A THEORY OF STRATEGY

DAVID E. LUNE
B.A.Sc., Systems Design Engineering,
University of Waterloo, 1988

A Thesis
Submitted to the School of Graduate Studies
Of the University of Lethbridge
In Partial Fulfilment of theg
Requirements for the Degree

MASTER OF ARTS

Department of Philosophy
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

For Declán, Ailís, and Sinéad.

Abstract

The notion of 'strategy' plays a central role in game theory, business, and war. This thesis offers an understanding of the term that can be rendered canonical for all three contexts. I argue first that rational behaviour is either complacent or non-complacent. Second, what makes non-complacent rationality distinct is reconnaissance and predictive deliberation. And so third, what we can count as 'strategic' behaviour is the employment of reconnaissance and deliberation in pursuit of alternative practices of higher utility.

Acknowledgements

There are people with whom I have had the privilege of proximity. Their encouragement, support and instruction have made this thesis possible:

Dr. Paul Viminitz – Supervisory Committee Chair;
Dianne Lune, Heather Vander Schaaf and Anthony Vander Schaaf – editing; and
The Staff of the Department of Philosophy, University of Lethbridge.

Figure 1 is used with permission from Pearson Education, Boston, Massachusetts.

Table of Contents

Introduction

Practical affairs are characterised by competition, success, boredom, excitement, envy, and fatigue. Most of us believe that we control – to some degree – these characteristics using deliberation, inferential reasoning and action. In all of this there is a presupposition that our rational decisions somehow emancipate us from the chains of nature.[1] Even so, we regularly pass up opportunities to engage in rationally deliberated decision-making. Instead we opt to base our choices on non-inferential intuition knowing full well that our intuitive decisions are prone to mistakes. For example, one's intuition may guide one to invest in a certain stock without having performed a reasonable investigation into the performance of the stock. And yet, choosing by intuition can be the most prudent option, however faulty, whenever conditions do not allow for a reasonable investigation.

For those not bothered by errors that result from intuitive choices, rationality still plays a role in the retro-justification of observations. Observation is, after all, theory-laden. What you perceive depends on the theory you hold. Rational retro-justification employs deliberation to contextualise observations and reasoning to accommodate these observations according to whatever view is currently held. But for an observation to count as meaningful it must be weighed according to its value in both reinforcing and disproving one's worldview. To retro-justify observations in a way that *only reinforces* whatever theory one currently holds can be dangerous. One may, for example, console oneself in believing that God has willed the loss of a child through sickness. Doing so may help deal with emotional distress, but such a theological rationalisation may prevent the investigation into the cause of the sickness or the administration of treatment. In certain cases where retro-justification is used solely to reinforce one's current worldview one's decisions may not be prudent. As Bertrand Russell has pointed out, most people would die sooner than think; in fact, they do so.

This thesis is an investigation into the cognitive virtues that separate rational retro-justification from strategic reasoning. Why should one be inclined to participate in such an investigation? The answer points back to the competition, boredom, envy, and fatigue associated with practical affairs. Mere retro-justification does not suffice under these conditions. Other conditions are preferred: conditions where one takes responsibility for the environment in which one interacts, creates an environment of integrity and trust, works to be a valued member among a group with common interests, achieves healthy relationships,

[1] Our reason, we believe, allows us to alter deliberately geological transformation or biological evolution.

is prosperous, and lives a balanced lifestyle. I will argue that these latter conditions result from thinking that is beyond that which is rational in the case retro-justification. These conditions result from thinking strategically.

Establishing that the aforementioned conditions result from strategic thinking is not a simple task. The term 'strategy' is used prolifically in describing military and political manoeuvring, in game theory, and among members of the business community. But while its use is frequent, those referring to – or counselling on – strategy often understand the term in either a vague or esoteric sense. It is not surprising, then, that the definition of strategy varies in the literature.[2] Further, prolifically used words (or terms) often become ornaments of the vernacular. The use of 'strategic' to predicate an agent's behaviour may simply be adding lustre to what is already straightforward rational behaviour. Still, there is an intuitive sense that 'strategic' somehow describes an additional cognitive virtue beyond that of rationality alone.

An investigation into the underlying nature of strategic behaviour will root out intuitive themes that distinguish strategic from rational behaviour. My goal, however, is to replace intuitive distinctions between rational and strategic cognitive virtues with those that follow from a sound depth-logic. In addition to identifying these cognitive virtues, what follows from a sound depth-logic is threefold: the intuitive notion that 'strategic' behaviour is somehow more shrewd than straightforward rational behaviour is validated; the term 'strategic' is recognised as an unambiguous and practically descriptive adjective; the cognitive virtues identified allow for the use of 'strategic' across a variety of fields (military, business, game theory and more). The problem, then, is developing a concept of strategy whereby the use of the adjective 'strategic' – in describing a property of an agent – has unambiguous applications in the decision-making aspects of military, game theoretic, and business interactions. While the definitions of 'strategic' vary in each of the aforementioned fields, I will show that these definitions do share a commonality that transcends the notions of strategic behaviour in the individual arenas.

In order to establish this transcendental concept I will adopt a reductionist approach through which I will describe the minimal properties of strategic behaviour under social interaction and determine whether the properties of rational behaviour, under these same minimal conditions, are the same as the properties of strategic behaviour. Whatever

---

[2] Army General André Beaufre claims that strategy is "the art of applying force so that it makes the most effective contribution toward achieving the ends set by political policy". See Army General André Beaufre, *An Introduction To Strategy*, Frederick A. Praeger Publishers, 1965, p13. Game theoreticians, on the other hand, take 'strategic' to describe the behaviour of an agent in a defined domain of interactivity that models social situations. See Andrew M. Coleman, *Game Theory and its Applications in the Social and Biological Sciences*, Butterworth-Heinemann Ltd., 1982, p3. And a recent vogue among members of the business community is to take 'strategy' as the synthesis of a short-term rational planning and long-term "social muddling". For a definition on social muddling see Henry Mintzberg, *Strategy Safari: A Guided Tour Through The Wilds Of Strategic Management*, The Free Press, 1998, p180.

differences exist between rational and strategic cognitive virtues will illuminate specific behaviours that are taken to be strategic in the three fields mentioned above.

Part I

Rational Decision-Making Agents

Chapter 1

Decision-Making Agents: The Foundational Concepts

The term 'agent' will be used throughout this paper. An agent is the author of an action where the action results from the author's ability to make choices. One might suppose that only humans are capable of acting as agents. But while humans often do possess the deliberative skills required to make choices, being human is neither a necessary nor a sufficient condition for agency. The status of agency can be dependent on the domain of interactivity in which the author of an action may be found. Consider, for example, the flyball governor shown in Figure 1. The domain of interactivity considered here is that under which the horsepower output of a steam engine is regulated. The domain of *regulation of horsepower output of the steam engine* defines the author of the action, i.e. the governor.[3]

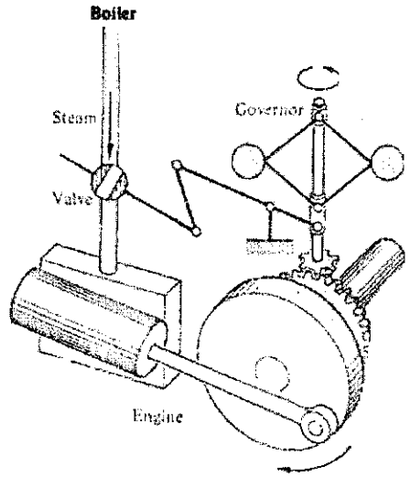

Figure 1. – A flyball governor assembly.[4]

Alternatively, consider the domain of moral behaviour. Intuitively one may claim that all humans are moral agents since moral behaviour is taken to be concomitant with human behaviour. There are counter examples, however: small infants or those inflicted with severe idiocy are not considered moral agents since they are not taken to have developed the capacity for moral behaviour.

---

[3] In the case of the flyball governor the term *deliberative capability* might seem inappropriate. It is possible, however, to view the flyball as authoring its behaviour according to the algorithms of centripetal forces.

[4] Richard C. Dorf, *Modern Control Systems*, Addison-Wesley Publishing Company, 1986, p4.

1.1 rationality: a property of agents or actions?

The conditions for rational agency that I propose follow from primarily two sources: John Rawls' "Justice as Fairness" and Peter Danielson's *Artificial Morality*. From Rawls, a rational agent has the following characteristics:

> they know their own interests more or less accurately; they are capable of tracing out the likely consequences of adopting one practice rather than another; they are capable of adhering to a course of action once they have decided on it; they can resist present temptations and the enticements of immediate gain; and the bare knowledge or perceptions of the difference between their condition and that of others is not, within certain limits and in itself, a source of great dissatisfaction.[5]

Notice that Rawls adds a fourth condition to the traditional description of a rational agent: a condition that allows for differences among agents within certain limits. Rawls' fourth condition takes into account that, independent of differences among agents, one agent is likely to imagine the restrictions that another would place on her if their circumstances were reversed.

According to Rawls', the principles of justice for the agents described above are twofold. First, each agent is to have an equal right of maximum liberty compatible with a like liberty for all. Second, unequal social and economic conditions among agents should be arranged to benefit the agents that are least advantaged where offices and positions are arranged to ensure they are open to all. Rawlsian agents, then, may observe differences among one another, but these differences are not so great as to eliminate the possibility that one agent would make the same choice as another, if their circumstances were reversed.

It is important to distinguish whether the property of rationality results from a domain of activity in which the agent is interacting or from the agent itself. According to Danielson, the property of rationality does not apply to the domain of activity. Instead rationality is a property of the author of the action, (i.e., the agent). Danielson's reasons for taking rationality to be – what I call – an *authoritive* property are as follows.

> First, testing different theories of rationality creates the possibility of mixed populations of players. In this environment, the question of the best action is poorly defined. One needs the best set of actions given many interactions, with several (kinds of) players. Players are sets (generators) of actions; this is the appropriate level for speaking of success or failure.[6]

The mixed population to which Danielson refers is based on David Gauthier's claim that an accurate characterisation of interacting rational agents must account for two dispositions: the traditional straightforward maximising agent and the responsive agent

---

[5] John Rawls, "Justice as Fairness", printed in King and McGilvray, *Social and Political Philosophy*, McGraw-Hill, 1973, p318. While Rawls' later work *A Theory of Justice* is more widely recognised, I take "Justice as Fairness" to be foundational to Rawls' later work and appropriate for a depth-logical analysis.

[6] Peter Danielson, *Artificial Morality: Virtual Robots for Virtual Games*, Routledge, 1992, p64.

disposed to conditional co-operation. For Gauthier, this responsive agent protects itself from exploitation by first scrutinising the disposition of the agent with which it is interacting. The responsive agent decides to co-operate only with other agents it deems to be co-operators.

Whereas Gauthier proposes a population comprised of two dispositions, Danielson proposes a population in which four dispositions are evident. The players Danielson constructs are disposed to 1) unconditional straightforward maximising, 2) unconditional co-operation, 3) conditional co-operation (following Gauthier), and – an interesting fourth – 4) reciprocal co-operation. This last disposition is characterised by an agent that co-operates when and only when co-operation is sufficient and necessary for the other's co-operation. As a result, the reciprocal co-operator exploits the unconditional co-operator whereas the conditional co-operator will not.[7]

Determining an agent's rational status according to outcomes (or states of affairs) is difficult given variability of time horizons (or epochs of interest). For example, eating chocolate cake for dessert tonight might be rational in today's time horizon but might not be rational if one's time horizon includes winning the annual sibling weigh-in at Christmas. But one can avoid the problem of varying epochs of interest by evaluating the success or failure of an action based on the success or failure of an agent's disposition.

Dispositions, while properties of agents, are characterised by sets of actions, or practices.[8] Unlike dispositions, practices are defined according to their constituents in a domain of interactivity. For example, the practice of baseball is defined by its constituents: a pitcher, a batter, a playing field, etc. Notice that the constituents of a practice can be so arranged as to neutralise potential conflicts arising from varying epochs of interest. Following the above example, an agent disposed to regular exercise can have chocolate cake tonight and still win the weigh-in at Christmas.

Measuring the success of the actions that agents generate provides Danielson with a second reason for focusing on rationality as an aspect of agents and not just the domain of interactivity:

> [a]ctions do not exist without players. To construct a generator of actions is to construct a player and evaluation should acknowledge this. The actions alone cannot be evaluated without the context of the player that generates them.[9]

---

[7] Ibid., p89.

[8] This notion of a practice is taken from Melinda Vadas' "A First Look At The Pornography/Civil Rights Ordinance: Could Pornography Be The Subordination of Women?", *The Journal of Philosophy*, Volume 4, 1987, pp492-497. Vadas distinguishes her own view from that of Alasdair MacIntyre's in *After Virtue* where, according to Vadas, "MacIntyre seems to regard practices as functionally related to the human virtues, and I do not" (page 493). The view presented in this thesis borrows Vadas' notion of a practice and, at the same time, holds the position that strategic behaviour results from a 'cognitive virtue' of rational agents.

[9] Op. Cit. Some might argue that claiming 'actions necessitate agency' is doing little more than re-describing a metaphysical necessity for existence. But while *cogito ergo sum* is generally taken to be an archetypal metaphysical

Following Danielson, then, the agents we will be considering have various dispositions according to the sets of actions they create and where the actions may also be the constituents of a practice.

1.2 rational agency

From Rawls and Danielson, rational agents can be defined as those that are capable of:

1. identifying relationships between two or more states of affairs,
2. identifying a preferred state of affairs and acting in pursuance of that preferred state,
3. re-describing sets of actions as practices, and, finally,
4. discovering their preferences among practices.

Both Rawls and Danielson see the rational agent as one that can resist the enticement of immediate gain in order to access greater utility at some time in the future. But Danielson's *rationality thesis* goes on to describe a co-operative disposition as *more rational* than the straightforward maximiser disposition. Says Danielson:

> a player capable of responsively constraining herself to pursue outcomes mutually beneficial to itself and other similar players is substantively more rational that a straightforward maximiser.[10]

Danielson uses ordered utility scores under various interactive conditions to substantiate the above claim. The co-operative dispositions do better by virtue of their ability to scrutinise another agent's disposition and adjust their choices accordingly.

The idea of more or less rational plays an important role in our intuitive notion of what it is to think strategically. Strategic thinking is somehow more cunning or shrewd than straightforward rational behaviour. Yet, Danielson does not call on the adjective 'strategic' to single out one disposition over another. Instead, for Danielson, each disposition is itself a strategy, and "[i]n less than completely transparent worlds, we should expect mixed populations of more or less sophisticated strategies".[11]

1.3 decision-making and the roles of utility, reconnaissance, and folk psychology

Thus far it is clear that the rational agent is one capable of discovering her preference for one practice over another. This preference can be measured in *utiles* where the utile is the basic unit of desirability. Rational agents are those disposed to maximise on utility. This

---

claim, the point applies equally well in non-metaphysical socially interactive matters: where there are thoughts there is a thinker. Both Hobbes and Russell have pointed out that the claim 'there are thoughts, therefore there is a thinker' is more appropriate than the Cartesian claim since Descartes presupposes that the thinker is himself. Most importantly, Danielson is not debating the agent's existence but instead whether rationality should be couched in terms of domain in which the agent is interacting, or a virtue of the agent itself.

[10] Peter Danielson, *Artificial Morality: Virtual Robots for Virtual Games*, Routledge, 1992, pp195-196.

[11] Ibid., p196.

disposition is often couched in folk-psychological terminology such as goals, objectives, missions, etc.[12] The notion of desirability is implicated in each of the aforementioned terms in a manner that follows from the eudaimonic tradition: a tradition beginning with the Epicureans and developed, incrementally, through the work of Jeremy Bentham, John Stuart Mill, and John Rawls.

The utility of a practice accounts for both the desirability of a practice and the energy expended in identifying, evaluating, and actualising the practice. The resources expended in identifying and actualising a practice may have a negative utility, or *disutility*. This disutility can be described folk psychologically in terms of identifying one's own beliefs and desires (which goes toward establishing various levels of consciousness) and determining the beliefs and desires of other agents (which goes toward predicting another agent's behaviour). Beyond the folk psychological framework, resources may be described in terms of calories, time, money, and in military context may include human lives.

The term *reconnaissance*, while peculiar to military context, will be extended to describe the energy expended in determining one's own beliefs and desires and the beliefs and desires of others. Knowledge, then, in the form of justified beliefs, results from reconnaissance. Knowledge plays a significant role in determining the functional relationships among states of affairs, including how another agent's behaviour might influence future states of affairs.

1.4 strategic agents

Having discussed what it is to be rational, let me make some preliminary remarks regarding what it might mean for an agent to be strategic. Intuitively, strategic behaviour is a subset of rational behaviour where the former is taken to be somehow more shrewd or cunning than the latter. Furthermore, one may claim that only humans possess the characteristics required for rational decision-making and thus only human beings can be strategic. But being human is neither a necessary nor a sufficient condition for being rational. Consider the following.

Automatically guided vehicles (AGVs) are used extensively in the auto industry for delivering parts to designated assembly points on a lengthy assembly line. These AGVs are equipped with sophisticated optical and electromagnetic sensory instrumentation, which allow the vehicles to identify among various states of affairs.[13] In addition, these AGVs carry an on-board programmable logic computer that allows the AGV to identify functional relationships among states of affairs. The vehicle's computer evaluates the consequences of

---

[12] It is important to note that term 'folk-psychological' is a technical term specific to philosophical investigation into the theory of mind. A folk-psychological approach involves discussing mental states in terms of beliefs and desires, and thus, intentional states (and of course emotions). The term should in *no way* be interpreted as implying that visions, missions, objectives and the like are 'pop-psychological' terms.

[13] The sensory equipment provides the AGV with information like when and where the AGV is located within the production facility at any given time, current inventory of parts on board, the speed at which the line is operating (in cars/hour), etc.

delivering certain parts to certain locations and identifies a sequence for delivery that will both maximise production line speed and allow for parts replenishment. Once the vehicle's computer has determined the optimal sequencing, it mobilises to deliver the parts against the program it has determined. In this example, these AGVs fulfil the requirements of a rational agent. Moreover, they are equipped with processing abilities such that they perform their task more efficiently than could a human being. So, while there may be those wishing to assert that only, if not all, humans can count as rational agents, there are certainly examples of non-human agents that qualify as rational according to the definition provided above.

Like 'rational', the term 'strategic' is increasingly being used to describe the behaviour of non-human agents. In the business community, for instance, banks and companies may be described as rational or strategic agents. Likewise, historical military battles are often presented in a way that both reveals the strategic decisions made by the chief commanding officer and expresses strategic behaviour as belonging to the vehicles of war.[14] And political agents such as nations are also ascribed strategic properties. For example, the United States of America is referred to as having a strategy against terrorism. On the view presented in this thesis, a complete picture of strategic behaviour must account for its use to describe businesses, nations, military forces, etc. Still, referring to a business, a nation or a military force as behaving strategically is problematic. Nations, teams, corporations, and the like are not usually taken to be rational individuals. Instead, nations, teams, and corporations are names that refer to a group of individuals all of whom share a posited affiliation[15]. The problem is that while rational individuals are taken to hold beliefs and desires, groups are not. So, although it is not uncommon to say that a team has a strategy, it would be incorrect to presuppose that a group holds the requisite beliefs needed to generate, and the desires required to evaluate, any number of possible outcomes.

---

[14] Examples of the former are Wellesley, The Duke of Wellington, and Napoleon in the case of Waterloo and The Bismarck and The Hood in the case of the latter.

