Utility Comparisons Naturalized

Vladimir Vlaovic

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements for the degree of

Master of Arts

Department of Philosophy

University of Manitoba

Winnipeg

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

Utility Comparisons Naturalized

BY

Vladimir Vlaovic

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

OF

MASTER OF ARTS

Abstract

Interpersonal comparisons of utility (ICUs) are often deemed to be problematic—perhaps especially problematic when construed as empirical, rather than ethical, claims or propositions. The metaphysical worry surrounding ICUs is that they may not be possible, and that there is nothing in the world that could ground such propositions. E.g., if I assert that my utility is lower than yours, then there must be something that makes that assertion or proposition true; but what could that be? I propose an answer, viz. that what metaphysically grounds ICUs are facts about what contemporary neuroscientists call "learning." Moreover, I make a case for why we ought not to doubt that true, reliable ICUs are indeed possible.

Acknowledgements

For producing the present work, I am greatly indebted to my thesis committee, consisting of my advisor, Tim Schroeder, as well as Rob Shaver and Irwin Lipnowski. I owe an additional intellectual debt to Tim Schroeder whose published work is foundational to this thesis. Another intellectual dept is to Bob Bright, who first got me interested in interpersonal comparisons of utility, and whose unpublished manuscript on the topic made it easier for me to untangle an unfamiliar literature. For crucial financial support, I thank the following: Minister Diane McGifford, the Department of Advanced Education and Training, and the Manitoba Government for endowing me with the Manitoba Graduate Scholarship, as well as the Social Sciences and Humanities Research Council (SSHRC) for awarding me their CGSM research grant. I also thank my mother, Jasmina Jovanovic, for moral support, and my good friend, J. Brendan Richie, who helped me proofread parts of the thesis that are well outside the scope of his own philosophical research interests.

This work is dedicated to my good friend and thesis advisor, Timothy A. Schroeder.

Table of Contents

Chapter 1

Background

This project constitutes an effort to solve the so-called problem of interpersonal comparisons of utility (ICUs). From the time of John Stuart Mill, Jeremy Bentham, Stanley Jevons and others, to more recent treatments by people like Donald Davidson, Amartya Sen, Jon Elster, John Rawls, John Harsanyi, R.M. Hare and many others, a fundamental worry amongst economists and philosophers—often philosophers who are concerned with utilitarian moral theory—has been the question of how we are to compare the level of one person's utility to that of another.

Various solutions have been proposed in the literature of both economics and analytic philosophy, but none have adequately addressed the now-infamous problem of ICUs. My intention is to look for new sources of information that I believe have fruitful insight to offer into this issue—viz. the natural sciences and contemporary philosophy of mind—in hopes of finally doing away with this problem by soliciting the help of cutting-edge empirical findings.

If utility is identical to the satisfaction of preferences, as many contemporary utilitarians and virtually all welfare economists concerned with the subject think it is, then there is a risk-based procedure—viz. the John von Neumann and Oskar Morgenstern (1944) synthesis of von Neumann and Morgenstern cardinal utility—which some think allows us to arrive at numerical representations of how much an individual prefers one outcome over another; a procedure some have tried deploying in their proposals for solving the problem of ICUs. The applicability of the von Neumann and Morgenstern theory to choice behaviour under risk is itself highly controversial.

However, what's not controversial is that whatever the merits of the von Neumann and Morgenstern construction may be, it still gives us no way to compare how much any given person values something compared to any other person without various supplementary formal manoeuvres, some of which rely on implausible assumptions.

There is, however, a vast amount of information about the biological basis of desire and the satisfaction of desire that has in the past few years been explored by neuroscientists, and which, on my view, has the potential to resolve issues concerning ICUs and measurability of utility in general. Specifically, my belief is that if we gloss utility in terms of desire satisfaction, and adopt a specific, biologically-founded theory of desire which lends itself to the findings of cutting-edge natural science, and which is neither hostile to common sense, nor to philosophy, then we can come up with a theory of utility such that ICUs, at least in principle, pose no special conceptual difficulty. My work is primarily concerned with the metaphysics of ICUs, and only suggestive of any epistemological implications. In other words, that reliable ICUs are *possible* (and may even become *actual* in the future) is my concern, whereas I only mention in passing questions about whether we humans can in fact gain epistemic access to them.

Given the explosion of information coming from late twentieth, and early twenty-first century investigations of the biological basis of mental states, the possibility of being able to quantify utility in purely physical terms is more real than utilitarians themselves have ever imagined.

The Remaining Chapters

Chapter 2 constitutes a close look at one utilitarian economist's effort to deal with the problem of ICUs using transformed von Neumann and Morgenstern utility

functions. I am referring to the work of John Harsanyi (1977; 1982) which has received much attention in the philosophical and economics literatures. I argue that Harsanyi's effort fails, but in a manner that is suggestive of the need to pay attention, when theorizing about ICUs, to the ways in which actual human psychologies function. Chapter 3 is essentially a summary of the theory of the nature of desire mentioned above, viz. Tim Schroeder's (2004) Reward Theory of Desire. Schroeder's theory, though simple and elegant in its formulation, is interdisciplinary in origin with highly complex and at times surprising implications. Chapter 4 explores whether some of those implications can be applied to the problem of ICUs; specifically, implications concerning the nature of desire strength. Chapter 5 is a summary of the conclusions drawn in previous chapters.

Assumptions

Before proceeding with any substantive work, it seems fair to the reader for me to explicitly state the assumptions that I operate under throughout the remainder of this project. The most important of these include: (i) that Schroeder's (2004) Reward Theory of Desire is true; and (ii) that some kind of physicalism about the mind is true. By (ii) I simply mean that I accept the view that mental states are physical in nature; that there is nothing supernatural or non-physical about them, and that they are either identical to or biologically realized by brain states and processes. A related assumption is: (iii) that logical behaviourism (about the mind) is false. By 'logical behaviourism' I have in mind the view that there is nothing more to mental states than dispositions to behaviour.

Having given the reader some background about the nature of the present project, and the assumptions it operates under, it is time to proceed with chapter 2, and examine the nature of von Neumann and Morgenstern utility functions, as well as John Harsanyi's attempt to use them in an attempt to solve the problem of ICUs.

Chapter 2

In this chapter, we shall examine John Harsanyi's attempted reduction of interpersonal utility comparisons (ICUs) to a certain class of *intra*personal comparisons of utility, paying close attention to his (1977: Ch. 4) and (1982) discussions. The class of intrapersonal comparisons that Harsanyi claims ICUs are reducible to are hypothetical in the sense that they are of the form 'If such and such *were* the case, then the utility I *would* assign to some outcome *would be* such and such'. Ultimately, Harsanyi's 'reduction' fails, partly in virtue of being grounded in an implausibly formulated assumption about the fundamental similarities of different individuals' psychologies called the "similarity postulate" (Harsanyi 1982: 50). Nevertheless, the reduction's failure—in particular, its reliance on the similarity postulate—is suggestive of what a successful solution to the problem of ICUs must seriously take into consideration, viz. facts about the psychology of pleasure, or desire satisfaction, or preference satisfaction (depending on how one defines an individual's utility[1]). For now, however, let us have a close look at Harsanyi's proposed solution to the problem of ICUs.

Harsanyi's attempt at tackling the problem of ICUs takes the form of a reduction of ICUs to *intra*personal comparisons of utility, via what is sometimes called the mechanism of 'extended sympathy'—or, as Harsanyi often calls it, "imaginative empathy"[2] —and with the use of von Neumann and Morgenstern utility functions. Both of these aspects of the supposed reduction are highly controversial, as is the reduction

[1] Discussion on this topic is postponed for chapter 4 below.
[2] E.g., in his (1977: 51); and (1982: 50).

itself.[3] But let us hold off on the criticism, and try to get a handle on what it is that Harsanyi claims to have accomplished. The place to begin is some clarification of the nature of von Neumann and Morgenstern cardinal utilities,[4] since these constitute an integral part of Harsanyi's attempt to solve the problem of ICUs.

von Neumann and Morgenstern Utility

Suppose that Jones prefers some outcome *A* to some other (mutually exclusive) outcome or event[5] *B*, and that she also prefers yet another (mutually exclusive) outcome or event *C* to *A*. If all the data we have to go on is a purely *ordinal* ranking of Jones's preference for *A* over *B*, and her preference for *C* over *A*, then on the face of it, there seems to be no logical foundation for claiming that, e.g., her preference for *A* over *B* is *greater than* or *exceeds in magnitude* her preference for *C* over *A* nor for claiming that Jones will enjoy a greater increase in utility or welfare from obtaining *A* instead of *B* than she would from obtaining *C* instead of *A*. Putting all general scepticism surrounding the meaningfulness of propositions about preference strength aside, suppose that Jones does indeed prefer *A* to *B* more than she prefers *C* to *A*. On a purely ordinal conception of preferences, there seems to be no way to mathematically represent this fact (we are supposing that it is indeed a fact) about her relative strengths of preference vis-à-vis the outcomes *A*, *B*, and *C*.

But suppose now that Jones has a definite preference for certain *risky prospects;* more specifically, suppose that she prefers *A* to a gamble *G* such that if she opts for *G*,

---

[3] I say 'supposed reduction', because, as pointed out in MacKay (1986: 319-22), it's not at all clear in Harsanyi's writings what *exactly* is being reduced to what, nor is it clear whether any of the possible candidate reductions are in fact successful. More on these issues below.

[4] Throughout this chapter, the name 'cardinal utility' denotes the concept of cardinal or measurable utility, as synthesized by von Neumann Morgenstern. It is not to be confused with other meanings of 'cardinal' utility.

[5] Or state of affairs; the correct ontology of outcomes is, however, *not* relevant to the present discussion.

she has a 50% chance of obtaining *B* (her least favourite outcome), and a 50% chance of obtaining *C* (her favourite outcome). Now, a proposition expressing Jones's preference for the certain outcome *A* over the gamble *G*, according to von Neumann and Morgenstern, "contains fundamentally new information." (1947: 18) This information about Jones's preference for obtaining *A* for certain over the 50-50 gamble *G*, "provides a plausible base for the numerical estimate" that Jones's preference for *A* over *B* exceeds her preference for *C* over *A*, according to von Neumann and Morgenstern (1947: 18).

If we use gambles or lotteries between pairs of outcomes with all kinds of probabilities (for obtaining each of the two members of those pairs; e.g., 60-40, 61-39, 74-26, or whatever), von Neumann and Morgenstern insist that an even "more direct" route to mathematical representation of the strengths of an individual's preferences relative to one another avails itself:

> Consider three events, C, A, B, for which the order of the individual's preferences is the one stated. Let α be a real number between 0 and 1, such that A is exactly equally desirable with the combined event consisting of a chance of probability 1 – α for B and the remaining chance of probability α for C. Then we suggest the use of α as a numerical estimate for the ratio of the preference of A over B to that of C over B. (1944: 18)

Hence, if von Neumann and Morgenstern are right, the above procedure of combining outcomes with probabilities offers us a way of numerically representing, e.g., the fact that, other things equal, my preference for tea over lemonade greatly exceeds my preference for coffee over tea. Why? Because I would rather have tea for certain, than risk drinking lemonade, even if my chances of ending up with lemonade are quite low, and my chances of ending up with coffee (my favourite of the three) are quite high.

In keeping with the above quoted passage, let us flesh out the example of coffee, tea, and lemonade a bit further: Again supposing that I prefer coffee to tea, and tea to

lemonade, if von Neumann and Morgenstern are right, we are also entitled to suppose that there is a lottery between coffee and lemonade such that I am *indifferent* between getting it on the one hand, and getting tea for sure on the other. And, again suppose (as happens to be the case) that I in fact much prefer tea to lemonade, whereas I am almost indifferent between tea and coffee. Given these, i.e. my actual preferences, I happen to be indifferent between getting tea for sure, and gambling between coffee and lemonade, if the probability of ending up with lemonade is 0.1, and the probability of getting the coffee, which I most prefer, is 0.9. Hence, according to the von Neumann and Morgenstern procedure that's under discussion, we are entitled to accept 0.9 as a numerical estimate for the ratio of my preference for tea over lemonade, to my preference for coffee over tea.

Now let us assume that the von Neumann and Morgenstern procedure as (very roughly) outlined so far is unproblematic in the sense that it gives us a method of mathematically representing facts about how much individuals prefer outcomes vis-à-vis one another.[6] That is, let us assume that the procedure does indeed give us a method for assigning *cardinal*[7] utilities to various outcomes, as distinguished from the merely ordinal rankings with which we started.[8] By this assumption, we have a way to mathematically represent that I prefer coffee to tea a lot less than I prefer tea to lemonade; are we now entitled to also assume that the von Neumann and Morgenstern procedure provides a basis for mathematically representing that, say, *my* preference for

---

[6] The entirety of the present chapter may be taken as operating under this controversial assumption. (If the assumption is false, the solution to the problem of ICUs defended by me in subsequent chapters is largely unaffected, but the same cannot be said for Harsanyi's proposal, examined in the present chapter.)

[7] Again, as a reminder, 'cardinal' here must be understood in the sense of von Neumann and Morgenstern's construction, and not in any other sense.

[8] Started the von Neumann and Morgenstern procedure under discussion, that is.

tea over lemonade exceeds *Jones's* preference for Cadillacs over Lincolns? Far from it, according to von Neumann and Morgenstern! Why? Because, at best, the von Neumann and Morgenstern procedure (vNM procedure) gives us a representation of the ratios of my preference strengths relative to one another. This procedure, which is designed for assigning cardinal utilities to outcomes from purely ordinal information, does not in any way whatsoever give us a means to represent, e.g., that I prefer coffee to tea exactly—or even *roughly*—three times as much, or one third as much (or whatever) as Jones prefers Cadillacs to Lincolns.

Even if we had mathematical representations of all the preference strengths (relative to one another) for all possible outcomes as well as gambles between outcomes, and the same kind of representations for Jones, Smith, and the rest of humanity, we still have no way to rule out the possibility that, e.g., Smith's preferences are all exactly double the strength of mine, and exactly one third the strength of Jones's.

The reasons for the above claim about the inefficacy of the vNM procedure as a standalone method for comparing preference strengths across individuals requires further explanation, but before offering such explanation, it is only fair to von Neumann and Morgenstern to point out that they are in *explicit agreement* with this (as far as I know, uncontroversial[9]) claim. In a footnote, they caution their reader to make note of the fact that they "have not obtained any basis of comparison, quantitatively or qualitatively, of the utilities of different individuals." (1944: 19). And on a subsequent page, von Neumann and Morgenstern again, rather explicitly, state the following: "We re-emphasize that we are considering only utilities experienced by one person. These

[9] There is a certain sense in which the still unpublished work of Bob Bright is in disagreement with this claim, but discussion of his reasoning on this matter would take us off course in the present context.

considerations do not imply anything concerning the comparisons of the utilities belonging to different individuals." (1944: 29) Abstracting from the technical details behind von Neumann and Morgenstern's assertion that their procedure for assigning cardinal utilities to outcomes (from purely ordinal information) gives no basis for conducting ICUs, a brief explication of the reasons supporting such claims is called for at this point.

Some of the underlying reasons behind scepticism about comparing the von Neumann-Morgenstern cardinal utilities (vNM cardinal utility/utilities) that one individual enjoys with those enjoyed by another can be summed up, in a relatively non-technical way, by the following quotation from Stanley Jevons, first written by him more than seventy years[10] before the concept of vNM cardinal utility appeared in the literature:

> The reader will find...that there is never, in any single instance, an attempt made to compare the amount of feeling in one mind with that in another. I see no means by which such comparison can be accomplished. The susceptibility of one mind may, for what we know, be a thousand times greater than that of another. But, provided that the susceptibility was different in a like ratio in all directions, we should never be able to discover the difference. Every mind is...inscrutable to every other mind, and no common denominator seems to be possible. (1911: 14)

Given that the concept of vNM cardinal utility is based entirely on *ratios* of an individual's strengths of preferences (relative to one another), it is easy to see that Jevons's worry would apply in full force to any straightforward or naïve attempt to compare vNM cardinal utilities across individuals.

