals may be formed by an individual either a) acquiring new signal sensitivity and pairing it with an existing strategy, or b) acquiring a new strategy and pairing it with an extant signal, or c) making a new connection between extant sensitivities and strategies. It would seem that the first and the third of these could be acquired by classical (Pavlovian) conditioning in which there is a repeated association between an unconditioned and a conditioned stimulus. It would seem that the second of these could be learned by operant conditioning in which an organism is rewarded or punished for behaviour until new strategies are acquired. Thus, higher-order plasticity would not appear to require high-order cognitive processes and sophisticated reasoning. (This is examined further in §2.6.) Further, second-order plasticity does not conflict with first-order plasticity. As Sober (1994) points out, if a belief is hard to learn, likely true and of significant consequence to fitness, then it may be favoured by natural selection. Accordingly, some conditionals that involve such beliefs might be predicted to be innate.

There is a great danger that organisms with second-order plasticity would learn (as opposed to inherit) conditionals that diminish inclusive fitness. Therefore, natural selection would tend to favour a learning mechanism for conditionals that is somehow oriented toward inclusive fitness. The selection pressure for new conditionals that enhance inclusive fitness persists for ever better conditionals – that enhance inclusive fitness ever more.[8] Therefore, it seems plausible that there is a mechanism by which a learned conditional may supplant another learned conditional. Relatedly, learned and inherited conditionals that have been activated by signals may include conflicting strategies: One strategy may undermine the effectiveness of the other or simply rule it out as possibility. One cannot (normally) fight and take flight at the same time. Therefore, we may employ a mechanism by which one conditional is selected over another in particular instances. (This is addressed is some detail in §3.1.) The behavioural flexibility of an organism would be enhanced if such a selection procedure could operate not only on learned conditionals but also on innate conditionals. If learned mechanisms tend to enhance inclusive fitness, then adaptive plasticity is enhanced by a mechanism that selects the most advantageous conditional whether it is learned or innate. If so, there may be conditions under which we may adopt learned strategies in preference to innate ones.

[7] 1998: pp. 25 – 26.

[8] I do not suggest that all extant psychological traits and behaviours are the product of natural selection alone. Genetic drift may be a factor in the evolution of some traits and natural selection may be subject to genetic, anatomical and developmental constraints. I bracket these issues in this paper and focus on predictions based on natural selection.

Let me briefly pull together key elements that have been presented in this subsection. Organisms that possess some sort of full-blown second-order plasticity are predicted to be sensitive to environmental signals, to adopt strategies on the basis of a schedule of signal-strategy conditionals, and be limited in their acquisition of new conditionals to those that tend over time to enhance inclusive fitness. In the approaches taken by Sober & Wilson, Batson, and Frank the signalling conditions that initiate cognitive processes and lead to motivation and action are carefully considered and described. They point to those poignant moments of engagement when the wellbeing of one's child is suddenly in question, when a classmate has lost both parents and a sibling in a car accident, when one discovers that one has been cheated or when one's reputation has been sullied. In the model of the processes that lead to behaviour that I develop in Chapter 3, initial stages involve this receptivity to signals and the activation of signal-strategy conditionals.

2.3 Signal Causal Relations

2.3.1 Introduction

Signal detection theory does not specify the nature of the causal relation between a signal and the distal world state to which it is correlated. Thus, it is possible that the chemical that cues bryozoans to adopt an alternative morphology in response to the presence of sea slugs could either be a substance that is a) produced by the sea slugs, b) produced by an undetected marine event that *lures* the sea slugs, or c) in itself a lure for sea slugs. Generalizing: organisms may depend upon signals that are either produced by a) that which produces benefit or harm (Type 1 Signal), b) that which is a common cause of that which produces benefit or harm and the signal (Type 2 Signal), or that which is in itself a prior cause of that which produces benefit or harm (Type 3 Signal) (see Figure 2). Let us look at the nature of the prior causes that constitute Type 3 Signals more carefully.

Let us say that the chemical signal that bryozoans rely on is in fact a lure for harmful sea slugs and that this chemical happens to be present in the vicinity of a colony of bryozoans. The presence of the chemical does not entail that there will in fact be sea slugs there because there may not be sea

slugs close enough to the chemical to be lured to it, or it may happen to be that there is another chemical nearby that repels sea slugs and so neutralizes the attraction of the first chemical. So, just as conditionals need only to tend to enhance inclusive fitness over time, so too these causes need only to tend to produce the relevant distal conditions over time for them to be subject to natural selection. I will refer to them as Probabilistic Causes and propose this more precise formulation:

> PC: An event E is a Probabilistic Cause of an important distal condition DC if the likelihood of DC given E and fixed relevant background conditions BC exceeds the likelihood of DC given not E and BC.

Thus, the fact that a signal is a Probabilistic Cause and that it is picked up by an individual does not entail that the signal will cause the sort of future event with which it is associated. This is notionally no different than counting smoking as a cause of lung cancer even if it does not cause lung cancer in a particular case.

2.3.2 Examples

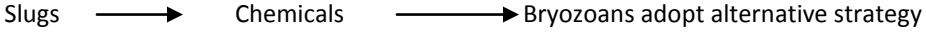
An example of a Type 1 Signal is the distress vocalizations of a child that are caused by the child becoming aware of a nearby predator (assuming that the parent detects the vocalizations but not the presence of the predator). The presence of the predator constitutes the relevant potentially harmful distal condition, and it is a cause of the distress vocalizations of the child that in turn constitutes a signal for the parent. This signal may then trigger a strategy that involves protecting the child.

Prior to the recognition that smoking is a cause of lung cancer, the diagnosis of lung cancer was often a Type 2 Signal. If the diagnosis was of an advanced stage carcinoma, the diagnosis signalled death. Both the diagnosis (the signal) and death (that which produces harm) were caused by smoking (the undetected common cause).

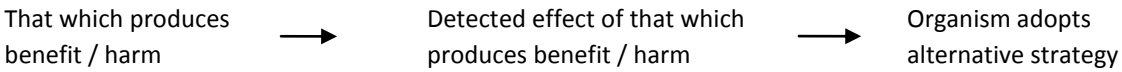
Figure 2 – Possible Causal Relations between Signals and Distal Conditions

Type 1 Signal: The signal is a detected effect of that which produces benefit or harm

Example:

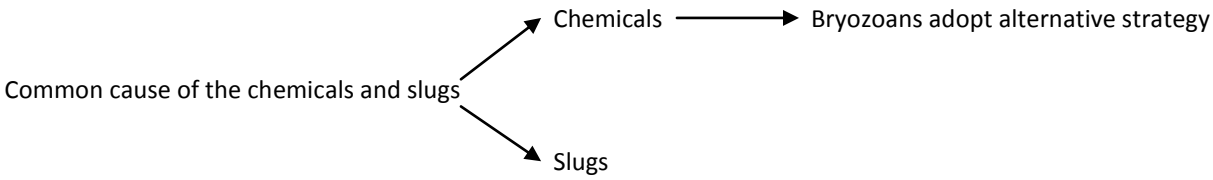


Model:

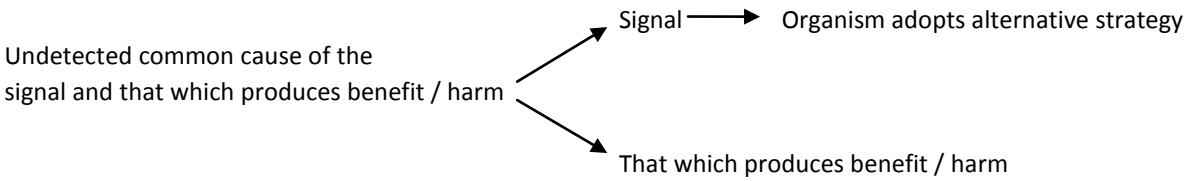


Type 2 Signal: The signal is a detected effect of an undetected common cause of the signal and that which produces the benefit / harm

Example:

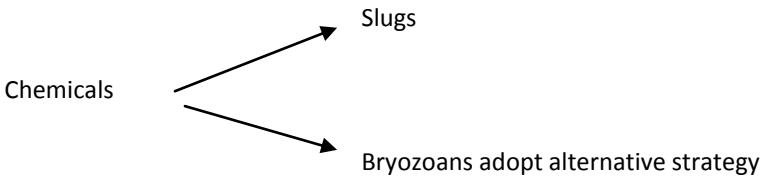


Model:

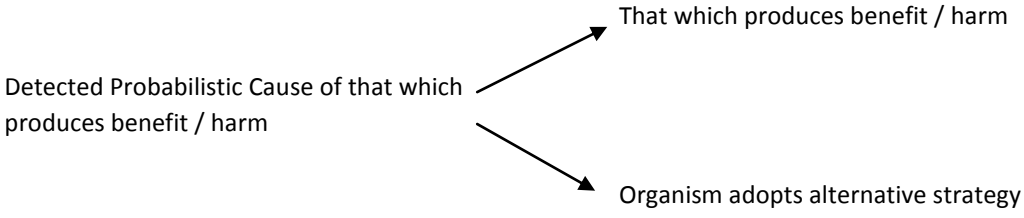


Type 3 Signal: The signal is a detected Probabilistic Cause of that which produces benefit / harm

Example:



Model:



The triggering events in Batson's tests and Frank's examples have the characteristics and functionality of Type 3 Signals, i.e., they are detected Probabilistic Causes of that which produces benefits or harms. These signals include the sorrow and distress communicated to Batson's test subjects, and certain of the perceptions to which Frank alludes, e.g., being harmed, insulted or cheated by others, or of perceiving that one's partners are in danger. These signals lead to motivations (perhaps sometimes reflected in reported empathy) and behaviours – the altruistic actions of Batson's test subjects and the retributive behaviour, refusals to accept unfair bargains, honesty, and self-sacrifice to which Frank alludes. They are detected Probabilistic Causes of important distal conditions in so far as they increase the likelihood of the relevant future states arising – unlike Type 1 and Type 2 Signals. Having perceived the expression of sorrow or distress of confederates, the subjects in Batson's studies are put in the position of being judged and potentially having their reputations altered on the basis of their response. From the point of view of the subjects, the studies are tests of their responses to these signals. Similarly, having been hurt by others or being put in a position of helping others in the cases Frank discusses, puts individuals in the position of being judged and potentially having their reputations altered on the basis of their responses. Regarding "soldiers who threw their bodies atop live grenades . . . [t]he payoff, if there is one, lies in such people being observably different – and more attractive—to others, which puts them in a better position to reap the material benefits of social cooperation" (p. 90).

Last, Type 3 Signals, as causes of important future events, must be inherently important. So long as a state or event has a causal influence on a future benefit or harm, it is rational for an organism to treat that state or event as being inherently important. Organisms may not need even to distinguish potential causes of harms from inherent harms in the process of determining its strategic behaviour. That is, it may be adaptive to adopt strategies that support or undermine those signals (as causes) directly. If the messenger is a cause of some harmful future event, it is rational to shoot the messenger. *Ceteris paribus,* adaptive pressure may have been applied on our forebears to a) support or protect the *causes* of future benefits and undermine or neutralize the *causes* of future harms, b) *maintain* the characteristics of benign causes or to *alter* the characteristics of malignant causes, and c) assume a preparatory stance in relation to the approaching conditions suggested by the detected causes.

2.4 Signal Reliability

2.4.1 Introduction

The concept of signal reliability is used to reference the degree to which an organism may safely depend upon a signal to forecast a world condition. It does so on the basis of the strength of the correlation between that signal and the distal world conditions to which it refers. Godfrey-Smith (1998) demonstrates why organisms need to address signal reliability with an illuminating example involving bryozoans. He constructs the following payoff matrix:

Table 1 – Payoff Matrix for Bryozoans

Strategies	Normal waters (Pr = .9)	Sea slug infestation (Pr = .1)
Normal morphology	10	0
Spined form	8	6

His example demonstrates the benefits of basing strategic choices on reliable signals.[9] Bryozoans determine whether they adopt the normal morphology or their spined form on the basis of a chemical signal. Table 1 indicates the value of signal reliability: If bryozoans adopt the spined form in normal waters, i.e., if sea slugs are not present, then their payoff will be 8 units instead of 10. If on the other hand, bryozoans do not respond to chemical signals for sea slugs, retain their normal morphology, and the signal correctly predicts sea slug infestations, then their payoff will be 0 units instead of 6.

In his analysis, Godfrey-Smith suggests that bryozoans possess a relatively sophisticated reliability measuring mechanism. It seems perhaps more sophisticated than that which humans deploy at times (such as when we are superstitious). It is suggested that bryozoans perceive the concentrations of the chemical signal in the water and that reliability is a function of the variable chemical concentrations. These organisms adopt one strategy when the concentration is below a

[9] Godfrey-Smith assumes that there are no mitigations available for the harm done by sea slugs or for the cost of adopting the spined phenotypic strategy. Perhaps once bryozoans adopt the spined phenotypic strategy, they can't revert back to their default morphology, or perhaps, if bryozoans face a sea slug infestation, that infestation remains for the rest of the bryozoans' reproductive life.

threshold and a different strategy when it is above that threshold. Certain of D. Harvell's (1986) findings are interpreted as suggesting that small colonies of sea slugs employ a higher threshold than do larger ones. This is explained by hypothesized differences between the payoffs that alternative strategies can produce for small colonies and for large colonies. Thus, signal threshold can depend upon world conditions which alter the expected payoffs associated with a strategy. So, bryozoans can deploy a multiple-level signal-reliability measuring device (MLMD) in order to determine the strategy that it will adopt.

Human behaviour seems often to be prompted without the deployment of MLMDs. It often seems that if certain signals are detectable at all, we act on them. (I provide several examples in the coming pages.) My discussion of signal reliability will focus on this subject. In particular, I will present an account consistent with signal detection theory that describes how some human behaviour can plausibly tend to enhance inclusive fitness even though it depends upon a signal merely being detected as opposed to being based upon the readings of a MLMD. (These findings will have implications for the cognitive modelling presented in Chapter 3: We will have the opportunity to present a simplified model in which individuals do not rely on MLMDs.)