[15] In the case of a nation, for example, each of the individual agents is inter-associated by affiliation as Canadians. What makes them Canadians is the agreement to posit upon all of those who reside within the geographical boundaries generally recognised as the landmass named Canada; the identifying adjective pertaining to the collective is a name usually based on the name of the landmass.

Chapter 2

The Non-Human Rational Agent

As we have seen above, a rational decision-making agent must be capable of identifying possible states of affairs and predicting which states of affairs are likely to come about. But groups of affiliated individuals do not possess the requisite abilities for prediction. Thus, we are faced with the problem of explaining how names of aggregated, affiliated individuals are frequently described and treated as if they possess rational decision-making characteristics.

2.1 the synecdochic option

There are a number of possibilities for resolving this issue. One might argue that since a corporation is comprised of rational agents, the corporation as a whole has the characteristics of that which it is comprised. On this view, a corporation's status as a rational agent is a matter of *synecdoche*: a figure of speech in which the collective whole is attributed the characteristics of one or more of its constituents. A nation, for example, can be said to hold the belief that the sun will rise tomorrow since the belief that the sun will rise tomorrow is taken to be held by one, many, or all of the affiliated constituents of which the nation is comprised.[16] In another example one might attribute to the United States the characteristic of wanting low oil prices since many U.S. citizens want low oil prices and would feel justified in claiming that 'the United States wants low oil prices'.

The synecdochic approach, however, is flawed. The use of synecdoche leverages a move in which one or more of the characteristics of one or more of the constituents is ascribed to the name that designates the collective. But a valid move where a *property of constituent* is posited as a *property of collective* may be jeopardised by the fallacy of composition. The fallacy of composition is an error associated with transposing characteristics from the constituent to the collective. An aeroplane, for example, may be comprised of parts, all of which are lightweight; it does not follow, however, that the aeroplane itself is lightweight. From the fact, then, that a corporation is comprised of constituents, many of which are rational individuals, it does not necessarily follow that the corporation itself will behave in a rational manner consistent with having a cognitive theatre in which logical inferences are made from beliefs and desires. And while there are many cases where transposing properties

[16] I say 'one, many, or all' here since an autocratic nation may be taken to hold the belief of its sovereign; a single party state like that of the former Soviet Union may be taken to hold the beliefs of its 'politburo'; a democratic nation may be taken to hold the beliefs that are shared by the majority of constituency.

from the constituent to the collective may result in an accurate characterisation of the collective, it is not necessarily so. Thus, one cannot count on the synecdochic approach as a means of accurately characterising a rational agent nor predicting its behaviour.

2.2 a second alternative: incorporation via internal decision structure

In the business arena it is often difficult to distinguish between the intentions of the individuals employed by the corporation and the intentions of the corporation itself. Says Peter French in "The Corporation as a Moral Person":

> [t]ypically, we will be told that it is the directors, or the managers, etc., that really have the corporate reasons and desires, etc., and that although corporate actions may not be reducible without remainder, corporate intentions are always reducible to human intentions.[17]

But French asserts that corporations can be distinguished as entities with intentions, responsibilities and obligations above and beyond those of the aggregate collection of biological persons of which the corporation is comprised. French's argument aims at driving a wedge between an individual's intentions and a corporation's intentions by showing that 1) corporations have reasons for doing things, and 2) these reasons are referentially opaque to an individual's reasons for doing things. French uses a Shakespearean context to provide an example of referential opacity. While it is true that 'Hamlet intentionally kills the person hiding in Gertrude's room', it is not true that 'Hamlet intentionally kills Polonius', since Hamlet is unaware that the person hiding in Gertrude's room is, in fact, Polonius. The figurative wedge that French describes exists in two forms. There is referential opacity among first, the subjects of decision-making and among second, the objects of decision-making.

In the first case – where individual decision-makers are referentially opaque to the corporate entity – opacity results from what French calls a *Corporate Internal Decision Structure*, or CID Structure. The CID Structure is comprised of an organisational chart, which shows relative authority in decision-making and a set of rules (policies and procedures) which describe how a corporate decision is to be reached.

> [The] primary function of the CID Structure is to draw experience[s and knowledge] from [the biological persons operating at] various levels of the corporation into a decision-making and ratification process.[18]

On French's view, the CID Structure *incorporates* the beliefs, desires, and acts of biological persons into that of the *corporate citizen*.

For example, a group of three individuals, all of whom are employed by a corporation, may believe that a new facility should be built and intend to justify their belief. The justification is described as a formal request for capital funds (RCF) where the RCF

---

[17] Peter French, "The Corporation As A Moral Person", found in *Business Ethics in Canada*, Edited by D.C. Poff & W.J. Waluchow, Prentice Hall Inc., 1991, p87.

[18] Ibid.

explains the costs associated with building the facility, the functionality of the facility, the return on the investment, etc. The justification, however, goes well beyond the knowledge and experience of the initial three individuals. Each aspect of the justification is formulated by drawing on the experience and knowledge of a number of individuals at various levels of the institution. These biological persons are chosen based on their expertise pertaining to the specific project. They are asked to express views that both expose the risks and illuminate the benefits associated with the endeavour. The RCF is then circulated among specific individuals in various positions of relative authority in order to gain formal approval to proceed with the project. In this way the CID Structure incorporates the acts of biological persons in such a way as to form a corporate intention versus that of the aggregate of individuals employed by the company. On approval of the project, it is accurate to claim that the company intends to build a new facility. Notice, however, that it is not necessary that a majority of individuals, each of whom work for the company, intend to build a facility. Without a CID structure, the aggregate intentions of the majority of individuals working for the company are referentially opaque to the company's intentions.[19]

As regards the objects of decision-making, French argues that policies and procedures exist within the CID Structure that subordinate an individual's ambitions to the needs of the corporation. Consequently, the objects of decision-making are referentially opaque between individuals and corporations. For example, consider that Executive X intends to hire the best person for Job Y. At the same time, Executive X intends to increase the company's profits. But it does not necessarily follow that the best person for Job Y will increase the company's profits. It is entirely possible that the best person for the job is, on Executive X's view, her son, since hiring her son will provide addition income into Executive X's household. But her son has little or no experience in Job Y and will likely burden the company rather than increase its profits. The CID structure ensures that the best person for Job Y is the one that will increase company profits.

On French's view, the referential opacity that exists between both the subjects and objects of actions results from the policies and organisational hierarchy that comprise the CID Structure. At the same time, this structure "provides the requisite devices to licence the predication of corporate intentionality".[20] French's corporate agent is irreducible since the CID Structure places the responsibilities of actions on the corporate subject and not on an aggregate of individuals. As such, French's argument does not involve the fallacy of composition and thus is more substantive than the synecdochic one. There are, however, socio-economic examples that are counter to French's proposal. For example, if a corporation fails to comply with government environmental policies, both the non-

[19] The CID structure that French proposes offers a solution to the error Mill makes in *Utilitarianism* where Mill claims that if all individuals are maximising their utility then the aggregate utility of all individuals is maximised.

[20] Peter French, "The Corporation As A Moral Person", found in *Business Ethics in Canada*, Edited by D.C. Poff & W.J. Waluchow, Prentice Hall Inc., 1991, p87.

compliant individuals and the corporation are held accountable.[21] So while French's approach manoeuvres around the fallacy of composition, the use of a CID Structure does not elevate the status of a corporation to a point where corporate intentions, in all circumstances, are referentially opaque with an individual and thus irreducible.

2.3 the socially interactive Dennettian agent

An alternative approach for characterising decision-making agents – one that avoids both the errors associated with the fallacy of composition and problems with reducibility – is the Dennettian approach known as *the intentional stance*.[22] According to Daniel Dennett, there are any number of methodologies for explaining and predicting the behaviour of a system S. One might, for example, decide to reduce the system's behaviour to the basic principles of physics. Explaining behaviour in purely physical terms is possible since macro-physical behaviour is reducible to micro-physical explanations. There are at least two worries however. First, there exists an, as yet, unresolved discrepancies among micro-physical theories. For example, the theory of general relativity is thus far incompatible with the current theory of quantum mechanics.[23] Second, and more importantly, even if we resolved theoretical discrepancies, describing how a rational agent might act in terms of microscopic particles would be an extensive and inefficient exercise under practical conditions.

Alternatively one may wish to explain and predict the behaviour of system S using an astrological approach. While the application of astrological principles may prove less complicated than, say, physical principles, one may find that astrological principles are ineffective for consistent and accurate predictions about specific states of affairs.

For Dennett the most successful means of characterising and predicting the behaviour of system S (which for Dennett includes as examples bats, computers and humans) is ascribing intentionality to the system by taking the intentional stance.

> Here is how it works: first you decide to treat the object whose behaviour is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires

---

[21] In Canada, for example, individual biological persons deemed to be directly responsible for environmental infractions are held accountable for their actions.

[22] Both problems are avoided by showing intentionality to be indivisible. Thus, the move where a property of a constituent is ascribed as a property of the whole is a valid one. One might object by claiming that beliefs and desires are the constituents of intentionality. But intentionality is not comprised of beliefs and desires. Instead, beliefs and desires are propositional: they are what the intentions *are about.*

[23] It should be noted that Brian Greene hypothesises that super-symmetrical vibrating string theory unifies general relativity with quantum mechanics. See Brian Greene, *The Elegant Universe,* Vintage Books, 1999, p14.

will in most instances yield a decision about what the agent ought to do; that is, what you predict the agent will do.[24]

Consider the following example where one ascribes intentionality to system S in order to predict its behaviour. A common thermostat is little more than a bimetallic element and an electrical contactor. The element is shaped like a clock spring. Sufficient temperature will expand the element and rotate the contact such that it makes or breaks an electrical connection. The workings of a thermostat are comprehensible for most people with a basic science background. But suppose we introduced an ancient ancestor, even one of great intellect, to thermostatic functionality. How would, say, Thales react to such a device? It is reasonable to assert that Thales, after some observation and deliberation, would explain the behaviour of this small box on the wall by attributing beliefs and desires to it. He might suggest that the box had the intentions of keeping the room temperature regulated, and as such, concludes that the thermostat held the requisite beliefs and desires from which to form intentions. Now, of course, one could counter that once Thales understood how the thermostat 'really'[25] worked that he would no longer attribute intentionality to the mechanical unit. This is likely true. But notice that, even when we fully understand how something works, indeed, even if we have designed the thing ourselves, there are still instances – usually under exigent conditions – where it is advantageous to attribute intentionality to non-rational agents.

Consider the programmer of a computerised chess opponent. The programmer writes, say, some 4000 lines of code. She understands each line of code and how the lines together form the subroutine architecture. Consider that the programmer decides to play the computerised opponent she has programmed. One might argue that the programmer could easily beat the computer. After all, she designed (and by way of memory has access to) the code. Then again, it would be extremely difficult to memorise 4000 ordered lines of code. Fair enough. So for this example we will allow the programmer access to a hard copy of the code. But a hard copy of the code is still not enough. We must also grant the programmer the time to work through the subroutine sequencing after each of the computer's moves in order for her to determine the proper counter move. Under these conditions, armed with a copy of the code and the time to work through the sequencing, the programmer could defeat the computer.

These are not, however, the conditions under which a normal match is conducted. In regular match play the time taken to decide on each move is cumulatively recorded in order to limit the game's duration. So, in regular match play, while the programmer may still have access to the code she would 1) have access only by memory, and 2) be constrained by the time she has to think about the code's structure. In the end, the programmer finds that, even though she fully understands how the computer program is structured, her chances of

[24] Daniel Dennett, *The Intentional Stance*, MIT Press, 1987, p. 17.

[25] This assumes that thermostats are not, in fact, rational agents.

beating the computer are better if she treats the computer as though it has the requisite beliefs to play the game of chess, the desire to win the match, and that it will behave accordingly.

The Dennettian approach allows us to ascribe beliefs and desires, not only to objects, but also to groups such as nations, teams, and corporations, so long as these names are taken to possess the property of being a system.[26] It would seem, then, that it is possible to avoid the problems of reducibility and the fallacy of composition and at the same time study the rational interaction between nations, corporations, or any system S, by ascribing intentionality to affiliated groups. Further, by taking decision-making agents to be Dennettian agents, we satisfy the game theoretic condition that players involved in strategic interaction are taken to be rational agents.

[26] One worry with Dennett's position is that of indeterminacy. Since Dennett is both a functionalist and an instrumentalist, he holds that beliefs and desires are posited as useful fictions for describing patterns of behaviour with no ontological connection to inner states of 'beliefs' and 'desires'. But this leaves Dennett with the problem of an innumerable ways of translating any type of behaviour. Using Quine's example, the behaviour of pointing toward a rabbit and uttering 'gavagai' could mean that 'gavagai' translates to 'rabbit', or 'rabbitlike', or 'undetached rabbit part', etc. But what counts as a system is exactly that which overcomes Quine's worry with indeterminacy. According to System Design theory, stable systems (like Dennett's) are characterised by a converging output even on the same input. In Quine's example, this would mean that while one could point and uttering 'gavagai' – and doing so could mean any number of possible things – a feedback signal allows one to compare input and output and hone in on, or converge toward, the specific meaning associated with 'gavagai'. So long as Dennett takes the system response from system S to be stable (i.e. converging), a case is made for determinacy.

Chapter 3

The Notion of Strategy

The term 'strategy' stems from the notion of generalship: a concept rooted in military affairs. Military definitions for what constitutes a strategy vary among commentators. Yet, on a survey of the literature, certain themes accumulate. Carl von Clausewitz, perhaps the most well known commentator on military strategy, claims the role of strategy (for purposes of war) is to determine the best use of the available resources to compel the enemy to do one's own will.[27]

While historically rooted in the military, the term 'strategy' is most prolifically, and least esoterically, used in the world of practical affairs and business. Henry Mintzberg, Bruce Ahlstrand, and Joseph Lampel have provided numerous accounts of strategic thinking in business organisations and have contributed substantive literature on the subject of strategy in the arena of business affairs.[28] According to Mintzberg and team:

> [a]sk someone to define a strategy and you will likely be told that a strategy is a plan, or something equivalent – a direction, a guide or course of action into the future, a path to get from here to there.[29]

Accompanying this normative notion that a strategy is a planned course of action toward a goal is the sense that certain short-term setbacks may be acceptable in the face of achieving long-term rewards. Notwithstanding that the distinction between long-term and short-term itself yields a sorites paradox, consider what Mintzberg et al. take to be a pervasive problem in defining strategy.

> Ask [the same] person to describe the strategy that his or her own organisation or that of a competitor actually pursued over the past five years – not what they intended to do but what they really did. You will find that most people are perfectly happy to answer the question, oblivious to the fact that doing so differs from their very own definition of the term.[30]

---

[27] Carl von Clausewitz, *On War*, edited and translated by M. Howard and P. Paret, Princeton University Press, 1976, p79.

[28] I have chosen to cite primarily from *Strategy Safari*. This work explains the various views held by academics studying business strategy. In comparison to other texts, which focus on theoretical details, *Strategy Safari* captures the theoretical views of the various schools in a manner that is accessible to practitioners. This is useful in that the final chapter of this thesis will focus on explanatory and evidential support from data gathered in the (business) field.

[29] Henry Mintzberg, Bruce Alhstrand, & Joseph Lampel, *Strategy Safari: A Guided Tour Through The Wilds Of Strategic Management*, The Free Press, 1998, p9.

[30] Ibid.

On Mintzberg and team's account, it is unclear whether a strategy is a normative set of instructions directing one toward achieving some goal or whether a strategy is a descriptive account revealing what choices agents have made and how they made the choices they did. The mainstream literature on business strategy (targeted at practitioners) asserts normative prescriptions for strategy formation. For example, formation of a strategy according to The Positioning School prescribes inductive reasoning techniques, including game theoretic modelling. On the other hand, strategy formation according to The Learning School prescribes methods for retro-justifying why things turned out the way they did. In the end, business oriented definitions of the term 'strategy', especially those targeted at practitioners, are ambiguous. This ambiguity makes it difficult to identify properties that distinguish an agent as strategic rather than merely rational. That being said, business literature dedicated to strategy is advantageous in that it provides numerous accounts of – what might be intuitively called – strategic behaviour. These accounts are useful for testing proposals for a transcendent concept of strategy.

The game theoretician definition of strategy is more rigorously defined than those of the other two arenas considered. Game theoreticians attempt to model social interactions, called *games,* among rational agents. According to Andrew Coleman social games are characterised by three properties:

1. there are two or more [rational] decision makers, called *players*;
2. the players have a well-defined preference among possible outcomes, so that numerical *payoffs* reflecting these preferences can be assigned to all players for all outcomes;
3. each player has a choice of two or more ways of acting, called *strategies*, where the strategy choices are governed by the players' preferences among outcomes.[31]

In the case of interactions among deliberative rational agents, games are usually modelled using numerical payoffs according to ordinal rankings. In this thesis, ordinal rankings will be used to describe the relationships among outcomes. When modelling behaviour under evolutionary conditions game theoreticians often opt for a payoff structure that is expressed cardinally.[32]

---

[31] Andrew M. Coleman, *Game Theory and its Applications in the Social and Biological Sciences*, Butterworth-Heinemann, 1982, p3.

[32] While ordinal rankings are expressed as 1st, 2nd, and so on, cardinal numbers are used in counting to indicate quantity. Note also that I refer to evolutionary conditions here and feel it is necessary to make some brief comments about Darwin's influence regarding what I take to be these conditions. First, Darwin recognised the importance of Charles Lyell's geological theories and in turn applied them to interaction among biological organisms. Lyell argued that the Earth's surface was not as it had always been but that it had changed over time: slowly in the case of wind and rain and at times abruptly as with volcanoes and earth tremors. Darwin argued that processes analogous to erosion were at work in the biological arena and it was these processes which resulted in differences among genera in both living organisms and those found in the archaeological record. Second, Darwin proposed an explanation for the actual process at work. Darwin adopted Malthus' view that the population would increase geometrically while the food supply only arithmetically, the result of which

The *outcome* of a game is dependent on both agents' choices. While outcomes depend on the choices each agent could make, the game theoretic definition of a player's *strategy* is based on the player's preference among possible outcomes. This is an important feature of game theoretic analysis since preference rankings often illuminate a dominant choice for each player. For example, suppose two players each have a choice between U for up and D for down. There are four possible outcomes: (U,U), (U,D), (D,U), (D,D), where the first letter represents the choice of the first agent and the second letter the choice of the second agent. Suppose that the rankings for each player are as follows:

Outcome	Player 1	Player 2
U,U	1st	2nd
U,D	2nd	4th
D,U	3rd	3rd
D,D	4th	1st

Notice that Player 1 always prefers to choose up over down. In this case, Player 1 has a *dominant strategy* and should always choose up to maximise on preference regardless of what Player 2 chooses. Player 2, on the other hand, always prefers to make the same choice as Player 1 rather than choosing the opposite of Player 1. Player 2 does not have a dominant strategy. Instead, Player 2's preference is dependent on Player 1's choice.

The game theoretic definition of a strategic agent requires only that rational agents interact such that the agents can discover their preferences among outcomes. *Prima facie*, the game theoretic definition of strategic is the leading candidate for a transcendental definition of strategy. But while the game theoretic definition is rigorous, the property 'strategic' applies to the domain of interactivity in which the players are interacting and does not rest on a cognitive virtue of the agent. This, however, is problematic.