To clarify, there is nothing in von Neumann and Morgenstern's procedure that rules out the possibility referred to by Jevons in the above quoted passage; that is, given the mathematical nature (much of which I am skimming over in the present discussion)

---

[10] The first addition of Jevons (1911) appeared in print in 1871.

of that procedure, an infinite number of numerical representations of an individual's preferences and cardinal utility differences are equally admissible. The reason is that the procedure is simply not designed to represent any kind of non-relational reading of preference intensity; all it represents is the cardinal utility that an individual assigns to outcomes in relation to the cardinal utility she assigns to other outcomes. Hence, a von Neumann and Morgenstern *utility function* (vNM utility function) that assigns numbers—i.e. vNM cardinal utilities—to outcomes, and represents an individual's preference strengths relative to one another is said to be unique *up to increasing linear transformation.* In other words, if U is a vNM utility function representing an individual's preferences over some set of outcomes, then so is any $U` = aU + b$ (such that $a>0$).[11]

We can also think of uniqueness up to increasing linear transformation in terms of the von Neumann and Morgenstern measurement procedure that we've been discussing yielding a scale that is unique up to the selection of a zero-point and a unit of measurement; selection of the positive constant (denoted by 'a' in the above equation) fixes the unit, while selection of b (denoted by 'b' in the above equation) fixes the origin.[12] In commonsense (i.e., even less technical) terms, we can take this to mean that one can add or multiply an individual's vNM cardinal utility function by any positive number, as long as the addition or multiplication is done across the board, so to speak—

---

[11] Thanks to Bob Bright, whose work—particularly chapter 3 of his unpublished *Foundations of Utilitarianism*—and input were crucial to my ability to conduct the present exposition of vNM functions in a non-technical and approachable fashion. I owe him a great deal of acknowledgement for much of this chapter. Compared to his discussion, mine only scratches the surface when it comes to vNM functions (in particular, their potential role in a formulation of philosophical utilitarianism, a topic not under discussion here).

[12] This exposition is indebted to Bright (chapter 3 of his unpublished *Foundations of Utilitarianism*).

as long as the addition or multiplication is done to the whole function (in the manner outlined in the equation above).

To help further illustrate what is meant by 'the uniqueness of vNM cardinal utility functions up to increasing linear transformation', it may be helpful to consider an example:[13] Suppose that Jones is indifferent between getting some outcome *A* for sure, and a gamble such that if she takes it, she has a 70% of obtaining some other (mutually exclusive) outcome *B*, and a 30% chance of obtaining yet another (mutually exclusive) outcome *C*. If we arbitrarily assign the value 1 to outcome *B*, and 0 to outcome *C*, then we must assign a value of 0.7 to *A* in order to explain Jones's choices in terms of her maximizing her expected utility (since 70%×1 + 30%×0 = 0.7). But notice that we could have chosen, e.g., the values 3 and 0 for *B* and *C* respectively, in which case *A* would have a value of 2.1 and the difference between *U(B)*[14] and *U(C)* would be 0.9; or 100 and -5 (in which case *U(B)* - *U(C)* = 100 - 68.5 = 31.5).[15]

To recap, the numbers—i.e. vNM cardinal utilities—assigned to outcomes by a person's vNM cardinal utility function are by definition without meaning, *except vis-à-vis other numbers assigned by the same function belonging to that same individual.* My vNM cardinal utility function assigns a certain cardinal utility to obtaining coffee instead of tea. But, on its own, that vNM cardinal utility (any particular vNM cardinal utility for that matter) means and represents nothing. It gets its representational content when considered vis-à-vis the strengths of all the other preferences that underlie the values (vNM cardinal utilities) my vNM utility function assigns to the alternatives specified by those other preferences. Moreover, the function itself is unique only up to

---

[13] The example is taken from chapter 3 of Bright's unpublished *Foundations of Utilitarianism.*
[14] '*U(α)*' here denotes the vNM cardinal utility assigned to outcome α.
[15] Another example borrowed from chapter 3 of Bright's unpublished *Foundations of Utilitarianism.*

increasing linear transformation, meaning that there is an infinity of equally admissible scales to measure an individual's vNM cardinal utilities such that all that's *in*variant between possible scales (for a given individual) are the ratios of preference strengths with respect to one another. These facts about possible and equally admissible scales for vNM utility functions preclude us from interpersonally comparing vNM cardinal utilities, or the strengths of the preferences underlying them. Hence, the numerical values (vNM cardinal utilities) assigned to outcomes by an individual's vNM utility function could be radically different from scale to scale, yet have exactly the same representational content, because in constructing these utility scales, we arbitrarily choose the zero-point and unit of measurement. The only way we could meaningfully compare the values assigned to outcomes by Jones's vNM utility function with those assigned by Smith's, is if we could fix a zero-point and a unit of measurement that's invariant across distinct individuals' scales; i.e., a *non*-arbitrary zero-point and unit of measure.[16] Absent some objectively fixed zero-point and unit of measure, on the face of it, different individuals' vNM utility functions seem un-comparable—vis-à-vis *one another,* they represent nothing.

Given only the rough and somewhat informal outline of the nature of vNM cardinal utilities and vNM functions that I've sketched in the above paragraphs, we should already find it at least prima facie puzzling why Harsanyi thinks he can (in part) rely on these mathematical constructions to compare different individuals' utilities. The next order of business is to introduce and discuss the nature of his attempt to do so via

[16] If we fix only the unit of measurement for distinct individuals' vNM functions, the best we can get is a comparison of utility differences; if we fix both that, and a zero-point, then utility levels are also comparable. This distinction, as shall become evident later in the discussion (chapter 4 especially), is not hugely relevant to the present project.

the behavioural-psychological technique of extended sympathy, and an a priori assumption he sometimes[17] calls the Similarity Postulate.

Harsanyi and Imaginative Place-trading Thought Experiments

With the above exposition of vNM cardinal utilities (their nature and genesis), and vNM utility functions in mind, suppose I have epistemic access to everyone's vNM cardinal utility functions, where 'everyone'[18] ranges over all members of whatever population or society I happen to be a member of. From what's been said in the preceding section, we know that this information is, on its own, insufficient as a logical basis for my claim that I would derive more utility from obtaining some outcome *A* than Jones would from obtaining some outcome *B*; or even for the claim that I prefer *A* to *C* more strongly than Jones prefers *B* to *D*.

But what if I knew exactly what objective circumstances Jones is in, and also knew all the psychological laws that govern human behaviour; would I then have a logical basis for making claims of the above form, and claims of any similar form (i.e. claims that are, incorporate, or are grounded in ICUs)? Harsanyi seems to thinks so, and seems to think that if I were to have all the above information, then I could effectively *reduce* any *inter*personal comparison of utility (e.g., of mine and Jones's utility vis-à-vis certain outcomes) to some kind of counterfactual, *intra*personal comparison involving the utility that my own vNM utility function *would assign* to some outcome, if I *were* in certain hypothetical circumstances. The remainder of this chapter is an

[17] E.g., see Harsanyi (1982: 50).

[18] Throughout the remainder of this section, unless explicitly noted otherwise, I use 'everyone' to denote all members of a population or society whose vNM utility functions enter into the definition of a social welfare function.

effort to explicate, in as much detail as needed, this so-called reduction of Harsanyi's, and then to criticize it.[19]

Harsanyi needs ICUs for a number of purposes, including arguments for his (1977: Ch. 4; 1982) case in support of the variety of utilitarianism that he favours as a plausible ethical theory. Moreover, he thinks that in general, individuals' moral value judgments concerning what he calls "social situations"—which include, inter alia, "alternative patterns of social behaviour (alternative moral rules), alternative institutional frameworks, alternative governmental policies, alternative patterns of income distribution..." (1977: 49)—can be mathematically represented by an additive, "social welfare function" (1977: 50). Any individual i's evaluation of the social merit associated with some situation A is represented by her social welfare function W_i, if i considers the values assigned to A by every individual's (including her own) vNM cardinal utility function, adds these values,[20] and takes their arithmetic mean. Constructing a social welfare function W_i for any particular individual i involves i conducting ICUs; Harsanyi is himself clear enough on this point:

> Our model is based on the assumption that, in order to construct his social welfare function W_i, each individual i will try to assess the utilities $U_j(A)$ that any *other* individual j would derive from alternative social situations. That is, he will try to make *interpersonal utility comparisons.* Moreover, we have assumed that i will attempt to assess these utilities $U_j(A)$ by some process of *imaginative empathy,* i.e., by imagining himself to be *put in the place* of individual j in social situation A. (1977: 51)[21]

When Harsanyi makes reference to one individual i being "put in the place of" a different individual j, he has in mind i imagining herself "to be placed in the objective

---

[19] My wording may be a bit imprecise here: it's not so much that Harsanyi claims to have reduced ICUs to intrapersonal comparisons, as that he thinks this is what we in fact do whenever we attempt to conduct everyday ICUs from the armchair. I read him as taking his theory to be an explanation of what he supposes are actual phenomena (i.e., implicit reductions of ICUs to intrapersonal comparisons), perhaps more so than an argument in favour of performing the reductions; see Harsanyi (1977: 293, Note 4).

[20] How can she do that, given the arbitrary nature of the zero-values and utility units of different people's vNM utility functions? The question is answered shortly.

[21] Italics are Harsanyi's.

conditions (e.g., income, wealth, consumption level, state of health, social position) that *j* would face in social situation *A*." (1977: 52) But also, *i* is expected to assess the aforementioned conditions "in terms of *j*'s own *subjective attitudes* and *personal preferences* (as expressed by *j*'s own utility function U_j)—rather than assessing them in terms of *i*'s own subjective attitudes and personal preferences (as expressed by his own utility function U_i)." (1977: 52) Can *i* really do what Harsanyi is asking of her here? That is, can some individual really assess some situation in terms of *another* individual's preferences (as represented by that person's vNM utility function) while pretending to be in the latter's objective circumstances, and do so "by some process of imaginative empathy"? Let us postpone the question, but only for a short while.

The postponed question aside, there is the distinct issue of how the additions of different individuals' vNM cardinal utilities are even made *possible*, on Harsanyi's model, given that the values (vNM cardinal utilities) assigned by any individual *i*'s vNM utility function U_i are meaningful *only in relation to other values assigned by* U_i (absent some non-arbitrarily chosen zero-value and unit of measure). Formally, how is an individual conducting a moral evaluation, on Harsanyi's model, supposed to add different individuals' vNM cardinal utilities? The short answer is that, on Harsanyi's utilitarian model, she must select a transformation for each individual's vNM utility function so that they are all expressed in the same utility unit:

> ... when an individual *i* is constructing his social welfare function W_i, the only way that he is really required to make interpersonal utility comparisons is by trying to compare the utility units of the different individuals' utility functions $U_1, \ldots, U_i, \ldots, U_n$. Suppose that he begins with individual utility functions $U_1, \ldots, U_i, \ldots, U_n$ expressed in arbitrary (and therefore in general presumably unequal) utility units. Then his basic task is to choose conversion ratios $q_1, \ldots, q_i, \ldots, q_n$, which in his best judgment will convert all these utility functions into the same common utility unit, by setting $U_1^* = q_1 U_1, \ldots, U_n^* = q_n U_n$. (Of course, he can always choose $q_i = 1$. That is, he need not change the utility unit in which he is expressing his own utility.) (1977: 57)

Now, given the prima facie difficulty of, in a sense, normalizing an entire population's vNM utility functions so that they assign values (vNM cardinal utilities) to outcomes expressed in the same unit, one expects wide disagreement with respect to the conversion ratios (transformations) that are supposed to allow moral evaluators to conduct such normalizations. And when there *is* disagreement between a moral evaluator h and another evaluator i about how to convert (perform transformations on) the population's vNM utility functions, one surely wonders whose social welfare function (W_h or W_i) to trust—i.e. whose transformations most faithfully reflect, or are consistent with, the facts about how much importance individuals actually assign to outcomes.

The question of how to choose the 'best' transformation—and what constitutes a 'good' (or 'the best') transformation for that matter—are important to the present discussion, because the selection of these transformations is precisely what's, formally, supposed to bridge the gap between ordinary vNM functions, and interpersonally comparable utilities, on Harsanyi's model. Harsanyi's discussion of what he calls "interobserver validity" (1977: 57-60) seems to be an attempt to flesh out some details concerning this bridging.

We can think of an ICU done by some individual Jones as being 'valid'[22]—in the sense of Harsanyi's interobserver validity—in terms of it being the case that, in constructing her social welfare function, Jones converts the utility units that the rest of society's vNM utility functions are expressed in using what one might call 'the right' conversion ratios. And for Jones's conversion ratios $q_1, \ldots, q_n$ to be *right* in the present

[22] It may be more accurate to say 'perfectly valid' here, since validity in the present sense admits of degrees (not to be confused with logical validity).

sense of 'right' (or, 'valid', in Harsanyi's terms), it must be the case that, once multiplied by $q_1, \ldots, q_n$ the vNM utility functions $U_1, \ldots, U_n$ of the rest of the population assign to outcomes numerical values that accurately represent how much importance individuals 1, … *n* actually attach to those outcomes.

Notice just how far beyond the representational content of ordinary vNM functions themselves Harsanyi aims to go. We know from the previous section of the present chapter that vNM functions numerically represent how much individuals prefer outcomes, *but only vis-à-vis other preferences for other outcomes.* An instance of some individual *i* successfully applying a 'perfect' conversion ratio q_j to some other individual *j*'s vNM utility function U_j would, on Harsanyi's model, amount to that now-converted (now-transformed) utility function representing, not just how much *j* prefers that some state of affairs *A* obtain[23] instead of some other (mutually exclusive) state *B* *as compared to her other preferences,* but instead how much she prefers *A* to *B* *simpliciter*. That is, Harsanyi thinks that we can get beyond the purely relational sense in which vNM functions represent the magnitude of an individual's preferences—i.e., that we can ascertain the magnitude of an individual's preferences in some *absolute* sense of preference strength.

Harsanyi's model actually suggests that he doesn't just think we can *in principle* get beyond the representational content of ordinary vNM functions, rather that we *in fact* do so (with varying success rates) every time we make a genuine moral value judgment. So how do we do that? Formally speaking, how does Harsanyi think we, at least sometimes, choose the right conversion ratios and ipso facto dramatically increase the representational content of individuals' vNM functions in such a way as to make

[23] Or that some event E take place, again, depending on the correct ontology of outcomes.

them interpersonally comparable? *In practice,* Harsanyi's answer is that we in fact conduct ICUs "by some process of *imaginative empathy*" (1977: 51). And if, on his utilitarian model, performing ICUs amounts to choosing conversion ratios, he must think that imaginatively empathising with one another is the process by which we choose the best ratios we can. But, we may now ask, how satisfactory of an answer is this? It may be rather hasty of Harsanyi to simply *assume* that imagining to be in someone else's place allows us to get beyond the representational content of her vNM function. That is, it may be implausible to suppose that the process Harsanyi calls "imaginative empathy" is of any help to us in trying to select the best transformation for an individual's vNM utility function (i.e., to select a conversion ratio for that individual) that we can.