2.4.2 Single-Level Signal-Reliability Measuring Devices

Godfrey-Smith's illustration of the importance of signal reliability described above involves signalling conditions that are simpler in some respects than those that humans often encounter. First, increased reliability of the chemical signal for bryozoans indicates increased likelihood of the sea slugs. In this case, signal reliability is positively correlated with the likelihood of the associated distal condition. However, signal reliability theory allows *ceteris paribus* for either positive and negative correlations between signal reliability and the likelihood of associated distal conditions. That is, where a signal in one instance is more reliable than a signal in another instance (assuming that both signals are associated with the same distal event), the signal that is more reliable could indicate either that the distal condition is more likely or that it is less likely – depending upon whether signal reliability is positively or negatively correlated with the likelihood of the associated distal condition. For example, warm air may be a negatively correlated signal of snowy conditions. So, as the

reliability of a warm air signal increases, the likelihood of snowy conditions decreases.[10] Second, for bryozoans chemical signals are correlated with a single world condition, the presence of sea slugs. However, a particular signal may be correlated with a number of distal conditions that are relevant to an organism's strategic choices. Clouds may be a signal for both rain and snow. Such a signal may be positively correlated with some distal conditions and negatively correlated with others. Third, bryozoans *commit fully* to one or the other of its available strategies. However, as Godfrey-Smith points out (p. 215), some organisms may possess strategies to which they are able to commit to varying extents. They might take lesser or greater risk, incur lower or higher costs and put less or more resources into their strategic responses. In such cases, it may be that the expected cost that individuals are willing to incur in the adoption of a strategy depends upon the level of signal reliability.

Let us consider an example to demonstrate the sorts of problems organisms face under such complexity. Say that an organism possesses two alternative strategies, S and T, that conflict so that it may only adopt one or the other. Bryozoans have two alternative strategies. However, unlike bryozoan strategies, strategies S and T need only to be adopted for a short period of time, after which the organism may either continue pursuing that strategy or *switch* to the other strategy. If it does switch, it may switch back to the first strategy following the same short period of time. Let us assume that under some conditions, the expected payoff of adopting S is higher than that of adopting T, while under other conditions, the organism is better off adopting T. Further, let there be four relevant world states (W1, W2, W3 and W4) that are the only four possible ways in which the world could be over the course of an individual's reproductive life. Say that the organism is able to perceive a particular signal that correlates differently with each of the four world states (positively, negatively, weakly or strongly), and the correlation strengths between the signal and the four world states depend upon the reliability of the signal. However, the reliability of the signal varies – just as the chemical concentration of the signal that bryozoans detect may vary over time. Let us say that sometimes the signal it detects is moderately reliable, R1, and at other times highly reliable, R2.

[10] Alternatively, these negative correlations may be viewed as positive correlations between the signal and the likelihood of the denial of a distal condition. Thus, warm air is positively correlated with non-snowy conditions. However, it is sometimes convenient for the analysis of the total expected payoff of strategies to reference negative correlations. This arises in cases where one stipulates a fixed number of possible distal world states so that the sum of the probabilities of the possible world states is one – an example of which is presented in Table 2. Under dynamic conditions of changing signal reliability, the probability of one or more of the possible world states may be increased if signals of higher reliability are detected. However, if such a likelihood were to increase and the sum of the probabilities remains fixed at one, then the probability of one or more of the other world states must decrease (which occurs for W4 in Table 2). In such cases, it is convenient to reference the underlying correlation as being negative.

Table 2 shows payoffs of adopting one the strategies, S, the probabilities of the world states given R1 signal reliability, the *expected* payoff of adopting S with an R1 reliable signal, the probabilities of the world states given R2 signal reliability, and the *expected* payoff of adopting S with an R2 reliable signal. When signal reliability increases, the total expected payoff of *S decreases* from 4.3 to 3.9 units. The decrease arises due to the negative correlation between signal reliability and the likelihood of W4 which decreases from .5 to .3 as signal reliability increases from R1 to R2.

Let us say in contrast that strategy T produces different payoffs under these same four world states, W1 . . . W4. In relation to T, when signal reliability increases, the total expected payoff *increases* from 3.5 with R1 signal reliability to 4.5 with R2 signal reliability. Such circumstances demonstrate the benefits of the ability to distinguish R1 from R2 signal reliabilities. Those benefits would arise when an organism a) adopts S if it detects R1 reliable signals (and expects 4.3 units instead of 3.5 units) and b) adopts T if it detects R2 reliable signals (and expects 4.5 units instead of 3.9 units). Thus, we may find that signal-strategy conditionals have the form, if signal X with reliability R1 is perceived, then adopt strategy S1, and if signal X with reliability R2 is perceived, then adopt strategy S2.[11] It would appear that an organism would require a MLMD in order to employ such conditionals – in order to distinguish R1 from R2 signals. In summary, increasing the reliability of a signal may change the expected payoffs of its available strategies and higher reliability enables organisms to choose the available strategy that will tend in fact to produce a higher payoff.[12]

Table 2 – Expected Payoff Analysis

	Payoff given S	Pr(Wi|R1)	Expected Payoff of S at R1	Pr(Wi|R2)	Expected Payoff of S at R2
W1	2	.0	0	0	0
W2	3	.2	.6	.4	1.2
W3	4	.3	1.2	.3	1.2
W4	5	.5	2.5	.3	1.5
Total		1.0	4.3	1.0	3.9

[11] S1 and S2 may either be two different kinds of strategy or two strategies of the same basic kind but that differ with respect to the risks or resources that they require.

[12] Godfrey-Smith measures signal reliability using this principle. The reliability of a signal S is determined by the probability that an organism having detected S will adopt the available strategy that produces the highest payoff for the organism for the distal conditions that actually arise.

However, organisms may deploy signal reliability detection mechanisms that are somewhat less functional and possibly less elaborate than MLMDs, but that are equally effective under an important range of conditions. Such a mechanism would *only* be able to detect whether or not a signal exceeds a single significant threshold. It would be a single-level signal-reliability measuring device (SLMD). Perhaps sensory perception mechanisms sometimes screen out superfluous data, providing just what is necessary to choose an effective strategy, which under certain conditions may consist of an indication that a threshold condition has been met and nothing more. To illustrate, smoke detectors and fire alarms indicate when air-borne particulates exceed the threshold at which it is wise to adopt the strategy of leaving a building. If our only concern is to know when to evacuate a building, we do not need to know what the air-borne particulate levels are on an ongoing basis, only whether or not that level is above a given threshold. SLMDs may require less information processing or simpler sensing mechanisms than MLMDs. This affords the possibility of them being more reliable and requiring fewer resources and less energy to operate. Thus under certain conditions, natural selection may favour SLMDs. In the following section I will identify distinct kinds of strategies that may be as effectively triggered by SLMDs as they would be by MLMDs and I will discuss conditions under which a set of signals picked up by a SLMD could achieve the functionality of a signal whose reliability was measured with a MLMD.

2.4.3 Conditions under Which SLMDs may be as Effective as MLMDs

2.4.3.1 Only-Choice Strategies

If an organism has only one available strategy associated with a particular signal above a given reliability threshold level (and the resources that are committed to that strategy do not vary with the level of signal reliability) then the organism does not benefit from using processing resources and energy to distinguish between levels of reliability higher than that associated with that threshold level. This applies both to conditions under which the threshold level remains constant or changes over the course of the life of an organism. Regarding the latter, bryozoans appear to take the size of their colony into account in determining a signal threshold level. Even if over the reproductive life of bryozoans their colony size can change, and so the optimal threshold level can change, a SLMD will be as effective as a MLMD provided that the SLMD's threshold level is able to

change in step with changes in its colony size. A SLMD with a variable threshold could allow bryozoans to have a different threshold when they live in small colonies from that which they have when they live in large colonies. I will refer to strategies such as the ones described in this paragraph as *Only-Choices*.

Human physiological reactions such as shivering or sweating may be considered strategies, as may some facial expressions associated with emotions such as surprise, disgust and anger.[13] They are Only-Choice strategies deployed by relatively inflexible systems. Additionally, when signals above a minimal threshold level are very-highly correlated with distal conditions, strategies that follow from those signals are predicted to being paired with a strategy that delivers the highest payoff for that distal condition – thus reducing the need for higher reliability and an MLMD. Such actions seem usually to be instinctive reactions. For example, an individual perceives that the trajectory of some projectile, say a rock, will intersect with the present location of her forehead. *Ceteris paribus*, that signal leads to an instinctive ducking action.

2.4.3.2 Harmless Strategies

The theory I rely upon predicts that SLMDs are as effective as MLMDs when individuals have all of the following attitudes in relation to a signal-strategy conditional: a) significantly higher levels of reliability for the signal are either unavailable or too costly to obtain regardless of the potential benefits that may follow from those signals, b) there are *no* significantly negative expected payoffs for adopting the strategy under any of the plausible alternative world conditions, c) there are positive payoffs under some even remotely possible world conditions, and d) adopting the strategy does not preclude adopting other strategies that may produce positive payoffs under the plausible alternative world conditions. I will refer to the strategies that are a part of such conditionals as being Harmless. [14]

[13] See for instance Schmidt and Cohn (2001).
[14] Decision theorists refer to a strategy that does at least as well in all states and better in at least one state as a dominant strategy.

The strategy suggested by Pascal's Wager under certain assumptions [15] (believe in God because there are potential benefits and no significant potential costs) is Harmless. Most superstitious behaviours reflect Harmless strategies. For example, it may be believed that there are no negative payoffs for avoiding walking on a crack and that there is a chance, however remote, that doing so might lead to the breaking of one's mother's back. Also, the freeze response to relatively unreliable signals of potential danger may be a Harmless strategy. Acting politely and respectfully to people who are not known well is also arguably a Harmless strategy.

MLMDs do not appear to produce adaptive advantages relative to SLMDs in the triggering of Harmless strategies. However, it would not seem necessary for strategies to be absolutely Harmless for there to be no significant adaptive pressure to have them triggered by MLMDs. It would seem that *nearly* Harmless strategies can be adaptively triggered by SLMDs. Potential benefits of greater SLMD efficiency might outweigh the small cost of adopting a nearly Harmless strategy. The small cost associated with the strategy may simply not produce sufficient adaptive pressure for the necessary adaptation. For instance, it may be that the helping behaviours Batson observed in his subjects were nearly Harmless strategies adopted by those subjects in response to the signals delivered to them over the course of their studies. Recall, that when many of Batson's subjects experienced empathy, they were willing to help confederates even though it did not appear as though they would receive subsequent self or social rewards. The costs of helping were surely not significant in all the studies – with the possible exception of Study 4. In that study, subjects viewed the ostensible suffering of a fellow student, Elaine, as she appeared to receive electrical shocks. Many subjects who felt empathy for the confederate Elaine were willing to take her place and receive those shocks. Batson intended that the study design include the condition that "helping is personally costly" (p. 65). However, there are a number of features of the study's design that suggest that it might not have been seen by subjects as especially costly. All the subjects were told that the electric shock was mild and all the subjects hear Elaine confess to having had a "traumatic experience with shock as a child" (p. 67). Batson wanted to "ensure that subjects would consider Elaine's extreme reaction to the shocks atypical and would not expect to find the shocks as unpleasant if they chose to take her place" (p. 67). Thus, it may be that the subjects who were willing to take Elaine's place did not anticipate a significant harm in doing so. Perhaps even in this case, subjects did not need a MLMD that detects varying levels of signal reliability. In virtue of the

[15] We need assume for instance that atheism is correct, that it is not psychologically harmful to believe in God, and there are is no Devil who punishes belief in God.

strategy (taking Elaine's place) being viewed as nearly Harmless by some subjects, once those subjects became aware of the situation, they had already met the signal threshold that was sufficiently reliable to activate the signal-strategy conditional.

2.4.3.3 Best Bet Strategies

Prior to describing what I call Best Bet strategies, I will need to establish a criterion for including a world state into the set of alternative world states that organisms theoretically take into account in order to determine total expected payoffs of strategies (the Relevant World States). That criterion follows from the principle that the payoff of adopting a particular strategy is an effect *both* of the strategy as a contributing cause and environmental conditions as a contributing cause. The payoff is an effect of the two causes at least counterfactually: If either had not occurred, then the payoff would not have occurred. Thus, a world state W is predicted to be a Relevant World State with respect to a strategy S if and only if W includes a unique condition C not included in any other Relevant World State such that the expected payoff of S for W with C is different from W without C. For example, bryozoans need to count a world state with the condition, sea-slug-infestation, as a Relevant World State because that condition is not included in any other Relevant World State and the expected payoff for adopting the spined morphology strategy is different for such a world state with the condition than it is for the world state without the condition. *Mutatis mutandis,* bryozoans need to include a world state with the condition, no-sea-slug-infestation.

I define a Best Bet as a strategy that produces the highest expected total payoff taking all of the Relevant World States into account and given signals with any plausibly achievable and affordable level of reliability. That is, even if the probabilities of the Relevant World States arising is altered by an increase in reliability above the threshold level that is plausibly achievable and affordable, a Best Bet strategy will continue to produce the best total expected payoff. Thus, Harmless strategies are a type of Best Bet. (But not all Best Bet strategies are Harmless, because some Best Bet strategies may be harmful.)

One context under which Best Bets strategies emerge is when levels of reliability higher than threshold are simply not plausibly achievable. This may occur when individuals encounter indeterministic Relevant World States for which increased levels of signal reliability are generally

unavailable. For instance, cigarette smokers receive signals that quitting smoking may prolong their lives. However, quitting smoking may in fact not prolong their lives and highly accurate signals indicating whether it would or would not are not available. Thus, smokers may have the attitude that quitting is a Best Bet in the absence of signals indicating that they would not contract cancer. Similarly, signals that are sufficiently reliable to indicate whether buying home insurance in a particular year or buying lottery tickets will actually pay off are not generally available. So, many people have the attitude that these strategies are a Best Bet under the indeterministic conditions they encounter.[16] If a signal, S, is detectable at all, the strategy that produces the highest expected payoff at the level of reliability that S possesses may be the Best Bet – even if that strategy has a cost.