Recall that, following Danielson's argument in section 1.2, the property of rationality should be attributed to agents and not to the domain of interactivity in which the agents are engaged. Danielson's reasons hold equally well when applied analogously to the property of strategy. First, actions must be evaluated in the context of the agents that generate the actions. Second, given the variety in kinds of agents, the appropriate means for articulating 'strategic' agency is based the success or failure of the agent's actions. Accompanying Danielson's two reasons is a third: ascribing a property to a domain of interactivity invites problematic speculation as to who is narrating the ascription. Under the game theoretic definition, strategic behaviour requires the interaction of two or more agents. These agents are taken to be rational and capable of discovering a preference among outcomes. But the

---

would force biological organisms to adapt over time as they competed for food. In terms of modelling these conditions game theorists are interested in replacing the notion of food supply with the idea that agents require energy, for 1) subsistence and 2) other higher level functions, which is expressed in terms of calories. Thus, cardinal values are used by those game theoreticians wishing to evaluate behavioural dispositions for agents under evolutionary conditions in terms of calories.

question that needs answering is 'rational for whom?' There are three possibilities to be considered.[33]

The first possibility answers the question from the first person perspective. In this case, an agent would take *itself* to be rational. Taking oneself to be rational requires self-knowledge. The agent would be aware of something like its own mental theatre in which cognitive states are played out, experiences are conceived, desires are felt, and intentions are formed. Consider a game in which two agents interact and where both agents take only themselves to be rational. Notice that there is no guarantee that either agent takes the other agent to be rational. So while these types of games meet the requirements of the game theoretic definition of strategy – both agents are taken to be rational – either agent might take the interaction to be parametric.[34] Alternatively, one agent, in order to conceal its disposition, may indeed wish to be taken as non-rational by the opposition but takes itself to be no less rational. In this case, for one agent the game is strategic but for the other player the game is parametric.

The second possibility answers the 'rational for whom?' question from the second person perspective. In this case both agents need to take *only* the other agent to be rational. But it is unclear how one secures a view where it is indeed possible to take another agent to be rational without taking oneself to be rational. In fact, according to José Luis Bermúdez such a view is paradoxical. The paradox follows from Bermúdez's account of self-consciousness. The paradox, argues Bermúdez, goes as follows:

1) The only way to analyze [any agent's] capacity to think in a particular range of thoughts is by analyzing the capacity for the canonical language expression of those thoughts;
2) [Any agent's] 'I' thoughts are canonically expressed by means of the first-person pronoun;
3) Mastery of the first-person pronoun requires the capacity to think 'I' thoughts.[35] Thus,
4) an agent could not analyse another agent's capacity to think in the range of which 'I' thoughts are a part without taking itself to have the capacity to think 'I' thoughts.

So, while it is logically possible for agents to take only each other to be rational – and such a possibility does meet the game theoretic definition of strategic – there are limitations placed on the range of thoughts to which a rational agent would have access. Specifically, strategic interaction would be that which falls outside the range in which either agent thinks 'I' thoughts. Of course, there might well be non-rational 'I' thoughts, but any preference for one state of affairs over another based on such thoughts could not arise as deliberative act.

---

[33] All of the possibilities are explored as second order predication.

[34] Parametric games are distinguished from strategic games in that the former is a game where one agent is taken to be rational and the other is taken to be non-rational.

[35] José Luis Bermúdez, *The Paradox of Self-Consciousness*, The MIT Press, 1998, p24.

An additional concern comes to bear under what game theoreticians call *zero-sum games*. Zero-sum games are strictly competitive games in which the outcome yields a winner and a loser. In the context of warfare, the number of dead or wounded in battle might determine the winner or loser among two opposing agents using force to resolve their dispute.[36] But it is rarely the case that the purpose of war is to dispose of the enemy's soldiery. The purpose of war is usually governed by a political agenda which calls for bending the enemy's will to one's own such that the enemy behaves in a way that maximises one's own utility.[37] The means by which the enemy's will is bent is a matter of establishing the material conditions, or appearance of the material conditions, that will create in one's enemy the belief that they have lost. An agent in such a doxastic state can be described as being in a state of *resignation*.

Consider that the taking of a fortress or establishing a secure position is rarely done solely for the economic value of the fortress or the land. Instead, these tactics are intended to evoke in one's enemy the belief that the fortress has been taken or a position has been lost. Under the doxastic state of resignation, the enemy is likely to surrenders their position. In most cases of conflict, the force employed to create resignation is far less than that required to dispose of the other agent (or all of the constituents of the agent). A siege, for example, employs a technique of economic isolation that reduces the supplies to those immured such that they are left with the choice between giving up and starvation. It is likely, however, that resignation will occur well before the effects of malnutrition take hold.

In answering the question 'rational for whom?' resignation is a reflexive condition where the resigning agent takes itself to be the loser (among the winner and loser) in a zero-sum game. But recall that we are discussing interacting agents that take only the other agent to be rational. Under this sort of interaction, an agent could neither win nor lose since, by not taking itself to be rational, it would be unable to establish a functional relationship among states of affairs where it would take itself to be either a winner or loser.

The third possibility responds to the '...for whom?' question from the third person perspective. The notion of the third party agent stems from *the ideal observer*. The ideal observer is disinterested in the outcome of the game but is at the same time keenly interested in being well informed and vividly aware of all the facts relevant to the two decision-makers.[38] Furthermore, a third party observer is an agent whose actions do not affect the possible outcomes of the game. But while the outcome of the game is independent of the third party observer's actions, on the game theoretic view, the interacting agents' status (i.e. rational or non-rational) *is* dependent on the third party observer. For example, a third party

---

[36] This is the notion of a battle within the framework of warfare according to Army General André Beaufre. See Beaufre, *An Introduction To Strategy*, p22.

[37] There have, however, been cases of a political agenda including genocidal policy – the Shoah, for example.

[38] This notion of an ideal observer is taken from Gilbert Harmon, *The Nature Of Morality*, Oxford University Press, 1977, p44. The ideal observer is also described by Jean Jacques Rousseau in *Social Contract*, edited by H.J. Tozer, Swan Sonnenschein & Co., 1897, p134.

observer may take Agent A to be rational and Agent B to be non-rational in which case the game is parametric. There are certainly cases, however, where Agent A takes itself and Agent B to be rational. Suppose Agent A is playing against a chess computer, Agent B. While the third party observer takes the chess computer to be non-rational, Agent A fares better by ascribing rational behaviour to the chess computer. Thus for Agent A, the game is strategic whereas for the observer the game is parametric.

One might argue that the game theoretician avoids the complications of answering the 'rational for whom?' question by stipulating that both agents must take themselves and the other agent to be rational. As with the chess computer example above, however, this still causes problems since it would be difficult to secure a position in which the chess computer took itself or even the other agent to be rational. Still, for the ascribing agent, the game would be game theoretically defined as strategic.

In all three possibilities considered the designation 'strategic' hinges on answering the question 'rational for whom?' Yet, depending on the perspective from which this question is answered, any one domain of interactivity could be strategic or parametric. As a result, game theoretically defined strategic agency is ambiguous: and, for the purposes of a transcendental definition of strategic behaviour, inappropriate. The game theoretic notion of strategic interaction does, however, offer the advantage of a more specific condition for strategic agency than that of business and military affairs. Further, using the term 'strategic' to describe choices of preferences among outcomes and not the outcomes themselves seems to bode well with our intuitions that strategic thinking somehow informs us of the best actions to take, independent of what the opposition might do. As a result, the game theoretic approach of ascribing 'strategic' to an agent in a certain domain of interactivity seems like a good candidate for the transcendental view.

But, following from Danielson, the appropriate language for describing the success or failure of a strategic decision is one in which behaviour is expressed as a disposition, or cognitive virtue, of the agent. While game theoretic analysis allows for this virtue in the form of identifying dominant strategies, the virtue is not linked to the property 'strategic'. Instead strategic agents are so predicated according to the conditions under which the agents are interacting. That being said, there is no intent here to dismiss game theoretic techniques from playing a role in determining a candidate for the transcendental view of strategic behaviour. Game theoretic analysis is effective in testing for preference maximisation among choices. Furthermore, a successful transcendental view of strategic behaviour must account for the use of the term 'strategy' in game theoretic circumstances.

Segue

Strategic versus Rational

The game theoretic approach for distinguishing between rational and strategic behaviour includes taking 'rational' to be a property of the agent and taking 'strategic' to be a property of the domain of interactivity. I am, however, interested in both 'rational' and 'strategic' as properties of the agent. At the same time, our intuitive notions tell us that being strategic is somehow more cunning or clever than straightforward rational behaviour. But these same intuitions do not inform a non-ambiguous distinction between rational and strategic agents. Nor do these intuitions explain what is indicated by the word 'strategy' when claiming that 'a hobbyist has a strategy for restoring a vintage vehicle' or 'a patient has a strategy for coping with the pain of treatments'.[39] While our intuitions tell us that strategic behaviour is a subset of rational behaviour no distinction between the two is clear. We must conclude then, that for lack of a depth-logical distinction, strategic is simply a synonym for rational.

[39] Here I am presupposing that neither the hobbyist nor the patient takes neither vehicles nor treatments to be rational agents, as would be required by the game theoretician in order to meet the game theoretic definition of strategic.

Part II

Choice Theories:
Rational and Empirical

Chapter 4

Rational Choice Theory

Intuitively, the adjectives 'strategic' and 'rational' describe certain shared behaviours but are still distinguishable. An investigation into rational decision-making may help us establish a distinction between the two. Rational choice theory is an appropriate starting place for this investigation. Rational choice theorists are concerned with determining what choices a rationally self-interested agent should (or will) make based on the expected utility associated with a certain state of affairs.[40] The development of modern rational choice theory is thought to have originated with Blaise Pascal's *Wager* and developed, in incremental stages, through Daniel Bernoulli, and John von Neumann and Oskar Morgenstern: the latter developments leading directly to game theory. The following section explores key developments contributing to current rational choice theory. What is provided below is by no means an exhaustive account of all contributions. Nor does it suggest that the significant contributions to rational choice theory were only subsequent to Pascal. For the purposes of this paper the developments in rational choice theory – as a formal theory – will be introduced starting with contributions from Thomas Hobbes.

4.1 developments in rational choice theory

In *Leviathan*, Hobbes draws a parallel between rationality and algorithmic processing. In his consideration "Of Man", Hobbes claims that reason "is nothing but Reckoning (that is, Adding and Subtracting) of the Consequences of generall names agreed upon, for the *marking* and *signifying* of our thoughts".[41] For Hobbes, reasoning results from one's computational ability to relate states of affairs (i.e. marking), and one's ability to demonstrate the significance of this computation to others using a common language (i.e. signifying).

In *La Logique*, Antoine Arnauld advances the significance of computational ability in rational behaviour by asserting that rational thought takes into account probabilistic computation.

> To judge what one must do to obtain a good or an evil one must consider not only the good and the evil in itself but also the probability of its

[40] There are some exceptions. De Souza argues, for example, that the normative/descriptive distinction is an ambiguous one. See Ronald De Souza, "Modelling Rationality: A Normative of Descriptive Task", printed in *Modeling Rationality, Morality and Evolution*, edited by Peter Danielson Oxford University Press, 1998, p119.

[41] Thomas Hobbes, *Leviathan* p18.

> happening or not happening, and view geometrically the proportion that all these things have together.[42]

Arnauld asserts that rational decisions should be made in the context of both the virtue of an outcome (couched in terms of its good or evil) as well as the likelihood of its occurrence.

Blaise Pascal's 'wager' stands as an archetypal argument in which choice is based on examining both the probability of an event's occurrence and the importance (or virtue) of the event. In *The Pensées*, Pascal uses the wager in arguing that it is rational to embrace Christianity by examining the outcomes associated with the following two sets of logical possibilities: 1) that God does/does not exist, and 2) that one may/may not hold the Christian doctrine. Pascal's examination yields the following four expected outcomes from most preferred to least preferred:

1) one holds the Christian doctrine and God exists whereby the stakes to be gained are infinite (in heaven);
2) one does not hold the Christian doctrine and God does not exist whereby the stakes to be gained are finite (since they can only be gained in earthly form);
3) one holds the Christian doctrine and God does not exist whereby the stakes to be gained are finite but less than that gained in 2), (presumably, holding a Christian doctrine reduces the payoff);
4) one does not hold the Christian doctrine and God exists whereby the stakes to be lost are infinite (eternity in hell).

While Pascal's intent is a rational justification for holding the Christian doctrine (the degree to which he is successful is itself an issue) his key contribution to rational choice theory is introducing a technique whereby outcomes are examined according to both importance and probability of expected outcome.

Daniel Bernoulli formalised Pascal's contribution by placing numerical values on expected outcomes. Bernoulli himself initially referred to these values as moral worth, and later coined the term *expected utility*. Bernoulli also suggested that incremental increases in utility diminished in relation to successive increments of a commodity acquired. Utilitarianists adopted Bernoulli's notion of expected utility along with the supposition that utility as a quantitative measure caused preference. Morgenstern and von Neumann, however, supposed the reverse.

According to Morgenstern and von Neumann, expected utility is only a description of relative preference and not a cause. By placing full emphasis on the former, Morgenstern and von Neumann argued that utility should be ordered according to an agent's choice among outcomes. Furthermore, they argued, only after understanding the ordered rankings (or preference) could the notion of utility be used for a meaningful analysis of what counts

---

[42] Antoine Arnauld, *La Logique : ou, L'art de penser* , excerpt from Richard Jeffrey's "decision theory" found in *The Cambridge Dictionary of Philosophy*, edited by Robert Audi, Cambridge, 1999, p207.

as a rational decision. This method of comparing preferences among outcomes forms the basis of game theoretic analysis.

4.2 rational choices in game theory

Recall from section 1.3 that game theoreticians attempt to model social interactions among rational agents. The canons of game theoretic analysis are adequately demonstrated by considering only games involving two players each having two choices. Consider the following example based on a problem originally presented by von Neumann and Morgenstern. In this game, The Final Solution, the players are Holmes and Moriarty, based on the characters created by Sir Arthur Conan Doyle.

		Moriarty	
		Canterbury	Dover
Holmes	Canterbury	Holmes detrains as Moriarty waits in ambush.	Holmes detrains safely but the chase continues.
	Dover	Holmes detrains and safely escapes.	Holmes detrains as Moriarty waits in ambush.

Figure 3: Outcome matrix for The Final Solution game.[43]

As usual, Moriarty intends to kill Holmes. Furthermore, Moriarty is equipped to do so unless Holmes can reach safe passage from England at Dover. Holmes boards a train in London headed for Dover with one stop at Canterbury. As the train departs, Holmes sees Moriarty on the platform and assumes, correctly, that Moriarty has secured passage on a faster train and will reach Dover before Holmes. Figure 3 describes the four possible outcomes: Holmes detrains in Dover with Moriarty waiting in ambush; Holmes detrains in Canterbury with Moriarty waiting in ambush; Holmes detrains in Dover with safe passage abroad; and Holmes detrains in Canterbury but the chase continues.

Game theoretic analysis distinguishes between an *outcome matrix*, shown in Figure 3, and a *preference matrix*, shown in Figure 4. Outcome matrices describe the functional relationships between choices and likely states of affairs. In the case of Figure 3, the outcome matrix describes what would happen if, for example, both Holmes and Moriarty chose to detrain in Dover. Preference matrices, on the other hand, describe the ordinal preference each player has for the states of affairs describe in the outcome matrices. The

[43] This description of the game is based on that of Morgenstern and von Neumann 1944, pp176-178.

preference matrix associated with The Final Solution game is shown in Figure 4 where a player's highest choice is ranked 1st to the lowest ranking, 4th.

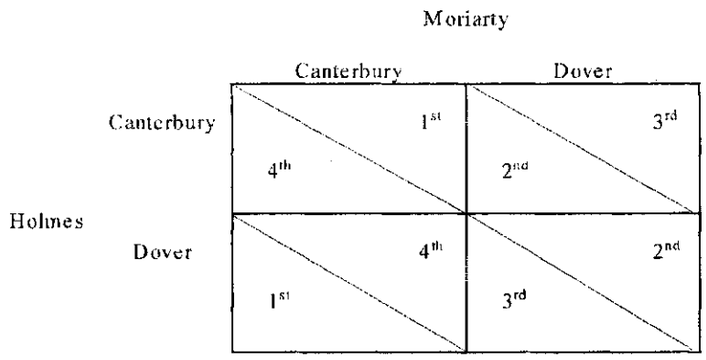

Figure 4: Preference matrix for The Final Solution game.

The numbers in the lower left-hand corners signify Holmes' ordinal preferences according to the outcomes described in Figure 3; the numbers in the top right hand corners are Moriarty's. Notice that if Moriarty chooses to detrain in Canterbury, Holmes' preference is to detrain in Dover over Canterbury. In fact, this particular outcome is optimal for Holmes. On the other hand, if Moriarty decides to detrain in Dover, Holmes prefers to detrain in Canterbury to avoid ambush. Presumably Holmes prefers ambush in Dover to ambush in Canterbury in the event that Moriarty's ambush fails. A similar analysis is possible for Moriarty.

Games in which the optimal preferences for both players coincide are termed *co-ordination games*. Games in which the players' optimal preferences are mutually opposed are termed *games of pure conflict* or *zero-sum games*. Little more will be said about these first two types of games. In all other circumstances, the optimal preferences for each player neither coincide nor are mutually opposed: these games are termed *mixed-motive games*. According to Rapoport and Guyer, there are seventy-eight distinct formulations of two player mixed-motive interactions of which only twelve are *symmetrical*. Symmetrical games are those where the preference structure remains unchanged if the agents switch positions. Notice that symmetrical games reflect the conditions that a Rawlsian agent would take to be just. Agents may observe differences among one another, but these differences are not so great as to restrict the choices available to either agent, should their circumstances be reversed.

Eight of these twelve symmetrical games have *optimal equilibrium points* in which case, both players' optimal choices coincide. As a result, neither player has any incentive to deviate from their current choice so long as the other player does not deviate. How players find their way to an optimal equilibrium, if in fact they do, is interesting. However, once both players' optimal choices coincide the game is no longer interesting. The four games remaining from the original twelve are those without optimal equilibrium points, which make them of great interest to those wishing to model social interaction. These four games have been deemed

worthy of special names; the games are Leader, Battle of the Sexes, Chicken, and Prisoner's Dilemma.[44]

Leader is usually explained in terms of two agents, both of whom are operating a vehicle and both are at a North American intersection intending to turn left. To co-operate, C, is to let the other driver go first; to defect, D, is to turn before the other driver. The possible outcomes are:

C,C – both drivers wait for the other to go first, missing all chances to advance;

D,D – both drivers go co-incidentally resulting in a collision;

C,D – one driver goes first and the other waits; and

D,C – the other driver goes first and one waits.

The preference matrix for leader is shown below in Figure 5.

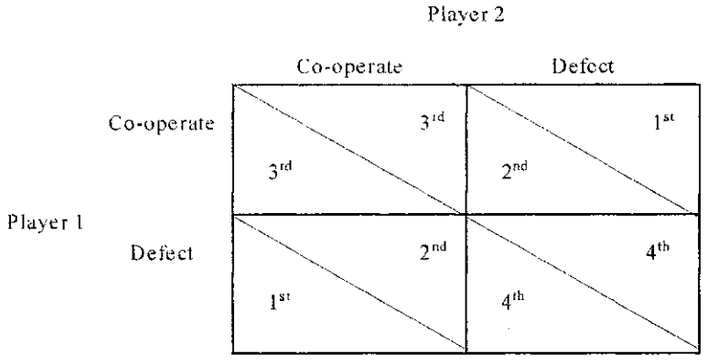

Figure 5: Preference Matrix for Leader.