One aspect of the worry is that it may simply be false to suppose that we do in fact ordinarily attempt to assess the utility an individual derives from some outcome by imagining to be in her objective circumstances with her subjective preferences. In other words, it seems far from obvious that we don't carry out such assessments by some entirely distinct process instead, in which case, Harsanyi's is a model of a non-existent phenomenon. But clearly Harsanyi thinks it *is* obvious:

> Simple reflection will show that the basic intellectual operation in such interpersonal comparisons is imaginative empathy. We imagine ourselves to be in the shoes of another person, and ask ourselves the question, 'If I were now really in *his* position, and had *his* taste, *his* education, *his* social background, *his* cultural values, and *his* psychological make-up, then what would now be *my* preferences between various alternatives, and how much satisfaction or dissatisfaction would *I* derive from any given alternative?' (Harsanyi 1982: 50)[24]

The above passage should be suggestive of why Harsanyi never *defends* his claim that the thought experiment he describes here (i.e. what he calls "imaginative empathy") is how we actually conduct everyday ICUs: he thinks "simple reflection" shows it to be

[24] Italics are Harsanyi's

the case that we do so. But why? I don't remember ever asking myself questions like the one Harsanyi says I ask myself all the time, and on reflection, I find no reason to believe that I do or have done so *implicitly* either. With Harsanyi, we must indeed admit that we do make quick, rough and ready ICUs, and do so very frequently, with varying success rates; but before familiarizing myself with the philosophical problems associated with ICUs and, specifically, Harsanyi's take on the surrounding issues, it had never so much as occurred to me to *substitute* myself for another individual in conducting a commonsense ICU; and, again, I do not believe that I do or have done so implicitly either. I for one have no idea how exactly I conduct quick, everyday ICUs; all I know is that it happens *somehow*. And simple reflection has never told me that it's by, implicitly or explicitly, imagining to be in someone else's shoes; when thinking of how much benefit an outcome would be to *Jones*, never have I thought about how much *I* would benefit from the same outcome *in whatever circumstances,* just so that I can estimate *Jones's* potential benefit. Is Harsanyi about to tell me that I am wrong about this, and that I haven't reflected enough? If he is, then so much the worse for his account, I am inclined to say. After all, it's supposed to be *simple* reflection doing the philosophical work needed to lend plausibility to imaginative empathy as a theory of commonsense ICUs. And if my own episodes of simple reflection failed to map onto something that Harsanyi assumes they would map onto—viz. that I actually conduct everyday ICUs by performing the Harsanyian thought experiment (imaginatively trading places with people)—it is his assumption that's wrong, at least with respect to one individual.

But such autobiographical observations do not, of course, constitute a knockdown argument against imaginative place-trading as a theory of how commonsense ICUs are *generally* done. Perhaps many people *do* conduct quick, everyday ICUs using something like Harsanyian imaginative empathy. But even if they do, *should* they? Individuals surely *ought* to imaginatively trade places with one another the way Harsanyi thinks they in fact do, if by so doing, they find themselves in a better epistemic position to assess others' utilities than they would be without the imagined exchange. But are we in any better an epistemic position to assess one another's utilities as a result of Harsanyi's place-trading thought experiment?

Harsanyi does indeed maintain that, at lest in some instances, imaginatively trading places with an individual is a reliable process for arriving at an estimate of the absolute[25] utilities that she assigns to various outcomes. According to Harsanyi, generally, "if we have enough information about a given person, and make a real effort to attain an imaginative empathy with him, we can probably make reasonably good estimates of the utilities and disutilities he would obtain from various alternatives." (Harsanyi 1982: 50) Now I think it is clear enough that the reason Harsanyi has any degree of confidence in the reliability of imaginative empathy when it comes to assessing different people's 'absolute' or 'true' utilities—i.e. the utilities their vNM functions would assign to outcomes, if the right transformations were applied to those functions[26] —must be his "similarity postulate" (SP). Consider his definition of SP, and the theoretical role he assigns to it:

---

[25] *Absolute* as opposed to 'in relation to other preferences for other outcomes' which, as we know, is the most that a vNM cardinal utility function can represent on its own.
[26] In Harsanyi's terms, if we selected the best *conversion ratios*.

> … any interpersonal utility comparison is based on what I will call the *similarity postulate*, to be defined as the assumption that, once proper allowances have been made for the empirically given differences in taste, education, etc., between me and another person, then it is reasonable for me to assume that our basic psychological reactions to any given alternative will be otherwise much the same. (Harsanyi 1982: 50)

Now if something like SP is right, inter alia there may be at least some measure of intuitive plausibility to the project of reducing ICUs to *intra*personal comparisons of utility (intraCUs): Deep down, we're all pretty much the same, so there's no reason to think that the satisfaction I derive from some outcome is any different in magnitude from the satisfaction you *would* derive from the same outcome *if* you instantiated all my superficial properties and stood in all my superficial relations, goes the intuition. But before we ask whether, at the end of the day, SP is helpful to the problem of ICUs (or even a plausible assumption in its own right), notice how central a role it plays in Harsanyi's insistence on the reducibility of ICUs to intraCUs: Not only is SP supposed to ground our commonsense ICUs—supposedly done via imaginative place-trading thought experiments—but also to serve as the foundation of the *explicit,* formal reduction of ICUs to counterfactual intraCUs undertaken by Harsanyi in his (1977: Ch. 4).[27] According to Harsanyi:

> … the possibility of meaningful interpersonal utility comparisons will remain, as long as the different individuals' choice behaviour and preferences are at least governed by the *same basic psychological laws.* For in this case each individual's preferences will be determined by the same general causal variables. Thus the differences we can observe between different people's preferences can be predicted, at least in principle, from differences in these causal variables, such as differences in their biological inheritance, in their past life histories, and in their current environmental conditions.
>
> … individual *i* will be able in principle to reduce any *inter*personal utility comparison that he may wish to make between himself and individual *j* to an *intra*personal comparison between the utilities that he is in fact assigning to various situations and the utilities that he would assign to them if the vector of the causal variables determining his preferences took the value R_j (which is the value that the vector of these causal variables takes in the case of individual *j*). (Harsanyi 1977: 58-9)[28]

---

[27] This is of course not surprising, given that Harsanyi's formal reduction is supposed to mathematically represent what actually takes place when we conduct quick, everyday ICUs using only the resources of commonsense, viz. a reduction to intraCUs.

[28] Italics are Harsanyi's

And, Harsanyi's ultimate conclusion is that:

> … given enough information about the relevant individuals' psychological, biological, social, and cultural characteristics, as well as about the general psychological laws governing human behaviour, *inter*personal utility comparisons in principle should be no more problematic than *intra*personal utility comparisons are between the utilities that the *same* person would derive from various alternatives under different conditions. (Harsanyi 1977: 58-9)[29]

What this conclusion and the remarks preceding it suggest is that the possibility of reducing ICUs to counterfactual intraCUs, on Harsanyi's view, indeed depends on the truth of SP or something similar—i.e. on interpersonal invariance of basic psychological laws. Formally, Harsanyi thinks that if we factor out a vector R_i that consists of all the causal variables needed to explain an individual i's preferences (*personal* preferences in Harsanyi's terms), what we'll be left with is something he calls an "extended" utility function: "Because the mathematical form of this function is defined by the basic psychological laws governing people's choice behaviour, this function V_i must be the same for all individuals i" (1977: 58). And, of course, if all individuals have a common utility function, then there are no formal problems associated with conducting ICUs that aren't also present in the case of intraCUs. All individuals, supposedly, *do* share a common utility function, viz. the one dependent on and only on the psychological laws that govern human behaviour; therefore, according to Harsanyi, there is no logical obstacle to comparing utilities across individuals, as long as something like SP is true, securing the existence of an interpersonally invariant utility function.

What Harsanyi's (1977: Ch. 4) and (1982) discussions—particularly their emphasis on the importance of SP—suggest, is that Harsanyi's view seems to be that the problem of ICUs is largely an *informational* one, rather than a *logical* one, *so long*

[29] Italics are Harsanyi's.

as something like SP holds: if we knew enough facts and laws, we could reduce ICUs to intraCUs and thereby avoid all the obstacles uniquely associated with the former. We ought to emphasize at this point that, by Harsanyi's own lights, imaginatively trading places with another individual turns out not to be necessary and may not even be relevant to conducting reliable ICUs, *given enough information*, since there is nothing that imaginatively trading places with a person can tell us that historical and psychological research into that person can't.

That extended sympathy is, in principle, a superfluous information gathering tool, if it *is* an information gathering tool at all (something I rather seriously doubt), brings to light a somewhat confusing aspect of Harsanyi's discussions, one that's been pointed out by MacKay: "How is Everyman's ability to put himself imaginatively in another person's shoes related to such esoteric matters as values of vectors of causal variables, basic psychological laws, and postulates of similarity?" (1986: 318) I have suggested an answer to MacKay's question, viz. that SP is supposed to be the foundation of our ability to put ourselves in others' shoes, and that it is also supposed to ground Harsanyi's formal reduction of ICUs to intraCUs in something like the following way: We in fact reduce ICUs to intraCUs by imaginatively trading places with people during the course of conducting any quick, everyday ICU; and, as long as something like SP is true, our everyday reductions are a fairly good way to assess one another's utilities, since SP assures us that we're all psychologically quite similar. And, according to the Harsanyian picture, the everyday reduction can be formalized, again, *as long as something like SP is true*, securing the existence of an interpersonally invariant utility function. Moreover, we wouldn't need the everyday reduction at all, if

we knew enough facts about the variables that causally determine an individual's preferences, since once these are accounted for, SP dictates that we'll all psychologically react in pretty much the same ways. In his commentary, MacKay seems to read Harsanyi similarly:

> We do have the ability to make some counterfactual judgments about what our psychological reactions would be in certain situations, often in situations very different from those which actually obtain. We don't, normally, do this by explicitly calculating the interaction of causal variables and environment according to psychological law. But, presumably, our capacities are based on (or reflect at an intuitive level) some such structures… There is nothing that performing the (self) manoeuvre[30] can tell us intuitively, that explicit, scientific calculation cannot tell us, in principle, if we know all the causal variables and psychological laws governing their interactions with various environmental situations, and are clever enough to work it out. But we don't and aren't, so we use the substitute, intuitive method instead. (MacKay 1986: 320-21)

If this is the right way to read Harsanyi, then we are correct to notice that, on his view, episodes of imaginative place-trading are not *prerequisites* to making ICUs, but are mere substitutes for the calculations that the limits on our knowledge base and information processing abilities preclude us from performing explicitly. So are they good or bad substitutes?

Putting aside my own intuitions on the matter, according to which imaginative empathy is *not* the everyday process by which we ordinarily conduct everyday ICUs at all, Harsanyi is misguided to think that such a process is an especially *good* one when it comes to assessing other people's utilities. I follow MacKay in this respect, who argues that when it comes to attempts at assessing others' utilities, there is no epistemic advantage to be had from imaginatively trading places with them:

> In making the total-objective-and-subjective-exchange supposition, I thereby lose touch with whatever it was about myself that offered the epistemic advantage in 'reducing' the interpersonal to the intrapersonal. And vice versa. If I do stay in touch with what it is about my relation to myself that provides epistemic advantage to performing the reduction, I do not totally make both the objective and subjective exchange suppositions. The 'me' that appears in these extreme hypotheses—a person in *his* position, with *his* taste, *his* education, *his* social background, *his*

[30] This is only MacKay's shorthand for saying 'imaginatively trading places with someone'.

> cultural values, and *his* psychological makeup—would indeed display *his* reactions, but wouldn't be a 'me' toward which I, the investigator, had any special epistemic authority. I, the investigator, would be no better off asking questions about this 'me' that I would about the other person. (1986: 322)[31]

MacKay's observations seem quite appropriate given the nature of the "extreme hypotheses" Harsanyi has us entertaining. Several paragraphs earlier, I asked whether it is reasonable to expect an ordinary individual with merely the resources of commonsense to even have the *ability* to assess a situation *entirely* in terms of another individual's tastes, while also pretending to be entirely in the latter's objective circumstances. My own inclination is to doubt whether human minds are in fact capable of genuinely making what MacKay describes as the *total*-objective-and-subjective-exchange, in light of, inter alia, the sheer quantity of information we are supposed to pretend is actually about us when we imagine to be entirely in someone else's shoes (not to mention, base new beliefs on that information). I have no doubt that we have *some degree* of ability to put ourselves in others' shoes in the sense required by Harsanyi's thought experiment; what I find implausible is the claim that we are capable of performing a *wholesale substitution* of our actual psychologies for imagined ones, and then forming new beliefs based *entirely* on the substituted, pretend psychology.

But even if we are capable of successfully making the wholesale imaginative exchange and basing beliefs on it, MacKay is right to point out, in the above quoted passage, that we have no reason to think that assessments of utilities based on these informationally extravagant thought experiments are any more *likely to be true* than assessments we might make without their 'help'. That is, granting that we are in fact *capable* of making the Harsanyian imaginative exchange, beliefs about others' utilities

---

[31] Italics are original to text.

that we form based on such a thought experiment don't seem as though they're likely to be true *in virtue of having been produced by the Harsanyian thought experiment.* After all, as MacKay notes (1986: 321-22), we have no reason to suppose that we retain the kind of special epistemic authority that we normally have towards ourselves and our actual mental states during the course of a Harsanyian thought experiment. In *pretending* that whatever properties and relations make Jones herself and not me are actually instantiated and stood in by me, I don't thereby gain Jones's first-hand knowledge of what it's like to instantiate those properties and stand in those relations. So, in trying to assess the absolute[32] utility Jones actually assigns to some outcome, why ask myself questions about me-with-Jones's-properties-and-relations, rather than simply ask myself questions about Jones?[33]

Assuming that we are right to conclude with MacKay that, at the end of the day, putting ourselves in others' shoes is, in itself, of no help to us in trying to solve the problem of ICUs, what about Harsanyi's insistence that SP secures the *in principle* possibility of interpersonally comparing utilities? The most straightforward answer is that SP, as construed by Harsanyi, is entirely unhelpful as part of a solution to the problem of ICUs. Let us take a closer look at SP itself, along with Harsanyi's grounds for accepting it as readily as he does.

Harsanyi and the Similarity Postulate

Recall that Harsanyi defines SP as "the assumption that, once proper allowances have been made for the empirically given differences in taste, education, etc., between

---

[32] Absolute in that it goes beyond the representational content of an ordinary vNM cardinal utility, in sense described above.

[33] MacKay makes similar observations; e.g. "You cannot actually satisfy conditions merely by imagining them satisfied. Does Harsanyi think you can?" (1986: 319)

me and another person, then it is reasonable for me to assume that our basic psychological reactions to any given alternative will be otherwise much the same." (1982: 50) So, e.g., if we take into account the superficial differences that make Bill Clinton who he is, and make me who I am, Harsanyi says it is reasonable to assume that Clinton and I will psychologically react in more or less the same way to any situation. Do we have any reason to accept SP, other than the fact (if it is one) that it's the key to reducing ICUs to intraCUs? Well, as Harsanyi construes it, SP "by its very nature, is not open to any direct empirical test" (1982: 51), so we certainly can't go peeking around the world in order to check whether it's true or not. But, as we shall see below, Harsanyi's grounds for classifying SP as an unverifiable assumption may turn out to rest on some evidently anti-physicalist intuitions about the nature of mental states, in which case, those grounds are suspect. Moreover, his reasons for thinking that SP is nonetheless a plausible assumption to operate under are also unconvincing. Consider this:

> I may very well assume that different people will have similar psychological feelings about any given situation, once differences in their tastes, educations, etc. have been allowed for. But I can never verify this assumption by direct observation since I have no direct access to their inner feelings.
> Therefore, the similarity postulate must be classified as a nonempirical a priori postulate…
>
> … Its intuitive justification is that, if two individuals show exactly identical behaviour – or, if they show different behaviour but these differences in their observable behaviour have been properly allowed for – then it will be a completely arbitrary and unwarranted assumption to postulate some further hidden and unobservable differences in their psychological feelings. We use this similarity postulate not only in making interpersonal utility comparisons but also in assigning other people human feelings and conscious experiences at all… When we choose the assumption that we actually live in a world populated by millions of other human beings, just as real and just as conscious as we are ourselves, then we are relying on the same similarity postulate. (Harsanyi 1982: 51)

Much has gone wrong here.