The strategies that Frank refers to in his Commitment Model seem to be Best Bets that are adopted under indeterministic conditions. I refer to strategies such as costly retaliation, restraint in relation to cheating, and self sacrifice. As Frank points out, the Commitment Model accounts for behaviours that arise when individuals are not able to predict associated material rewards or punishments. That is, these individuals do not receive highly reliable signals, ones that are highly correlated with future states or events. Further, the benefits of the strategies to which Frank refers may tend to accumulate over the long term. Reliable signals for such a string of events are all-but impossible to achieve. Thus, costly retaliation, restraint in relation to cheating, and self sacrifice may produce the highest total expected payoff given the reliability of the signals that are available.

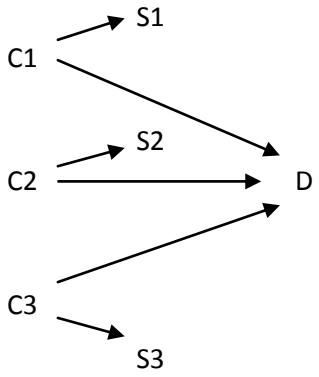
2.4.3.4 Multiple Levels of Reliability are measured by Multiple Signals that all Cue a Particular Strategy

Up to this point, we have focused on cases in which a particular signal is associated with a particular distal world state. However, there would appear to be conditions under which adaptive pressure would arise for organisms to be receptive to more than one signal for a given distal world state. The adaptive pressure for additional, alternative signalling may arise if there were some propensity for signal transmission and apprehension problems. Signals may arise in noisy environments and be

[16] At a given time, individuals may have a number of Best Bet strategies that they have adopted. These may be complementary in so far as adopting any one of them does not preclude adopting any of the others. One may buy home insurance and try to escape a flood early. Under such circumstances, when an appropriate signal is detected at all, the strategy may be employed regardless of whether or not any other complementary strategy has been adopted.

obscured or weak; and individuals may miss signals because they are not sufficiently attentive to some part of their environment. Thus, organisms may possess signal-strategy conditionals in which alternative signals are paired with the same strategy. Further, such alternative signals may be picked up simultaneously. If so, each additional simultaneous signal that points to the same distal world condition may increase signal reliability. To demonstrate such a possibility, let S1, S2 and S3 be three distinct (Type 2) signals for distal condition D (see Figure 3). Let those three signals have independent causes C1, C2 and C3 respectively. Assume that S1, S2, S3 and D have no other causes besides C1, C2 and C3 respectively; and that each cause C1, C2 and C3 is sufficient for S1, S2 and S3 respectively. Let us assume that the three causes are individually neither necessary nor sufficient for D – so that all the causes may each be attributed a probability of causing D regardless of whether any other cause is present or absent. Thus, relations between D and the three signals follow: $0 < Pr(D|S1) < 1$; $0 < Pr(D|S2) < 1$; $0 < Pr(D|S3) < 1$. Under these assumptions, C1, C2 and C3 are all independent contributing causal factors of D and the only causal factors of D and they may arise in any permutation or combination.

Figure 3 – The Causes of S1, S2, S3 and D



Let us consider the probability of D from the perspective of an individual who has simultaneously detected all three signals, S1, S2 and S3. This concerns the probability of D given the conjunction of S1, S2 and S3. With the set of assumptions described here, we may conclude that the total probability of D will equal the sum of the probabilities of D given each signal thus: $Pr(D|S1 \wedge S2 \wedge S3) = Pr(D|S1) + Pr(D|S2) + Pr(D|S3)$. The total probability will not add up to 1 unless S1, S2 and S3 are in combination sufficient for D. If any two of the three signals is detected but not sufficient for D, the *probability* of D would be subsequently increased by the detection of the third signal if that third

signal is positively-correlated with the likelihood of D; and the *reliability* of all the *perceived* signals *taken as a group* (which may consist either of one, two or three perceived signals) is increased with each additional, positively-correlated or negatively-correlated, independent signal of D.

Under such conditions, an organism may be *more* confident that D will either occur or not occur given a group of independent signals than with a single signal. Accordingly, either a SLMD that was able to measure such a multiplicity of such signals or a battery of specialized SLMDs could achieve the functionality of a MLMD—with the number of levels of reliability equal to the number of perceivable signals that are all correlated with the same distal condition. Thus, an individual that depends upon such a battery of SLMDs could possess a signal-strategy conditional in the following form: If signal 1 and signal 2 and signal 3, then strategy S. For example, say that there were a number of distinct, causally independent, physical features of members of the opposite sex that were each positively or negatively correlated with the ability to produce healthy offspring (such as clear complexion, wide hips, or high degree of symmetry). Then, these physical features, each of which is detected with a SLMD, could be taken as a group of signals with a reliability that is a function of the sum of the probabilities of being able to produce healthy offspring given each feature taken separately. For example, the presence of three such features may indicate greater reliability than the presence of two such features.

2.4.5 Conclusion

The domain in which organisms are predicted to need MLMDs may be much smaller than one might imagine. They are only predicted for strategies that are not Only-Choices, not Harmless, and not Best Bets. Nor would MLMDs be predicted for signal-strategy conditionals if organisms are receptive to an appropriate number of independent signals that all correlate with the particular relevant distal condition. Thus it would seem that flexible organisms may very often adopt strategies based merely on the detection of signals (above a threshold that correlates with a given level of reliability), without further assessment of signal reliability (except in so far as they might take into account a multiplicity of such signals), and without compromising inclusive fitness. Nonetheless, it may well be that organisms possess both SLMDs and MLMDs.

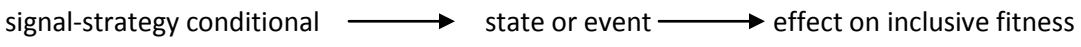
2.5 Grading the Effect of States and Events on Inclusive Fitness

2.5.1 Introduction

It does not appear as though first-order plasticity (see §2.2) requires that individuals possess an "onboard" mechanism to grade the effects of various kinds of states or events on inclusive fitness. Innate signal-strategy conditionals evolve largely by natural selection and individuals automatically adopt that strategy that is paired with the signals it perceives. Finally, the *effect* on inclusive fitness of the states or events that follow from those strategies may be reflected in the further natural selection of signals, strategies or signal-strategy conditionals.

The situation is entirely different for organisms that possess second-order plasticity, that are able to learn new signals, strategies or signal-strategy conditionals. To insure that such potential acquisitions enhance inclusive fitness, organisms must be able to grade the effect that the states or events that follow from the adoption of such acquisitions have on inclusive fitness (see a representation of the flow of causal influence in Figure 4). Without that ability, organisms may adopt signals or strategies that diminish inclusive fitness. Such happenings tend to go against the evolution of second-order plasticity by natural selection. Thus, we may predict that adaptive pressure for second-order plasticity is associated with adaptive pressure for the ability to grade the (actual or potential) payoffs of the strategies organisms potentially acquire (or at least the ability to rank those payoffs) so that learned signal-strategy conditionals may replace others in a process that tends to enhance inclusive fitness.

Figure 4 – Grading the Effect of States or Events on Inclusive Fitness



A behaviour that follows from the activation of a signal-strategy conditional may produce a stream of benefits or harms over time associated with the states or events that follow from a conditional being activated. Thus, the relevant total effect of particular conditionals on inclusive fitness is the

sum of these benefits or harms over organisms' reproductive life. For conditionals that are used repeatedly and not discarded after a single use, the relevant effect is the sum of such sums that are produced with their repeated use. Thus, states or events that produce relatively immediate enhancements and diminishments of inclusive fitness (e.g., having a child or becoming infertile) are not the only relevant effects. Downstream effects (e.g., having grandchildren) are also relevant.

I have suggested that a purpose of grading the effect of states and events that follow from behaviour is to create a basis for acquiring new signal-strategy conditionals. It follows from this purpose that adaptive pressure would be brought to bear for organisms not to wait for the realization of all the benefits or harms that may follow from the adoption of a strategy – which could take the rest of an organism's life. The sooner they assess the utility of alternative strategies, the sooner they will be able to either confirm that conditionals may be reused or not. So in addition to the ability to grade the *realizations* of benefits or harms, it would be useful for organisms to be able to grade states and events on the basis of whether they causally influence possible subsequent benefits or harms. They may thus grade states and events that are *not inherently* beneficial or harmful. For example, friendly interaction with a *potential* mate or incurring a significant injury that *may lead* to infertility may be graded for their respective causal influence on subsequent states or events that are inherently beneficial or harmful.

Last, when states or events are relevant in part because they involve certain conspecifics, the effects on inclusive fitness of those states or events depends upon individuals' relationships with those conspecifics. Say that an individual becomes aware of the suffering of others or that the social status of others has been raised. In such cases, the relationship that the individual has with those others is a critical factor in determining the effect that those states or events have on the individual's inclusive fitness. For instance, c*eteris paribus* the effect on inclusive fitness of one's offspring being healthy and happy is positive, whereas *ceteris paribus* the effect on inclusive fitness of an enemy with whom an individual does not have a close genetic relation being healthy and happy is negative. Therefore, organisms that possess second-order plasticity are predicted to be sensitive to those significant features (whatever they might be) of conspecifics that would alter the effect of states and events (that are relevant because they involve those conspecifics) on their inclusive fitness.

2.5.2 Pleasure and Pain

Consider the possibility that the process by which humans grade the effect of states and events on their inclusive fitness involves pleasure and pain.[17] These affects[18] possess between them certain attributes that are consistent (if not necessary) for such a grading mechanism. They have an acknowledged causal role in motivation and action, vary in intensity, have between them a positive and a negative valence, are fundamental to classical and operant conditioning, and arise in a wide range of motivation and action events. However, we need to consider whether pleasure and pain arise at the times and intensities that would be consistent with a mechanism whose purpose was to grade the effect of states and events on inclusive fitness at the time that such a mechanism evolved. We need to consider whether the *causes* of pleasure and pain are the kinds of states or events that would have tended to correlate with an enhancement or diminishment of inclusive fitness at the time that these traits evolved.[19]

It would seem that Sober & Wilson hold the view that pleasure and pain do not provide a good mechanism by which to grade the effects of states or events on inclusive fitness. One of their arguments against psychological hedonism is that it depends on individuals anticipating an *appropriate* level of pleasure or pain in relation to future events involving the wellbeing of their offspring (see §1.1). They suggest that the anticipated pleasure of "eating chocolate ice cream and from having one's temples gently massaged to the sound of murmuring voices" (p. 315) may exceed the anticipated pain of not helping one's child. Let us consider the *kind* of pleasure and pain that Sober & Wilson reference in this context more carefully.

---

[17] The evolutionary theories that I have depended upon do not require that individuals be aware of the nature or purpose of this grading process or perhaps even aware of the process at all. Although, it is necessary that valuations of states or events be based, even by wholly unconscious processes, on such a grading process. However, if that grading process were to include conscious reasoning, then we would obviously have some awareness of the process but not necessarily know its true purpose.

[18] The admittedly broad and controversial psychological category "affect" will be useful in this paper. Affects are traditionally distinguished in psychology from behavior and cognition. They include emotions, moods, suffering of any kind (e.g., that caused by tissue damage or mental distress) and pleasure of any kind (e.g., feelings of achievement or the pleasure of chocolate).

[19] I don't know exactly when they evolved. Nonetheless, it seems worthwhile to consider the plausibility of some very basic traits being useful to our distant forebears (who were not even necessarily *Homo sapiens*). We might keep in mind that the enhancement or diminishment of inclusive fitness by states or events depend on chronic environmental conditions, such a chronic food shortages, persistent weather patterns, chronic predators or infectious disease and other competitive factors.

The pleasures of ice cream, gentle massage and the sound of murmuring voices are caused by physical stimulation. External stimulation of pleasure may be chemical (usually smells and tastes), thermal (cold is pleasurable for hyperthermic subjects and warmth is pleasurable for hypothermic subjects), or mechanical in nature (e.g., pleasure of massage).[20] Clearly pleasure produced by such external stimuli is not always plausibly correlated to states or events that would have tended over time at the time that the relevant human responsiveness evolved to be associated with an enhancement or diminishment of inclusive fitness. For example, ingesting sugars and fats *in excess* may not have tended to enhance fitness and the pleasure of massage may exceed its value in terms of inclusive fitness. So, I agree with Sober & Wilson that the intensities of such pleasures (pleasurable tastes and massage) may not be well correlated with the effect of the associated events on inclusive fitness.

Sober & Wilson also argue that pain mechanisms do not seem to correlate with effects on inclusive fitness. They support their view by pointing out that physical pain seems to be a poor indicator of corporeal injury. They explain that "tricky engineering" (see §1.1) would be needed in order for pain to be a good indicator. Recall that tricky engineering issues arise under conflicting adaptive pressures. For example, it may have been useful to have sensitive fingertips for fine manipulation of objects and, as a result, physical trauma to finger tips can be exceedingly painful even though that event may *not* be well correlated with a diminishment of inclusive fitness. Thus, adaptive pressure for sensitive fingertips conflicts with adaptive pressure for pain to grade for inclusive fitness. The physical pain that Sober & Wilson consider in this context seems to fall under the definition of pain suggested by *The International Association for the Study of Pain*: "An unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage." There are several other recognized forms of physical suffering besides the pain associated with tissue damage, e.g., the pain associated with feeling cold, overheated, hungry, dizzy or nauseous. The onset and intensity of these other forms of physical suffering seem also to be poorly correlated with inclusive fitness. For instance, the intensity of the pain associated with hunger and cold diminishes over time while its significance to inclusive fitness increases. Neither starvation (provided one remains stationary) nor hypothermia (after the point at which uncontrolled shivering ceases) is painful *per se*.[21]

[20] See for instance, Michel Cabanac (1979).

[21] Perhaps, these responses to starvation and hypothermia follow from adaptive pressure to conserve energy.

In addition to these points that Sober & Wilson make about physical pleasure and pain, note that there are many kinds of states and events that would seem to impact inclusive fitness and yet fail to cause either pleasure produced by external stimuli or the kinds of physical suffering described above: the wellbeing of a child, the failure to secure resources necessary for life, and rejection by a potential mating partner. Thus, these kinds of pleasure and pain are not associated with as wide a range of motivation and action events as one might expect from general measures of inclusive fitness.