Battle of the Sexes is usually described in terms of two players deciding between going to a romantic movie, the preference of Player 1, and going to a boxing match, the preference of Player 2. The four possible outcomes are described below:

C,C – both players go to their respectively preferred venue but go alone;

D,D – both players go to their respectively non-preferred venue and go alone;

C,D – one player agrees to go to their non-preferred venue but the players go together; and

D,C – the other player agrees to go to their non-preferred venue but, again, the players go together.

The preference matrix for Battle of the Sexes is shown in Figure 6.

---

[44] Andrew M. Coleman, *Game Theory and its Applications in the Social and Biological Sciences*, Butterworth-Heinemann Ltd., 1982, pp107-108. The four games are outlined here and will be further explored in Chapter 7 during an examination into the explanatory force of a transcendental concept of strategy.

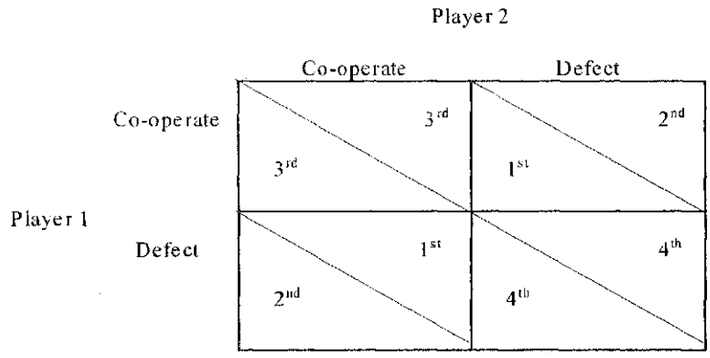

Figure 6: Preference Matrix for Battle of the Sexes.

The game of Chicken is characterised by two players driving cars headed toward each other: to stay the course is to defect and to swerve out of the way is to co-operate. The four possible outcomes associated with Chicken are:

C,C – both players swerve, missing each other;

D,D – both players stay the course resulting in a head on crash;

C,D – one player stays the course and the other swerves;

D,C – the other player stays the course and one swerves.

The preference matrix associated with Chicken is shown in Figure 7.

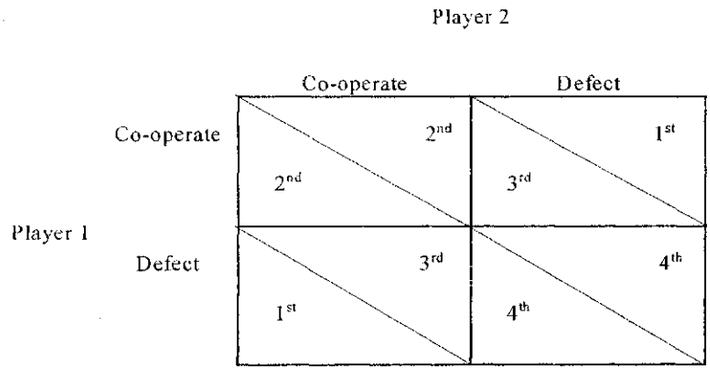

Figure 7: Preference Matrix for Chicken.

Finally, the game of Prisoner's Dilemma is characterised by two players both charged with committing a crime and held in custody under interrogation. To co-operate in Prisoner's Dilemma is not to lay guilt on the other player; to defect is to claim the other player is guilty. The four outcomes are typically as follows:

C,C – both players do not 'rat out' the other, thus receiving a sentence of 4 years each;

D,D – both players do 'rat out' the other, thus receiving a sentence of 7 years each;

C,D – one player 'rats out' the other while the other does not, thus the first gets 2 years while the other gets 10 years; and

D,C – the other player 'rats out' the first, while the first does not, thus the first gets 10 years while the other gets 2 years.

The preference matrix for Prisoner's Dilemma is shown in Figure 8.

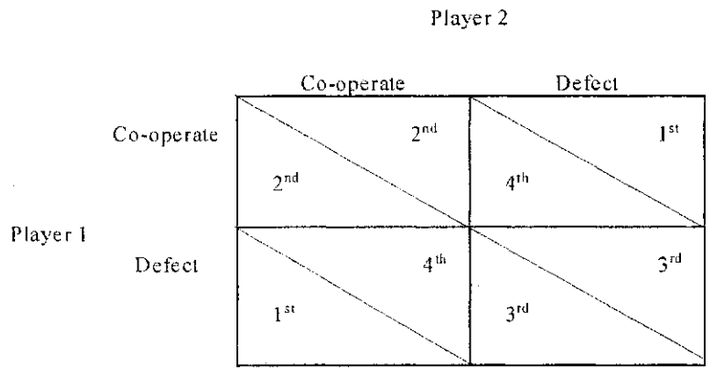

Figure 8: Preference Matrix for Prisoner's Dilemma.

An analysis of a Prisoner's Dilemma reveals an interesting challenge for the players involved. Recall that earlier in the chapter we discussed the notion of a dominant strategy. What is interesting about a Prisoner's Dilemma is that defection is a dominant strategy for both players. According to the Prisoner's Dilemma preference matrix, a rational agent should always defect regardless of the other player's choice. If, for example, the other player co-operates, a rational agent's 1st choice dominates over the 2nd. Conversely, if the other player defects, a rational agent's 3rd choice dominates over the 4th. Our intuitions, however, run counter to what appears to be the rational choice. Many people, when faced with possibility of being an agent in a Prisoner's Dilemma, choose to co-operate, often based on the claim that 'it's the right thing to do' rather than choose the dominant strategy of defection.

4.3 considering all possible outcomes and the role of ceteris paribus

Under game theoretic analysis it is preferable to present choices as logical complements. Presenting choices as logical complements ensures that the decision-maker views the alternatives in an idealised fashion where all possible outcomes are taken into account prior to making any decision. That being said, the Holmes/Moriarty case above does not present choices as logical complements. Had this been the case, the choices available to each player would have been *Dover* and *Not Dover*, for example. By presenting Dover and Not Dover as the available choices, all possibilities (including, for example, a locomotive engine failure) are accounted for in the analysis. Still, it is not always practical to present choices as logical complements. The possible and likely circumstances that would fall

under a Not Dover/Not Dover choice are extensive. In fact, the possibilities are so great that there is insufficient detail from which to describe a practical outcome of a Not Dover/Not Dover choice. As a result, even though the use of logical complements accounts for all possible states of affairs, providing an appropriate description of an outcome requires greater detail than logical complementary choices would allow. In order to overcome the impracticality of using logical complements, game theoreticians often treat non-complementary choices as logical complements by stipulating the caveat of *ceteris paribus*. In the Holmes/Moriarty case both players have the choice to detrain in Canterbury or Dover where Dover and Canterbury are treated as logical complements, all other things being equal. The use of *ceteris paribus* captures the notion that choices take into account all possible outcomes by eliminating those outcomes that are taken to be inconsequential to either player.

4.4 the implications of reconnaissance in rational decision-making

Decisions are rational when the preference for one state of affairs is identified among a number of likely future states of affairs. But settling on what states of affairs are likely to come about requires an agent to expend energy. The term reconnaissance describes the activities that are required to determine likely future states of affairs, including the energy that is expended in predicting another agent's behaviour.

Consider a situation where a rational agent, the protagonist, attempts to determine the functional relationship between states of affairs when interacting with another agent, the antagonist. On the Dennettian view, the protagonist's best chance of predicting the antagonist's behaviour is by adopting the intentional stance: that is,

1) treat the antagonist as a rational agent, then
2) figure out what beliefs the antagonist ought to have, given its place in the world and its purpose, then
3) figure out what desires the antagonist ought to have, on the same considerations, and finally,
4) predict how the antagonist will act to further its goals in the light of its beliefs.

In order to follow through on Dennett's instructions the protagonist must expend energy to 'figure out' the antagonist's beliefs and desires. The *amount* of energy, however, is left unspecified.

One would expect that the amount of energy expended in figuring out the other agent's beliefs and desires would depend on the situation. Suppose, for example, that two agents encounter each other in front of a sports arena. The protagonist wishes to purchase tickets for tonight's game; the antagonist wishes to sell tickets for tonight's game. The protagonist needs only to hear the antagonist bark out, "tickets, who needs tickets?" to infer that the antagonist desires to sell the tickets and believes that he will find a buyer by yelling 'tickets, who needs tickets?' while standing in front of the arena. The protagonist could

certainly expend further energy to figure out other beliefs and desires the antagonist might have. For example, the protagonist could pursue a line of questioning aimed at determining if the ticket seller was a theist or an atheist. Or the protagonist could enquire into the antagonist's political beliefs. In fact, the protagonist could expend significantly more energy in figuring out the antagonist's beliefs and desires than is required to purchase tickets for tonight's game.[45]

Of course it seems absurd, or at least uneconomical, to pursue an extended series of questioning in order establish all the beliefs and desires of another agent. It must be the case, then, that when Dennett tells us to 'figure out...given the other agent's place in the world and its purpose', Dennett is presupposing that the energy expended in figuring out the other agent's beliefs and desires must be finite. Dennett is suggesting, then, that it only makes sense to adopt the intentional stance according to the *principle of economy*, where agents make decisions that favour maximising the attainment of certain ends on finite means.

Let us consider the impact of the principle of economy as it relates to total expected utility. Recall that total expected utility is a function of the utility of the expected state of affairs plus the disutility of reconnaissance. According to the principle of economy, then, one can maximise on utility by 1) maximising the consumption of whatever commodity delivers the state of affairs, 2) minimising on the disutility of the cost of the commodity (in this case reconnaissance), or 3) both. In the case under consideration, then, the protagonist can maximise on the utility of buying the tickets by minimising the energy used in determining the antagonist's beliefs and desires.

Suppose another situation in which a human protagonist wishes to minimise the amount of energy expended determining the beliefs and desires of a non-human antagonist (say, a canine). In this case, the protagonist:

1) wishes to predict the behaviour of the canine antagonist,
2) takes her own beliefs and desires to be characteristically human,
3) takes herself to behave according to functional relationships among certain states of affairs (for example, when a human desires food, she goes to a location where it believes food exists),
4) ascribes to the antagonist characteristically human beliefs and desires. The protagonist may then
5) conclude that the antagonist will, in a characteristically human fashion, identify functional relationships among certain states of affairs. As a result,

[45] What does 'significant' mean here? As we saw earlier, according to System Design theory, stable systems (like Dennett's) are characterised by a converging output even on the same input. The amount of energy that could be expended on interactive reconnaissance is determined by the comparing input and output, and hone in on, or converge toward, the specific meaning associated with 'ticket's, who needs tickets?'. What is at issue here is how much introspection can be used to converge on what is meant (and believed and desired) by the other agent.

6) the protagonist may predict that, when the dog desires food, it will go to a location where, the dog believes food exists.

Whether or not the dog actually experiences beliefs and desires played out in a mental theatre is not significant. Instead, let us focus on the ascription of characteristically human beliefs and desires to the canine. Notice that the protagonist used introspection to determine the beliefs and desires to be ascribed. Notwithstanding that the protagonist could have asked the dog what its beliefs and desires are, or constructed an elaborate laboratory experiment to test theories about what beliefs and desires are the best for ascription, the protagonist finds that introspection minimises the energy expended on reconnaissance. The energy required to ascribe mental states to the dog is only the calories needed to establish premises 2) through 5).

In the example above, the protagonist predicted the antagonist's behaviour by running a mental simulation of how the antagonist would act if it had typically human beliefs and desires. Simulation theorists such as Robert Gordon and Alvin Goldman both hold that by simulating another's mental process one can anticipate another's behaviour. Gordon and Goldman differ, however, on what aspect of simulation theory should be emphasised. Gordon suggests that mental simulations are based on empirical data, which have been mentally catalogued. Goldman, on the other hand, asserts that imaginative projections into another's situation can occur solely as a result of *a priori* introspection. In the above situation, both positions play a role. On the one hand, notice that the protagonist leverages her empirical knowledge of typically human beliefs and desires. On the other hand, she requires no previous contact with dogs in order to ascribe characteristically human mental states. Setting aside the debate among simulation theorists, what is clear is that introspection does play a role in ascribing beliefs and desires.

Predicting, or explaining, the behaviour of a non-human agent by ascribing characteristically human beliefs and desires to that agent is called *anthropomorphism*. But what if the protagonist holds no notion of what is 'characteristically human beliefs and desires'? Furthermore, what if the ascribing agent, with no notion of characteristically human beliefs and desires, wishes to *minimise* the amount of energy spent on figuring out the other agent's beliefs and desires? These two questions will be addressed in the sections 4.5 and 4.6.

4.5 authoromorphism and the law of marginal utility

Anthropomorphism is a view in which characteristically human beliefs and desires are ascribed to a non-human agent. The case I wish to account for, however, is where a not-necessarily-human-rational-agent predicts the behaviour of a not-necessarily-non-human-rational-agent by ascribing its own beliefs and desires, or algorithm for rationality, to the other agent. And since these agents are neither necessarily human nor non-human they can only be thought of as *the author of an action*. It is necessary, then, to establish a word which captures 1) the notion of agents as authors of actions and 2) behaviours which are analogous to anthropomorphism with the exception that the beliefs and desires being ascribed are

those of the ascribing agent and not necessarily characteristically human ones. To do this I will introduce the term *authoromorphism:* the view that one agent predicts another agent's behaviour by the ascription of the former's own beliefs and desires to the latter.[46]

Prima facie, authoromorphism is an effective means for predicting another's behaviour. Many people, after all, choose to act according to the dictum 'do unto others as you would have them do unto you'.[47] Like authoromorphism, the golden rule's effectiveness as a means of predicting another's behaviour relies on both agents holding the same belief set. But when both agents do not hold similar beliefs and desires, authoromorphic predictions are problematic. Consider a situation where the protagonist holds beliefs and desires consistent with masochistic behaviour. Consider further that the antagonist 1) threatens physical harm on the protagonist, 2) employs the dictum 'do unto others as you would have them do unto you', and 3) concludes that the protagonist will move out of harm's way since the antagonist herself would do the same. But remember that our protagonist is a masochist and will enjoy experiencing pain. In this case the antagonist's prediction is incorrect since the protagonist prefers circumstances in which pain is inflicted. In this case authoromorphism is an ineffective means of predicting behaviour.

But while authoromorphism is prone to mistakes, it may be rational to risk making these errors. To illustrate this point, we will consider the relationship between total expected utility and reconnaissance. Expected utility has, thus far, described states of affairs where the term 'state of affairs' suggests a distinct spatio-temporal domain. On such a view, utility indicates the value of a distinct set of material conditions at a distinct time. But utility as a discrete measure is limiting. A measure is required that accurately reflects an agent's current beliefs and desires but, at the same time, allows utility to vary as one's beliefs and desires change.

The variability requirement is accomplished by viewing total expected utility as a dependent variable and knowledge as an independent variable. The practice of walking illustrates this point. One can certainly imagine that a rational agent is walking and is initially unaware of its behaviour. The agent may reflect, from time to time, on its behaviour and, in doing so, establish beliefs and desires associated with its practice. The agent may come to believe a great many things about the practice: when I move my legs in this manner I move from point A to point B; I will call this walking; the more I extend my legs the more ground I cover; I feel hungry after I walk for a long time; etc. Eventually, the agent is able to develop a functional relationship between its justified beliefs about walking and the expected utility associated with the states of affairs that come to pass from the behaviour. Notice also that the total expected utility associated with walking changes as the agent acquires more beliefs about the *practice* of walking.

---

[46] The term 'authormorphism' may be more accurate. An additional 'o' has been added for epenthetic reasons.

[47] 'Do unto others …' can be taken as a moral dictum. This example assumes, however, that moral behaviour is reducible to rational behaviour.

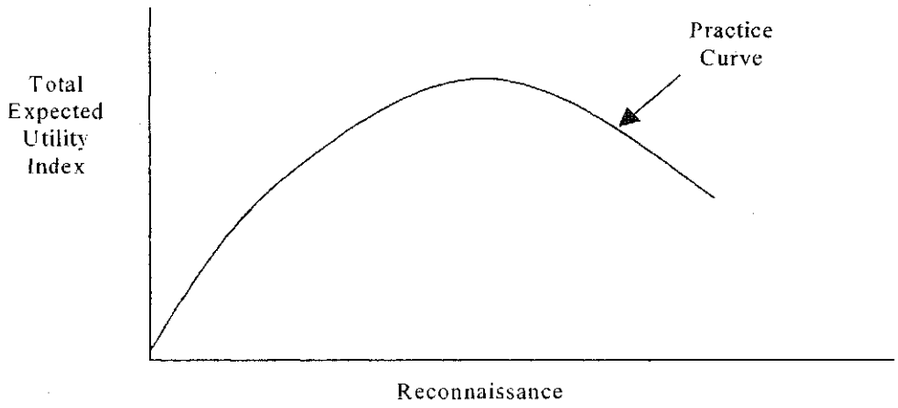

Figure 9: Relationship between total expected utility of a practice and reconnaissance.

The term practice, first introduced in Chapter 1, will be leveraged at this point.[48] The meaning of a practice is now extended to that which maps a functional relationship between the total expected utility and the energy expended in ascertaining certain beliefs and knowledge. Figure 9 shows the relationship between the total expected utility of a practice and the reconnaissance associated with that practice. The curve relating the two variables is called a *practice curve*. For each practice curve there exists a *marginal utility* where the marginal utility for a practice is the change in expected utility (on the practice curve) associated with expending one unit more or one unit less of energy on reconnaissance.

The curve shown in Figure 9 does not typify every possible relationship between total expected utility of a practice and reconnaissance. The key messages, however, are twofold. First, at some point, additional energy spent on reconnaissance will not yield additional total expected utility and, in fact, additional reconnaissance will reduce the total expected utility. Second, the marginal expected utility derived from successive units of energy expended on reconnaissance diminishes as the total consumption of energy associated with reconnaissance increases. Consider an example.

4.6 soldiers in Stockpileland: implications for Dennett

Suppose that soldiers from Force A are sent to find out how many missiles Force B has in Stockpileland. Suppose also that each soldier is capable of entering Stockpileland, estimating the number of missiles, and transmitting the information back to Force A's headquarters. There is a risk, however, that by transmitting the information the soldier will be found out, caught, and executed. Dispatching the first soldier will yield an initial estimate

[48] From section 1.2, rational agents are defined as those that are capable of identifying relationships between two or more states of affairs, identifying a preferred state of affairs and acting in pursuance of that preferred state, re-describing sets of actions as practices, and, finally, discovering their preferences among practices.

of the missiles in Stockpileland; every subsequent soldier dispatched will improve on the accuracy of the initial estimate. There is a point, however, at which the utility of another estimate will not improve with each additional soldier dispatched. In fact, there is a point where dispatching additional soldiers begins to reduce the total expected utility since the soldiers' lives must be taken into account in determining Force A's total expected utility. The 'soldiers in Stockpileland' example illustrates how the relationship between total expected utility and reconnaissance follows the *law of diminishing marginal utility.*[49]

The law of diminishing marginal utility states that for each additional unit of commodity consumed, the additional unit of utility realised is reduced. In the case of the Stockpileland example, each additional unit of knowledge garnered via reconnaissance will result in an incremental unit of utility. The increments of utility diminish as the number of units of knowledge increase. The point at which the incremental utility is zero is the point at which the practice is maximised. This relationship is shown in Figure 10.