First, there is Harsanyi's[34] claim that we already rely on SP in order to avoid some kind of scepticism about other minds, or about physical reality external to our own minds; this amounts to Harsanyi missing the point. If someone is sceptical about ICUs being grounded in facts (presumably facts about mental states), pointing out that we generally operate under the assumption that people *have* mental states, as Ilmar Waldner has observed (1972: 92-3), will not alleviate her concern. Waldner notices that sceptics about the existence of factually-grounded ICUs are worried because "they do not believe there is any empirical evidence that could support interpersonal comparative judgments about the intensities of certain kinds of mental states—even though there is evidence for other kinds of judgments about other persons' mental states." (1972: 92) Hence, I can operate under the assumption that, e.g., Bill Clinton is a physical object with mental states, and so basically like me, and still think there is no factual basis for claims about which transformation of his vNM function best represents his actual interests.

Second, there is a serious problem with what Harsanyi takes to be the "intuitive justification" of SP. According to Harsanyi, it is "completely arbitrary and unwarranted" to suppose that two people exhibiting exactly similar observable behaviour have notable differences between their respective mental states (1982: 51). We would be perfectly justified in assuming that two people displaying the same behaviour are in type-identical mental states, *if logical behaviourism were true*.[35] But if the possibility of conducting factually grounded ICUs depends on the truth of behaviourism about mental states, then we who think that ICUs are perfectly

---

[34] This aspect of Harsanyi's acceptance of SP is borrowed from an earlier discussion of his (1955), but ultimately from I.M.D. Little's (1950: 56-7) arguments.

[35] Thanks to Bob Bright for pointing this out before I ever had a chance to notice it on my own.

respectable empirical claims are in hot water, in virtue of being committed to one of the more implausible theses about the nature of minds.

Finally—and this I believe to be at the core of what's wrong with Harsanyi's proposed reduction as well as what's right about the proposal that I defend in subsequent chapters—there is Harsanyi's insistence that SP "by its very nature, is not open to any direct empirical test." (1982: 51) In principle, there is no reason to think that we can't observe one another's mental states, including psychological reactions to outcomes, as long as *physicalism* is a true thesis in the philosophy of mind. By 'physicalism' I mean the view according to which *mental* states are *physical* states, either identical to, or biologically realized by brain states. If physicalism so construed is true, then it follows that, rather than a "nonempirical a priori postulate" (Harsanyi 1982: 51), SP is a perfectly ordinary *empirical hypothesis,* open to empirical test. Yet, if we have the (in principle) ability to directly observe any individual's psychological reaction to a given alternative, where is the theoretical payoff for adhering to a sweeping, generalized assumption like SP? If physicalism is true, we could, in principle, devise elaborate tests to discover whether SP is a reasonable assumption, but why on earth would we bother if what we're interested in are ICUs? Under the physicalist assumption that mental states (or in the very least, their biological underpinnings) are observable, rather than investigating whether SP is right, we can isolate the relevant psychological reactions themselves, and measure these directly. Our only problem seems to be that we don't seem to as of yet know which mental states to measure, where in the brain they're located, or what to look for.

Perhaps our ignorance of the nature and biological basis of certain mental states is what leads Harsanyi to say of ICUs that they "do pose important philosophical problems" which he describes as "the problem that they require us to use what I have called the similarity postulate." (1982: 50) But if we knew which brain states and processes to look for, we would have no reason to bother with questionable postulates like SP, and no reason to bother with questionable 'reductions' based on it. We could instead accept a far more plausible assumption, viz. that physicalism is true, and then, based on *that* assumption, directly measure whatever brain states or processes are identical to, or realize the psychological reactions that individuals have to various outcomes.

But notice that by assuming the truth of physicalism, we have effectively traded one a priori assumption (SP) for another (physicalism). Hence, the question becomes which of the two assumptions—physicalism or SP—is the easiest to swallow. If we must choose, the answer is surely that the assumption contemporary philosophy of mind has already given us many good reasons to accept ought not to be traded for the one whose sole theoretical payoff seems confined to facilitating Harsanyi's reduction. Yet, there is a worry with such a line of argument, since there may be a plausible reading of Harsanyi that's consistent with SP, as well as physicalism. In other words, rather than as a behaviourist, Harsanyi may be read as himself committed to physicalism, but not necessarily to a behaviourist interpretation of physicalism. On such a reading, SP can be taken to mean something like the following:

> SP`: If two individuals were to have exactly similar casual histories and exactly similar physical compositions, and were in exactly similar objective

circumstances, then they would be in type-identical mental states when in those circumstances.

But, even if such is the right way to interpret Harsanyi's assumption, accepting its truth in the present context amounts to little more than assuming that scepticism about the logical possibility of conducting ICUs, construed as empirical claims rather than evaluative statements, is unfounded. On its own, vis-à-vis the problem of ICUs, all SP` guarantees is that, as long as there are some facts of the matter about the magnitudes of one individual's mental states, then there are some facts of the matter when it comes to claims about the magnitudes of different individuals' mental states. And, if we assume that any individual *i*'s vNM function to some extent mathematically represents the magnitudes of certain mental states that *i* is or may be in, then we have also assumed that that there are indeed facts about the magnitudes of *i*'s mental states, viz. the ones being represented by her vNM function.

The problem with reading Harsanyi in the above fashion, however, is that such a physicalist interpretation seems philosophically thin (in the context of the problem of ICUs) for lack of an adequate theory of preference. Without a theory about the *nature* of the psychological attitudes individuals direct towards outcomes—be these preferences, desires, or whatever—we have no way to empirically investigate how accurate our mathematical representations of their magnitudes are.[36] And so, if Harsanyi's a priori assumption is best interpreted as some roundabout affirmation of physicalism, there is a sense in which I am in limited agreement with his claim that this assumption helps to secure the metaphysical or logical possibility of conducting empirically meaningful

[36] This is true of both the relational reading of these magnitudes that we find in ordinary vNM functions, *and* the 'absolute' sense in which these magnitudes may be represented by an appropriately transformed vNM function.

ICUs. But, if we are to get beyond the representational content of ordinary vNM functions, simply assuming physicalism will not take us to that content on its own; all physicalism will get us is the comfort of knowing that there is some such content to be had. I would go so far as to say that merely accepting physicalism while assuming that individuals' attitudes towards outcomes do have magnitudes (perhaps in virtue of assuming that vNM functions do represent some facts about these magnitudes, to some extent) at best guarantees that there *is* a real solution to the problem of ICUs out there, but says little about the nature of that solution. Hence, assuming physicalism and taking ordinary vNM functions to have some representational content about the magnitudes of certain mental states, at best, constitutes a partial solution to the problem of ICUs. The reason for considering such a proposal a merely *partial* solution is that, while it may answer scepticism about the mere *existence* of facts comparing different individuals' utilities, it tells us little about the nature of these facts.

Ultimately, the conclusion that I believe we ought to draw from the discussion of the preceding chapter is that neither an a priori assumption resembling Harsanyi's SP, nor episodes of imaginative place-trading are likely to be successfully combined with the brilliant formal accomplishments of von Neumann and Morgenstern as a satisfactory solution to the infamous problem of ICUs. One of, if not the most thoroughly worked out attempts to do so, viz. that of John Harsanyi, comes up short in a number of ways. But what I have suggested is that the most instructive of these failures consists of Harsanyi's recognition that the problem of ICUs must incorporate some reference to the ways actual human psychologies function. But rather than looking for a solution in unhelpful postulates such as Harsanyi's SP, the places to look are the

experimental data of psychology and especially the neurosciences, as well as conclusions drawn from contemporary philosophy of mind. The next chapter outlines the nature of the data we need to look at in order to arrive at a satisfactory and empirically respectable solution to the problem of ICUs.

Chapter 3

We ended the previous chapter by suggesting that any plausible candidate for a solution to the problem of interpersonal comparisons of utility (ICUs) ought to be sensitive to the ways in which human psychologies actually function. More specifically, we saw that under the physicalist assumption that mental states are in principle *observable*—in virtue of being *physical* states—there is nothing (in principle) barring us from directly observing psychological reactions that individuals display in reaction to various outcomes. The problem of course is that, lacking an adequate account of the nature of preference, we have no idea what to look for, if what we're interested in is exactly how an individual psychologically reacts to a given outcome; specifically, the nature of her psychological reaction to her own preference satisfaction and frustration.

Yet if we had a theory *of* preference—or perhaps some clue as to how preferences are biologically realized in the brains of organisms like us—then we might be able to devise methods of precisely ascertaining the strength of some individual's preference by directly observing and measuring the magnitude of the activity of its neural basis. Without knowing exactly what preferences *are,* however, the chances of identifying and measuring the activity of the neurophysiological structures realizing them directly through empirical observation seem poor.

But what if contemporary philosophy of mind were to produce a plausible theory of the nature of *desire*, built atop a foundation of empirical data from the neurosciences; could such a theory prove useful to those of us who are troubled by the problem of ICUs? What if, with the intellectual purchase of this theory, came for free an empirically measurable and interpersonally comparable account of *desire strength*?

That sounds like it may indeed be helpful. Fortunately, such a theory exists, and the business of this chapter is to explain its inner workings, with special attention paid to what it says or entails about the nature of desire strength. The theory in question is Tim Schroeder's (2004) *Reward Theory of (intrinsic) Desire,* and if true, this theory can indeed be deployed in a satisfactory solution to the problem of ICUs.

But why are we talking about desire all of a sudden, and not about preference? Are the two equivalent, or perhaps coextensive? If we *have* abandoned preference talk, why *aren't* we talking about *pleasure*? We postpone these and related topics[37] for Chapter 4, turning now to exposition of Schroeder's (2004) Reward Theory.

The Reward Theory of (intrinsic) Desire

Traditionally, philosophers who have thought about desire—including philosophers of mind and utilitarians—have largely thought about it in terms of either *motivation* or *pleasure*; both of these are contingently related to desire, but neither nor both constitute its essence, according to the Reward Theory of intrinsic Desire (RTD). Instead, the essence of desire is *reward;* but not merely the commonsense understanding of 'reward' or the sense that the 1950s behaviourist psychologists toiled with.[38] The sense of 'reward' that constitutes the essence of desire is the one used by contemporary neuroscientists when they talk about the neurological effects of what we intuitively are inclined to call 'rewards' and 'punishments'.[39] Moreover, we must note that RTD is a theory of *intrinsic,* as opposed to *instrumental* desire, meaning that its scope is confined to explaining the nature of desiring outcomes for their own sake, and

---

[37] More specifically, the relationships between desire, preference, utility, and welfare are analyzed throughout chapter 4 below, including arguments in favour of approaching the problem of ICUs using the language of desire rather than preference.

[38] For discussion on reward and behaviourism, see Schroeder (2004: 43-8).

[39] For discussion on reward and common sense, see Schroeder (2004: 39-43).

not as a means to some other end or ends.[40] Finally, 'desire', in the context of RTD (as well as the present work), is meant to include all mental states that we commonly call 'desiring', 'wanting', or 'wishing', as long as they are desires, wants, or wishes directed at an outcome (or outcomes) *for its* (or their) *own sake,* and not as a means to obtaining some further, distinct outcome (or outcomes) (Schroeder 2004: 5).[41]

Having thus outlined the scope of RTD, we may introduce it as follows:

> Reward Theory of Desire (RTD): To have an intrinsic (positive) desire that *P* is to use the capacity to perceptually or cognitively represent that *P* to constitute *P* as a reward. To be averse to it being the case that *P* is to use the capacity to perceptually or cognitively represent that *P* to constitute *P* as a punishment. (Schroeder 2004: 131)

Before turning to detailed exposition of RTD, it could prove useful to the reader to get a sense of the motivation behind identifying the essential nature of desire with reward (and of aversion with punishment).

Presently, there are only two serious candidates for the essence of desire that are alternatives to reward: pleasure on the one hand, and motivation on the other. Desires tend to move us to act so as to get what we want, and avoid what we don't want; desires also set us up to experience pleasure when we get what we want, and to experience displeasure when we don't. Motivational and hedonic theories of desire each classify one of the above features (or some variation thereof[42]) as *essential* to desire, whereas, according to RTD, they are typical though contingent *effects* (*not* constituents *or* essential properties and relations) of desires. It happens (by evolutionary chance) that in organisms like us, desires have these motivational and hedonic effects, but in principle

---

[40] See Schroeder (2004: 4-5).

[41] In the present context, we may take the term 'desire' as being meant "in a broad but not limitlessly broad sense, encompassing things naturally thought of as wishes, wants, goals, desires, ends, and so on, but not intentions, plans, episodes of trying, or beliefs about what is good. In particular, 'desire' is meant to include wishes about how the past might have been as well as desires for the present and future, and to include sensuous along with intellectual goals." (Schroeder 2004: 132)

[42] See Schroeder (2004: Ch. 1) for various formulations of hedonic and motivational theories of desire

there could be, according to RTD, strange creatures with desires that are quite incapable of both action, and pleasure.

So why think that the two older varieties of desire theory—certainly in the case of motivational theories, far more widely accepted ones than RTD—should be discarded in favour of the rather surprising RTD? What, to be more specific, makes reward a plausible candidate for the essential nature of desire—more plausible than the two familiar candidates?

For present purposes, the above questions are best answered with emphasis on the positive sides of the issue: rather than pointing out the long list of problems plaguing various formulations of motivational theories of desire, and the also large inventory of those that hedonic theories suffer from,[43] the focus shall instead remain on certain parts of what makes the identification of reward with the essence of desire a plausible one. We begin with some intuitive remarks on the topic from Schroeder:

> In favour of the theory that rewards and punishments are constituted by an organism's desires are the commonsensical thoughts that if you want to reward someone, the best way is to give that person something she really wants, and if you want to punish someone, the best way is to find out what that person would hate done and then do it. Likewise, it is widely believed that a common failure in rewarding a person is to give him what the giver would like to be given, rather than what the recipient actually *wants*. (2004: 67)[44]

Hence, if Schroeder's survey of commonsense is more or less correct, we see, even without much in the way of philosophical theorizing, that there are some remarkable *extensional* parallels to be drawn between the objects of desire on the one hand, and rewards on the other.

We commonly believe that giving an individual what she actually wants, all things equal, counts as rewarding her; in other words that a reward for an individual is

---

[43] Schroeder (2004) is a work doing, inter alia, precisely that.
[44] Italics are Schroeder's.

also an object of a desire that the individual has. Suppose, e.g., that we set out to reward Sam by giving her the opportunity to drive an expensive and energy inefficient luxury SUV for a year at no cost. However, and unbeknownst to us, one of the things Sam really wants is that she and others living in her country take steps to reduce their country's dependence on foreign oil. It is surely in accordance with our commonsense beliefs about reward that Sam has *not* been rewarded by the gift of a year's use of the SUV, all else being equal, since she sees her countrymen indulging in that kind of product as part of an ongoing problem, one she herself wishes *not* to contribute to. Had she been given something she really wants to drive—an object of one (or more) of her desires, say a gas-electric hybrid—then surely commonsense dictates that she would have in fact been rewarded, all things equal.