Thus, physical suffering and the pleasure produced by external stimuli alone do not appear to be the basis for a workable grading or ranking mechanism of the effect of states and events on inclusive fitness. However, we are not yet in a position in which we may reasonably conclude that it is not plausible that we possess such a mechanism. As I suggested earlier, it is not clear how second-order plasticity could tend to enhance inclusive fitness without such a mechanism. I will suggest in the following section that we *plausibly* possess a species of pleasure and pain that is more promising and that does not originate in tissue damage or physical stimulation.

2.5.3 Psychological Pleasure and Mental Distress

Let us consider pleasure and pain that does not depend upon external physical stimuli or involve tissue damage. Consider purely psychological reactions to perceived, assumed or imagined states and events. I will refer to these broadly as psychological pleasure and psychological pain. I include here states that are variously described as *feeling* good or bad, satisfied or dissatisfied, happy or sad, gratified or disappointed, rewarded or punished, and feelings of accomplishment or failure. *Psychological* pain is sometimes referred to as mental distress, mental pain, anguish, or in medical terminology, psychalgia. Is it plausible that this collection of states involving psychological pleasure and mental distress are the basis (or partial basis) of a (possibly imperfect) mechanism for grading the effect of states and events on inclusive fitness? Is it plausible that the formative adaptive pressure that led to the evolution of these various psychological affects relates significantly (or even solely) to grading and ranking the effect that states or events have on inclusive fitness?

Psychological pleasure and pain share the attributes that made physical pleasure and pain initial candidates for measures of inclusive fitness: They have an acknowledged causal role in motivation and action, serve as

reward and punishment in classical and operant conditioning, vary in intensity, have between them a positive and a negative valence, and arise in a wide range of motivation and action events. Unlike physical pleasure or pain, the tricky engineering problem to which Sober & Wilson referred also does not seem to apply to them (or if it does, it may not do so to the same extent), i.e., competing adaptive pressures may not have been as great. Last, the notion that such psychological affects are associated with adaptive behaviour has a certain consistency with the assumptions that underlie Frank's suggestions about the pleasures and pains that reward and punish us (see §1.3). Recall that the reward mechanism produces pleasant and unpleasant feelings in response to material incentives. Frank holds that this mechanism is a product of natural selection, although imperfect: "There may be sharp limits on the extent to which nature can fine-tune the reward mechanism" (1988: p. 90). I will refer to this hypothesis that psychological pleasure and pain are the basis upon which humans grade the effect of states and events on inclusive fitness as the Psychological Affect Grading Hypothesis (PAGH). Let us attempt to pump our intuitions about the appeal or lack of appeal of PAGH by considering the sorts of predictions PAGH (together with other evolutionary theories) makes about the conditions under which feelings of satisfaction and mental distress ought to arise and the relative intensities of those affects. [22]

I argued earlier (§2.5.1) that the theory upon which I have depended here predicts that we would grade states and events that are both *inherently* beneficial or harmful *and* states and events that merely have a *tendency to cause* events that are beneficial or harmful. Therefore, psychological affects ought to arise in response to either of these conditions or to signs of these conditions (e.g., the distended abdomen associated with one's pregnancy).[23] Thus, feelings of satisfaction ought to accompany a) the event of the birth of one's offspring, b) the causal antecedents of that event, and c) the signs of either the event or its causal antecedents. In addition, the intensity of these psychological affects will be proportionate to the extent to which the relevant states either enhance or diminish inclusive fitness or proportionate to the anticipated likelihood that an event will produce benefits or harms.

If PAGH is true then the closeness of our genetic relatedness to others ought to be an important factor in the intensity of the psychological pleasure or mental distress that is triggered by the

[22] I will assume in what follows that the significance that we normally attribute to our own reputation, standing in the community and opportunities for social cooperation are all evolved adaptations.

[23] Recall that Type 1 Signals are detected effects of benefit or harm (see §2.3.1) and that Type 3 Signals are probabilistic causes of benefits or harm. Thus, we may predict that we grade Type 1 and Type 3 Signals and that they trigger psychological pleasure or displeasure.

wellbeing or ill-being of those others. Offspring and siblings are closer than cousins and PAGH predicts affective responses to the propagation of genetic structure to the extent that it matches one's own genetic structure. That which would increase the likelihood that one's genes are passed along will be pleasant while that which decreases the likelihood will be distressing. Hence, individuals ought to be sensitive to their attractiveness to those they would have as a mate and that which individuals predict will *cause* them to be more or less attractive is predicted to produce good or bad feelings respectively.

PAGH predicts that the perceived loss and subsequent recovery of resources that support inclusive fitness will cause disappointment and relief respectively. The intensity of these affects ought to vary according to the assumed contribution of those resources to inclusive fitness. Thus, the intensity of distress associated with the loss of a mountain of financial resources which does not contribute to inclusive fitness ought to be less than the distress associated with the loss of a pittance that could potentially enhance inclusive fitness.

PAGH also suggests that one is apt to feeling satisfied or dissatisfied about one's health. Thus, the level of *psychological* pain that ought to be triggered by *physical* suffering (stemming from say tissue damage or hunger) should not depend upon the intensity of that suffering *per se*, but rather the anticipated diminishment to inclusive fitness of which the suffering is a signal. For example, the great pain that normally accompanies childbirth ought not to cause mental distress provided that the health of the mother and child are not threatened. If the pleasure that is caused by external stimuli (such as chocolate cake in one's mouth) is *not* taken to be a sign that one's inclusive fitness has been enhanced, then that physical experience ought not to produce a rewarding psychological experience. However, eating the cake could be deeply satisfying to someone who believed that eating it could save his life. Thus, PAGH predicts that under normal circumstances Sober & Wilson need not worry that under hedonism parents could forsake their children for a piece of cake.

If PAGH were to be true, individuals ought to feel satisfaction or dissatisfaction in relation to the status of their associations with those they may cooperate with – especially given the myriad ways in which practices involving social cooperation are able to enhance inclusive fitness. Let us define a *Partner* as an individual with whom one freely cooperates in the pursuit of one's own interests (irrespective of the Partner's interests) with respect to either a particular transaction or a set of transactions, or with respect to an ongoing range of transactions. Let us also define a *Partnership* as

an association between two or more mutual Partners. Thus, Partnerships exist between friends, spouses, strangers involved in a particular reciprocal transaction, members of a group or society that promote social action, members of a political party, employers and employees, shopkeepers and their customers, and parents and their children. So, PAGH predicts affective responses to a) finding potential Partners, b) establishing Partnerships, and c) dissolving Partnerships that continue to retain some potential to enhance inclusive fitness. Such good or bad feelings are predicted to arise both when individuals perceive that those events have taken place and when they perceive a sign that they may have taken place. Thus, PAGH predicts that satisfaction follows from signs that a potentially beneficial Partnership has been formed and dissatisfaction follows from signs that such a possibility has become unlikely. For example, some selfish or insincere acts (that are inconsistent with Partnership) by those who would potentially make a good Partner ought to be unpleasant; while generous acts and an open and honest demeanour (that are consistent with Partnership) by those who would potentially be a good Partner may be pleasant.

Some individuals may possess qualities that allow them to contribute more to Partnerships than others. Partnerships with these individuals are predicted to be more desirable than Partnerships with others. Therefore, PAGH predicts that satisfaction or distress ought to be triggered by the belief that one's *desirability* as a potential Partner has been increased or decreased respectively. We ought to be sensitive to signals that we are liked, admired, or respected; and distressed about being viewed as untrustworthy or unproductive (or whatever quality is viewed negatively in particular Partnerships).

2.5.4 Conclusion

The plausibility of the evolution of second-order plasticity by natural selection depends on organisms being able to grade the effect that states and events have on their inclusive fitness. Sober & Wilson have persuasively argued against *physical* pleasure and pain being well correlated with the effects of states and events on inclusive fitness. However, the set of *psychological* pleasure and pain states possess several characteristics that are consistent with their functioning as a measure of the utility of states and events; their evolution may not have been subject to conflicting adaptive pressures (the tricky engineering problem) to the same degree as that of physical pleasure and pain states; and the PAGH seems intuitively to predict reasonably well the conditions under which feelings of satisfaction and mental distress ought to arise and

the relative intensities of those affects.[24] Being based on a presumed product of natural selection, PAGH allows the possibility that grading based on psychological pleasure and pain states could lead us astray from time to time or that they might produce poor rankings of the importance of events occasionally. It even allows patterns in which certain types of events are consistently graded inappropriately. I do however suggest that, at the least, it is *not* implausible that we possess a mechanism for grading effects on inclusive fitness. For all these reasons, I will adopt as a *working hypothesis,* for the purpose of the present investigation, the position that this *psychological* pleasure and pain has this grading function.

2.6 A Brief Remark on Reason and the Acquisition of New Signal-Strategy Conditionals

We found that organisms that possess second-order plasticity could acquire new signal-strategy conditionals by classical or operant conditioning (see §2.2). (Perhaps new conditionals can be acquired by means that require even less cognitively.) Thus, human second-order plasticity only requires that level of cognitive sophistication that is necessary for such conditioning. Nonetheless, it would seem that natural selection would favour alternative, cost-effective mechanisms that facilitate the development of new, effective and efficient conditionals. So it may favour an ability to reason that could have this function.

Some of the ways in which reason could facilitate the development of new conditionals are evident. It would be helpful to be able to predict a) correlations between extant signals and their associated distal conditions, b) new signals, c) the effects of states and events on inclusive fitness, d) new strategies, and e) new conditionals from combinations of extant signals and strategies. (An additional function of reason is discussed in §3.2 involving the prediction of distal states or events that may follow from the adoption of particular strategies.)

However, adaptive plasticity theory predicts adaptive pressure for new conditionals to be constrained so that those new conditionals enhance inclusive fitness (see §2.2). We also found that such a constraint would require that the effect of states and events on inclusive fitness be graded

[24] I am not suggesting in this that the physical affects that I have described necessarily have wholly distinct etiological and functional characteristics from the psychological affects.

(see §2.5). Thus, if reason is involved in the acquisition of new conditionals, then we may predict that it may not do so in a way that contradicted this grading of effects. If my working hypothesis about the role that psychological pleasure and pain play in this process is true, then reason may not be able to form a signal-strategy conditional that contradicted good or bad feelings that are produced in relation to potential new conditionals.

2.7 The Expression of Innate Signal-Strategy Conditionals

The model of behaviour that arises out of adaptive plasticity theory suggests that we are from time to time under the influence of signal-strategy conditionals. So it is possible that we become aware of some sort of expression of those signal-strategy conditionals. Let us consider this possibility with respect to the *innate* conditionals that go along with first-order plasticity.

Their being innate suggests that the expression of these conditionals may be universal or may be recognizable even with cultural influences impacting their expression. We may be able to infer what that expression looks like. It happens that the rather obvious inferences that I make about innate signal-strategy conditionals correspond to the characteristics that Paul Ekman (1999) uses to distinguish "basic emotions" from each other.[25] For example, the existence of innate conditionals would suggest near universal receptivity to the signals that form part of these conditionals. This corresponds to Ekman's criterion for basic emotions that he calls "universal antecedent events" (p. 53). One may predict (see §2.2) an automatic engagement with the signals that are part of conditionals. This corresponds with the "automatic appraisal mechanism" that Ekman identifies in basic emotions. He describes a "mechanism which selectively attends to . . . stimuli . . . which are the occasion for . . . one or another emotion" (p. 51). If there are innate conditionals, one might reasonably expect universal strategies (that might have distinctive cultural expression) that follow from the universal signals. This corresponds with Ekman's suggestion that emotional reaction communicates important messages to others. They produce what he calls "distinctive universal signals" (p. 47). He takes it to be "central to the evolution of emotions that [these communications] inform conspecifics, without choice or consideration, about what is occurring" (p. 47). The adoption

[25] Not all of these characteristics are exhibited by each of Ekman's basic emotions.

of universal strategies is suggested in another of Ekman's characteristics of basic emotions. These affects induce "emotion-specific physiology" because "there should also be physiological changes preparing the organism to respond differently in different emotional states" (p. 48). The adoption of universal strategies is also suggested in a characteristic of basic emotions Ekman proposes that involves the production of distinctive thought. His view is that emotion is associated with "biological constraints put on the cognitive system" (p. 55). Further, being innate, one would expect to find a pattern in the developmental appearance of conditionals and possibly their presence in other primates. These are also characteristics that Ekman uses to identify basic emotions. Finally, if the activation of conditionals is independent of other behaviour control devices (except each other – a matter examined at length in Chapter 3), then the activation should be automatic. This corresponds to Ekman's "unbidden occurrence" criteria of basic emotion.[26]

Thus, it may be that such basic emotions may be viewed as being, in part, expressions of innate conditionals. If so, then Ekman's empirical findings may support the hypothesis that humans are influenced by innate signal-strategy conditionals. Further, this relation between emotion and innate signal-strategy conditionals seems to support the motivational functionality that Frank suggests is provided by emotions: Emotion can lead to adaptive behaviour – even if that behaviour seems irrational and costly over the near term. Emotion may produce an innate responsiveness to conditions. If those emotions are activated, standard strategies may be adopted. However, Frank qualifies the influence of an *innate* reactiveness: Moral sentiments

> are almost surely not inherited in any very specific form. Definitions of honesty, notions of fairness, even the conditions that trigger anger, all differ widely from culture to culture. If people inherit anything at all, it is a receptiveness to training about the attitudes that are likely to serve them in life (p. 93).

This view may conflict with Ekman's concept of basic emotions, having such characteristics as "universal antecedent events." Perhaps, we can allow the possibility that the cultural influences that Frank observes relate less to the initial activation of an emotion than they do to subsequent

[26] I have referenced here nine of Ekman's 11 characteristics. I left out "quick onset" and "brief duration" that are intended to distinguish emotion from mood and disposition. Psychologists generally distinguish these three states by the time frames in which they impact behaviour. Emotions come on suddenly and last short periods of time, minutes or hours. Moods persist for days or months. And dispositions continue for a long term, can last a life time and are often considered personality attributes.

cognitive processes that determine whether the emotion leads to action. In chapter 3, we will find that this latter possibility accords better with the motivational model that we will construct on the basis of adaptive plasticity theory.