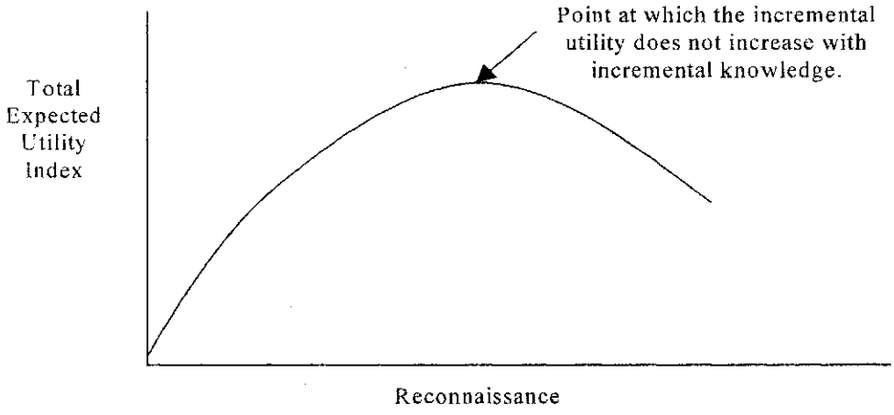

Figure 10: Law of diminishing returns and maximising when marginal utility is zero.

Recall Dennett's instructions for adopting the intentional stance. On these instructions, one is required to determine the beliefs and desires of the system whose behaviour is to be predicted. But we know that reconnaissance plays a role in 'figuring out' the beliefs and desires of other agents. So in following Dennett's methodology, one can assert that the amount of 'figuring out' that goes into the determination of the beliefs and desires of the agent will only be the amount that maximises on total expected utility on a certain practice.

Also recall that in section 4.4 we asked the question, what about the agent who wishes to minimise the amount of energy spent on figuring out another agent's beliefs and desires? Intuitively, one might assert that an agent whose reconnaissance involves ascribing beliefs and desires solely on introspection would spend less energy than an agent whose

[49] Richard G. Lipsey, Douglas D. Purvis, Gordon R. Sparks, Peter O. Steiner, *Economics*, Fourth Edition, Harper & Row Publishers, 1982, p154.

reconnaissance requires physically gathering empirical data. As it turns out, this intuition is correct. A 70 kg human, for example, expends only 77 kcal/hr while lying undisturbed yet consciously engaged in deliberation. That is to say, an agent would expend 77 kcal/hr of energy engaged in authoromorphic activity. The same agent would expend 100 kcal/hr sitting at rest but engaged in verbal intercourse, 140 kcal/hr while typing notes, 280 kcal/hr engaged in sexual intercourse, and up to a maximal activity expenditure of 1440 kcal/hr.[50] Therefore, the minimum energy expenditure required to engage in reconnaissance is that required to reconnoitre authoromorphically. Even with the inherent potential for errors, it may be rational to risk making authoromorphic errors should the conditions be such that minimising on the energy expenditure associated with reconnaissance is the best means of maximising total expected utility.

[50] Arthur J. Vander, James H. Sherman, Dorothy S. Luciano, *Human Physiology: The Mechanisms Of Body Function*, McGraw-Hill Book Company, 1985, p536. I have taken the liberty of assuming that sexual intercourse is a means of reconnaissance. This assumption is based on the exploits of Margaretha Geertruida Zelle (also known as Mata Hari). Of Dutch ancestory, Mata Hari was a professional dancer in Paris who, during World War 1, was accused of spying for Germany, arrested and executed by the French.

Chapter 5

Empirical Decision Theory

Empirical decision theory seeks to explain human judgement and decision-making based on observation, data collection, analysis and inference. Unlike the normative presuppositions of rational decision theory, an empirical explanation of human judgement follows from a descriptive account of rational behaviour. Empirical decision theorists concentrate on establishing correlations between the circumstances under which decisions are made and the decisions themselves. Empirical research focuses on understanding the limitations of inductive reasoning. In an effort to model real world conditions, the empirical decision theorist subjects decision-makers to conditions of uncertainty.[51] Under such conditions, heuristic algorithms (or rules of thumb) often emerge as the basis for making choices. Daniel Kahneman and Amos Tversky are pioneers in the field of empirical decision theory. Their work identifies judgement biases resulting from the limitations of a decision-maker's processing speed, capacity to hold information, etc. Kahneman and Tversky observe that humans who attempt to make decisions with only limited knowledge rely on heuristics to make their decisions. Further, Kahneman and Tversky have classified the judgement biases that result from using heuristics in making decisions with limited knowledge.

The idea that knowledge is limited is certainly not a new one. The limits of knowledge can be thought in terms of availability of knowledge and on processing limitations of a decision-maker's physical system.[52] This incompleteness of knowledge, and the extent to which incompleteness limits of rationality, is relevant for agents interacting in the world of practical affairs and business. For Herbert Simon, author of *Administrative Behaviour*, the central concern is "the boundary between rational and non-rational aspects of human social behaviour" in practical situations.[53] Of chief interest for Simon is the extent to which, 'administrative man[sic]'[54] behaves rationally. Administrative man, for Simon, acts on the principle of efficiency discussed earlier, according to which agents make decisions that

---

[51] Here, 'uncertainty' is specific as per Andrew M. Coleman, *Game Theory and its Applications in the Social and Biological Sciences*, Butterworth-Heinemann Ltd., 1982, p23.

[52] The tradition of epistemic scepticism flourished as early as 360 BCE in the Pyrrhonian School. Later, Rene Descartes explored methodological scepticism and David Hume took metaphysical scepticism to its logical conclusion. More recently W.V.O. Quine and Richard Rorty have challenged the notion that there are any firm foundations for knowledge. Like sceptics, empirical decision theorists assert that there are limits on what we can know. But empirical decision theory places less emphasis on what is, or can be, considered knowledge.

[53] Herbert A. Simon, *Administrative Behaviour: A Study of Decision-Making Process in Administrative Organization*, The Free Press, 1957, pxxiv.

[54] We will find throughout Simon's work that he writes in the tradition of his time.

favour maximising the attainment of certain ends on finite means. Simon takes 'administrative man' to follow from classical 'economic man' where "the correctness or goodness of a decision is a matter of the extent to which it maximises an individual's interest in a certain end".[55] Simon's look at the individual decision-maker includes asking what limits are placed on 1) the individual's ability to decide, and 2) the individual's ability to make correct decisions.[56] The former focuses on attributes outside of the realm of consciousness, which includes restrictions on strength, dexterity, reflexes, etc. The latter focuses on the limits within the decision-maker's conscious realm: on the incompleteness of knowledge and on the performance attributes for making effective inferences. The question of correctness in decision-making, says Simon, can be understood by "placing limits on the mass of knowledge that human minds can accumulate and apply".[57]

Simon's view on decision-making is in contrast with what he calls the *idealised picture* of those who hold the view of rational decision theory. According to the idealised view, the decision-maker views the behavioural alternatives in a panoramic fashion prior to making any decision: the agent considers the whole complex of consequences that would follow on each choice, and by using a system of values as criterion, singles out the appropriate set of alternatives.[58] But, argues Simon, under practical circumstances this idealised picture is not realisable for three reasons:

1. the knowledge of the consequences of any single decision is always fragmentary;
2. the consequences of any decision lies in the future and thus imagination must supply the lack of experience, a lack of which can only be imperfectly compensated;
3. the scope of behavioural possibilities is vast, even infinite, of which only a few possibilities ever come to mind.

"Rationality, then, does not determine behaviour...instead, behaviour is determined by the irrational and non-rational elements that bound the area of rationality".[59] The upshot for Simon is that "two [agents], given the same possible alternative, the same values, the same knowledge, can rationally reach only the same decision".[60] For any one agent, then, having a specific set of beliefs and desires, only one rational behaviour, or practice, is available. And, according to bounded rationality, a rational agent will adopt whatever practice maximises

---

[55] Op. Cit., p39.

[56] It may appear that Simon is running the risk of holding a position of Cartesian dualism. Simon's position is, however, compatible with a physicalist theory of mind.

[57] Op. Cit., p40.

[58] The panoramic view can be achieved, as discussed in the preceding chapter, through the use of *ceteris paribus* condition.

[59] Ibid., p241.

[60] Ibid.

total expected utility according to the specific level of knowledge available, and will not change its practice until the boundaries for rationality change.

5.1 systematic errors

According to Simon, under practical conditions, rational agents – due to limitations in data gathering, processing speeds and information capacity – are unable to view behaviour alternatives in panoramic fashion. Instead, says Simon, real behaviour, even that which is ordinarily thought of as rational, possesses many elements of *disconnectedness* not present in the panoramic picture. Kahneman and Tversky have studied this notion of disconnectedness and conclude that disconnected decision-makers inevitably make systematic errors. The methodology for discovering these errors involves observing how rational agents make choices when provided with descriptions of various situations where the degree of information available and the time required to answer are varied. Consider the following illustration:

> Description: Steve is very shy and withdrawn, invariably helpful, but with little interest in people, or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.
>
> Question: Is Steve likely to be a farmer, salesman, airline pilot, librarian or physician?

What is of interest for Kahneman and Tversky is, in this particular case, how people order the possible occupations from most likely to least likely and, in general, the errors associated with the various methodologies used to determine their ordering. An assortment of systematic error types is documented in Kahneman and Tversky's *Judgement under uncertainty: Heuristics and biases*. Systematic errors are classified as errors of representedness (also known as category mistakes), errors of availability (the ease of which occurrences can be brought to mind), and errors in adjustment and anchoring (adjusting initial conditions to yield the final answer). Within these error categories Kahneman and Tversky identify twelve sub-classifications, each a contributing factor to systemic error. Here I want to focus on the specific problem of *insensitivity to predictability*.

Insensitivity to predictability is an error in which a prediction is made without considering the reliability of the method used in making that prediction. The example used by Kahneman and Tversky is one in which an agent is asked to predict a company's future profit based solely on company descriptions. The descriptions themselves vary in reliability, a fact known to the decision-makers. Invariably, however, decision-makers predict future profits that correlate with the favourability of the description and not with the reliability of the information itself. According to Kahneman and Tversky: "[t]he degree to which the description is favourable is unaffected by the reliability of that description or by the degree

to which it permits accurate prediction".[61] Under pressing conditions, decision-makers rely more on the way the message is delivered rather than on the reliability of the information itself. Insensitivity contributes to, what Kahneman and Tversky call, the *fallacy of planning*. Fallacies of planning are

> a consequence of the tendency to neglect distributional data and to adopt what may be termed as an internal approach to prediction, in which one focuses on the constituents of the specific problem rather than on the distribution of outcomes in similar cases.[62]

The *internal approach* to which Kahneman and Tversky refer is that which leverages categorisation, the ease in which occurrences come to mind, and the anchoring of initial conditions to overcome the incompleteness of knowledge.

5.2 the fallacy of planning and alternative practice opportunities

Consider that committing the fallacy of planning can, in and of itself, *ensure* that an agent takes itself to be acting rationally, regardless of the practice in which it is engaged. How can this be so? Rational behaviour is characterised by maximising on a certain practice given finite knowledge. Agents operating under the fallacy of planning focus on the constituents of the specific problem and are predisposed to overcome incompleteness of knowledge by adopting an internal approach to prediction. Furthermore, the agent committing the fallacy of planning neglects the success rate of using an internal approach. And since – on the internal approach – whatever prediction is justifiable as maximising total expected utility, decision-making agents are justified in believing that they are maximising on their total expected utility, and therefore acting, albeit subjectively, rationally. For an agent disposed to the fallacy of planning, rationality amounts to justifying whatever practice in which the agent is engaged. We saw an example of this type of justification in the introductory comments where an agent consoled himself in believing that God has willed the loss of a child through sickness. Doing so helped in dealing with emotional distress, and as such is rational.

At the same time, the fallacy of planning can be problematic. In the above case, the fallacy creates a potential barrier to investigating the possible causes of the sickness and discovering a treatment. By avoiding the fallacy of planning, however, it is possible to access other practices that have greater total expected utilities. Figure 11 shows the practice curve p0 which represents the practice of prediction using the internal approach. According to Kahneman and Tversky's research, with additional reconnaissance delta k – specifically

---

[61] Amos Tversky and Daniel Kahneman, "Judgement under uncertainty: Heuristics and biases", taken from *Judgement under uncertainty: Heuristics and biases*, edited by Kahneman, Slovic, and Tversky, Cambridge University Press, 1982, p8.

[62] Amos Tversky and Daniel Kahneman, "Intuitive prediction: Biases and corrective procedures", taken from *Judgement under uncertainty: Heuristics and biases*, edited by Kahneman, Slovic, and Tversky, Cambridge University Press, 1982, p415.

reconnaissance which overcomes the fallacy of planning – it is possible to discover an alternative practice, p1, that improves on the accuracy of prediction. In doing so, one is able to access a greater total expected utility on an alternative practice line p1.[63]

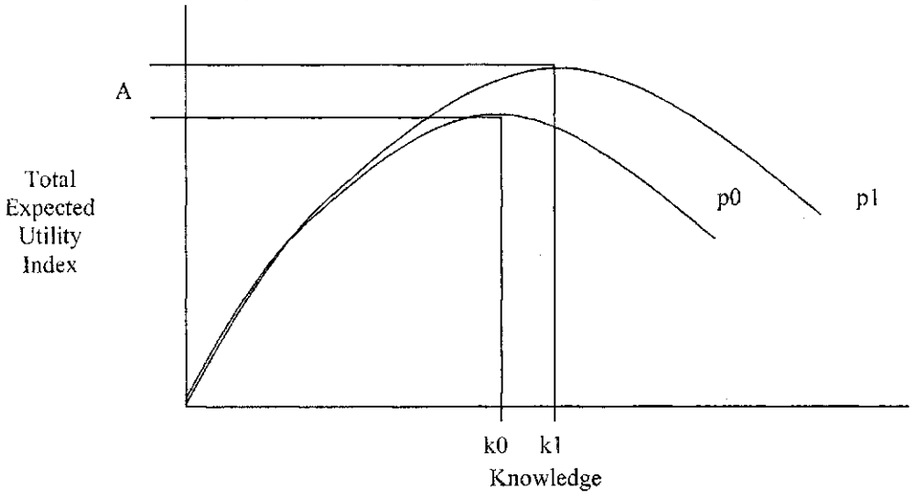

Figure 11: Alternative Practice Opportunities.

5.3 alternative practice opportunities

Rational behaviour, according to the empirical decision theorists, is characterised by practices where the total expected utility associated with specific knowledge is maximised. And by way of the fallacy of planning one can rationalise any practice. Still, the question remains as to *what practice* delivers the maximum total expected utility. To that end, rational behaviour must also be characterised by a comparison of total expected utilities among practices. On this view, the *alternative practice opportunity* is the incremental utility or disutility associated with the adopting an alternative practice. The alternative practice opportunity is not merely the difference in utilities between two practices at a specific level of knowledge. The alternative practice opportunity takes into account that a rational agent may maximise the total expected utility of a practice according to different levels of knowledge to which the agent has access. Figure 11 above depicts an alternative practice opportunity A (the difference in utility between where practice p0 is maximised and p1 is maximised). The alternative practice opportunity, then, measures the difference in utility between acting rationally on practice p0 and acting rationally on practice p1 and takes into account that incremental knowledge is required to access the alternative practice.

This is a useful concept. One can identify the change in the total expected utility associated with alternative practices along with the reconnaissance required to adopt the alternative practice. Using this concept, one can assert that, if the alternative practice

[63] Alternatively, additional reconnaissance delta k (specifically reconnaissance that overcomes the fallacy of planning) may reveal an alternative practice, say p2, that both improves on the accuracy of prediction and reveals the possibility of a practice that has a lesser total expected utility than p0.

opportunity is *positive* (then the total expected utility of an alternative practice is higher than current), it is rational to adopt the alternative practice over the current practice.

Figure 12 accords with the following example. Suppose an agent makes and sells shoes. Suppose also that the agent has achieved a level of knowledge k1 about shoemaking and, at this level of knowledge, adopts a rational practice p1. This practice involves completing one set of shoes before moving on to start the next. On the knowledge available, the shoemaker's total expected utility is maximised. After a few years of employing this practice, the shoemaker visits another village and happens across another shoe shop. The shoemaker decides to investigate. This is an important point since the shoemaker has decided to engage in reconnaissance. The shoemaker is now expending energy to learn more about other shoemakers and evaluate his current practice against alternative practices. He goes into the competitor's shop, acts as a customer, and strikes up a conversation with the proprietor of the shop. During the course of the conversation the first shoemaker learns that the other completes all the uppers first, then builds the soles, then stitches the two together in three separate stages. While the proprietor of the shop says he sells no more shoes, by adopting this practice over making the shoes one at a time, it allows him to spend more time with his family.

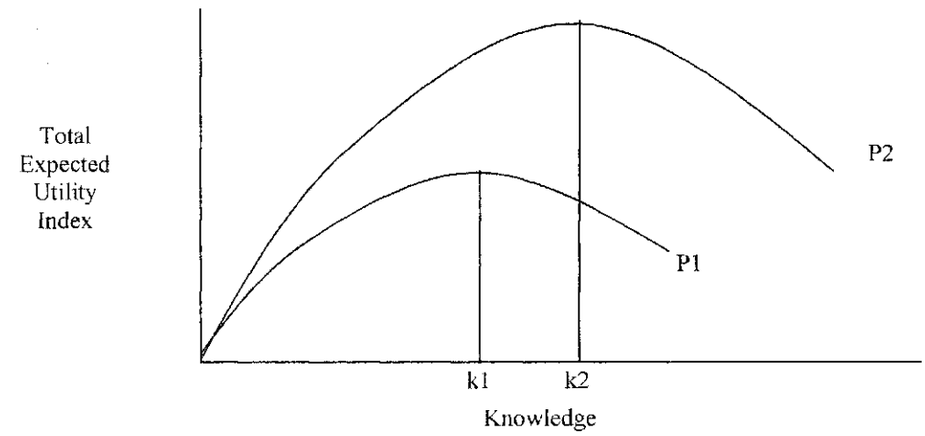

Figure 12: Alternative Practice Opportunity for The Shoemaker.

The first shoemaker, armed with this new knowledge – represented by k2 in Figure 12 – assesses the total expected utility of adopting an alternative practice. Based on the principle of building uppers, then soles, then assembling the two, the shoemaker decides that this new practice would allow him to have more spare time, perhaps to be spent with his own family. Therefore, for the first shoemaker, the total expected utility associated with practice p2 is greater than the total expected utility of practice p1. The alternative practice opportunity is the additional total expected utility associated with one practice over another from increasing one's knowledge from k1 to k2.

5.4 rational bankruptcy: distinguishing between deliberation and non-deliberation

Finally, we will consider one last factor before concluding the discussion on empirical decision theory. We can certainly imagine circumstances in which such pressing and urgent demands are imposed on an agent that rationality is no longer bounded but indeed eliminated as a characteristic of the agent's behaviour. The distinguishing factor is a matter of *deliberation*. While a rational agent may have the capability to deliberate (i.e. to identify functional relationships between states of affairs, etc) the conditions imposed upon an agent may be so pressing that the agent acts without deliberation. In the vernacular, this is often referred to as 'acting out of instinct'.