In addition, notice the perhaps even more remarkable *causal* parallels between desire and reward, also pointed out by Schroeder:

> Furthermore, according to common sense, the causal consequences of getting what one wants are by and large the same as the causal consequences of being rewarded. Rewards cause people to feel pleased and dance for joy; getting what one desires does the same. The opportunity to get a reward is motivating; the opportunity to get what one wants is the same. People decide what to do by thinking about how they were rewarded in the past; they likewise think about how they were given what they want in the past. Rewarding a child can influence her future thinking and actions through unconscious processes; giving a child things she wants does the same. (2004: 68)

In other words, having one's desires satisfied and being rewarded seem to produce the same types of events in an individual. E.g., suppose Lee wants a certain young puppy with a specific and impressive genetic lineage. Lee's desire that he own the dog will, all things equal, motivate him not to spend excess money so that he can afford to buy her from the breeder, motivate him to go to the breeder in order to buy her, and so on. If Lee should be fortunate enough to procure the dog, he will experience pleasure, and his long-term thoughts and feelings will also be influenced (e.g., he will likely take steps in

the future to take good care of the dog, he will feel good when she is around, he will be disposed to experience great anguish when she dies, and so on). Much the same can be said about Lee and the dog by substituting desire for reward: The opportunity to be rewarded by the procurement of the dog will motivate Lee not to spend money, to go to the breeder, etc.; being rewarded by finally getting the dog will cause Lee to experience pleasure, to take future steps in caring for her, to feel good when she is around, to feel anguish when she is dead, etc.

What the two preceding paragraphs suggest is that while quite surprising indeed, the identification of the essential nature of desire with reward is still far from being radically counterintuitive. The resources of common sense, in other words, suffice to lend a fair measure of prima facie plausibility to RTD, in virtue of the causal and extensional parallels observed between desire and reward.

Incorporating neuroscientific findings with those of common sense, we can find deeper justification for RTD than the simple observations we've made so far. As it turns out, crucial aspects of the neurological underpinnings of motivation as well pleasure have a common cause in organisms like us, viz. 'reward signals'.[45] And, since desires are commonly thought to motivate us and cause us to feel pleasure when satisfied and displeasure when frustrated, it's far from implausible to hypothesize that desires are realized in our brains precisely by the signals that constitute the unique common cause of both pleasure and motivation:

> The neural basis of reward is the normal cause of pleasure and an important cause of motivation, while pleasure and motivation have much less influence upon one another and neither exerts a dominating influence upon the reward structure. This striking finding is a fairly recent revelation of scientific work, and it has not yet been given an interpretation by the scientific community in terms of desire. Yet some such interpretation seems called for. (Schroeder 2004: 37)

[45] Much more on these below.

To elaborate on Schroeder's remarks without going into a lot of technical detail (we postpone doing so only momentarily, however), in normal humans, the neurophysiological structures realizing reward exert a powerful influence on the structures responsible for the production of voluntary movement (and therefore behaviour), and on the structure that realizes pleasure and displeasure. Realizing reward is a set of twin structures within the brainstem, known as the *ventral tegmental area* (VTA) and *substantia nigra pars compacta* (SNpc), while pleasure seems to be realized in the *perigenual anterior cingulate cortex* (PGAC) (Schroeder 2004: 36). The VTA/SNpc send their chemical signals to many parts of the brain, including the areas responsible for producing behaviour, as well as the PGAC. Moreover, our voluntary behavioural circuitry and our pleasure 'centre' (PGAC) are connected to one another, but not so to any remarkable extent: "It may come as a surprise to learn that the connections between the neural basis of pleasure, in the PGAC, and control of the voluntary muscles appears to be fairly modest. Instead of pleasure dominating motivation, motivation appears much more influenced by the neural basis of reward, in the VTA/SNpc." (Schroeder 2004: 37) As for pleasure itself, it turns out to be only one amongst many causal influences affecting the VTA/SNpc, whereas the latter are normal, and extremely powerful causal contributors to experiencing pleasure (to activation of PGAC in humans).[46]

In short, pleasure and motivation seem to only have a moderate influence upon one another, whereas both are strongly affected by the chemical signals sent by the twin reward structures (VTA/SNpc). If desire is best thought of in terms of reward—i.e., if

[46] See Schroeder (2004: Ch. 3) for details on pleasure, and the PGAC.

RTD is true—then the common sense observations that desires tend to move us, dispose us to pleasure or displeasure, and that their objects are coextensive with rewards is strongly supported and illuminated by the neuroscience of reward, behaviour, and pleasure. We know desire to be accompanied by motivation and pleasure; the common cause of the neural basis of motivation and pleasure is the reward signal released by the VTA/SNpc; therefore, the hypothesis that reward signals realize desires is, on first blush, far from unreasonable.[47]

Yet there is much about RTD that requires clarification beyond the prima facie (positive)[48] support for the theory outlined in the preceding paragraphs; let us begin such clarification with the notion of representation employed in the definition of RTD, as it happens to be rather straightforward. For some *X* to qualify as a mental representation in the sense required by RTD, *X* must be "a content-bearing thing, making up some perceptual or cognitive attitude, localized in or distributed through some perceptual or cognitive centre of the brain, capable of passing output to the reward system." (Schroeder 2004: 134) Ignoring for the moment what the reward system itself consists of, we can see that the constraints placed by RTD on what counts as a bona fide perceptual or cognitive representation are rather minimal: as long as *X* is located or realized[49] in a perceptual or cognitive part of the brain and is *about* something, *X* is a mental[50] representation in the sense required by RTD. Hence, as long as one concedes that there are at least *some* genuine contents of one kind or another that are involved in

---

[47] There are, however, many more arguments in Schroeder (2004) supporting RTD, and it is not the business of the present work to defend them per se.
[48] Again, nothing has been nor will be said here about what is wrong per se with alternatives to RTD.
[49] If mental states just are brain states, then mental representations are (literally and straightforwardly) spatiotemporally located in parts of the brain; if mental states are biologically realized by brain states, then mental representations are realized in certain spatiotemporal locations in the brain.
[50] Throughout the discussion I use the term 'mental representation' as shorthand for 'perceptual or cognitive representation'.

perception and/or cognition, one need not be hostile to RTD in virtue of finding the notion of mental representation suspect.[51] Moreover, Schroeder points out that neuroscientists are generally of the view that some representation of the world (both internal and external to one's body) *does* take place in the brain throughout the *sensory cortex* and a large part of the *association cortex*, as well as the *hypothalamus* (2004: 49). Representation taking place in sensory cortex and association cortex is that "corresponding to familiar sense perception and everyday cognition" while some of the representational functions of the hypothalamus include things like monitoring blood glucose and electrolyte levels, body temperature, etc. (Schroeder 2004: 49)[52]

Assuming that the above paragraph constitutes a sufficient characterization of the meaning of mental 'representation' (as it applies to RTD), we may now inquire into a further substantive feature of RTD, viz. the distinction between *positive* and *negative* desire-like attitudes—i.e., between desires proper and aversions. The distinction is meant to capture the (supposed) fact that having an aversion towards some outcome "is not the same as having a positive desire or appetite for its contrary" (Schroeder 2004: 132); and, it is a distinction that we shall be forced to return to in chapter 4 below.[53]

---

[51] We may note that Schroeder rightly denies that RTD is "committed to the view of mental representation found in the works of, say, Dretske (1988, 1995), Fodor (1990, 1998) or Millikan (1984, 1993)." (2004: 134)

[52] For details, see Kandel, Schwartz and Jessell (2000), whose neuroscience text is often referred to by Schroeder "when describing well-known facts about the brain." (2004, Ch. 2: Note 9). (Perhaps *well-known* to neuroscientists, their assistants, and students.) Also, it may be worth mentioning that Schroeder notes that the things represented by the hypothalamus may not be things we access *directly* through consciousness, though we can experience a kind of *indirect* conscious awareness of them (2004: 49). E.g., the activity of the hypothalamus may make one feel lethargic or fatigued when one's blood sugar is low. In any case, "the hypothalamus engages in some activity that neuroscientists typically call 'representing', and that counts as tokening a representation or intentional icon according to some well-known theories of intentionality (Cummins 1989, 1996; Dretske 1995; Fodor 1990; Millikan 1984; Sterelny 1990)." (Schroeder 2004, Ch. 2: Note 10)

[53] By attempting to incorporate RTD in our solution to the problem of ICUs, we shall sooner or later have to address the issue of how to weigh the strengths of an individual's aversions against the strengths of her desires in calculating how well she's doing on balance. Notice the stark contrast on this matter between

The idea is this: if individual *i* has an aversion towards some outcome *A*, and another individual *j* has a desire directed at that outcome's contrary ~*A*, then mentally representing that *A* is the case will contribute to a different psychological reaction in each of the two individuals *i* and *j*. E.g., if Smith desires that a Republican candidate not be elected to office, and Jones is averse to it being the case that a Republican is in office, then realizing that a Republican candidate has in fact been elected will contribute to one kind of psychological reaction in Jones and another kind in Smith. As for these psychological reactions themselves, in commonsense terms, Schroeder's view is that being averse to some outcome "sets one up for anxiety or relief" whereas desiring some outcome "makes possible joy or disappointment." (2004: 132) Hence, in the above case of Smith and Jones, mentally representing a Republican in office will contribute to feelings of disappointment in Smith (who desires that there not be a Republican in office), and feelings of anxiety in Jones (who is averse to there being a Republican in office). And conversely, if Smith and Jones both mentally represent a non-Republican in office, Smith's psychological reaction to that representation will be one of joy, whereas Jones's may be better described as relief.

Just how committed we are to the aversion/ desire proper distinction in our commonsense discourse (and our commonsense understanding of desire for that matter) is not at all obvious to me,[54] but in any case, there are empirically known phenomena that do seem to support it, at least if RTD is the right way to think about desire. The empirical support for the distinction comes from differences between what RTD says is

---

the language of preference and that of desire. Again, we put off discussion on these and related issues for chapter 4.

[54] This is not to say that I am of the view that we have reason to doubt our supposed commitment to the distinction, rather that, as far as I'm concerned, the matter is unclear as far as our pre-theoretic understanding of it goes.

the biological basis of desire on the one hand, and what it says is the biological basis of aversion on the other—i.e. from differences between the neuroscience of reward and that of punishment. Having said that, it seems natural for us to now have a closer look at the neural bases of reward and punishment, and thereby get a sharper image of the essence of desire (and of aversion).

The Neuroscience of Desire and Aversion

Let us begin with desire proper. An individual *i* having a desire for some outcome *A* is analyzed by RTD as *i* having the capacity to mentally represent *A*, and using that capacity to constitute *A* as a reward; but what does it mean for *A* to be 'constituted as a reward'? According to Schroeder, to say that *i* constitutes *A* as a reward is equivalent to saying that when *i* mentally represents *A*, that mental representation tends to result in the production of a certain signal (in the case of humans and animals, chemical[55] signal):

> Contingency-based Learning Theory of Reward (CLT): For an event[56] to be a reward for an organism is for representations of that event to tend to contribute to the production of a reinforcement signal in the organism, in the sense made clear by computational theories of what is called 'reinforcement learning'. (Schroeder 2004: 66)

Hence, according to Schroeder, *i* using her representational capacities to constitute some outcome *A* as a reward (*i* desiring *A*, in other words) amounts to *i* having a certain *disposition*. The disposition in question is the tendency of *i*'s mental representations of *A* to result in the release of a "reward signal" from *i*'s brain.[57] What *exactly* is a reward signal? Consider this description from Schroeder:

---

[55] The chemical in question is the neurotransmitter dopamine.

[56] Nothing in the present context rests on the ontology of rewards; e.g., rewards might turn out to be better analyzed as states of affairs with no consequence to RTD, or CLT. (Schroeder 2004, Ch. 2: Note 33)

[57] For the remainder of the present chapter, I assume that when we discuss some individual, that individual is human, or something close enough—i.e. some organism with a brain that's organized in

> A reward signal is an event which causes a characteristic, mathematically describable form of learning, and a punishment signal is an event which causes an opposing form of learning... this is learning in a very specific sense: it is a change in the connectivities of units which are themselves describable at an appropriately abstract level. These units are neurons in the case of creatures like us, and their form of connection is very specific to their biological character, but nothing in the nature of learning theory forbids the existence of very different, even inorganic, units playing the roles that neurons play in us. Hence, if something is a causal system which is mathematically describable as instantiating contingency-based learning, then it is the site of such learning. (2004: 134-35)

As commented in the previous section of the present chapter, in humans and similar organisms, reward signals are released in the form of the neurotransmitter *dopamine* by a set of twin structures inside the brain stem, viz. the VTA/SNpc (Schroeder 2004: 50).

There are several good reasons to single out the VTA/SNpc as the unique set of output structures responsible for releasing reward signals, according to Schroeder (2004: 49-52). One is that the "VTA and SNpc have the reach necessary to distribute a reward signal, and the signal can be received." (Schroeder 2004: 50) The neurons of the VTA/SNpc reach out to the rest of the brain, "including almost the entire cortex, and many sub-cortical structures" all of which are equipped to receive the dopamine signal sent by the VTA/SNpc (Schroeder 2004: 50).[58] Recall that amongst the standard recipients of reward signals are the voluntary muscle control (and hence behavioural) structures, as well as the PGAC (the structure thought to be the neural basis of pleasure in humans). That desires are closely associated with motivation and pleasure is explained partly by the fact that the VTA/SNpc do send signals to the pleasure and action parts of the brain on a regular basis.

In addition, "the pattern in which dopamine is released by the VTA/SNpc is exactly the pattern required to transmit information about reward that is useful as a

---

much the same way that a human brain is, with cortical and subcortical structures carrying out more or less the same functions that they do in the case of a human.

[58] For details, see Haber and Fudge (1997), and KSJ (2000).

learning signal, as a signal that can cause useful change at the neural level." (Schroeder 2004: 50) On this point, Schroeder cites the experimental work of Wolfram Shultz and colleagues,[59] whose "finding is that the dopamine-releasing neurons of the VTA/SNpc have a baseline level of activity (fairly low), and that these neurons sometimes fire more rapidly than this baseline, and sometimes more slowly." (2004: 50) What lends plausibility to the claim that the VTA/SNpc have the kind of dopamine release pattern needed for a useful learning (neural modification) signal are the facts that (1) "VTA/SNpc neurons fire in a pattern that carries information about the difference, at time *t*, between the rewards received and expected at *t* versus those rewards the organism was predicting (at *t*-1) it would receive or expect at *t*"; and (2) "such a signal is exactly what is most computationally useful if a system is going to modify itself adaptively on the basis of rewards received." (Schroeder 2004: 50) Both of these facts have, as Schroeder notes, been shown to be the case through empirical investigation.[60]

There is also the study conducted by Bao, Chan, and Merzenich (2001), which, according to Schroeder constitutes a "decisive demonstration of the power of the reward signal observed by Schultz and colleagues." (2004: 51) Bao, Chan, and Merzenich actually observed the VTA/SNpc "changing neural connection strengths, in exactly the sort of fashion required by the computational theory of reward-based learning." (Schroeder 2004: 51) Though the study was conducted on rat auditory cortices, there is "no reason to expect this effect to be restricted to rats, or to auditory regions in the

---

[59] Specifically, Schroeder (2004: 50, Note 13) references the experimental conclusions of Romo and Schultz (1990), Schultz and Romo (1990), Ljungberg et al. (1992), Schultz et al. (1993), Waelti, Dickinson and Schultz (2001), which are reviewed in Schultz (2000).

[60] Schroeder (2004: 50) cites the following experimental works in relation to the latter (2): (Houk, Adams and Barto 1995; Montague, Dayan and Sejnowski 1996; Schultz, Dayan and Montague 1997). The former (1) comes from the above mentioned work of Shultz and colleagues.

brain" and Schroeder (quite reasonably) speculates that conducting the experiments on rats rather than primates was simply a function of convenience (2004: 51).[61]

In short, it seems quite reasonable to accept the experimental conclusions drawn by neuroscientists, and from these conclusions infer that the VTA/SNpc are indeed *the* output structures of the reward system.[62] If Jones mentally represents an event, and then categorizes that event as a *reward*, we can be reasonably sure that, all things equal, her VTA/SNpc will release a reward signal in the form of dopamine to other parts of her brain. That signal causes what is called (by neuroscientists) 'learning' to take place, which is understood as the strengthening of neural connections in Jones's brain; and conversely, if Jones perceptually or cognitively represents an event, and then categorizes that event as a *punishment*, then neural connections in her brain are weakened.