3 THE SIMPLIFIED HEDONISTIC MODEL

3.1 Modelling for Incompatible Strategies

A functional cognitive model of motivation or action limited to dealing with circumstances under which there are *no* conflicting strategies can be rather simple. If activated signal-strategy conditionals (activated by signals whose reliability is above an associated threshold) have strategies that are either Best Bets, Only-Choices or Harmless (see §2.4.3), then individuals may simply adopt those strategies. However, the situation is very different when a) two or more signal-strategy conditionals have been activated, b) two or more of these conditionals are not Harmless, and c) one or more of the warranted signal-strategy conditionals is incompatible with any other conditional. Under these circumstances, two or more incompatible conditionals are activated. This incompatibility may be due to matters such as resource scarcity or incompatible effects of strategies. For instance, offering N quantity of resources to person X is incompatible with offering any quantity of resources to person Y if the total available quantity of resources is only equal to N and the effects that follow from the adoption of fight and flight strategies are incompatible. I shall refer to such incompatible conditionals as strategy-incompatible conditionals (SICs). On the basis of the predictions made in Chapter 2, I will attempt in this chapter to model a cognitive architecture that would allow individuals to choose between SICs – or more precisely, chose between the strategies which SICs contain. This will provide an opportunity to account for the seeming contradiction in a) hedonistic and altruistic behaviour addressed by Sober & Wilson, b) self-oriented behaviour and other-oriented behaviour marked by empathetic reactions addressed by Batson, and c) behaviour that reflects rational self-interest in material incentives and behaviour that tends to produce long-term benefits in social interactions addressed by Frank.

I assume that organisms would face adaptive pressure to solve the problem of choosing between the strategies contained in SICs in such a way as to optimize inclusive fitness. Thus, if natural selection is the only relevant force and there are no adaptive constraints, organisms would tend to choose that strategy whose total expected payoff in terms of inclusive fitness is greatest. The determination of this payoff would appear to require individuals to deploy mechanisms that grade the effect of states and events on inclusive fitness. We adopted as a working hypothesis (see §2.5) the possibility that psychological pleasure and pain performs such a grading function. However, in that context the grading followed from the assumption that we possess second-order plasticity and that grading was required *for the purpose* of acquiring new signal-strategy conditionals. It seems

plausible that the grading mechanism, whatever it may be, that is used for that purpose may also be used here to determine the total expected payoff of competing strategies in order to chose between them. So, I shall extend my working hypothesis somewhat and suggest that the grading of effects of states and events on inclusive fitness by good and bad feelings may serve *both* purposes.

The working hypothesis then suggests that psychological affects provide individuals with *valuations* of those states and events that are expected to follow from the adoption of alternative strategies. In this regard, *psychological* pleasure and pain may play a role in the control of *intentional* behaviour that is analogous to the role that *physical* pleasure and pain plays in the control of *homeostatic* behaviour. A theory about the latter is advanced by the physiologist, Michel Cabanac (1992):

> At present as physiologists studying various homeostatic behaviors, such as thermoregulatory behavior and food and fluid intake, we have no common currency that allows us to equate the strength of the motivational drive that accompanies each regulatory need, in terms of how an animal or a person will choose to satisfy his needs when there is a conflict between two or more of them. Yet the behaving organism must rank his priorities and needs a common currency to achieve the ranking . . . A theory is proposed here according to which pleasure is this common currency. The perception of pleasure, as measured operationally and quantitatively by choice behavior (in the case of animals), or by the rating of the intensity of pleasure or displeasure (in the case of humans) can serve as such a common currency. The tradeoffs between various motivations would thus be accomplished by simple maximization of pleasure (p. xxx).

Further, this idea that choice is determined by the most favoured affect is found in Robert Frank's model. There, pleasant and unpleasant feelings (realized or merely anticipated) that are caused by rewards and punishments (realized or merely anticipated) compete with moral sentiments. However, Frank allows that pleasant and unpleasant feelings produced by the reward mechanism compete with each other when individuals make choices between material incentives.

The optimization of inclusive fitness does not require that every strategy contained in every SIC be graded for inclusive fitness (in accord with our working hypothesis or by any other means). For example, if that grading is encoded in psychological pleasure and pain, then those affects may be used to *rule out* competing strategies. Let us say that an organism has a choice of acting either on

SIC1 or SIC2. It may be that a) a negative affect is attached to SIC1, b) no affect is associated with SIC2, and c) the organism can choose SIC2 because SIC1 has been *ruled out*. Oppositely, good and bad feelings may identify the only *positive* strategy in a set of possible strategies. All but one in the set may trigger a negative affect or no discernable affect while one strategy is associated with a pleasant feeling.

3.2 How Feelings might Influence Strategic Choices

Whatever mental states are employed in choosing between SICs, those states must issue *prior* to such choices being made and often *prior* to the future world states to which the relevant signals are correlated. However, in my account of the acquisition of new signal-strategy conditionals (see §2.5), I argued that the mechanisms that grade for effects on inclusive fitness respond to states or events after the fact. Let us turn to our working hypothesis for a moment and ask whether it is plausible that psychological pleasure or mental distress grade events *in advance* of their occurrence as would appear to be necessary if they are employed to chose between SICs.

I think that is plausible. I will suggest two possible means by which feelings could be aroused in advance of future states in order to influence strategic choices so that if those feelings in fact grade for effects on inclusive fitness, then those choices will reflect (to some extent at least) that grading. First, it is at least possible that previously established affective reactions are firmly associated with some SICs so that when these SICs are activated by signals, so also are those associated feelings. These feelings could then compete with those of other activated SICs. Individuals could then choose a strategy on the basis of a competition between those feelings. Under this approach, the intensity of those feelings may depend upon the individual's prior experiences or they may be innately established. I am imagining a possibility in which for example individuals have prior feelings attached to the alternative responses that they can take when mildly abrasive or obnoxious behaviour is encountered (signalled in some way). They might feel somewhat bad about confrontation responses in general or about simply ignoring such signals in general. On the other hand they might be positively predisposed to courteously informing others about their hurtful behaviour. Such previously established affective reactions could then compete in a process for

choosing the strategy that will be adopted. One disadvantage in having previously established and fixed levels of intensity in the pleasure or pain that is associated with the adoption of a strategy is that it is somewhat inflexible in so far as it does not take relevant proximal conditions into account.

A second more flexible means by which affects could influence choice involves anticipating the psychological pleasure or mental distress that would be aroused by future states or events. These anticipated affects could be the basis upon which individuals chose the strategy that they will adopt when two or more SICs have been activated. This idea of anticipating future pleasure or pain is considered in the Sober & Wilson analysis (and found unreliable in relation to motivating care for one's children[27]); it is utilized in Frank's analysis; assumed to be sound in Batson's experimental design; and it is a widely accepted theory in psychology, being fundamental to operant conditioning.

In order for the anticipation of pleasure or pain to be aroused by future events (for the purpose of choosing between incompatible strategies), those events must be predicted. Thus, it would seem that after a SIC is activated, individuals predict one or more alternative future world states that would plausibly follow given proximal conditions and the strategy contained in that SIC. However, the theory that we employ does not specify the form that these predictions can take, nor the nature or level of sophistication of the cognitive processes involved in producing them. For example, laboratory-conditioned responses involving reward-seeking behaviours (e.g., a chimpanzee pulling a lever to receive a treat) seem to involve a kind of prediction on the part of the conditioned individuals. We may find more sophisticated predictive processes that take the form of imagined outcomes of particular behaviours, the synthesis of which may primarily involve unconscious processes. Predictions that precede anticipated pleasant or unpleasant feelings may arise from a conscious inferential process or calculation, as is perhaps suggested by Frank's term "rational calculations." (Thus, reason may play a role not only in the acquisition of new conditionals (see §2.6). It may also play the role described above.) The *psychological* purpose of those rational calculations according to Frank is to determine the pleasant or unpleasant feelings that would follow from the adoption of strategies. However, they may have the additional biological function of grading the effect of future states or events on one's inclusive fitness.

[27] Sober & Wilson argued against the reliability of anticipated affects in motivating adaptive behavior by assuming that such anticipations are (in the relevant ways) like typical beliefs about the world that are readily subject to revision. One could simply learn that helping one's offspring is not as pleasurable as eating chocolate ice cream just as one could learn that touching a red glowing stove top can be painful. I strongly disagree with their argument (see §3.3.3).

The total payoff that may be expected to follow from the adoption of a particular strategy is determined not only by the single most likely world state that may be caused by the adoption of that strategy. Rather, the total expected payoff is the sum of the expected payoffs (see §2.4.2) taking into account all the Relevant World States (see §2.4.3.3) and the likelihoods of each such state. This suggests that a rather cumbersome and perhaps implausible cognitive exercise may be required in order to properly grade the effect of a single strategy: It would seem that individuals need to predict all the Relevant World States, consider the likelihood of each, anticipate various pleasures or pains that would be produced by each, and to anticipate an affective reaction from adopting a particular strategy that is proportionate to the sum of the various pleasures and pain. That final reaction would then constitute a grade for the effect that a strategy will have on inclusive fitness.

On one hand, we have this seemingly implausible exercise that we are required to undertake if we are going to grade properly the effects of a strategy on inclusive fitness. On the other, we have the force of adaptive plasticity theory pointing to the adaptive pressure that would be presented to take all these factors into account. I am more impressed by the latter and allow the possibility that there may have evolved means for dealing with this multiplicity of factors in grading. However, the scope of my thesis does not include taking an empirical position on this issue. So, I will suggest here only that it seems plausible that natural selection produced a grading process that includes short cuts and highly efficient sub-processes. For instance, humans may only assess those Relevant World States that have a high likelihood of coming to pass. It would be too cumbersome and costly to do otherwise. Perhaps we have an innate talent for focusing only on the most likely states and events that will follow from the adoption of a strategy and perhaps the assignment, addition and subtraction of anticipated good or bad feelings are highly efficient, automatic and speedy processes. Finally, perhaps we *sometimes* refrain from availing ourselves of some such shortcuts when the choices we make are especially important, with the result that such choices become difficult to make, costly and time-consuming but nonetheless worthwhile.

The two possible means by which feelings can influence strategic choices that I have proposed are compatible. Thus, it may be that we sometimes (perhaps in relation to certain kinds of SICs) utilize one of those means and sometimes the other. Alternatively, it may be that both means may be utilized in making a particular choice: It would seem that individuals could have an affective predisposition attached to particular SICs and yet be subject to having those feelings modulated as a

result of proximal conditions. For instance, the individual who is positively predisposed to the strategy of courteously informing others about their hurtful behaviour might predict on a particular occasion that such a response would be greeted with a cold unpleasant rebuke. As a result, the individual's initial affective response may be modulated in a negative direction.

3.3 The Simplified Model

A partial model emerges to this point, which is represented in Figure 5. It is simplified in several respects while one component has been added to facilitate the anticipation of psychological pleasure or pain. I discuss the simplifications and the addition immediately below. Following this, I will add components, complete the model (shown in *Figure 6*), and conclude my thesis by demonstrating how the model works with detailed examples.

3.3.1 The Simplifications

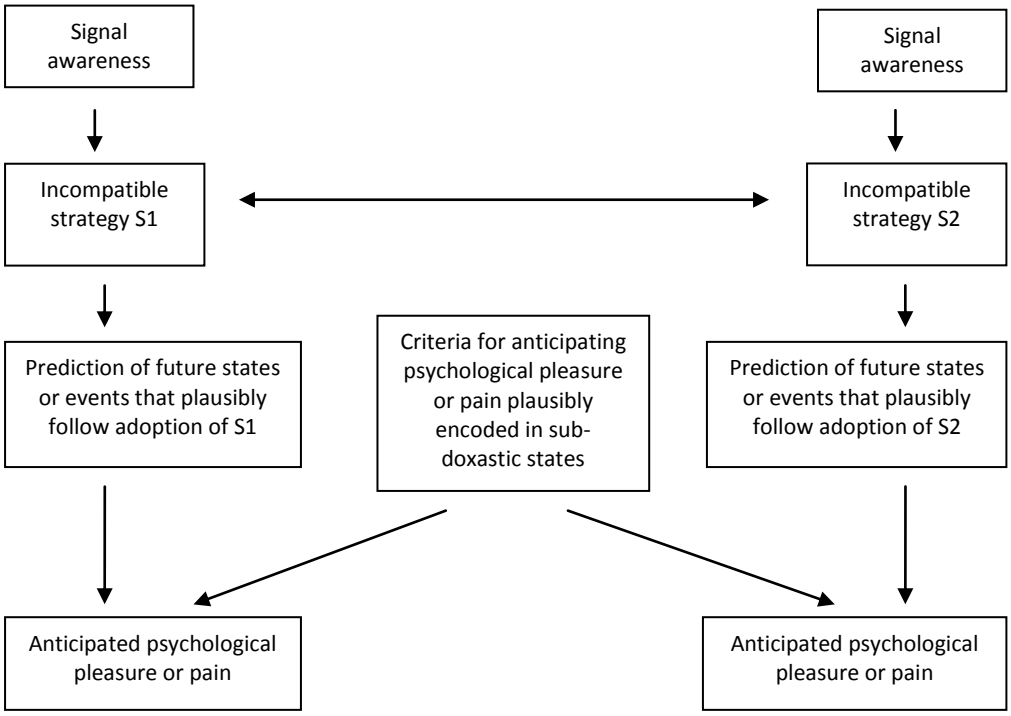
I do not reference in this simplified version of the model the possibility that SICs (two or more signal-strategy conditionals whose strategies are incompatible, thus strategy-incompatible conditionals) have firmly associated feelings that influence strategic choices (the more inflexible possibility suggested in §3.2). I only present the possibility that after SICs are activated individuals predict future world states and then anticipate psychological pleasure or pain. Second, in complex cases, the determination of the total expected payoff of a particular strategy may require taking more than one reasonably likely Relevant World State into account and factoring in the likelihoods of the states. The simplified model does not include processes that could factor the likelihood of each such Relevant World State in that determination. Third, the model does not explicitly reference either the exercise of reason or the arousal of emotion. As I suggested earlier reason may be deployed at a number of stages in the process of developing signal-strategy conditionals (see §2.6) and reason may be involved in the prediction of future world states that may follow from the adoption of a strategy (see §3.2). However, reason does not appear necessary at any particular stage. I also suggested earlier (see §2.7) that from the viewpoint of adaptive plasticity theory,

emotion looks like innate signal-strategy conditionals. So, the inclusion of emotion in this model appears redundant. Last, the simplified model does not explicitly reference the processes that measure signal reliability. Let us examine this last simplification more carefully by reviewing relevant findings presented in §2.4.