Instinctual behaviour is illustrated by the following. Imagine that you are walking along a quiet corridor in a hotel. You are thinking about what you will do tomorrow. Suddenly a colleague leaps out from around a corner and calls 'Boo'! You jump back, gasp aloud, and draw your hands up in defence. Such behaviour would be instinctive. Alternatively, consider something that most have experienced at the doctor's office. The mallet strikes one's patellar tendon located just below the kneecap, the tendon is depressed, stretching the attached muscles and exciting the receptors within the muscle spindles. This excitement signals the motor-neurons controlling the muscles. The neurons fire, the muscles shorten, and the foreleg raises to give the knee jerk response. Few would assert that, in either of the examples above, the agent's behaviour resulted from deliberation. This is not to say that the brain plays no role in the scenarios described. The brain, of course, receives signals via the spinal cord which indicate what is happening. But the brain plays no causal role in controlling the behaviour exhibited. Similarly, deliberation plays no role in blinking, breathing or digesting food. On the empiricist's view, then, in order for an agent to act rationally the conditions must not be so pressing as to invoke *rational bankruptcy*. That is to say, the conditions must be such that an agent's acts are causally related to deliberation.

Notwithstanding that the term 'rational' is absent from its title, empirical decision theory, as we have seen, does leave room for ordinary 'panoramic' rationality. But under pressing conditions, disconnectedness arises. Agents are unable to view alternatives in an 'ideal panoramic' fashion as suggested by the rational choice theorist who employs logically complementary choices and the *ceteris paribus* condition. According to the empirical decision theorist, other than the conditions resulting in rational bankruptcy, an agent uses heuristics to fill in the gaps in knowledge in order to entertain possibilities, identify causal relationships, and discover a preference for one state of affairs over another. It is advantageous to view empirical decision-making on a deliberative continuum with the ideal panoramic rationality at one extreme and rational bankruptcy at the other. On such a view, the extent to which a decision is made empirically depends on the conditions that bound rationality. Even under bounded rationality, the rational agent, given whatever knowledge is available to her, can identify a practice where the marginal expected utility is zero at which point total expected utility is maximised.

Chapter 6

Distinguishing between Rational and Strategic Behaviour

If the difference between rational agents and strategic agents is only the domain in which they interact, then our intuition that strategic agents hold some cognitive virtue beyond mere rationality is an unfounded one. This thesis, however, seeks to explain our intuitions by discovering the relevant cognitive virtue. Thus, to enable discovery, an analysis of the minimal social conditions for rational interaction will be conducted. The conditions for a *minimal social situation* are not so demanding as to evoke rational bankruptcy but are, at the same time, difficult enough to expose differences between rational and strategic behaviour. By determining whether rational agents can interact non-strategically, the criterion for what counts as strategic behaviour will present itself. This criterion may then be tested empirically to ascertain its value in establishing a transcendental concept of strategy: one that will hold across the field of military investigation, game theoretic analysis, and in the world of business affairs.

6.1 the game theoretic minimum social situation

Early game theoretic investigations into the minimum social situation are described by Sidowski, Wyckoff, and Tabory.[64] The minimal social situation used in this thesis is similar to the Sidowski, Wyckoff, and Tabory experiments in that the two interacting agents are (1) aware of their own beliefs and desires but (2) unaware that they are interacting with another agent. The agents' knowledge of the situation is limited by physically separating them in adjacent rooms. The agents have access to an apparatus on which is mounted a pair of buttons labelled L and R for left and right respectively. The agents are not given instructions on the role of the apparatus other than they may push either of the two buttons but only one button at a time. The agents are unaware that pressing the right button corresponds to *a reward to the other agent* in the form of points and pushing the left button corresponds to *punishment to the other agent* in the form of electric shock.

In the original experiment two types of interactions were tested: sequential interactions and simultaneous interactions. Sequential interactions have no purposeful delay

[64] Sidowski, Wyckoff, and Tabory. "The Influence of Punishment and Reinforcement in a Minimal Social Situation", *The Journal of Abnormal and Social Psychology*, 52, pp115-119.

between the time a button is pushed and the time the other agent receives its reward/punishment. The simultaneous interaction is wired in a manner that both agents must choose a button before any reward/punishment is delivered. Only after both have pushed a button is the appropriate reward/punishment simultaneously doled out. It is also important to distinguish the *strictly* minimal social situation from the *non-strictly* minimal social situation. In the strictly minimal social situation

> the players are ignorant even of one another's existence: they know they are making decisions under uncertainty, but they do not know the uncertainty arises [in part] from their involvement with [another player].[65]

The non-strictly minimal social situation occurs when both agents are informed of the other's existence. For now we will only consider the strictly minimal social situation and later we will look at some interesting characteristics of the non-strictly minimal social situation.

Sidowski and team found that agents who adopted a *win-stay/lose-switch* methodology for choosing between buttons were able to access co-operative behaviour that maximised payoffs for both agents. The term 'win-stay/lose-switch' describes an agent's disposition to choose the same button in the current round as the button chosen in the last round provided that choosing that button in the last round delivered a payoff, or win. Likewise, agents using the win-stay/lose-switch methodology are disposed to choose a different button in the current round provided pushing the other button in the last round delivered a punishment, or loss. There are empirically based explanations for adopting the win-stay/lose-switch methodology. E. L. Thorndike's *Law of Effect*, for example, observes the win-stay/lose-switch methodology.[66] According to Thorndike,

> of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond.[67]

While this is an interesting observation it hardly counts as a knockdown argument for why one should choose the win-stay/lose-switch methodology. The question remains then as to whether Thorndike's claim can be supported by way of logical reduction. But before exploring this reduction I would like to make a modification in terminology.

---

[65] Andrew M. Coleman, *Game Theory and its Applications in the Social and Biological Sciences*, Butterworth-Heinemann Ltd., 1982, p41. Brackets mine.

[66] See Kelley et al 1962.

[67] E.L. Thorndike, *Animal Intelligence*, MacMillan, 1911, p244.

6.2 no-lose-stay/ lose-switch versus win-stay/lose-switch

The win-stay/lose-switch terminology allows for the possibility of a draw where an agent takes an outcome to be neither win nor lose and, therefore, choice neutral. In order to account for an agent's indifference to an outcome, I will adopt a view where win is equivalent to no-lose. I will call this methodology the *no-lose-stay/lose-switch practice*. We can assume, then, that agents engaged the no-lose-stay/lose-switch methodology would, on any given choice, not take indifference to be a win but will nonetheless take indifference to be a no-lose situation. Thus, should an agent be indifferent to an outcome, it follows that the agent would choose *no-lose-stay*. Sequential no-lose-stay choices can be thought of as a *no-lose-stay-loop*. In the vernacular, this type of loop is often referred to as 'doing nothing'. While, strictly speaking, it is not possible to do nothing, I will refer to those agents engaged in the no-lose-stay-loop as 'doing nothing'.

Finally, before moving on to the reduction, it should be pointed out that, as we saw with Gauthier and Danielson earlier, an agent's disposition towards a practice can be characterised. An agent disposed to a no-lose-stay-loop is said to be *complacent*. Let us explore this notion of complacency further. Suppose, as we saw in Chapter 5, that an agent is satisfied that its current practice maximises total expected utility. And recall that, on the empirical decision theorist's view, an agent is rational when maximising total expected utility on available knowledge even when it is unaware of potential deficiencies – relative to other practices – in its current practice.[68] Agents, then, are *rationally complacent* when they are satisfied that their current practice maximises total expected utility even if other practices, yielding positive marginal total expected utilities, are available to them. We will now move on to the reduction.

6.3 the reduction

We shall see that the no-lose-stay/lose-switch methodology can be reduced to i) rational complacency (characterised by the no-lose-stay-loop practice) and ii) a behaviour that we will come to describe as strategic. The reduction of the no-lose-stay/lose-switch methodology is achieved by removing an errant presupposition found in Kelley et al., and Sidowski et al. These experiments, all of which happen to use human agents, are described presupposing that agents are compelled to act out of innate curiosity or motivated by a substantive improvement to their utility. Fair enough. But in the strictly minimal social situation the terms 'curiosity' and 'substantive improvement' are too vague to justify an agent's behaviour. To claim, for example, that an agent is predisposed just to push a button to see what happens is a misleading representation of the strictly minimal social situation. Presumably, someone who decides just to push a button and see what happens has faith that pushing either button will not result in self-destruction. The agents described in the Kelley

[68] A result of the fallacy of planning, for example.

and Sidowski experiments have this faith. They are, after all, subjects in a controlled experiment and are likely to believe that the lab-coated experimenter can be trusted not to kill them. But this is an unacceptable assumption.

In the strictly minimal social condition we cannot assume that the interacting agents will just push one of the buttons to see what happens. Consider the following re-description of the strictly minimal social situation in the absence of this 'curiosity' presupposition. Let us assume instead that the two agents are far more complacent, or, better yet, cautious. Suppose the subjects – destined to be agents in the experiment – are going about their day to day business unaware that a lab-coated experimenter is stalking them. Unannounced, the stalker blankets each subject's upper body, physically restrains them, and transports these hooded subjects to the laboratory conditions. The subjects are secured under experimental conditions and the hoods are removed. For the first time since being subdued the subjects are able to see their surroundings. Under these conditions, the subjects will not be so eager just to push a button and see what happens. The subjects will more likely be very cautious and distrustful of the surroundings.

Some may argue that I have merely replaced the term 'curious' with 'distrustful'. Agreed. This alternative scenario makes an equally unfair supposition as that made in the original scenarios. This dialectic has been chosen, however, to emphasise that the minimal social situation must be one where the agents are complacent with respect to their surroundings. So, rather than 'curious' or 'cautious', the agents in the strictly minimal social situation are rationally complacent as regards the practice in which they are engaged. Initially, then, the rationally complacent agent will 'do nothing' when placed in the strictly minimal social situation. Furthermore, not only are these agents rationally complacent they are symmetrically positioned with one another. Both maintain the *no-lose-stay-loop;* they are both content to sit and 'do nothing'. For a rationally complacent agent, then, engaged in a no-lose-stay-loop it must be the case that, at current knowledge k1, the practice has marginal expected utility equal to zero (as shown in Figure 13).

Given that these agents are rationally complacent, we can simplify the conditions of the experiment by recognising that the agents have a choice between 'doing nothing' or pushing a button. It makes little difference if there is one or more buttons. So for now we will assume there is only one button and that the agents are, for the moment, satisfied to sit and do nothing.

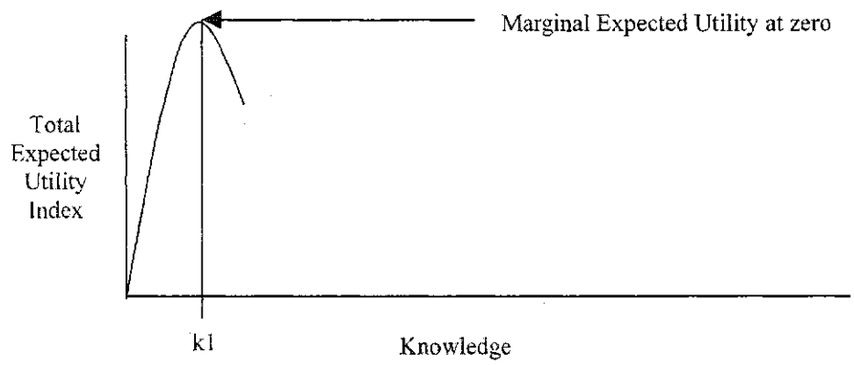

Figure 13. The no-lose-stay (or 'do nothing') practice.

Notice, however, that for any agent, 'doing nothing' still requires the consumption of calories to run the base no-lose-stay-loop algorithm. And one would assume, as do Kelley, Sidowski, et al., that – as the subjects eventually run low on calorie reserves [69] – one of the two agents will ascertain the knowledge that their calorie reserves will, in future, be depleted. But the conditions for rationality do not necessarily require that an agent reach this conclusion. The conditions for rationality require only that the agent is satisfied, on knowledge k1, with its current practice. It is not necessary that the agent establish a causal relationship between its current practice and its energy reserves in order for the agent to behave rationally.

Let me provide an example. Suppose a human agent is driving her car from point A to point B where there are no fuel filling stations between A and B. The trip – one this agent has made often – usually requires a quarter tank of fuel. On this particular day, the agent has only a quarter tank of fuel and is driving into a headwind sufficient to impact fuel efficiency to the extent that a quarter tank of fuel will not be enough. The car-driving agent makes no causal connection between fuel efficiency and headwind and, as a result, runs out of fuel before reaching point B. Notice that, in this scenario, the agent is still acting rationally, given the knowledge she has (which does not include a causal relationship between headwind and fuel efficiency). She is satisfied with the practice of driving from point A to point B on a quarter tank of fuel; for this agent the fact that she ran out of fuel must be an inexplicable anomaly.

The strictly minimum social situation, then, must have two rationally complacent agents both engaged in the no-lose-stay-loop. And, as discussed above, the two agents are in symmetrical positions. But in real-life situations this kind of symmetry is usually broken.

---

[69] The very reason for adopting the *win-stay/lose-switch* methodology points out the error in its appellation. Agents do nothing until they are compelled to reconnoitre when their calorie reserves are low. And pushing a button is the only action available to them. Strictly speaking they are not choosing a win-stay/lose-switch approach but a *no-lose-stay/lose-switch* approach where the *no lose* condition accounts for both winning and doing nothing. The logically complementary no-lose-stay/lose-switch name is, I believe, an accurate representation for either agent's choice to maintain the symmetry under no-lose or break the symmetry due to loss.

6.4 breaking the symmetry

As represented by the practice curve in Figure 14, beyond k1 the marginal expected utility associated with the no-lose-stay practice is negative. At any point beyond k1 an agent may believe that its current practice will deliver a negative total expected utility. This is the point at which the agent is no longer satisfied with the current practice.

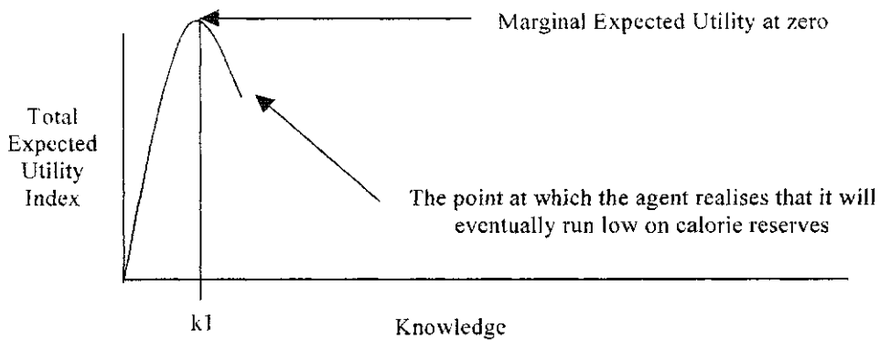

Figure 14: The no-lose-stay (or 'do nothing') practice

Since, in the minimal social situation, to not 'do nothing' is to switch, the agent decides to switch. An agent may choose to switch any time the agent falls under the belief that their calorie reserves will eventually run low. Suppose that the agent 'switches' and pushes the button, holding the button down. Further suppose that doing so delivers a reward. The delivery of the reward provides the agent with new knowledge and a greater total expected utility. Still, the agent continues to sit with his finger on the button. Some time latter the agent once again falls under the belief that the current practice of holding the button down will no longer deliver reward. And what is worse, holding the button down is expending energy at a faster rate than 'doing nothing'. Once again, he realises he will eventually run low on calorie reserves and releases the button. Still, energy is being expended – even in doing nothing – and the agent 'switches' again to pushing the button.

Notice that by 'switching' the agent is no longer engaged in the no-lose-stay-loop practice. Instead the agent has now adopted the no-lose-stay/lose-switch practice where to lose is to come to the belief that calorie reserves will eventually run low. Figure 15 shows the impact of, not only the first 'switch', but, every switch that occurs each time the agent believes that its calorie reserves will eventually run low. Also, notice that a single function is shown in Figure 15 where, in fact, there are two separate practice curves: the no-lose-stay-loop practice and the no-lose-stay/lose-switch practice. The same pattern repeats until the agent realises that pushing and releasing the button repeatedly delivers reward. At a certain point, the agent recognises a functional relationship between pushing the button and receiving reward, and realises that that relationship exists has a total expected utility.

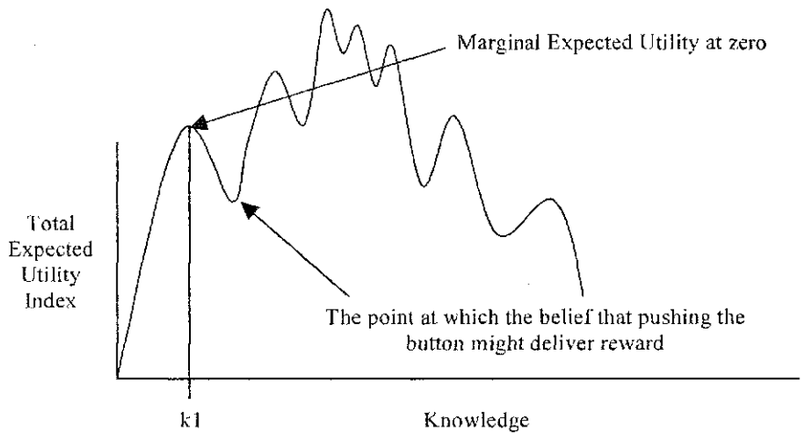

Figure 15: The no-lose-stay / lose-switch algorithm

The function shown in Figure 15 is re-described as two practices in Figure 16. P1 illustrates the no-lose-stay practice and p2 the no-lose-stay/lose-switch practice.

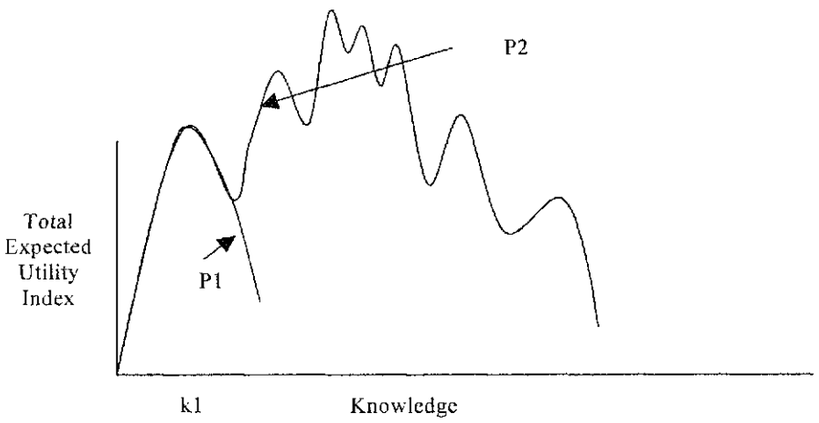

Figure 16: The no-lose-stay practice, and no-lose-stay / lose switch practice

While it is possible to show the impact of every 'switch' that occurs in the no-lose-stay/lose-switch *practice* (as in Figure 16), a 'smoothed', or *meta-practice*, curve can be used to accurately describe the no-lose-stay/lose-switch practice. The use of meta-practice curves is particularly useful in comparing among total expected utilities of practices without the distraction of each switch that may occur. Figure 17 illustrates how the no-lose-stay/lose-switch practice is represented using a meta-practice curve. Notice that even in terms of a meta-practice curve, there is still a point at which the agent believes no further knowledge is required to maximise on the meta-practice as represented by k2.