There is a structure responsible for categorizing representations as either rewards or punishments and it is the *orbitofrontal cortex* (OFC): "the OFC responds selectively to stimuli which we, as observers, would intuitively call 'rewards' and 'punishments', doing so after the rewards or punishments (or signs of coming rewards or punishments) have been represented as more particular things elsewhere in the brain." (Schroeder 2004: 53) The OFC takes specific representations of varying content and complexity (often *greatly* varying content and complexity) as inputs, categorizes them as either rewarding or punishing to the organism it belongs to, then outputs this information to the VTA/SNpc: "By sending output to the structures responsible for

---

[61] See Schroeder (2004: 51, Note 13) as well.
[62] There is more experimental work referenced by Schroeder than the studies I mention above; for details see his (2004: 48-57).

generating the reward signal, it allows representations of [e.g.] tastes, fractal images and money to cause the release of reward signals."[63] (Schroeder 2004: 53)

In addition, there is another structure (or structures) responsible for *predicting* the occurrence of rewards and punishments, and keeping track of the information corresponding to how rewarding or punishing a representation of the world will be to the organism in question. This structure seems to be the *nucleus accumbens*, and it may work in conjunction with the *caudate nucleus* (Schroeder 2004: 53).[64]

The modification of neural connections caused by the transmission of reward signals (i.e. learning signals) affects one's emotions and dispositions. As noted earlier, the structures responsible for pleasure and motivation are important recipients of reward signals, and so, according to RTD, it is unsurprising that desires are so often linked to motivational tendencies, and dispositions to feel pleasure or displeasure. More specifically, the effects of the release of a reward signal influence "the short-term operation of the brain" as well as "its long-term dispositions, effects that, in organisms like us, affect our feelings and modify dispositions to act, think and experience, all in ways that tend to increase the acquisition of rewards and the avoidance of punishments." (Schroeder 2004: 49) In typical humans, desires tend to modify connections in our brains so as to better satisfy them, and to be pleased when we are successful. 'Learning' in the neuroscientific sense *just is* this modification of neural connections (connections between *neurons* in the case of humans), caused by the

---

[63] Schroeder's examples of tastes, fractal images, and money are taken from the experimental work of Rolls (2000), and Shultz et. al. (2000).

[64] According to Schroeder, "however this calculation is performed, it is clear that some version of it must be performed, since the dopamine neurons of the VTA/SNpc express the results." (Ch. 2). We must note that Schroeder's speculation on this matter is far from arbitrary, stemming from the experimental work of Knutson et al. (2001), Pagnoni et al. (2002), Berns et al. (2001), and Schultz et al. (1995).

transmission of information about how rewarding or punishing features of the world are to an organism equipped with a reward system.[65]

We should now be in a position to summarize the sequence of events that takes place in an individual's brain when some intrinsic desire of hers is satisfied.[66] Suppose, e.g., that Khan desires (for its own sake)[67] that the death penalty be abolished from United States law (Khan desires that *P* for short). And suppose at time *t*, Khan's desire that *P* has indeed been satisfied (an anti-death penalty Bill is passed, say). Assuming Khan's desire that *P* is not a *fleeting*[68] one—i.e., assuming her desire that *P* persists through time—at time *t*-1 (before she learns of the abolishment of the death penalty), Khan's nucleus accumbens (and, perhaps, caudate nucleus) had already predicted to what degree (if any), her desire that *P* will in fact be satisfied (that she will be rewarded) in the future, i.e., at some time after *t*-1. Let '*x*' denote the degree to which Khan's brain predicts (at *t*-1) that her desire that *P* will be satisfied in the future (that she will be rewarded in the future); in other words, *x* = *predicted* reward information. Now suppose, at time *t*+1, Khan learns that *P* is in fact the case; this fairly complex fact is mentally represented at *t*+1 in her sensory cortex and association cortex. At time *t*+2, Khan's OFC receives the information (from sensory cortex and/or association cortex) that *P* is the case; it then categorizes this information as either rewarding or punishing (rewarding and therefore desire-satisfying in the present example), to some degree or other. Let '*y*' stand for the degree to which Khan's OFC classifies *P* as rewarding for

---

[65] Much of chapter 2 in Schroeder (2004) includes discussion of what 'learning' in this sense involves, as well as descriptions of the experiments which have identified and described the processes. There are many details which I omit in the interests of economy.

[66] For further details concerning the sequence of events in question, see Schroeder (2004: 49-54).

[67] Unless otherwise noted, for the remainder of the discussion, when speaking of desires, I shall ignore instrumental desires.

[68] See Schroeder (2004) for discussion on fleeting (i.e. short-lived desires). (Basically, his view is that there aren't any genuine intrinsic desires that are also fleeting.)

Khan; in other words, y = *actual* reward information. At time t+3, Khan's nucleus accumbens (and, perhaps, caudate nucleus) calculates the difference between y and x. Let 'z' denote the difference between y and x; in other words, z = difference between predicted and actual reward information. At time t+4, the information z is received by Khan's VTA/SNpc, and at time t+5, sent to multiple structures throughout her brain in the form of a chemical (dopamine) signal—i.e., *reward* or *learning* signal. At time t+6, the signal sent by Khan's VTA/SNpc (at t+5) is received by structures underlying pleasure and displeasure (PGAC)[69] emotion, motivation, and movement, with the result of modifying (in this case, *strengthening*) neural connections in the process (i.e. with the result of *learning*[70] taking place in Khan's brain).

The learning process briefly outlined above is well documented in the neuroscience literature, much of which is referenced throughout Schroeder (2004, especially Ch. 2). The experiments conducted by Wolfram Schultz and colleagues throughout the 1990s,[71] as well as the above mentioned study conducted by Bao, Chan and Merzenich (2001), are especially illuminating on the topic of reward-driven learning, and if the reader is interested in reviewing the empirical evidence for her or himself, these are excellent places to start. But it is *not* the business of the present work to review such evidence, since we are operating under the assumption that Schroeder's (2004) analyses and interpretations of it are correct. Specifically, we have assumed that the identifications made by RTD and CLT are correct, and that, therefore, our above

[69] For detailed discussion of the PGAC, see Schroeder (2004: 76-83).
[70] Recall that this is 'learning' in a technical sense of the term.
[71] Romo and Schultz (1990), Schultz and Romo (1990), Ljungberg et al. (1992), Schultz et al. (1993), Waelti, Dickinson and Schultz (2001). Reviewed in Schultz (2000).

sketch of reward-driven learning is an accurate (though somewhat rough and ready) construal of the neural processes that underlie desires and their satisfaction.[72]

Unfortunately, the neuroscience of punishment-driven (contingency-based) learning, and hence of aversion, is not nearly as complete as that of reward-driven learning (and hence desire). Schroeder unhappily notes the following:

> No investigation of the neural basis of punishment has put forward evidence for a punishment centre with the persuasive force of the evidence for the reward centre. But given the apparent fact that the OFC and nucleus accumbens generate both reward and punishment information (Rolls 2000), there surely must be a structure somewhere in the brain that makes use of this information to produce a punishment signal, a counterpart to the VTA/SNpc reward signal. The leading candidate for a punishment centre is found in the *dorsal raphe nucleus*, or *DRN*, located high in the brainstem, near the reward centre. (See, e.g., Wise, Berger and Stein, 1973; Deakin 1983; Deakin 1998.) A distinctive feature of the DRN is that almost all of the cells in the brain which release serotonin are found in it or in related neighbouring cell clusters (Kandel, Schwartz and Jessell 2000). Hence, serotonin has been thought to be the messenger carrying the neural punishment signal, just as dopamine is the messenger carrying the reward signal. (2004:)

In other words, the *dorsal raphe nucleus* (DRN) may be the biological seat of negative desire-like attitudes (aversions). The matter is not certain, however, and neither is the suggestion that the chemical signal carrying punishment information takes the form of the neurotransmitter *serotonin* in organisms like us. Nevertheless, Schroeder's hopefulness that a punishment counterpart to the VTA/SNpc will sooner or later be identified by neuroscientists is far from unfounded, given the above quoted point regarding the OFC and nucleus accumbens: The OFC and nucleus accumbens *do* generate reward as well as punishment information; how plausible is it to suppose that one kind of information (reward) is received, processed, and transmitted, while the other (punishment) goes unused? Not very. If there were no punishment centre, it seems reasonable to suppose that the OFC would have the much easier task of categorizing merely to what extent a mental representation is rewarding, rather than its actual task of

---

[72] We could easily modify the above sketch in such a way that it describes the neural processes involved in having one's desires *frustrated,* rather than satisfied. The reader is again referred to Schroeder (2004: Ch. 2) for details.

categorizing *whether* a representation is rewarding *or* punishing, *and to what extent.* Nevertheless, by comparison to our knowledge of how exactly reward information gets used, we are presently somewhat in the dark when it comes to parallel details concerning the use of punishment information. Hence, when our method of conducting ICUs on the basis of reward and punishment information is formulated in chapter 4 below, we shall be forced to leave open the matter of where exactly in the brain to look for the neurological basis of aversion strength.[73]

Now, before proceeding any further, it is crucial to the plausibility of RTD that we emphasize the fact that the theory does not require of desires that they involve *actual representations*; e.g. *actual* sense data, *actual* thoughts, imaginings, or whatever. Even though Schroeder's view is that representations are literally *proper parts* of desires, "desires need not involve tokened representations, need not involve actual episodes of representing" (Schroeder 2004: 134). For it to be the case that, say, Ali has an intrinsic desire that his wife be elected Prime Minister of Canada, RTD does not require that Ali actually see her being inaugurated, or hear that she's been elected on the news, or even that he actually imagines his wife's victory. Ali's desire that his wife win can exist without the actual tokening of such mental representations, since, according to RTD "to desire is to be so organised that tokened representations… *if* they occur, will contribute to the production of reward signals."[74] (Schroeder 2004: 135). Schroeder is explicit in pointing out that RTD "requires a link between representational *capacities* and reward signals, rather than a link between occurrent representations and reward

[73] The details and consequences of this concession are discussed in chapter 4 below.
[74] Italics are mine.

signals."[75] (2004: 135) This feature of RTD allows us to say that, e.g., Ali desires that his wife be elected Prime Minister even when he is in a deep sleep or coma, and representing nothing whatever to do with his wife.

So far, the present chapter has been dedicated to discussion of the *nature* of desire (and, to a lesser extent, aversion); we still have not said anything about desire *strength,* its measurability, or its interpersonal comparability. Therefore, desire strength is our next order of business.

Desire Strength

Reward (learning) signals are not an all-or-nothing affair; in organisms as sophisticated as humans, they can be strong or weak, and so too can the desires that they neurologically underlie:

> What does desire strength amount to, on the reward theory? If to desire is to constitute a state of affairs as a reward or punishment, then a strong desire is one that constitutes a state of affairs as a substantial reward or punishment, whereas a weak desire would constitute the same state of affairs as a minimal reward or punishment. If contingency-based learning in a given organism is not incredibly crude, then some representations will normally contribute to very powerful learning signals—signals with great power to change neural connections—while other representations will have much less influence over the learning process. The strength of a desire will thus come to the relative power of the desire to change neural connections, and so modify its owner's mind. (Schroeder 2004: 138-39)

Thus, desire strength is *not* to be measured in terms of how much pleasure an individual experiences, or how powerfully he or she is motivated to act (as has traditionally been supposed by many philosophers committed to variations of hedonic or motivational theories). These things may or may not be indicative of desire strength, since it is only contingently true of human beings that desire tends to have a close association with feelings of pleasure, and our dispositions to behave. The PGAC (the neural seat of pleasure) is, after all, entirely distinct from the structures realizing desire (VTA/SNpc),

---

[75] Italics are Schroeder's.

and so are voluntary muscle control centres.[76] If we are in interested in knowing actual facts about how much an individual wants something to be the case, the place to look is the VTA/SNpc of that person's brain; specifically, the strength of the reward (learning) signal it produces when that outcome is mentally represented (or the strength of the learning/reward signal that *would* be released, *if* the outcome *were* represented). Hence, if RTD is the right way to think about desire, and CLT the right way to think about reward, then desire strength can, in principle, be read from the strength or magnitude of learning (reward) signals.[77] The next chapter of the present work includes a theory of how the aforementioned reading may be accomplished.

We may conclude this chapter with a summary of what has been said so far. If the theory of desire assumed to be true in the present work (viz. RTD) is in fact true, then the essential nature of desire is independent of both pleasure and motivation, but not reward. The dopamine signals released by our biological reward systems exert powerful and regular influences over both pleasure and motivation, while the latter two are connected with one another to only a moderate extent. Therefore, if RTD is true, our common sense observations that desires both motivate us to act and dispose us to feel pleasure is rendered intelligible and illuminated by the interpretation of the brain's reward system in terms of desire. Moreover, neuroscientific investigations of the effects of rewards reveal them to be primarily characterized by a specific neurocomputational learning that modifies the strength of neural connections. The consequences of that modification on pleasure and behaviour are both immediate, and long-term. Some of the long-term effects include physical changes in the brain that dispose the organism to

---

[76] See Schroeder (2004: 36-7).
[77] Though, as we shall see throughout chapter 4, there are complications.

seek out rewards—i.e., to satisfy desires—and to experience pleasure when successful. The way our reward structures—viz. the VTA/SNpc—exert their influence upon the rest of the brain is via the reward or learning signals that they release. These signals can be strong, resulting in relatively large changes in neural connectivities, or weak, causing relatively weak changes. What is of principal interest to us is that reward signals are the neural basis of desires, that their strength is the neural basis of desire strength, and that their strength can in principle be *empirically measured.* The next chapter constitutes my own effort to employ RTD in what I shall argue is the most satisfactory solution to the problem of ICUs proposed to date; a solution that rests on the empirical measurability of the intensity of reward signals.

Chapter 4

The previous two chapters have concentrated first (chapter 2) on what is *not* likely to prove helpful to the formulation of a satisfactory solution to the problem of interpersonal comparisons of utility (ICUs), then (chapter 3) on what I have suggested *is* likely to help, viz. Schroeder's (2004) Reward Theory of (intrinsic) Desire (RTD); the purpose of the present chapter is to put that suggestion into practice and defend what I argue is the best solution to the problem of ICUs proposed to date.

The solution in question is desire-based, in the sense that utility is defined in terms of net intrinsic desire satisfaction, where 'desire' is to be understood in terms of RTD; that is, in terms of reward- and contingency-based learning. Moreover, the view defended does not, on its own, constitute a method of *mathematically approximating or representing* the utilities individuals assign to specific outcomes. The view is instead best taken as a theory of the *nature* of what is being represented by mathematical approximations such as von Neumann and Morgenstern (vNM) utility functions, viz. the 'absolute' importance individuals actually attach to outcomes—*absolute* as opposed to simply *in relation to other outcomes* (which, as was discussed at length in chapter 2, is the extent of the representational content of an individual's vNM function).

Desire Strength, Measurability and Interpersonal Comparability

Because the solution to the problem of ICUs defended below depends entirely on the measurability and interpersonal comparability of individuals' intrinsic desires, a natural starting place is a suggestion of how one might go about empirically measuring the precise strength of a single intrinsic desire within a single mind.

Recall (from chapter 3) that, according to the theory of desire assumed to be true in the present work—i.e., Tim Schroeder's (2004) RTD—desires are realized by *reward* or *learning signals*[78]. What's of primary interest to us in the present work, however, is that, like the desires they realize, learning signals come in varying *strengths*; strong desires are realized by strong signals, and weak desires by weak signals. Moreover, we know that in organisms such as ourselves, learning signals are chemical in nature, consisting of the neurotransmitter *dopamine.* So, if desires are realized by a chemical signal, could we not simply have an individual mentally represent something she really wants—something she constitutes as a reward, in other words—and measure the quantity of dopamine subsequently released by her brain's reward system in order to ascertain the precise degree to which she wants it? The natural sciences, e.g., chemistry or various biological sciences, have after all become pretty good at measuring quantities of things like dopamine—certainly better than economists are at measuring things like the strengths of preferences.