Taking the reliability of signals into account was predicted to produce an ability to choose strategies with higher total expected payoffs. However, it is plausible that single-level reliability measuring is reflected in the threshold at which a signal is perceived and at which awareness arises. The signal reliability associated with that threshold might be sufficient to warrant the activation of a signal-strategy conditional. For example, it may be that such a threshold is reached at the moment at which one becomes aware that one's offspring is in need, at the moment that empathy for others is triggered, and at the moment at which the moral sentiments to which Frank refers are aroused. Thus, it may be that a level of signal reliability is reflected in signal awareness and the activation of a conditional. Accordingly, the first elements in my simplified model are instances of signal awareness associated with the activation of two respective SICs.

Nonetheless, when strategies are neither an Only-Choice, nor Harmless, nor a Best Bet (defined in §2.4.3.1, §2.4.3.2 and §2.4.3.3 respectively) higher levels of reliability of a signal may change the expected payoff matrix of its associated strategy. Thus, higher levels of reliability for signals that activate SICs may increase the likelihood of adopting strategies that would tend over time to produce higher payoffs. So it may be that when individuals sense that signals have a higher reliability than that which is associated with their threshold levels, they factor this higher reliability into their determination of the total expected payoff of strategies. I have simply not reflected this possibility in the simplified model that I have represented in Figures 5. That is, I have not shown that anticipated psychological pleasure or pain may be separately influenced by levels of signal reliability that are higher than that which arises at those signals' threshold levels.

Figure 5 – Simplified Model of How Signals that Are associated with Incompatible Strategies may Lead to Anticipated Psychological Pleasure or Pain



3.3.2 How Abstract Are the Criteria for Anticipating Pleasure and Pain?

For individuals to anticipate their future psychological pleasure or pain on the basis of their predictions about future world states, they would seem to need criteria upon which to base those anticipations. They need to know that future event X will produce an affect Y. Additionally, where varying degrees of pleasure or pain are produced according to variable features of states (such as the size of a hunter's catch or the deadliness of a predator), then it would seem that individuals need to know pleasure-producing and pain-producing parameters of predicted future states and what the ranges of those parameters might be. Sober & Wilson address this matter when they describe how hedonistic behaviour depends upon beliefs about the pleasure or pain that follow from hedonistic actions. Thus, an individual has to know both that chocolate cake and helping one's offspring will produce pleasure, and have criteria (that may be based on previous experience) that would suggest the degree of the pleasure. This matter is implicit in Frank's model where he suggests that individuals calculate future rewards and punishments that produce pleasure and pain. Those

calculations require a set of assumptions about what constitutes a reward or punishment. Let us consider how abstract these criteria that are applied when we anticipate pleasure or pain are likely to be.

It follows from my working hypothesis about the function of psychological pleasure and pain and the inferences that I have drawn in this chapter that anticipated psychological pleasure or pain arises in causal chains that lead to action. Additionally, the adaptive plasticity theory upon which we rely requires that the model allow flexible behaviour. Thus, the set of criteria for anticipating pleasure or pain should allow those anticipations to be flexible. For example, the criteria should allow individuals to anticipate affect response F given the prediction of future world state X, and affect response G given prediction of future world state Y.

I believe that this necessity for flexibility allows us to make inferences about the relative abstractness of these criteria. In particular, if the criteria were just abstract enough that criteria could combine then individuals could have the variability possible with the permutations and combinations of these criteria. To illustrate, the two relatively concrete terms "red cube" and "yellow sphere" provide between them less descriptive flexibility than the four more abstract terms "red," "cube," "yellow," and "sphere." The latter allow the identification not only of red cubes and yellow spheres, but also of yellow cubes and red spheres. We find the utilization of abstract criteria in a number of perceptual systems. For example, Stich (1978) cites "perspective, size, surface texture, the perception of edges and corners, occlusion, illumination gradients and stereopsis" (p. 502) as criteria used to make judgements about depth perception. And findings suggest that we use abstract criteria to determine facial attractiveness. For example, Gillian Rhodes (2006) concludes that "[a]verageness, symmetry, and sexual dimorphism are good candidates for biologically based standards of beauty. A critical review and meta-analyses indicate that all three are attractive in both male and female faces and across cultures" (p. 199).

Now let us contrast the use of relatively concrete criteria with that of combinations of relatively abstract criteria in the anticipation of pleasure or pain that will follow from predicted events. Consider examples involving predictions that someone will suffer. A relatively *inflexible* organism may employ the relatively concrete criterion, "if I predict that X will suffer, then anticipate Y level of pain on the occurrence of that event." Consider the relatively concrete and inflexible criterion that Sober & Wilson suggest in their portrayal of psychological hedonism: They say an individual must believe that pleasure will follow from helping one's child. The relatively inflexible responses that may be predicted to follow from such criteria may be contrasted with the relatively flexible responses that may follow from the use of a multiplicity of criteria such as a) the intensity of suffering, b) whether the sufferer is a friend or foe, c) the closeness of the

genetic relation between the subject and the sufferer, and d) whether or not there will be competing uses for the resources that would be required to help the sufferer. Using such criteria, individuals may anticipate varying levels of pleasure or pain depending upon the particular proximal conditions. Adaptive plasticity theory predicts that natural selection would favour this flexibility over the less flexible alternative that would seem to follow from the deployment of relatively concrete criteria. So the motivational model that I advocate includes the utilization of such abstract criteria in the anticipation of pleasure and pain. That model is therefore expected to allow greater behavioural flexibility (that may tend to enhance inclusive fitness) than the hedonistic model that Sober & Wilson argue against.

3.3.3 The Mental States that Encode the Criteria for Anticipating Pleasure and Pain

One of Sober & Wilson's (1998) arguments against psychological hedonism is that the belief that helping one's child produces pleasure is readily susceptible to being corrupted by evidence to the contrary. A parent need only discover that helping their offspring does not produce pleasure. So we have, as Armin Schulz (2009) puts it, "the possibility of maladaptive updating" of beliefs.[28] If Sober & Wilson are correct, then the model we are building here has a serious problem. Adaptive plasticity predicts that these anticipations of pleasure or pain are causal factors in the adaptive choice between incompatible strategies. If these anticipations can be readily corrupted by experience, then inclusive fitness would be diminished in the process. So, it would seem that the criteria for anticipating pleasure or pain need to be both relatively abstract (as discussed in previous subsection) and now resistant to change caused by contrary experience.

Stephen Stich (2007) addresses Sober & Wilson's concern about corruptibility: Hedonistic motivations plausibly depend upon "sub-doxastic" mental states that are resistant to change. Stich (1978) provides an example of how such states operate:

> There is, of course, a mechanism of some complexity mediating between [a] subject's hearing [a] sentence and the formation of the belief that it is grammatical. And, while we know little in detail about the workings of this mechanism, it is plausible to speculate that the mechanism exploits a system of psychological states, which serve to store information

[28] Sober & Wilson argue that altruistic ultimate desire does not depend on such beliefs and this contributes to making psychological altruism more reliable than psychological hedonism.

> about the grammar of the subject's language. If this speculation proves accurate, then these states which store grammatical information are a prime example of subdoxastic states (pp. 501-502).

Stich (1978) explicates the property of being resistant to change: Sub-doxastic states "are largely inferentially isolated from the large body of inferentially integrated beliefs to which a subject has access" (p. 507). This is possible because they are "not open to conscious awareness or reporting" (p. 505). In summary, Stich (2007) says that these states are "'stickier' belief-like states [that] are harder to modify . . . They are also typically unavailable to introspective access" (p.12).

Returning to the matter of anticipated pleasure or pain: What is it exactly that may be "inferentially isolated" and "not open to conscious awareness"? Is it a) the abstract criteria for anticipating pleasure or pain (such as the closeness of the genetic relationship between the subject and sufferer) or is it b) particular anticipations of pleasure or pain (such as "helping my child who is suffering a great deal will alleviate the pain I now experience and produce pleasure")? We are *not* ordinarily aware of the former, nor do we tend to attempt to integrate our concern for such criteria as the genetic relatedness of sufferers into what Stich refers to as the "large body of inferentially integrated beliefs to which a subject has access." Yet, we are often aware of our anticipations of good or bad feelings that are prompted when we imagine what might follow from our behaviour. Thus, it *seems* as though it is the abstract criteria for anticipating good feelings and mental distress that is encoded in sub-doxastic states (and not the anticipations of pleasure or pain that may be ordinary beliefs). How this *seems* is consistent with Stich's (1978) observation that sub-doxastic states are a mechanism "mediating [a perception] and the formation of [a] belief" (p. 501). Sub-doxastic states may mediate *predictions* of future world states and a *belief* about the anticipated pleasure or pain that may follow from the adoption of a strategy. Sub-doxastic states thus "play a role in the proximate causal history of beliefs, though they are not beliefs themselves" (1978, p. 502). Last, the utilization of sub-doxastic states in making judgements on the basis of a number of abstract criteria is consistent with the other examples of the deployment of sub-doxastic states provided by Stich (1978). For example, the formation of beliefs involving depth perception, such as object X being in front of object Y, requires the deployment of a number of abstract criteria that Stich claims are encoded in sub-doxastic states. Stich further cites findings that beliefs about female attractiveness to men depend upon such a multiplicity of abstract criteria in relation to female appearance.

3.4 A Prediction of Monistic Psychological Hedonism

Adaptive plasticity theory predicts that intentional actions take inclusive fitness into account. So I argued that second-order plasticity and the ability to chose between two or more activated strategy-incompatible conditionals seem to require special mechanisms in order to take inclusive fitness into account – particularly, mechanisms that allow them to grade and rank the effects of states or events on inclusive fitness. I have agreed with Sober & Wilson that *physical* pleasure or pain that may be produced by physical stimulation or tissue damage would not seem to be a good metric of inclusive fitness. Further, I adopted as a working hypothesis the plausibility that psychological pleasure and mental distress, good and bad feelings that are purely psychological responses to states and events, grade effects on inclusive fitness. If the working hypothesis is true, then in order for motivations and intentional actions to take inclusive fitness into account they would have to follow from these affects – either directly, by means of anticipations of those affects, or via a process by which competing strategies are *ruled out* on the basis of those affects.

In other words, adaptive plasticity theory in combination with the working hypothesis predicts that human motivation will tend over time to be adaptive to the extent that it follows from the psychological affects to which I refer. Further, *if* adaptive pressure is against motivations that do not tend to enhance inclusive fitness, *and* natural selection has in this instance maximized inclusive fitness (and discounting phenotypic aberrations), then our theories in combination with the working hypothesis predicts that *all* motivations take inclusive fitness into account and follow from psychological pleasure or pain (in the various ways in which they may so follow). If these assumptions are correct and the working hypothesis about the role of psychological affects is true (and there are no other measures of inclusive fitness), then all instrumental desires that lead to action must follow from psychological pleasure or pain and that intentional action cannot disregard the way in which a state or event has been graded by those affects. This prediction that all motivation follows from such psychological affects constitutes a particular kind of hedonism, one that is specifically based on psychological pleasure and mental distress, and specifically *not* on physical pleasure and pain.

Let us consider the implications of this finding for the standard "practical reasoning" model – in which an ultimate desire combines with belief to produce an instrumental desire, and in which an instrumental desire combines with belief to produce a further instrumental desire. In the accounts

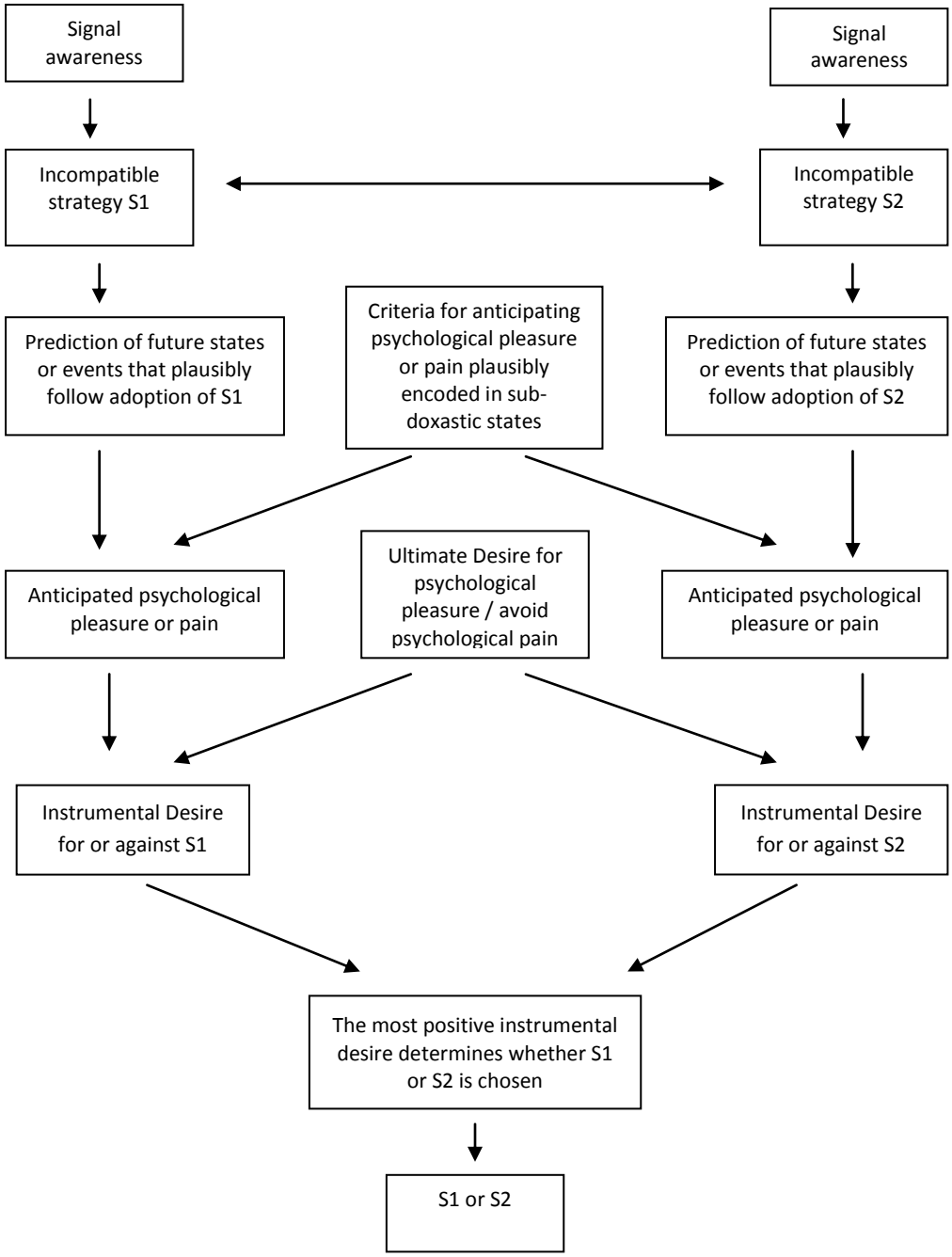
given to us by Sober & Wilson (their pluralistic account), Batson and Frank, motivations may track back to different ultimate desires. In Sober & Wilson's and Batson's account, some motivations track back to an altruistic desire while others may track back to an egoistic desire. In Frank, some motivations may track back to desires for a good reputation or sincere-manner, and others track back to desires for material rewards. However, the prediction that all motivation follows from psychological affects implies that all motivations track back to a desire for psychological pleasure or to avoid mental distress. Thus, the theory and the working hypothesis suggest that the desire for psychological pleasure and to avoid mental distress *is the ultimate desire*, the only ultimate desire.