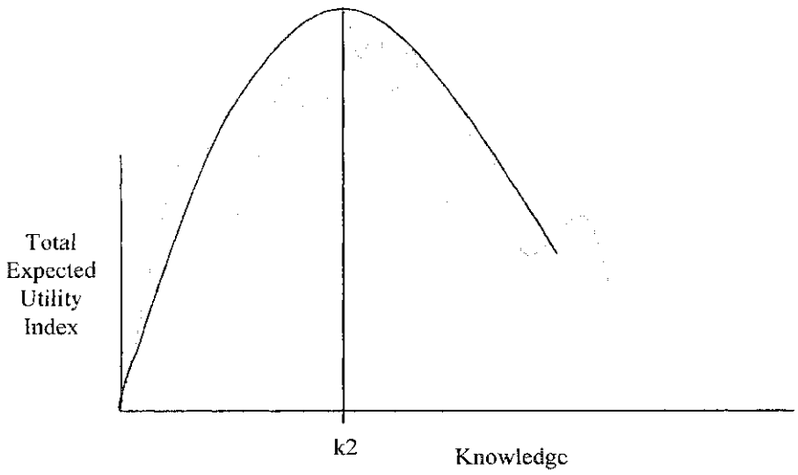

Figure 17: The no-lose-stay / lose-switch meta-practice.

The key points here are twofold. First, like Kelley, Sidowski, et al, I concur that the no-lose-stay/lose-switch methodology characterises rational behaviour under the minimal social situation. Second, unlike the Kelley and Sidowski experiments, I do not take the no-lose-stay/lose-switch methodology to be one consciously adopted by interacting agents at the onset of interaction. As shown in Figure 18, rationally complacent agents will engage in the no-lose-stay-loop practice until such a point where they believe that their energy reserves will eventually be depleted at which point the agents will then adopt the no-lose-stay/lose-switch meta-practice.[70]

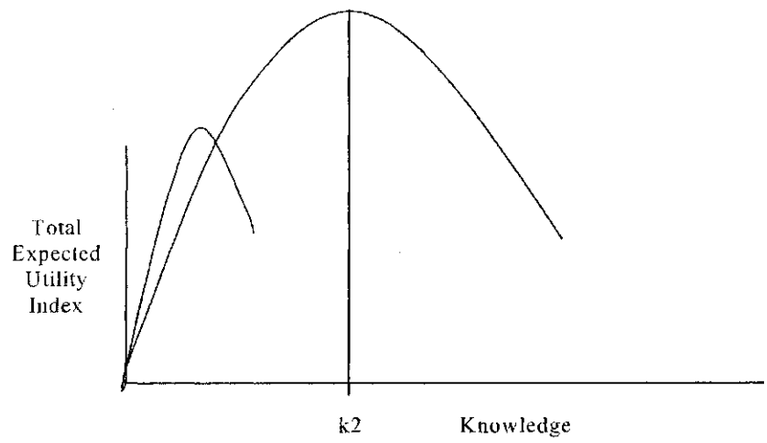

Figure 18: Moving from no-lose-stay to no-lose-stay/lose-switch.

[70] Some may take the term 'reserves' to be too restrictive. Certainly we can imagine a case where an agent can function only when it has a direct and continuous energy supply. Without capacitance, however, such an agent will become inactive once the energy source is removed. As a result, agents without capacitance are more likely to have limited mobility than those equipped with some level of on board capacitance: whereas the former is spatio-temporally reliant on an energy source the latter able to act even when it is not in direct contact with an energy source.

But what occurs between engaging in one practice and simulating another is important. We cannot assume that rational agents, in the strictly minimal social situation, are simply disposed to a no-lose-stay/lose-switch practice from the onset. According to empirical decision theory, the necessary and sufficient conditions for rationality require maximising expected utility on the current practice, even if doing so results in one's death. As I noted earlier, Russell has claimed that many people die sooner than think. We can add to Russell's claim that 'indeed, many do' and doing so is rational.

At this point the importance of Danielson's rationality thesis comes to bear. Certain agents, says Danielson, are disposed to be 'more rational' than others. In the minimal social situation, agents that are dissatisfied with the no-lose-stay-loop are likely to survive. Survivors, on this view, are disposed to making the causal connection between their current state of affairs and the disutility of running out of reserves. The survivors are those rational agents that make this causal connection, exhibiting a cognitive virtue 'more rational' than rational complacency.

There is still a worry here, however. It is the case that the surviving agents described above construct functional relationships among states of affairs. And it is indeed rational to model relationships between inputs and outputs, or stimulus and response. Isaac Newton, for example, constructed a mathematical model in which the force of gravity between two bodies M_1 and M_2 is a function the distance between the two bodies.[71] Still, while Newton provided the mathematical model for this functional relationship, he himself acknowledged that he could not explain gravity's mechanisms.

> Gravity must be caused by an agent acting constantly according to certain laws; but whether this agent be material or immaterial, I have left to the consideration of my readers.[72]

While Newton asserts that there are mathematical laws governing gravitational behaviour, he acknowledges that the mechanism of gravity, in his own time, remained insoluble.

In this thesis we have been considering functional agents whose behaviour is modelled according to an algorithm for rationality. The question remains, however, as to whether practices can be treated in the same fashion: is it rational to model practices functionally?

Before answering this question, recall section 1.1 where we saw that an agent's disposition, while defined functionally, is only characterised by practices (where the practices are defined according to their constituents). The example that was provided referenced the practice of baseball in which the constituents were a pitcher, a batter, a playing field, etc. But it seems that practices can also be explained in terms of algorithmic relations. In fact, the

---

[71] Force of Gravity = GM_1M_2/R^2 where G is the gravitational constant, M_1 and M_2 are the masses of the two bodies, and R is the distance between the bodies. Thus, Newton's model shows that the force of gravity between two bodies is, in fact, inversely proportional to the square of the distance between them.

[72] Isaac Newton, *Sir Isaac Newton's Mathematical Principle of Natural Philosophy*, University of California Press, 1962, p634.

meta-practice shown in Figure 17 seems to be algorithmically related to the no-lose-stay and lose-switch practices. It would be incorrect to say, however, that the no-lose-stay/lose-switch meta-practice is *defined* by a functional relationship with the no-lose-stay and lose-switch practices. Any practice, or meta-practice for that matter, is defined by its practice constituents and not by a functional relationships. So, while it might be advantageous for an agent to think of practices in functional, or causal, terms, for the purposes of our analysis, practices are only characterised, and not defined, by algorithmic modelling. There will more to say on this in chapter 6.

Let us now return to the question 'is it rational to model practices functionally?' The short answer is 'it is!', but, of course, this is too quick. We will need to investigate the types of interactions that occur among rational agents, and the patterns that exist during these interactions, in order to assert that these agents have good reasons to model practices according functional algorithms. First we will look at rational behaviour under simultaneous interactions.

6.5 no-lose-stay/lose-switch methodology: rational under simultaneous interactions

Recall the circumstance under which agents in the strictly minimal social situation are provided feedback simultaneously. This feedback is achieved by waiting until both agents have pushed a button and then simultaneously doling out the appropriate consequence to each. Sidowski and team concluded that under conditions of simultaneous feedback that the agents will learn to co-operate, usually after three iterations. The agents were observed to achieve co-operation by adopting the no-lose-stay/lose-switch methodology for decision-making. The no-lose-stay/lose-switch methodology in the simultaneous situation can be described in game theoretic terms according to the outcome matrix in Figure 19 followed by the preference matrix in Figure 20.

		Agent B	
		No-lose-stay	Lose-switch
Agent A	No-lose-stay	In the preceding round both agents were rewarded. Both agents are rewarded in this round also.	In the preceding round A was rewarded and B punished. In this round A is punished and B is punished.
	Lose-switch	In the preceding round A was punished and B rewarded. In this round A is punished and B punished.	In the preceding round both A and B were punished. In this round both A and B are rewarded and this outcome leads to reward/reward.

Figure 19: Outcome matrix for no-lose-stay/lose-switch game.

From the outcome matrix we can create a preference or choice matrix. For both agents, no-lose-stay is the top preference since it delivers two rounds of reward. Next preferred is the outcome where both agents choose lose-switch. When both agents receive a reward and a punishment, this outcome leads directly to the reward/reward outcome in the next round and thus is the next preferred outcome for both agents. The remaining two outcomes are rated according to whether the outcome delivers a punish/reward or a punish/punish. The resulting preference matrix is shown in Figure 20.

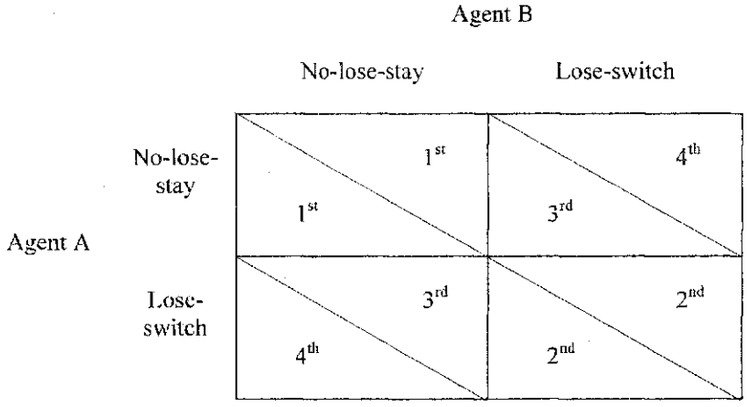

Figure 20: Preference matrix for no-lose-stay/lose-switch game.

The preference matrix points out that in simultaneous interactions both agents prefer the no-lose-stay choice. Unless both agents initially choose no-lose-stay they will have to work through the alternative outcomes until they arrive at the no-lose-stay/no-lose-stay equilibrium. The situation described is one in which both agents are acting rationally.[73]

6.6 no-lose-stay/lose-switch methodology: rational under sequential interactions

I noted above that two types of interactions were studied: simultaneous and sequential. We have looked at the simultaneous situation. The second type of interaction to be analysed involves sequential feedback. In this case, both rewards and punishments are delivered directly to the agents upon the agents' pressing the respective buttons: no delay beyond that of the 'natural' physical delivery of the signal is imposed. In this type of experiment an interesting phenomenon occurs. If both agents happen to reward the other

[73] First they demonstrate that they are capable of identifying functional relationships between two or more states of affairs; they have identified that a relationship exists between pushing one of the two buttons and either getting points or being punished. The agents demonstrate that they are able to evaluate the consequences of their actions relating states of affairs; a push of the 'reward button' results in a reward of calories. The agents demonstrate that they possess the ability to discover their preferences among possible states of affairs; pressing the 'reward button' indicates their preference for calories over electric shock. And the agents demonstrate that they are capable of identifying a preferred state of affairs and acting in pursuance of those preferred states; they continue to push the 'reward button' to maximise on calories.

agent on the first round – and assuming they follow the no-lose-stay/lose-switch methodology – then the two agents will engage in ongoing co-operation. But if either agent punishes the other on the first round then the no-lose-stay/lose-switch methodology results in a cyclical exchange in which both agents will choose a repeating pattern of *punish, reward, punish; punish, reward, punish; punish, reward, punish;* etc.[74] Let us consider the two possible outcomes associated with the sequential, or directly wired, game. These two possibilities are the reward/reward scenario and the repeating pattern scenario.

In the reward/reward possibility one agent rewards the other and the rewarded agent then returns the favour. As long as both agents choose reward on the first round and do not deviate in subsequent rounds, they can completely avoid punishment. Based solely on the number of possible outcomes, the chances of initial mutual reward is 25%. In fact, if we consider rational complacency, the chances of ongoing mutual rewards are lower than 25%.

Suppose that the first agent rewards the second. Notice that the second agent is rewarded when 'doing nothing'. The second agent could respond to reward by reciprocally rewarding the first agent. To respond at all, however, runs counter to Thorndike's law of effect. According to Thorndike, the second agent should continue to do nothing since she is rewarded for doing nothing and has no incentive to push either button. The first agent may even push the reward button three or four times in order to test for a pattern. Eventually, however, the first agent will switch away from reward and push the punish button. Punishing the second agent is likely to motivate her to switch from doing nothing. However, regardless of what button the second agent chooses, the two agents will fall into the second possibility: i.e., the repeating pattern scenario of *punish, reward, punish; punish, reward, punish; punish, reward, punish;* etc.

There are two variations of this second possibility. In the first variety, *reward outweighs punishment.* Let us assume, then, that the utility associated with one unit of reward outweighs the disutility of two shocks.[75] The total expected utility can be illustrated according to Figure

---

[74] The algorithm for explaining each agent's behaviour may look something like this:

1. If there is a positive marginal expected utility with finding out what will happen if you push one of the buttons go to 3, else go to 2.
2. If stimulated go to 3, else go to 2.
3. Flip a coin, if heads then go to 4, else, go to 6.
4. Press the button on the left.
5. If not negatively stimulated then go to 4, else go to 6.
6. Press the button on the right.
7. If not negatively stimulated then go to 6, else go to 4.

The above algorithm will yield a steady state result where each agent is pushing the button that results in reward. Also, notice that steps 1) through 3) are a matter of breaking the symmetry condition.

The resulting algorithm required is merely:

1. Press the button on the left.
2. If not negatively stimulated then go to 1, else go to 3.
3. Press the button on the right.
4. If not negatively stimulated then go to 3, else go to 1.

[75] In this case, either the shocks are relatively mild or the reward is relatively high.

21 where the practice curve shows the total expected utility of participating in the punish, reward, punish; punish, reward, punish; punish, reward, punish; etc. practice.

As the agents increase their knowledge about the practice they will undoubtedly recognise that a pattern has emerged. And while they are none too happy about receiving two successive shocks for each reward, the utility associated with the points outweighs the shocks. Thus, the practice yields a positive total expected utility. Further, the point where the marginal expected utility is zero is the point at which no additional energy need be expended on reconnaissance to understand the utility of the practice. The agent, then, simply engages in the practice expending only the minimal energy on running the basic algorithm. Consequently, an agent having access to knowledge k1 will retro-justify the existing practice as one that maximises total expected utility on the grounds that one reward is better than two shocks. There is no need for this rational agent to change its practice.

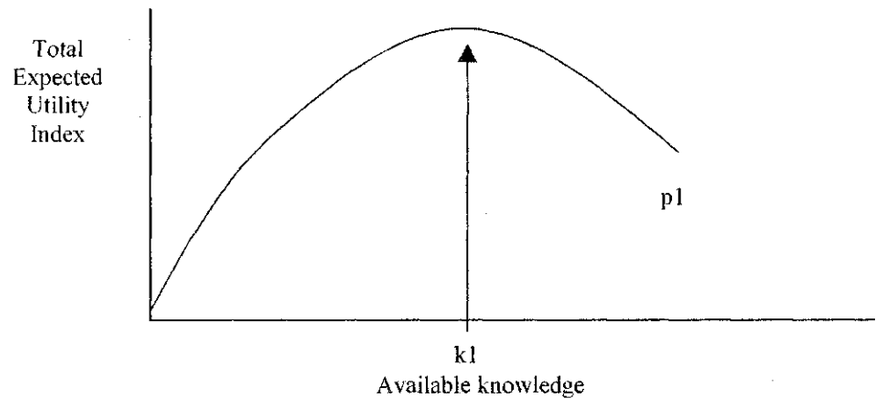

Figure 21: The reward outweighs punishment scenario.

The above might describe the way a teenager learns to drive a car. Initially, the driver must spend energy on both learning and executing the hand-eye-foot co-ordination algorithm in order to manoeuvre the car according to the rules of the road. Eventually, however, the algorithm is developed to the point where the driver needs only to expend energy on executing the algorithm. While it is true that learning more about the physiology of hand-eye-foot co-ordination or about automobile mechanics may improve the teen's driving skills, she deems attaining this additional knowledge as unhelpful work which reduces the total expected utility rather than improving it.

In the second variety of the *punish, reward, punish* outcome, *punishment outweighs reward.* In this case, the utility associated with one unit of reward does not outweigh the disutility of two shocks.[76] The total expected utility portrayed in Figure 22 where the practice curve shown represents the no-lose-stay/lose-switch practice. Initially the agent is content to do nothing and the practice delivers a maximum total expected utility at k1. Somewhere

[76] Presumably, in this case, the shocks are quite painful or the reward is small.

between k1 and k2 the agent decides to 'switch' and pushes the other button. As the agents exchange the rewards and punishments an agent will realise, at k2, that one reward is not worth two punishments. From then on the game has an infinitely low total expected utility since following the no-lose-win/lose-switch methodology results in infinitely repeating shocks that just are not worth it!

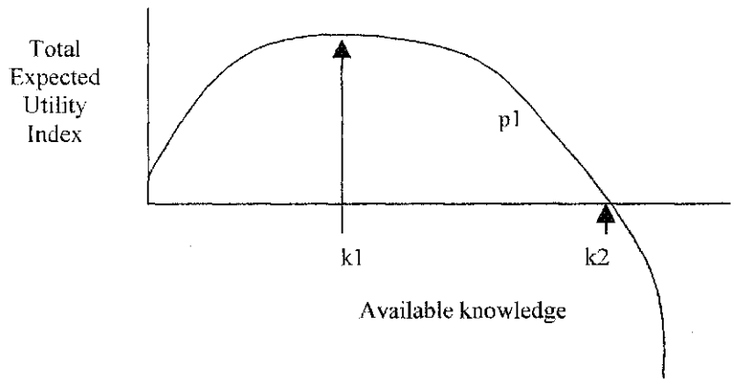

Figure 22: Punishment outweighs reward.

In this case, the no-lose-stay/lose-switch methodology is rational since the agent is, given the limits of knowledge available, maximising on total expected utility at the knowledge point k1. An agent that does not move beyond the knowledge point k1 but continues in the practice must find some justification for infinite punishment to behave rationally.

6.7 shocking new developments: an evolutionary analogy

Consider what would occur if the electric shocks were of sufficient voltage to kill the interacting agents. In this case, the only way both agents would survive would be if the initiating agent, by chance, chose the button that rewarded the other agent and the rewarded agent, also by chance, happened to choose the reward button. Now suppose that these experiments are occurring among a large population of agents. Of the total number of agents participating, only those pairs that are predisposed to pushing the right hand button would survive. Under any other circumstances both agents would perish: one due to electrocution, the other due to depleted energy reserves. So the surviving agents are those that push the reward button. These agents, disposed to push the right hand button, might somehow transmit this disposition either genetically – through reproduction – or socially, through teaching others what button to push.[77] Both the genetically disposed progeny and socially disposed students would survive and, indeed, would flourish. On this view, the minimal social situation can unfold in a way that models the evolutionary principle of natural selection. The practices that become dominant are those to which agents are predisposed for

[77] This position reflects Nicholas Rescher's in *Induction, An essay on the justification of inductive reasoning*, University of Pittsburgh Press, 1980, pp80-88.

whatever reason. In the case where both agents choose the right-hand button, conditions of interaction do not block the social or genetic transmission of the disposition. In the case where an agent has the disposition to push the left-hand button, the disposition is blocked from transmission, since doing so will result in the agent's death, a state from which it is not possible to transmit the disposition either genetically or socially.

6.8 changing practices

It would not be rational for an agent to continue to engage in any practice once recognising that the practice resulted in a negative total expected utility. One would expect that the rational agent would adopt an alternative practice that, in this case, did not result in death by a thousand electric shocks. And that is just what the agents in the Sidowski experiment did! Sidowski's agents dismissed the no-lose-stay/lose-switch practice and instead adopted a random testing approach. Purely random testing in this experiment amounts to an even chance between being rewarded and being punished. With sufficient tests, patterns begin to emerge. Eventually agents recognise these patterns and, as a result, adopt still another practice where they are able to achieve a nearly flawless reward/reward cycle. The three practices described are shown in Figure 23 with random testing, p2, delivering better results than p1, and the reward/reward practice, p3, delivering a higher total expected utility than random testing.