While the preceding observation is not meant, by any means whatsoever, to downplay the efforts or formal accomplishments—in some cases, even genius—of individual economists, it is indeed meant to showcase the fact that the accuracy of the best economists' estimates of the strength of a single preference within a single mind pales in comparison to the accuracy with which a neuroscientist could measure the quantity of a well known neurotransmitter, using the right equipment. So, if we can read desire strength from the strength of chemical signals, perhaps we can simply measure

[78] Learning signals *are* reward signals, and the two terms may be used interchangeably; see chapter 2 of the present work or for a more detailed discussion, Schroeder (2004). I use the former label more frequently than the latter.

the amount of the chemical in question, in this case dopamine, upon its release by the reward system.

With the above remarks in mind, consider the following proposal for measuring the strength of one individual's (single) desire, e.g., Sam's desire that the world's endangered species not become extinct; *s*'s desire that *P* for short. According to RTD, desires have content-bearing, mental representations of outcomes[79] literally as *proper parts*. So, in the case of *s*'s desire that *P*, *s*'s mental representation of *P* is literally a part of that desire. Moreover, because (we're assuming) *s* indeed desires that *P*, mental representations of *P* will, by our definition of desire, tend to produce learning/ reward signals which are released by the twin reward structures called (recall from chapter 3) the *ventral tegmental area* and the *substantia nigra pars compacta* (VTA/SNpc) within *s*'s brain. And, since we are generally confident in our ability to measure and compare quantities of chemicals such as dopamine, it seems as though we could in principle have *s* represent that *P*, and then measure the quantity of dopamine released by *s*'s VTA/SNpc using brain imaging technology. Hence, we might think, the larger the quantities of dopamine released by an individual's brain when she represents the object of her desire, the more she desires that particular outcome.

The above proposal is, however, too naive and deeply flawed, the principal reason being the following: Though reward/learning signals are indeed chemical in nature when it comes to organisms such as ourselves,[80] simply measuring the quantity of the relevant chemical at the time of the release of the reward signal is a bad way to

---

[79] These being events, or states of affairs, or situations, depending on the right ontology, which, as has been said earlier, is irrelevant for present purposes.

[80] Perhaps there are alien life forms that have learning signals of a non-chemical nature.

measure the strength of the signal itself, and hence a bad way to measure strength of desire in general.

Recall that empirical studies conducted by Wolfram Shultz and colleagues cited in chapter 3 of the present work, which Schroeder (2004: 50) appeals to, indicate that the neurons of the VTA/SNpc have a baseline level of firing. When representations of something constituted as a reward—something desired in other words—are tokened, these neurons briefly fire at an above-baseline level, releasing a reward signal in the form of dopamine. The quantity of dopamine released in a reward signal is thus always relative to the quantity of dopamine involved in VTA/SNpc baseline firing. Therefore, just measuring how much dopamine gets released by the VTA/SNpc when an individual mentally represents something she wants simply won't get us a straightforward measure of desire strength, since that quantity of dopamine must be taken in relation to the quantity involved in baseline firing.

There are more complications still to holding the naïve theory I just sketched. To return to our earlier example of *s*'s desire that *P*, suppose that *s* comes to know that *P*, the object of her desire, is indeed the case. When *s* recognizes this, her VTA/SNpc will, all else being equal, release a powerful reward signal. But, even if we relativised the quantity of dopamine released by that reward/learning signal to the baseline level of firing in *s*'s VTA/SNpc, we are *still* not guaranteed an accurate measure of *exactly how much* she wants *P* to be the case. For one thing, her *expectations* regarding whether *P* will or will not be the case in the near future are a crucial factor in determining just what information we *can* read off the reward/learning signal released upon her coming to know that *P*. If *s fully predicted* that *P* will be the case in the near future, and then

shortly thereafter learned that *P* is the case, her reward system may very well be entirely unresponsive to her coming to know that *P*. In reference to the above mentioned experimental work of Wolfram Shultz and colleagues, Schroeder writes:

> Thus, the elevated firing rate of VTA/SNpc neurons corresponds to the time at which information is received that an *unpredicted* reward has been received *or* is coming. A fully predicted reward, however, causes no deviation in VTA/SNpc activity… Absence of reward when no reward is predicted by the organism has no effect, but when a predicted reward fails to materialize, VTA/SNpc activity immediately drops. Hence, VTA/SNpc neurons fire in a pattern that carries information about the difference, at time t, between the received and expected at t versus those rewards the organism was predicting (at t-1) it would receive or expect at t. Or, to put things in a more intuitive if less precise manner, VTA/SNpc neurons signal the difference between how good the world was predicted to look at t and how good it in fact looks at t. (2004: 50)[81]

So, if we want to read an individual's desire strength off the strength of her reward signals, we have to take her expectations into account. The reason being that the strength of a reward signal does not indicate the strength of desire straightforwardly, but realizes the difference between actual and expected reward information. In other words, a reward signal carries the information that consists of the difference between *actual net intrinsic desire satisfaction,* and *expected net intrinsic desire satisfaction.* Simply put, expectations *must* be taken into account if we are to have any hope of reading desire strength information from the strength of the reward/learning signals that realise desires.

The above point about expectations furthers the case against any kind of naïve attempt to measure how much, e.g., I desire the evolutionary advancement of the human species, by simply asking me to think about us evolving, and then measuring, say, how many microlitres of dopamine my reward system releases. Not only do we have to keep track of the quantity of dopamine it takes for my reward system to fire at its baseline, we have to keep track of what my expectations are; in this example, expectations vis-à-vis the evolutionary advancement of humans. So, if the naïve method won't work, is

[81] Emphasis original to Schroeder.

there one that will? Specifically, is there an, in principle, viable and practicable method of reading desire strength from the strength of reward signals? The answer is yes, though we may not be able to carry it out with our present level of brain imaging technology. This answer requires clarification, to which we now turn.

Since the learning signal carries information about the difference between actual and expected net intrinsic desire satisfaction, we can write

(1) LS = actual nIDS-expected nIDS

where 'LS' refers to learning signal and 'nIDS' refers to net intrinsic desire satisfaction. The first step in a plausible method of getting a reading of the strength of a single desire is therefore to create conditions such that actual nIDS-expected nIDS = 0, and that expected desire satisfaction regarding the desire being measured is 0; then to surprise the individual with a mental representation of the object of the desire we wish to measure. E.g., suppose Jack desires that his stamp collection be complete; imagine he is one rare stamp short of completion. It seems we could, in principle, convince Jack that his stamp-desire is not likely to be satisfied in the immediate future, and then give him the stamp. All things equal, mental representations of acquiring the missing stamp will tend to produce a strong learning signal in Jack's brain under such conditions. But, because Jack's expectations vis-à-vis the stamp-desire, we are supposing, are at 0 when he mentally represents acquiring the stamp, the learning signal produced will carry information about how much of a difference having the stamp-desire satisfied makes to his net intrinsic desire satisfaction (actual nIDS). We would have to measure the quantity of dopamine released in the learning signal that was produced when Jack represented the stamp; as we saw earlier in the present chapter, this quantity would have

to be relativised to quantities involved in the baseline firing of Jack's VTA/SNpc. Once the aforementioned quantities are compared, it seems at least prima facie plausible to maintain that what we have as a result is a good, empirically derived approximation of how much Jack desires that his stamp collection be complete.

The above procedure is anything but simple or straightforward, but no one could seriously deny that it is conceivable, possible, and even practicable, under the right experimental conditions. Were such a procedure to ever be attempted in order to measure individual desires, those conditions would have to be fine-tuned, so to speak, in order to prevent possible experimental contamination, but the general idea behind the procedure itself is not all that far-fetched. Whether we are presently capable of successfully measuring an individual's desires using the procedure (call it 'DP') or something like it, I am not sure. But, if the scientific community were to accept an interpretation of the reward system in terms of *desire,* such as the one offered by Schroeder (2004) that we have here assumed to be true, it is certainly foreseeable that experimenters might then attempt something like DP in order to measure desire strength.

So far we have only discussed what I have insisted looks like a plausible method of measuring individual desires within a single mind; but what about interpersonal comparability? That, after all is a major concern in the present work, since if the magnitude of, say, Jack's desire that his stamp collection be complete cannot be compared to the magnitude of Mary's desire that she have a post-graduate education, then we have failed, and there is no sense in proceeding with the present proposal any further.

The proposal for measuring desire strength within a single mind discussed above—DP—will not get us our most sought after goal, viz. ICUs, because, like the naïve attempt preceding it, DP still depends on measuring dopamine. Consider, e.g., a small child on the one hand, and an American football player who weighs over 150 kilograms on the other. Presumably, everything about the football player is bigger, including his skull and brain. It is hence conceivable that even a desire minute in strength belonging to the football player may be realized by a reward signal involving a larger quantity of dopamine release than that realizing even the child's very strong desires. In the same spirit, we can imagine intelligent aliens of a similar construction as us that happen to be far larger or far smaller than us, with far larger or far smaller brains that release great or minute quantities of dopamine, depending on the sheer size of those brains. In other words, that *x* amount of dopamine is released when individual *j* mentally represents something she wants is no measure of how much she wants it, for *j*'s brain itself may be massive or minute compared to some other individual *k*. If *j*'s brain is one hundred times the size of *k*'s and they are of similar construction with similar neurofunctional properties, then *of course j*'s brain will have to release a far larger sheer quantity of dopamine than *k*'s.

If we amend DP by changing our focus from measurement of dopamine levels per se, to measurement of *learning* (in the neuroscientific sense) *as such,* then we can indeed end up with an experimental technique of conducting ICUs. This is not too surprising because, for one thing, not all learning happens with dopamine, and what we are interested is indeed measuring learning signals since we have assumed that they realize desires. Moreover, we can add to the list of problems with DP that the fact that

learning signals involve dopamine is not essential to their nature. As was mentioned in chapter 3, there may be creatures unknown to us that are subject to learning and whose minds are therefore affected by learning signals, but who don't have so much as a trace of dopamine within them; if our proposal cannot even in principle be extended to conducting ICUs between us and them, then it seems like the metaphysics of that proposal are somewhat unsatisfactory.

Therefore, instead of measuring dopamine, perhaps we should instead measure the potential of a learning signal to modify its owner's neural connections—its potential to cause learning to take place. Recall that, according to RTD, "The strength of a desire will thus come to the relative power of the desire to change neural connections, and so modify its owner's mind." (Schroeder 2004: 139) Also recall from the same chapter that the study conducted by Bao, Chan, and Merzenich (2001) involved the experimenters actually observing the VTA/SNpc changing neural connection strengths; so, it seems that, in principle, experimental observation of learning taking place is quite possible. How then do we modify DP so as to watch for learning as such, and not simply dopamine levels?

To begin answering the above question, DP is right with respect to expectations in that we have to take expected desire satisfaction into account in the sense of our earlier example of Jack and his stamp collection. In other words, we have to keep expectations of the owner of the learning signal the strength of which we are trying to measure into account. If we're measuring how much Jack wants his stamp collection to be complete, we can't let him think that he's about to be rewarded with the missing stamp. So, what we said about expectations in sketching DP holds true for the theory—

call it 'LM'—being sketched now. Like DP, LM also involves relativization, but in the following, more complex sense: I suggested, in discussing DP, that we relativise quantity of dopamine involved in the release of a learning signal to that involved in VTA/SNpc baseline firing; instead, what has to be relative is the amount of learning the signal being measured is able to produce, to the amount the strongest and weakest signals can produce—in other words, the ones that can cause the most and least amount of learning, respectively. To find out how much an individual *s* wants *P* to be the case, we need to know how the learning signal released when she represents *P* (under the conditions of expected desire satisfaction described above) compares in the potential amount of learning caused as a consequence of its release, to: (i) the signal or signals that, were they ever released, would have the potential to cause the lowest amount of learning out of any that s's brain is capable of releasing; and (ii) the signal or signals that, were they ever released, would have the potential to cause the highest amount of learning out of any that s's brain is capable of releasing. In addition to putting Jack in the right expectation conditions before rewarding him with the stamp and then checking to see how much learning can take place, we have to, in a manner of speaking, know the specifications of his reward system. We have to know them in order to put the fact that rewarding Jack with the missing stamp is capable of producing *x* amount of learning into a neurofunctional context.

Therefore, according to the proposal LM, under present consideration, to experimentally measure exactly how much Jack wants his stamp collection to be complete, we need to know *a lot.* It seems as though we aren't, at this point in time, capable of epistemically accessing the relevant information with our current level of

brain imaging technology, but learning can be observed and has been. LM requires a lot more detailed and extensive experimental observation, but there doesn't seem to be a relevant enough difference to throw the proposal into metaphysical disrepute. So, how do we develop it further and compare, e.g., how much Jack wants the stamp to how much Mary wants a post-graduate education?

In order to conduct interpersonal comparisons of desire strength along the lines of LM, we must extend the proposal to take into consideration the fact that the upper and lower bounds for how much learning a signal is capable of causing vary across individuals. The amount[82] of learning the strongest signal belonging to Jack is capable of causing may be less than the amount of learning the strongest signal that belongs to Mary can cause. And the same goes for the weakest signals; how little learning they can cause also varies across individuals. Thus each individual has her own range of potential learning; in order to make an interpersonal comparison between the strength of a desire belonging to Jack with the strength of one belonging to Mary by conducting a procedure along the lines of the one sketched in LM above, we have to take into account the different individuals' ranges of potential learning. How is that to be done?

Imagine two straight objects such as two metre sticks; suppose one stick is a metre long, but the other 1.5 metres long. Suppose, in addition, that each stick is marked with 10 equidistant lines that are numbered 1 to 10. Now, for the sake of analogy, suppose that Jack's desire that his stamp collection be complete is in fact equal in magnitude to Mary's desire that she have a post-graduate education, and that both of these are very strong desires. Moreover, suppose that the strongest learning signal that can be produced by Mary's reward system is capable of causing 1.5 times the learning

[82] 'Amount' here refers to how much a signal can strengthen or weaken neural connections.

Jack's strongest signal can cause; and likewise for the ratio of potential learning between Mary and Jack's weakest signals. In measuring the maximum potential of learning that can take place in Jack's mind as a result of having his stamp-desire satisfied and measuring the same for Mary and her education-desire, we shall have to use an analog of the metre stick for Jack, and the 1.5 metre stick for Mary. Nevertheless, to continue with the stick analogy, each of the two individuals' desires will be equally close to the line marked 10 on their respective sticks, because (we have supposed) their desires are in fact equal in strength. If Mary's desire were stronger than Jack's, then it would be closer to the 10 mark on *her* stick than Jack's would be to the 10 mark on *his*.

If we did not use some kind of analog to the numbered sticks, and simply compared how much learning Jack's stamp-desire can cause with the amount that Mary's education-desire can, then we would not have a means to compare the two. Acquiring the neurofunctional specifications for each individual's reward system, in the sense of determining the maximum and minimum amounts of learning that each is capable of causing, lets us know which stick to use for which individual.

As far as what the real-life counterpart to the numbered sticks from the above analogy might look like, I am not exactly sure. Hence, my proposal for interpersonally comparing desire strength does not include a method of mathematically representing desire strength. And though I have left many concerns regarding how we might epistemically access the real-life analogs to the sticks, I have specified what those analogs are metaphysically grounded in, viz. the potential of individuals' learning signals to cause changes in neural connection strength (i.e., to cause learning), relative

to the specifications of each individual's reward system. What I have said in the present chapter, combined with the assumptions that physicalism[83] about the mind is true and that Schroeder (2004) is right about the nature of desire, it doesn't seem plausible that interpersonal comparisons of desire strength can be dubbed *metaphysically* problematic.