Hedonism and an ultimate desire to feel good may suggest a shallowness of character or temperament that may not be warranted by the model we are considering. First, this ultimate desire is a practical psychological mechanism and not a biological end goal. We do not merely strive to feel good. Our hypothesis is that feeling good or bad are mental states that were adapted *to grade inclusive fitness*. They are not more than measuring sticks according to this model. Second, it does *not* need to seem as though our ultimate desire is to strive for psychological pleasure or to avoid psychological pain. There is no theoretical necessity that I can discern for these affects to *seem* to be the end goal. If my working hypothesis is true and there are no other measures of inclusive fitness, then the only necessity is that psychological pleasure or pain be *used*, even unconsciously, as a kind of metric of inclusive fitness. It is consistent with my approach that good feelings and mental anguish be viewed by individuals as merely incidental to states and events that are valued and cared about – if and only if we have unconsciously inferred the import of those events on the basis of those affects. Further, the plausibility that anticipations of pleasure or pain are subserved by sub-doxastic mental states that are not open to awareness fits neatly into this picture.

I have added elements that follow from this analysis to the model in Figure 5 and presented them in Figure 6. I will refer to the configuration of cognitive devices shown there as the Simplified Hedonistic Model (SHM) because it is still based on the simplifications reflected in the Figure 5 model. In the SHM, anticipated psychological pleasure or pain (as a belief or belief-like state based on predictions and criteria for anticipating psychological pleasure or pain plausibly encoded in sub-doxastic states) combine with the ultimate desire and produce instrumental desires for or against particular strategies. Finally, two or more instrumental desires compete in a process that leads to

the adoption of a strategy. In the following three subsections, I will contrast the SHM with accounts provided by Sober & Wilson, Batson and Frank.

Figure 6 – The Simplified Hedonistic Model with Activated Strategy-Incompatible Conditionals



3.5 The Simplified Hedonistic Model

3.5.1 The SHM in Contrast with Sober & Wilson's Pluralistic Model

In §1.1, I pointed out that Sober & Wilson's Type-Two Pluralism involves cognitive mechanisms that may produce inflexibility (without their accommodation for competing desires) – by not limiting the conditions under which individuals opt to help their offspring – and (with their accommodation for competing desires) fails to regulate some behaviours in a way that may be predicted to tend to enhance inclusive fitness – by allowing a multiplicity of conflicting desires with no adaptive means of prioritizing them. In contrast, the SHM allows flexible responses (that depend upon the conditions organisms encounter) and includes a means by which individuals chose between competing desires that may be predicted to tend to enhance inclusive fitness (via "sticky" criteria for anticipating future pleasure or pain).

Further, their support for psychological altruism rests primarily on their prediction that it is more reliable than psychological hedonism, at least at motivating individuals to attend to the wellbeing of their children. They suggest four causes of the putative inferior reliability of psychological hedonism. *First*, they suggest that beliefs about future pleasure or pain are subject to maladaptive updating. However, this worry was based on the assumption that the criteria for anticipating pleasure and pain are readily subject to maladaptive revision (see §3.3.3). As Stich (1978 and 2007) suggested, this assumption is unfounded.

Their *second* assumed cause of the unreliability of hedonism involves the necessity for an additional belief mechanism that is not required by psychological altruism: "For the future oriented hedonist to be motivated to help, the belief that her children need help must trigger a further belief – that helping will bring pleasure, or that not helping will bring pain" (p. 318). This is a concern about an increased risk of system failure or creating an additional source of error. Altruistic models do not have as many component parts as hedonistic models. However, in the light of the astonishing complexity of the neurological systems that subserve these models, their worry seems odd. Both altruism and hedonism require perceptual systems, acquired or innate strategies and behavioural routines, the ability to understand the nature of present circumstances, the ability to make predictions about future states given that the individual does not adopt a helping strategy, and the

ability to predict future states given that the helping strategy is adopted. Predicting against hedonism on the basis that it requires an additional belief seems a little like predicting that all airplanes have fixed landing gear on the basis that fixed landing gear have fewer moving parts than retractable landing gear.

Their *third* concern is that when pleasure or pain is produced, it will not be of an appropriate intensity. This follows from their position (on which I agree) that some kinds of pleasure and pain (particularly, tissue damage and pleasure that derives from *physical* stimulation) do not correlate well with inclusive fitness (see §2.5). However, my model is based on what seems to me to be a plausible working hypothesis that *psychological* pleasure and pain correlated to some significant degree with inclusive fitness at the time that the relevant traits evolved. If so, the basis for their concern will be appreciably weakened.

The *fourth* of their worries is that hedonism depends only upon the motivation of pleasure and pain, whereas the pluralism that they support may rely on motivation produced either by pleasure and pain or by ultimate altruistic desires. This is an argument for a biological redundancy for the purpose of increasing reliability. Whatever benefit this pluralism provides must be weighed against the cost it introduces in connection with a) diminished flexibility and b) the problem of prioritizing between competing desires. These costs may exceed those benefits.

3.5.2 The SHM in Contrast with Batson's Altruistic Model

Batson's model may be seen as a proposed solution to a theoretical problem arising from empirical evidence that individuals sometimes help others when there are no apparent subsequent rewards for doing so or punishments for not doing so. The problem is that these findings seem to conflict with the view that we employ cognitive mechanisms that often cause us to *only* act in anticipation of reward or punishment. Batson's pluralistic model provides distinct psychological paths leading to "other-oriented" and "self-oriented" behaviour. The altruistic path includes the arousal of empathy (see §1.2).

The SHM advances a very different solution to the problem Batson addresses. We may contrast it with his by seeing how SHM accounts for findings in one of their studies. In Batson's Study #6, subjects see Elaine's distress and anxiety caused by the electric shocks she is receiving and are given the opportunity to take her place or not to do so. Thus, the subjects face two incompatible strategies: Elaine's distress and anxiety signal a helping, altruistic strategy, S1; and viewing Elaine while she receives the shock and then being offered the opportunity to take her place together signal a possible future state in which the subject is on the receiving end of the electric shocks and this may activate a strategy of trying to avoid these electrical shocks, S2. S1 and S2 are incompatible in that under the study conditions, S1 may only be pursued by not adopting S2 and *vice versa*.

The SHM suggests that the test subjects may then make predictions about future world states that could follow from the adoption of the two alternative strategies. A prediction that may have followed from S1 (taking Elaine's place) is that Elaine will find out that the subject was willing to take her place, feel indebted, and reciprocate at some point in the future. A second prediction may have been that others will find out about the subject's willingness to suffer the electric shocks for Elaine, they will find out about their generosity and kindness, and as a result hold them in higher esteem, enhancing their reputation. (In Frank's view, altruistic behaviour is often conspicuous so as to allow the possibility for reputation to be positively impacted.) A third prediction, may have suggested how uncomfortable the electric shocks will be. Predictions that may have followed from S2 (avoiding the electric shocks), include that a) Elaine will find out that the subject declined the offer to trade places and will become unwilling to help the subject if that help should ever be required, and b) others will learn about the subject's unwillingness to help Elaine and their view of the subject's character will be dimmed.

In the account provided by SHM, subjects then evaluate their predictions according to criteria for anticipating psychological pleasure and pain. For example, predictions that the shock would be very uncomfortable would lead to an anticipation of mental distress if S1 were to be adopted. Predictions that one's reputation would be adversely effected, would lead to an anticipation of mental distress if S2 were to be adopted. Thus a certain anticipated psychological pleasure or pain builds around both S1 and S2. The resulting anticipated good or bad feelings may have then combined with an ultimate desire for psychological pleasure and to avoid mental distress to cause instrumental desires for S1 and S2. It may have been that those who chose to help Elaine put a high value on their

reputations and a low value on avoiding the electric shocks because they felt that the latter would not have had a significant impact on their wellbeing.

According to the SHM, the *behaviour* (helping Elaine) depends upon the corresponding *instrumental desire* (the instrumental desire for helping Elaine, S1). In particular, according to SHM a) when the instrumental desire for S1 is low and the instrumental desire for S2 is high, then subjects would *not* help Elaine, and b) when the instrumental desire for S1 is greater than the instrumental desire for S2, then test subjects would help Elaine. The same relations and formal structure is found between helping behaviour and a prior arousal of empathy that Batson describes: "when empathy is low, helping drops dramatically if escape is easy. When empathy is high, however, helping remains high even if the empathically aroused individuals can easily reduce their arousal by escaping exposure to the suffering victim" (p. 52). Thus, in the context of the SHM, the instrumental desire to help others is a *functional* replacement for empathy. Accordingly, predictions about empathy may be based on the SHM. For instance, we may predict general conditions under which the arousal of even high levels of empathy does not lead to altruistic behaviour. This may occur when subjects experience an even *stronger* instrumental desire in relation to an incompatible strategy. Perhaps, empathically aroused individuals fail to act on their empathy when that affect competes with say a powerful fear response. The SHM predicts that a low-empathy condition arises when an individual fails to detect the payoff in helping – explaining, in a fashion, indifference to the suffering of others.

The empirical data that Batson suggests support his pluralistic hypothesis consists primarily of the altruistic behaviour of empathetically-aroused test subjects in the purported absence of anticipated self or social rewards or punishments – including praise, honour, guilt and shame. However, Batson does not seem to consider the possibility that the mental states that he has identified as empathy, and that apparently preceded that altruistic behaviour, involved, included or masked the anticipated psychological affects to which I have referred. For instance, it is entirely plausible that the mental states that he identified as empathy involved the anticipated affects and instrumental desire that are included in the SHM. These may have led to the prosocial behaviour Batson found. That is, it is plausible that the behaviour that followed from empathetic reactions was already hedonistic and that Batson's subjects did not need additional hedonistic motivation, did not need additional self or social rewards or punishments, in order to act.

3.5.3 The SHM in Contrast with Frank's Pluralistic Model

Recall that Robert Frank's pluralistic account consists of a "self-interest model" and a "commitment model" (see §1.3). The self-interest model includes a reward mechanism and pleasant and unpleasant feelings that reward and punish agents respectively. Also, the self-interest model includes *rational calculation* that is employed to predict costs and benefits and anticipate future pleasant or unpleasant feelings. It deals with the pursuit for predicted material benefits. However, according to Frank, the self-interest model does not account for an important range of social behaviours that are covered in the commitment model. These behaviours are mediated by moral sentiments – including guilt, anger, envy and love. As opposed to producing foreseeable rewards, they may merely tend to produce benefits over the long run. Thus, it may be that altruistic behaviour, moral behaviour, the maintenance of values, acting according to conscience, and exercising self control tend over the long term to be beneficial in a social context. When the moral sentiments that mediate such behaviour conflict with anticipated pleasure or pain produced by the self-interest model, those sentiments compete with and often overcome the pleasure or displeasure produced by the reward mechanism.

Let us consider the account that may be provided using SHM of a paradigm case that Frank uses to introduce his pluralistic model. The case involves the behaviour of members of the American McCoy family in response to the murder of their two children, Alifair and Calvin, committed by members of the Hatfield family. The McCoys had as one strategy, retaliation on some meaningful scale, S1; and, as Frank suggests, in the alternative, they could have not retaliated and perhaps attempted to bring the long-standing feud with the Hatfields to an end, S2. On Frank's account, S1 would have been mediated by emotion while S2 would have been mediated by rational calculation that leads to anticipated pleasure or pain. Finally, according to Frank, the emotion would have competed with the anticipated pleasure or pain, and won that competition, producing the retaliatory behaviour.

There are parallels between Frank's modelling in relation to S2 (not retaliating) and that presented in the SHM. He proposes that the pleasure or pain anticipated to follow from the contemplation of S2 is caused by rational calculation. I suggest that the process that may reasonably be characterized as rational calculation involves individuals making *predictions* (that may either employ conditioned responses or higher-order cognitive processes as suggested in §3.2) and that these combine with criteria for anticipating psychological pleasure or pain that are resistant to revision (plausibly

encoded in sub-doxastic states). Their combination produces anticipated pleasure or pain that follows from the contemplation of S2. Where there is substantial disagreement between Frank and me around S2, is in Frank's suggestion that rational calculation only involves events that may well occur in the foreseeable future. Thus, we both agree that the McCoys may have predicted the possibility that not retaliating may have lead to an end of the feud and lead to the safety and security of their family. However, I propose that they may also have predicted (employed "rational calculation") events that might not occur for a very long time, or that involves a series of small payoffs that accumulate over long periods of time. (The issue from the perspective of adaptive plasticity theory is the total accumulated payoff of a strategy over the life of the organism.) For example, they might have imagined a) it becoming known that the McCoys did not retaliate for the murder of their children, b) over the course of many years, many circumstances would arise in which individuals may have something to gain in acting in a way that may harm the McCoys, and c) that knowing that the McCoys had not previously retaliated, would assume that they could undertake those actions with impunity. A reason for thinking that the McCoys might have made such a calculation is that it is plausible that the McCoys would have realized that in their experience they regulated their own social behaviour on the basis of their particular predictions about how others would respond to those behaviours.

Thus, it may be that imagining the first prediction (the McCoys enjoying safety and security) caused the anticipation of pleasantness, and imagining the second prediction (others acting against them because they could do so with impunity) caused the anticipation of suffering. These two may have combined to produce an affective response that more or less reflected the total expected payoff of adopting strategy S2. If so and given the opposite valences of the two component affects, the McCoys may have been neutral about the adoption of S2. The SHM would then predict a weak or non-existent instrumental desire for S2.

Frank's model of the cognitive devices that are activated in connection with S1 (retaliate) is very different than that of the SHM. The SHM employs the same processing sequences for both S1 and S2. Thus, the activation of the signal-strategy conditional associated with the S1 strategy leads the McCoys to make predictions about future world states that would follow from the adoption of S1. Perhaps they imagined others in their community learning that they retaliate. And the prediction in combination with the criteria that they hold for anticipating pleasure or pain caused them to anticipate feelings of satisfaction. This affect may then have combined with the ultimate desire for

psychological pleasure to produce an instrumental desire for S1. According to SHM, that instrumental desire was stronger than the instrumental desire for S2 and lead to retaliation.

Referencing the *emotional* response of members of the McCoy family is not necessary in my model– even though it is a fundamental element in Frank's model. The anger that may have accompanied the McCoy retaliation may be an *expression* of an innate signal-strategy conditional (see §2.7): The receptivity to the appropriate signals and the strategy S2 (retaliate) are already referenced in the model. However, I have neither referenced the physiological responses nor facial expressions associated with anger. Nonetheless, the purpose of SHM was to attempt to capture key components in the causal chain that runs from the signal to the adoption of a strategy, and these physiological changes do not appear to me to be on that path.

Furthermore, unlike Frank's model, SHM suggests that an emotional response was not necessary for the adoption of the retaliation strategy. SHM leaves open the possibility that individuals acquire signal-strategy conditionals that include retaliatory behaviour that is not accompanied by emotion. (Although, SHM does not address the possibility that emotional responses impede cognitive processes in such a way as to negatively bias the evaluation of incompatible strategies.) And unlike Frank's model, SHM suggests the conditions under which emotional responses may fail to lead to action even in the absence of a rational calculation that makes an alternative strategy look good. This could arise if individuals rationally calculated that the strategy associated with the emotional response looked worse than the alternative. Thus, the McCoys might not have retaliated if they predicted that their communities took an especially dim view of such actions.

3.6 Summary

I have attempted to derive a general pattern in the sequencing of human mental states that would tend to maximize behavioural flexibility as a means of maximizing inclusive fitness. This approach accords with Godfrey-Smith's views that "[t]he function of cognition is to enable the agent to deal with environmental complexity" (p. 3) and that adaptive plasticity is the principle adaptive response to environmental complexity taken by many organisms. Thus, he predicts that an important function of cognition is to

facilitate adaptive plasticity. I have contrasted my approach with that of writers who have attempted to provide psychological accounts for certain kinds of contradictory behaviour, viz., Sober & Wilson (1998), C. Daniel Batson (1988) and Robert Frank (1988 and 1990). They examined, respectively, hedonistic and altruistic behaviour; self-oriented behaviour and other-oriented behaviour marked by empathetic reactions; and behaviour that reflects rational self-interest in material incentives and behaviour that tends to produce long-term benefits in social interactions (see §1.1, §1.2 and §1.3 respectively).

My Simplified Hedonistic Model (SHM) (represented in Figure 6) is the product of a series of predictions about particular adaptive pressures that relate to plasticity, i.e., about the particular potentialities of organisms to realize higher payoffs in terms of inclusive fitness through flexibility. Organisms can *potentially* do better employing a multiplicity of behaviours or strategies than by adopting a generalized strategy in a heterogeneous environment (given certain assumptions about the cost of adaptations and the reliability of signals). Notionally, specialized strategies may be more finely adapted to the changing particular conditions in which organisms may find themselves. And most of the particular arguments that I presented involved deducing the abilities, sensitivities, and behavioural characteristics that would be necessary in order to realize the potential benefits of behavioural phenotypic plasticity. The theory of natural selection predicts an evolution of mechanisms that subserves such abilities, sensitivities, and behavioural characteristics. It is these mechanisms that constitute my hedonistic model.

In order to appreciate higher payoffs from a multiplicity of strategies specialized for dealing with particular world states, organisms need to be attuned to environmental conditions that signal those world states; and these signals need to be firmly associated with those specialized strategies (see §2.1 and §2.2). Godfrey-Smith refers to these connections between such an awareness of a world state and a behavioural predisposition as conditionals because they may be stated in the form “if the world is in state A, then adopt strategy B.” To reduce ambiguity, I often referred to these as signal-strategy conditionals.

However, organisms may possess specialized strategies that may deliver benefits under some conditions and harms under others; and organisms depend upon environmental signals when they choose which strategy they will adopt. Thus it would seem that organisms would encounter adaptive pressure to take the reliability of those signals into account when they make such choices (see §2.4). Organisms would seem to need signal reliability detection devices. I distinguished two general kinds: those that detect a single level of reliability and those that detect multiple levels of signal reliability. My analysis showed that the domain in which organisms are predicted to need the latter multiple-level devices may be much smaller than one might

imagine. Thus, flexible organisms may very often adopt strategies based merely on the detection of signals that are above a particular reliability threshold and without the further assessment of reliability. That is, simple signal awareness (above a reliability threshold) appears very often to be sufficient to adaptively activate signal-strategy conditionals. Accordingly, the first psychological state with which the SHM is concerned is signal awareness, and I omitted multiple-level reliability detection from my simplified model.

The adaptive plasticity of many animals may be *potentially* enhanced by possessing second-order plasticity, the ability to learn new signal-strategy conditionals. They offer the possibility of outfitting organisms so that they are receptive to more reliable signals and so that they may employ more effective strategies. However, how is an individual organism to "know" whether a new conditional is an improvement over what they already have? Thus, in §2.5, I argued that the plausibility of the evolution by natural selection of second-order plasticity requires that organisms be able to grade the effect on inclusive fitness of the states or events that follow from the adoption of new candidate signal-strategy conditionals. Without that grading ability, organisms may adopt signals or strategies that diminish inclusive fitness.

Quite clearly, the mechanism that could subserve that grading ability may be based on pleasure and pain. However, I found these categories of mental states to be overly broad; and I distinguished *physical* pleasure and pain from *psychological* pleasure and pain. I counted as physical pain (or physical suffering) the pain that originates in tissue damage, and pain associated with feeling cold, overheated, hungry, dizzy or nauseous; and I counted as physical pleasure the pleasant feelings produced by external chemical, thermal and mechanical stimuli, e.g., smells, tastes, heat, and cold. In contrast, psychological pleasure and psychological pain (or mental distress) are purely psychological reactions to perceived, assumed or imagined states and events. Further, I suggested that psychological pleasure and pain states may have different etiological and functional characteristics than physical pleasure and pain states. And particularly, Sober & Wilson's persuasive arguments against pleasure and pain being well correlated with the effects of states and events on inclusive fitness seem to apply to *physical* pleasure and pain and not to the *psychological* affects. Finally, I established as a working hypothesis that psychological affects function to grade the effects of world states and events on inclusive fitness.

I considered the role that reason may play in the acquisition of new conditionals (see §2.6). It is consistent with the theory upon which I depend that reason is employed to help predict a) correlations between extant signals and their associated distal conditions, b) new signals, c) the effects of states and events on inclusive fitness, d) new strategies, e) new conditionals from combinations of extant signals and strategies, and f)

distal states or events that may follow from the adoption of particular strategies (see §3.2). However, that theory also predicts that reason may not play any sort of role that contradicted the grading of effects – that contradicted good feelings or mental distress, if my working hypothesis is true.

However, most of the sequencing of mental states in the SHM is predicted on the basis of adaptive pressure to identify the most beneficial strategy when two or more strategy-incompatible conditionals have been activated (by signals) (see §3.1). Thus, the SHM is largely oriented toward the resolution of such conflicts. Particularly, we reason that the ability to choose between incompatible strategies in order to maximize inclusive fitness requires the antecedent ability to *predict the states or events* that would follow from the adoption of each of such strategy. And further, we reason that this choice requires the further antecedent ability to grade those states and events (with respect to their effect on inclusive fitness) that had been predicted to follow from the incompatible strategies (see §3.2). (Thus, it may be that we grade both predicted and realized states and events and we may grade states and events in order to acquire new signal-strategy conditionals that are more beneficial than extant conditionals and in order to choose between incompatible conditionals that have been activated.) If the working hypothesis is true, then the ability to grade predicted world states is realized in an ability to *anticipate the psychological pleasure or pain* that would be produced by those predicted states and events.

Motivation and behaviour are at or near the end of a causal chain of psychological states (as represented in Figure 6). Therefore, the flexibility of those final motivations and behaviours depend on flexibility of the intermediate elements in that causal chain. And because the anticipation of affects arises prior to those motivations, I argued that adaptive pressure ought to arise for those anticipations to have a flexibility, variability, and responsiveness to the environment that would not limit the motivational and behavioural flexibility of organisms that follow (see §3.3.2). I suggested that possessing a number of abstract criteria for anticipating psychological pleasure or pain – much like the abstract criteria that are employed in depth perception – confers greater flexibility in those assessments than would be achieved with relatively concrete criteria such as "taking care of my children will produce pleasure" (provided that they are not significantly more costly) (see §3.3.3).

The mental states to which I have referred may be incorporated into the traditional standard reasoning model in which an ultimate desire combines with belief to produce an instrumental desire, and in which an instrumental desire combines with belief to produce a further instrumental desire. First, I reasoned that psychological pleasure and pain – that in my analysis serves only as a metric for inclusive fitness – is elevated

by the standard reasoning model into a monistic ultimate desire, to have psychological pleasure and avoid mental distress (see §3.4). This result follows from the prediction – based on adaptive plasticity theory in combination with my working hypothesis about psychological affects – that human motivation will tend to be adaptive to the extent that it follows from psychological pleasure and pain. Thus in the end, adaptive plasticity theory, second-order plasticity, and my working hypothesis about psychological pleasure and mental distress, combine to predict monistic psychological hedonism based on psychological pleasure and pain and not at all on physical pleasure and pain. Second, anticipated psychological pleasure or pain combines with the ultimate desire to produce instrumental desires for or against particular strategies. Third, two or more instrumental desires compete in a process that leads to the adoption of a strategy.

My simplified model provides an account for the paradigm cases advanced by Sober & Wilson, Batson and Frank who all advance pluralistic models of human motivation (see §3.5). And the SHM does not have the worrisome elements that I pointed out in those pluralistic models – such as, not providing a means by which conflicting strategies are resolved, or by disregarding the plausible affective aspect of empathy, or by making the resolution of strategies involve a competition between *incommensurable* psychological states. And in addition to being free of those problems, the SHM provides a simpler, monistic account of intentional action in which the most diverse human actions may all follow from similar sequences of psychological states.

REFERENCES

Batson, C. D., Dyck, J., Brandt, J., Batson, J., Powell, A. McMaster, M. and Griffitt, C. (1988): "Five Studies Testing Two New Egoistic Alternatives to the Empathy-Altruism Hypothesis," *Journal of Personality and Social Psychology* 55, pp. 52-77.

Cabanac, Michel (1992): "Pleasure: the Common Currency," *Journal of Theoretical Biology* 155, pp. 173-200.

Cabanac, Michel (1979): "Sensory Pleasure," *Quarterly Review of Biology* 54, 1, pp. 1-29.

Daly, M. and Wilson, M. (1988): *Homicide,* Hawthorne, N.Y., Aldine De Gruyter.

Ekman, Paul (1999): "Basic Emotions," in T. Dalgeish and M. Power (eds.), *Handbook of Cognition and Emotion*, Sussex, U.K., John Wiley & Sons Ltd.

Frank, R. H. (1988): *Passions Within Reason*, New York, W. W. Norton & Company.

Frank, R. H. (1990): "A Theory of Moral Sentiments," in J. J. Mansbridge (ed.), *Beyond Self-Interest,* Chicago, University of Chicago Press, pp. 71 -96.

Godfrey-Smith, Peter (1998): *Complexity and the Function of Mind in Nature*, Cambridge, U.K., Cambridge University Press.

Harvell, D. (1986): "The Ecology and Evolution of Inducible Defences in a Marine Bryozoan: Cues, Costs, and Consequences," *American Naturalist* 128, pp. 810-23.

Rhodes G. (2006): "The Evolutionary Psychology of Facial Beauty," *Annual Review of Psychology* 57, pp. 199-226.

Schmidt, Karen L. and Cohn, Jeffrey F. (2001): "Human Facial Expressions as Adaptations: Evolutionary Questions in Facial Expression Research," *Yearbook of Physical Anthropology* 44, pp. 3–24.

Shultz, A. W. (2009): "Sober & Wilson's Evolutionary Arguments for Psychological Altruism, a Reassessment," *Biology and Philosophy*, Published online: 30 August 2009.

Sober, E. and Wilson, D. S. (1998): *Unto Others: The Evolution and Psychology of Unselfish Behavior*, Cambridge MA, Harvard University Press.

Stich, S. (1978): "Beliefs and Sub-Doxastic States," *Philosophy of Science* 45, pp. 499-518.

Stich, S. (2007): “Evolution, Altruism and Cognitive Architecture: A Critique of Sober & Wilson's Argument for Psychological Altruism,” *Biology and Philosophy 22,* pp. 267-281.