Utility and ICUs

In the preceding section, we have, inter alia, explored what seems to be a plausible procedure (LM) for experimentally measuring the strength of individual desires, that I claim can in principle be extended to comparing desire strength across individuals; but our overall goal is to compare *utility* across individuals, hence, that is the next topic. And it seems reasonable to start by asking: On my view, what exactly is utility?

> Utility $=_{df}$ The extent to which an individual's intrinsic desires are satisfied, counterbalanced with the extent to which that individual's intrinsic aversions are satisfied.

First, there may be a sense in which the wording of the above definition—call the definition 'UTD'—could invoke some confusion, stemming from the term 'satisfied'. For some individual *j*, the more desires *j* has satisfied and the stronger those desires are, the *higher* her level of utility is; the more aversions *j* has satisfied and the stronger those aversions are, the *lower j*'s level of utility is. The word 'satisfied' often refers to a certain emotional state a person may be experiencing, but here, it does not. It simply describes a mind-world relationship: if a desire is satisfied, then what the mind wants and how the world is with respect to that desire, are the same. If an aversion is satisfied, then what the mind is averse to and how the world is with respect to that aversion, are

[83] Some kind of functionalism, type-identity theory, or something in between.

the same. A person with many strong aversions that are satisfied will experience anything but a *feeling* of satisfaction or pleasure. The wording of UTD having been clarified, let us proceed with unpacking the definition further.

Assuming that, through something like LM, we can in principle ascertain an empirical and experimentally derived measure of the strengths of individual desires, there are epistemological issues concerning how we are to take an inventory of *every* intrinsic desire that a human individual has in order to arrive at a measure of her level of utility. We cannot simply ask her to state all her intrinsic desires, for there is no guarantee that she could do this; desires are not *in* consciousness,[84] but are sometimes indirectly accessed by our conscious minds. Our emotions and actions give us some degree of insight into what it is that we desire, but that insight is far from infallible.[85] But, since the present work is primarily concerned with the metaphysics of ICUs—and only suggestive regarding informational or epistemological issues—I shall only say that, given the plausibility of something like LM, and under our assumptions (see chapter 1 above), it would be astounding if no way of inventorying an individual's desires turned out to be possible.

Going on, according to UTD, net intrinsic desire satisfaction is only half the story about an individual's level of utility, since net intrinsic desire satisfaction must be counterbalanced by net intrinsic aversion satisfaction. This is somewhat of an obstacle for the proposed solution to the problem of ICUs being defended here, as it is to

---

[84] See Schroeder (2004).

[85] I, e.g., have often thought that I no longer desired to be in the presence of a certain ex-girlfriend, but have often been proven wrong. Another good example, given to me by Tim Schroeder during a (2006) informal discussion, is a desire that many people seem to have, but often don't realize it; a desire that they verbally understand what is being said around them. They may come to know of the presence of such a desire upon spending a prolonged period of time in a foreign country surrounded by individuals speaking a language they cannot understand.

Schroeder's (2004) RTD, on which the solution is based. But, it is surely not a fatal obstacle; when our neuroscientific understanding of the punishment system catches up to that of the reward system, a more detailed philosophical analysis of the nature of punishment and punishment signals will follow. Hence, my theory of how to conduct ICUs remains incomplete, since half the story about utility, as I have defined it in UTD, is still incomplete. Once the neurological underpinnings of aversion are more thoroughly analyzed, the analysis can be deployed to fill in the blank, as it were, left in my theory by our insufficient understanding of the neuroscience of aversion. Again, I do not see how this incompleteness could be fatal to the overall project of metaphysically grounding ICUs that I have here undertaken.

So, let us assume, for the sake of clarifying UTD, that once we have an understanding of the nature of aversion that's parallel to our understanding of the nature of desire, an analog of LM can be developed for measuring the strengths of individuals' aversions. To get an accurate assessment of an individual's level of utility in accordance with UTD, each desire and aversion must be weighted according to its strength; stronger desires get weighted more heavily than weaker ones, and the same goes for aversions. Satisfied aversions have a negative impact on level of utility (and frustrated aversions have a positive impact on level of utility), whereas satisfied desires have a positive impact on level of utility (and frustrated desires have a negative impact on level of utility). If, e.g., a desire *D* and an aversion *A* are weighted the same (because they are of equal strength) and both are satisfied, then the two cancel out with respect to the impact they have to an individual's level of utility; taken together they, in other words, have no *net* impact on level of utility. If, e.g., a desire *D* and some other desire *F* are weighted

the same, but *D* is satisfied while *F* is frustrated, then *D* and *F* jointly have no impact on level of utility; and the same goes for any pair of equally weighted aversions such that one is satisfied while the other is frustrated. If, e.g., a desire *D* is weighted more heavily than an aversion *A* and both are satisfied, then the two have a positive net impact on level of utility.[86] How exactly the weighting is to be done, I leave to someone with a better understanding of mathematical approximations, but the general schema of how to ground measures of utility in accordance with UTD should be fairly clear by now.

As for ICUs, they should present no special problems, as long as my theory of how to compare desire strength across individuals is plausible. To elaborate, according to my theory of how to compare desires interpersonally and my definition of utility (UTD), necessary conditions for the possibility of conducting ICUs include these:[87] (i) that it's possible to put an individual whose desires we wish to measure in the right expectation conditions;[88] (ii) that it's possible to measure the potential amount of learning a signal causes under the right expectation conditions; (iii) that it's possible to measure the amount of learning the weakest signal that can be released by a given individual's reward system has the potential to cause, and the amount of learning that the strongest signal that can be released by that individual's reward system can cause; (iv) that (iii) is possible for all individuals equipped with reward systems; (v) that it's possible to inventory all the desires an individual has; (vi) that (v) is possible for all individuals; (vii) that what is said in (i) to (vi) about desires is true of aversions—that

[86] I won't point out all the logical relationships involved; the reader can surely deduce them from what I have said on an as-needed basis.
[87] They include others, viz. my background assumptions; see chapter 1.
[88] See the previous section of the present chapter.

'desire' can, more or less,[89] be substituted for 'aversion' with no loss of truth value. Conditions (i) through (vii) along with the assumptions outlined in chapter 1 are jointly sufficient to secure the possibility of conducting ICUs. That is, if the modal claims made in (i) through (vii) are true, and if my assumptions about the nature of mind and desire are true, then there seems to be nothing metaphysically odd about ICUs. We, in other words, have no good reason to doubt that ICUs are possible.

Concluding Remarks

To summarize what has been said in the present chapter, I have suggested an experimental procedure for measuring the strength of individual desires, called DP, and that it or something similar is a plausible means to compare the strength of desires interpersonally, as long as we take into account the interpersonally variant quantities of dopamine involved in the baseline firing patterns of the neurons belonging to different individuals' reward systems. DP, however, turns out to be, on balance, a bad way of measuring desire strength in virtue of relying on dopamine quantity measurement. Therefore, DP was discarded in favour of the more plausible LM that is not limited to measuring dopamine, relying instead on measurement of learning as such. Complications surrounding how LM is to be extended in order to interpersonally compare desire strength were also spelled out. I, in addition, offered a definition of utility (UTD) that construes an individual's level of utility in terms of net intrinsic desire satisfaction counterbalanced with net intrinsic aversion satisfaction. The proposal remains limited, however, in virtue of our highly incomplete understanding of the neurological underpinnings of aversion.

---

[89] More or less in the sense that when referring to aversions, rather than reward signals (as in the case of desires), we are dealing with *punishment* signals.

Chapter 5

This chapter is a summary of the sub-conclusions that have been drawn in previous chapters, along with a statement of the overall conclusion drawn from the present project. There is little in the way of novel philosophical work over and above that which is conducted in previous chapters, but this chapter does put things into context.

Summary

To recap what has been said in the way of philosophical argument in previous chapters, chapter 1 is an introduction and an inventory of background assumptions. Chapter 2 argues that John Harsanyi's (1977; 1982) attempt at reducing ICUs to a class of counterfactual *intra*personal comparisons of utility (intraCUs) is unsatisfactory for a number of reasons. Chapter 3 is a summary of Schroeder's (2004) work on desire. Finally, chapter 4 constitutes my effort to deploy Schroeder's (2004) theory of desire in a proposal for conducting ICUs, and comparing desire strength interpersonally. In chapter 4, I offered a theory of what ICUs and interpersonal comparisons of desire strength are metaphysically grounded in, viz. facts about learning signal strength and different individuals' reward system specifications.

Conclusion

The conclusion being highlighted can be summed up with the following remarks: If utility is construed in terms of intrinsic desire (and aversion) satisfaction, then interpersonal comparisons of utility (ICUs) pose no metaphysically interesting problems. I have argued that ICUs are possible, and I have given a theory of what they,

and interpersonal comparisons of desire strength, are metaphysically grounded in, viz. facts about learning.

The aforementioned conclusion, however, requires qualification in the following sense: The conclusion is plausible, only insofar as one accepts Tim Schroeder's (2004) theory of intrinsic desire (and aversion), viz. the Reward Theory of Desire (RTD) and physicalism about the mind, along with rejecting logical behaviourism.[90] The aforementioned three assumptions, along with the philosophical work done in previous chapters, secure the possibility of conducting reliable, empirically derived and verifiable ICUs.

However, for those intent on defining utility in terms of the satisfaction of preferences—as, e.g., some contemporary proponents of utilitarian moral philosophy are—the preceding work offers nothing in the way of argument beyond presenting what I take to be a more *metaphysically attractive* alternative. By 'metaphysically attractive' I mean that the proposal I offer, along with acceptance of the above mentioned assumptions, leaves little doubt about the *possibility* of conducting reliable and accurate ICUs, and is perhaps even suggestive of how that possibility might be actualized based on recent empirical research in the field of neuroscience.

Of course, if one has philosophical quarrels with assuming that physicalism and RTD are true, and that logical behaviourism is false, then I have no argument to the contrary. But, since I take it that physicalism is attractive for many well known reasons, that RTD is attractive for the reasons spelled out in Schroeder's (2004) work, and that logical behaviourism is highly implausible for reasons that no longer need to be

[90] There may be logical relationships between these assumptions (specifically between assuming RTD and the negation of logical behaviourism), but they are not interesting in the present context. Hence, there is no harm in thinking of the assumptions separately.

rehearsed, the fact that I have no arguments for accepting the aforementioned assumptions I take to be of limited philosophical import to the preceding proposal for solving the problem of ICUs.[91]

[91] At least the metaphysical aspects of that problem, such as the worry that accurate and reliable ICUs are impossible.

References

Bao, S., Chan, V. and Merzenich, M. 2001. "Cortical Remodelling Induced by Activity of Ventral Tegmental Dopamine Neurons." *Nature* 412, 79-83.

Berns, G., McClure, S., Pagnoni, G. and Read Montague, P. 2001. "Predictability Modulates Human Brain Response to Reward." *Journal of Neuroscience* 21, 2793-98.

Bright, B. *Foundations of Utilitarianism*. Unpublished

Cummins, R. 1989. *Meaning and Mental Representation*. Cambridge, MA: MIT Press.

Cummins, R. 1996. *Representations, Targets, and Attitudes*. Cambridge, MA: MIT Press.

Davidson, D. 1980. *Essays on Actions and Events*. New York, NY: Oxford University Press.

Deakin, J. 1983. "Roles of Serotonergic Systems in Escape, Avoidance and Other Behaviours." In Cooper, S. (ed.) *Theory in Psychopharmacology, Volume 2*. New York, NY: Academic Press. 149-93.

Deakin, J. 1998. "The Role of Serotonin in Depression and Anxiety." *European Psychiatry* 13 (Supplement 2), 57s-63s.

Dretske, F. 1988. *Explaining Behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Dretske, F. 1995. *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Fodor, J. 1990. *A Theory of Content: And other essays*. Cambridge, MA: MIT Press.

Fodor, J. 1998. *Concepts: Where cognitive science went wrong*. New York, NY: Oxford University Press.

Harsanyi, J. 1955. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *The Journal of Political Economy* 63 (No. 4), 309-321.

Harsanyi, J. 1977. *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press. 48-83.

Harsanyi, J. 1982. "Morality and the theory of rational behaviour." In Sen, A. & Williams, B. (eds.) *Utilitarianism and Beyond*. Cambridge: Cambridge University Press. 39-62.

Jevons, W. S. 1911. *The Theory of Political Economy: 4th Edition.* London: MacMillan and Co., Ltd. (1st Edition: 1871).

Kandel, E., Schwartz, J. and Jessell, T. 2000. *Principles of Neural Science: 4th edition.* New York, NY: McGraw-Hill.

Knutson, B., Adams, C., Fong, G. and Hommer, D. 2001. "Anticipation of Increasing Monetary Reward Selectively Recruits Nucleus Accumbens." *Journal of Neuroscience* 21, 1-5.

Little, I. M. D. 1950. A Critique of Welfare Economics. Oxford: Oxford University Press.

Ljungberg, T., Apicella, P. and Schultz, W. 1992. "Responses of Monkey Dopamine Neurons During Learning of Behavioral Reactions." *Journal of Neurophysiology* 67, 145-63.

MacKay, A. 1986. "Extended Sympathy and Interpersonal Utility Comparisons." *The Journal of Philosophy 83*, 305-322.

Millikan, R. 1984. *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.

Millikan, R. 1993. *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.

Morillo, C. 1990. "The Reward Event and Motivation," *Journal of Philosophy* 87, 169-186.

Pagnoni, G., Zink, C., Read Montague, P. and Berns, G. 2002. "Activity in Human Ventral Striatum Locked to Errors of Reward Prediction." *Nature Neuroscience* 5, 97-8.

Rolls, E. 2000. "Orbitofrontal Cortex and Reward." *Cerebral Cortex* 10, 284-94.

Rolls, E., Critchley, H., Mason, R. and Wakeman, E. 1996. "Orbitofrontal Cortex Neurons: Role in Olfactory and Visual Association Learning." *Journal of Neurophysiology* 75, 1970-81.

Romo, R. and Schultz, W. 1990. "Dopamine Neurons of the Monkey Midbrain: Contingencies of Response to Active Touch During Self-Initiated Arm Movements." *Journal of Neurophysiology* 63, 592-606.

Schroeder, T. 2004. *Three Faces of Desire*. New York: Oxford University Press.

Schultz, W., Apicella, P. and Ljungberg, T. 1993. "Responses of Monkey Dopamine Neurons to Reward and Conditioned Stimuli During Successive Steps of Learning a Delayed Response Task." *Journal of Neuroscience* 13, 900-13.

Schultz, W., Dayan, P. and Read Montague, P. 1997. "Neural Substrate of Prediction and Reward." *Science* 275, 1593-99.

Schultz, W. and Romo, R. 1990. "Dopamine Neurons of the Monkey Midbrain: Contingencies of Response to Stimuli Eliciting Immediate Behavioral Reactions." *Journal of Neurophysiology* 63, 607-24.

Schultz, W., Tremblay, L. and Hollerman, J. 2000. "Reward Processing in Primate Orbitofrontal Cortex and Basal Ganglia." *Cerebral Cortex* 10, 272-83.

Sterelny, K. 1990. *The Representational Mind: An introduction.* Cambridge, MA: Blackwell Publishers.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behaviour.* Princeton, NJ: Princeton University Press.

Waelti, P., Dickinson, A. and Schultz, W. 2001. "Dopamine Responses Comply with Basic Assumptions of Formal Learning Theory." *Nature* 412, 43-48.

Waldner, I. 1972. *The Journal of Philosophy* 69, 87-103.

Wise, C., Berger, B. and Stein, L. 1973. "Evidence of *a*-Noradrenergic Reward Receptors and Serotonergic Punishment Receptors in the Rat Brain." *Biological Psychiatry* 6, 3-21.