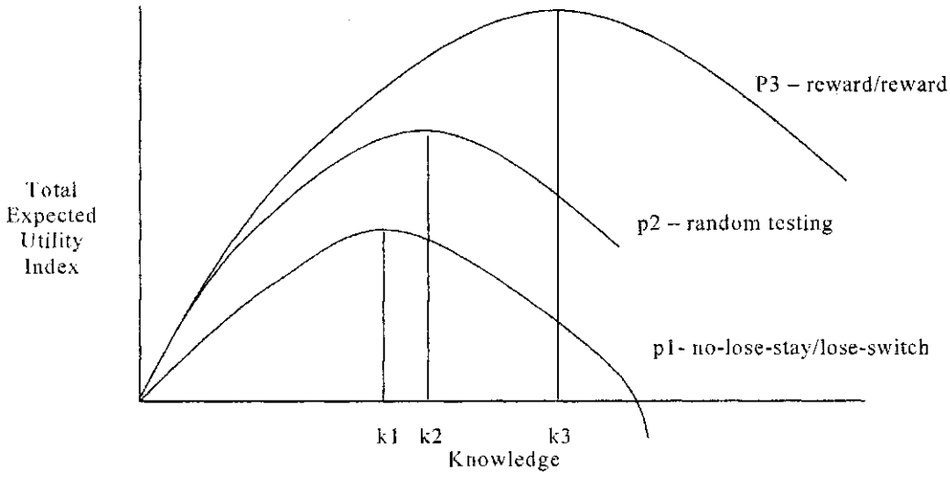

Figure 23: The utility of random testing.

The increase in knowledge from k1 to k2 represents the energy spent in reconnaissance: the energy spent in first, determining that practice p1 would ultimately deliver a negative total expected utility and second, that the random testing practice p2 needs to be adopted. The reconnaissance engaged during the trial and error practice is represented by knowledge

gained from k2 to k3 and includes recognising emerging patterns.[78] Notice that these agents qualify as rational since each practice curve is maximised given the available knowledge; but would they be considered strategic? Changing practices when facing death is certainly rational. And there is something intuitively strategic about expending energy to deliberate over what other practices are available.

6.9 authoromorphically informed agents

Recall the distinction between the strictly and non-strictly minimal social interactions, the latter being distinguished from the former by the players' awareness of each other's existence. Also recall that when faced with a negative total expected utility, agents adopt an alternative practice to the no-lose-stay/lose-switch practice. In the Sidowski experiment, agents opted for a trial and error approach where random trial and error delivers an even chance between reward and punishment for both agents. After a number of trials, however, agents adopted another practice that ultimately delivered better than chance reward/reward outcomes. Empirical testing has been carried out by Kelley et al. to determine how quickly interacting agents are able to adopt a practice that delivers better than chance results. According to Coleman, *informing* agents that they are interacting with another rational agent results in significant performance improvement. When pairs of players are informed about each other's existence, they are able to achieve a mutually rewarding practice faster than when they are uninformed.

> Mutually rewarding outcomes were much more frequent in informed pairs. When choices were made simultaneously, the relative frequency of rewarding choices in informed pairs rose to 96 per cent [from 75] after about 150 trials. Even under the alternating procedure, the frequency of rewarding choices gradually increased over trials and greatly exceeded chance expectations. [In uninformed pairs under the alternating procedure the rewarding choices was no greater than chance and showed no tendency to increase when the game was repeated over 140 trials].[79]

The upshot is that interacting agents tend to adopt practices with a higher total expected utility when they are informed than when they are not informed (see Figure 24).

Coleman takes informed to mean that the agents are told of the other's existence by the overseer of the experiment. But consider that an agent – rather than being informed by some lab-coated overseer – may be informed by adopting the authoromorphic stance. Let me explain. Suppose, for example, that an agent ascribes her own beliefs and desires to the console in front of her. The disposition to take the console as a rational agent may be a

---

[78] I show a discrete jump from p2 to p3 when in fact there may have been a number of trial practices between p2 and p3. To entertain these trial sub-practices would invite a sorites paradox which, while an interesting dilemma on its own, is not of primary concern in this paper.

[79] Andrew M. Coleman, *Game Theory and its Applications in the Social and Biological Sciences*, Butterworth-Heinemann Ltd., 1982, p47.

socially or genetically transmitted one. Or, the agent's disposition to treat the console as a rational agent may have occurred *a priori*. Regardless of how she came to hold the belief, the ascribing agent takes the console to be a rational agent and, thus, is informed. Furthermore, informed pairs – even authoromorphically informed pairs – will achieve the same kind of improved performance that Coleman described above. Thus, agents disposed to authoromorphising are likely to access a higher total expected utility than agents that are not so disposed.

One might argue that the authoromorphic requirement goes too far. For example, an agent may be informed by way of the intentional stance and would likely move from practice to practice in the same way as the authoromorphically informed agent. In the minimal social situation, however, Dennettian ascription does not go far enough. Authoromorphism is distinguished from intentional ascription by the former arising solely via introspection. Recall that in the minimal social situation, the conditions are not so demanding as to evoke rational bankruptcy. The conditions are, however, demanding enough to minimise the resources available for reconnaissance according to the principle of economy. Also, the minimal resources expended on reconnaissance are the 77 kcal/hr expended during authoromorphic introspective deliberation. As a result, the *minimal* resources required for an agent to expend in order to become informed are those resources associated with authoromorphic reconnaissance. In fact, the minimal distinction between the rationally complacent agent and the *more rational* agent is that the latter expends energy on introspective authoromorphic reconnaissance.

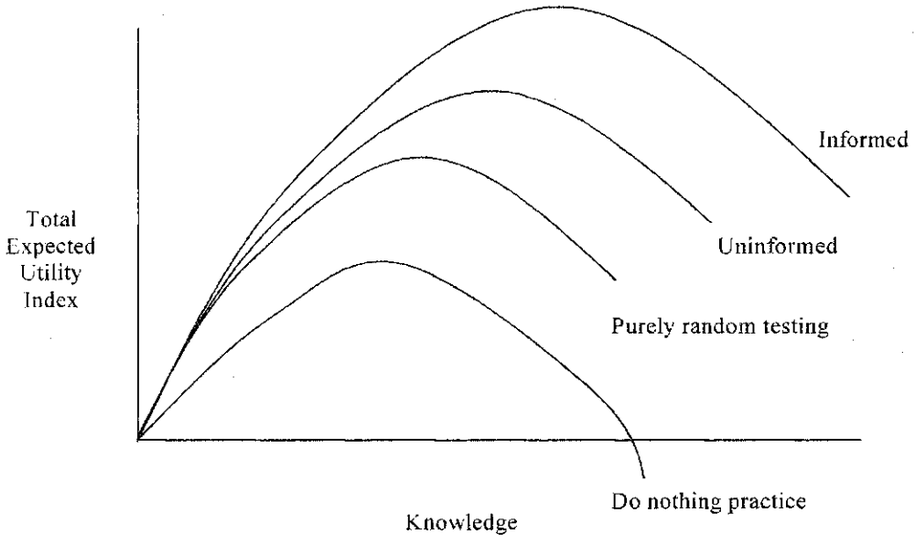

Figure 24: Distinguishing 'Informed' and 'Uninformed' performance.

Let us look at one last situation. Again, the situation is one in which the shocks could kill the agents. As with the previous case, if the agents knew the dangers of their situation, it would be advisable for them to be certain of their method before pushing any buttons. Of course, these agents do not know that they are in a social situation where an electric shock could kill them. So in this case we will give them more information. We will treat these agents as authoromorphically informed agents: both agents will ascribe their own beliefs and desires to the console in front of them and in doing so take the consoles to be rational agents. We will also assume that, as Rawlsian agents, these agents will take the conditions of their interaction to be symmetrical: agents may observe differences among one another, but these differences are not so great as to alter the choices available to either agent should their circumstances be reversed. Finally, we will also assume that these agents are predisposed, via social or genetic selection, always to push the reward button on the first round.[80] They can, however, choose from either button in successive rounds. Finally, we will add to the conditions initially described that each of these informed agents is told 'pushing one of the buttons will reward the other agent, pushing the other button will punish them'.[81] By providing the agents with this information, they find themselves in what Hobbes describes, in Chapter 13 of *Leviathan*, as the first of two natural conditions of human interaction. Says Hobbes:

> Nature hath made men so equall, in the faculties of body, and mind; as though there bee found one man sometimes manifestly stronger in body, or of quicker mind then another; yet when all is reckoned together, the difference between man, and man, is not so considerable, as that one man can thereupon claim to himselfe any benefit, to which another may not pretend, as well as he. For as to the strength of body, the weakest has strength enough to kill the strongest, either by secret machination, or by confederacy with others, that are in the same danger with himselfe.[82]

It has been suggested that Hobbes is claiming that all humans are equally able to kill each other. But this is not Hobbes' point. Instead Hobbes asserts that all humans are equal, not in that they can kill each other, but in that *they may be killed* by each other. The question then is how this *vulnerability condition* plays into the minimal social situation.

Consider that an agent may be initially satisfied with the 'do nothing' practice. After deliberating over the situation, the agent comes to believe that it will eventually run low on calorie reserves. Assuming the agent does not wish to die, the agent will reject the no-lose-stay practice and adopt the no-lose-stay/lose-switch practice where to switch is to push one of the buttons. The agent is predisposed to push the reward button and does. What is

---

[80] This condition is added in order to avoid the *'punish, reward, punish'* repeating pattern at the onset of interaction which leads to certain death for both agents.

[81] This information eliminates the condition that pushing either button is outcome neutral.

[82] Thomas Hobbes, *Leviathan*, p60.

especially interesting, however, is how the other agent reacts after the button has been pressed under the sequential situation.[83]

Suppose the first agent rewards the second.[84] It is rational for the second agent, who has been rewarded for doing nothing, to continue with its current practice. The first agent could reward the second indefinitely without receiving any points. For the first agent, the circumstances are unchanged. The first agent still faces a negative total expected utility. Thus, the first agent, like those in the Sidowski experiment, is likely to adopt a trial and error practice, the result of which will kill the other agent. It is at this point that the vulnerability condition plays an important role for the second agent. The second agent is being rewarded for doing nothing and it is rational to continue to do nothing. Upon deliberation and inference, however, the second agent realises that even though she is currently being rewarded, she may, at any time, be punished by the other agent.[85] In order to avoid punishment, once rewarded, the second agent's best chance of success is to switch from doing nothing to reciprocally rewarding the first agent. From this point on, the two agents will continue to reward each other on the no-lose-stay practice.

Notice that it would have been entirely rational for the second agent to be complacent, to maintain the 'do nothing' practice and continuing to receive rewards. Of course, there's a good chance that she will be electrocuted (but at least life was good while it lasted). The second agent can, however, choose a practice of higher utility than the 'do nothing' practice. Of higher utility for the second agent is to reward reciprocally the first. The rationale for this action turns on Hobbes' vulnerability condition. The second agent is informed authoromorphically; she knows that she has the ability to reward or punish the other agent; she believes the situation is symmetrical; thus, she believes that she is vulnerable to being punished by the other agent. In order to avoid punishment she needs to reward the first agent's behaviour before the first agent decides to switch.

6.10 a proposal for the distinction between rational and strategic

The no-lose-stay loop is characterised by complacency where an agent is satisfied, on available knowledge, that its current practice maximises total expected utility. It can be

---

[83] Under simultaneous interaction both agents will continue to press the same button and continue to be rewarded.

[84] We'll assume that the first agent has come to realise the gravity of the situation and decides to switch practices first.

[85] The conclusion would follow from:

1) The console has the same beliefs and desires I have;
2) If I were rewarding the console and it did not respond then I would switch buttons;
3) If the console was rewarding me and I did not respond then the console will switch buttons;
4) If the console switches buttons I will be punished;
5) I do not want to be punished;
6) The console is currently rewarding me;
7) I need to respond before it switches;
8) I should reward the console before it punishes me.

argued that complacent agents are not maximising on total expected utility if a positive alternative practice opportunity exists. Recall, however, that on Simon's view, for any given state of affairs, when an agent has access to knowledge k, there is only one rational practice to adopt. So for any complacent agent, even one having a positive alternative practice opportunity, it is still rational to engage in practice p1 at knowledge k1 since the agent is unaware that an alternative practice of higher utility is available. Nevertheless, it is still rational to engage in additional reconnaissance in order to identify a positive alternative practice opportunity even when doing so less than maximises utility on the current practice.

Identifying a positive alternative practice opportunity is characterised by the following.

1) The agent is expending additional energy beyond k1 by engaging in reconnaissance from which it establishes additional beliefs and desires used in characterising the current practice;

2) Part of expending additional energy beyond k1 includes envisioning alternative practices and deliberating on the total expected utilities associated with these alternative practices;

3) Inherent in a total *expected* utility is the condition that an agent must not only identify what possible practices are available but also assess the likelihood of those practices being adopted. Embedded in an assessment of likelihood is an investigation into both the energy required to adopt a practice and the expected payoff of the practice;

4) From 1) through 3) it follows that by expending additional energy on reconnaissance and investigating alternative possibilities – and at the same time continuing to engage in the current practice p1 – the agent engages in practice p1 in a manner which less than maximises total expected utility for p1;

5) Thus – notwithstanding that from 4) it follows that the agent is not acting rationally since it is not maximising on total expected utility in its current practice – from 3) and 4) it follows that the agent may appear to be acting irrationally with respect to its current practice when in fact the agent *is* acting rationally by expending energy on reconnaissance and deliberation to realise a positive alternative practice opportunity.

Point 5) highlights an important characteristic of complacency.

Also, recall that strategic behaviour can be couched in terms of one agent tricking another agent into believing that the former is doing one thing when indeed it is doing another. Specific to Figure 25, we can imagine that an agent appears to be engaged in practice p1 at k1 when, in fact, the agent has knowledge k2 and can engage in practice p2 but if, and only if, the second agent continues on practice p1.

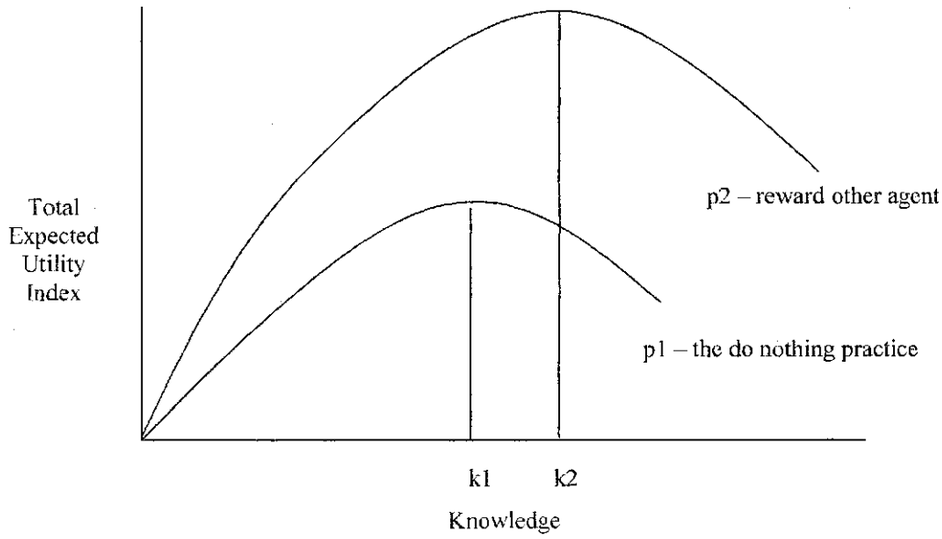

Figure 25: Distinguishing 'Informed' and 'Uninformed' performance.

Figure 25 is useful in depicting the practices of the agents in the minimal social situation. The first agent initially switched from the 'do nothing' practice p1 to the 'reward the other agent' practice p2 on the knowledge that she would eventually run out of calories. The second agent, on the other hand, must overcome rational complacency in order to switch from the do nothing practice, even when doing nothing resulted in reward.

Finally, recall that earlier we discussed Hobbes' natural condition of humankind where agents are equal in that they may be killed by another agent. But Hobbes offers a second condition of equality: "a greater equality amongst men, than that of strength".[86] I take this greater equality to which Hobbes refers to characterise the condition of rational complacency that was available to the second agent described above.

> That which may make such equality incredible, is but a vain conceipt of ones owne wisdome, which almost all men think they have in a greater degree, than the Vulgar; that is, than all men, but a few others, whom by Fame, or for concurring that howsoever they may acknowledge others to be more witty, or more eloquent, or more learned; Yet they hardly see their own wit at hand, and other mens at a distance. But this proveth rather that men are in that point equall, than unequall. For there is not ordinarily a greater signe of the equall distribution of any thing, than that every man is contented with his share.[87]

Hobbes' second equality condition characterises rationally complacent agents.

It is at this point that a case can be made for a cognitive virtue that distinguishes strategic from rational behaviour. Rationally complacent agents are those satisfied with their

[86] Ibid.

[87] Ibid. p61.

current practices, unaware of alternative practices of higher total expected utility, and unwilling to invest in reconnaissance to identify alternative practice opportunities. Fortunately, rational agents are not bound to rational complacency. Non-complacent rational agents explore alternative practices and deliberate over alternative practice opportunities. Rational agents, then, are either complacent or non-complacent. Suppose that we replace the term 'non-complacent' with the term 'strategic'. What follows is that the set of all rational behaviour consists of two subsets:

1. that which is rationally complacent, where an agent maximises total expected utility on available knowledge, and
2. that which is rational but not complacent – that is, strategic – where strategic behaviour is taken as the cognitive virtue of identifying positive marginal total expected utilities among alternative practices.

Hobbes himself alludes to this distinction in (what he calls) *Deliberation*. Says Hobbes:

> the whole summe of Desires, Aversions, Hopes, and Feares, continued till the thing be either done, or thought impossible, is that what we call DELIBERATION...And it is called *Deliberation*; because it is a putting an end to the *Liberty* we had of doing, or omitting, according to our own Appetite, or Aversion...Every Deliberation is then sayd to End, when that whereof they Deliberate, is either done, or thought impossible.[88]

On Hobbes' view, deliberation continues so long as the belief exists that one might attain an alternative practice that accords with one's own appetites. Deliberation ends when the practice is realised (or dismissed). Recall that rationally complacent agents engage in deliberation. But these agents engage in deliberation as a means of retro-justifying a current state of affairs or a current practice. Strategic behaviour follows from a non-complacent, or *predictive*, deliberation with the belief that one is free to discover an alternative practice which offers a positive alternative practice opportunity.

At this point I believe it is appropriate to articulate a proposal for a transcendental concept of strategic behaviour where the behavioural property is a cognitive virtue of the agent and not a condition of the domain of activity. The proposal is a recapitulation of 5) above:

> a strategic agent is that which expends energy on reconnaissance and deliberation so as to less than maximise the total expected utility of a current practice in order to discover an alternative practice which offers a positive alternative practice opportunity.

The statement above can be somewhat shortened for purposes of efficiency. I propose a concept of strategic behaviour as a cognitive virtue characterised by *the methodological critique of a practice*: specifically that which enables one to access positive alternative practice opportunities.

---

[88] Ibid., p28.

6.11 objection and reply: the recursion problem

In section 6.4 we explored the worry that there was insufficient reason for agents to model practices causally. We can now consider the full impact of this concern. We have, from section 6.10, good reason to assert that strategic behaviour is that which less than maximises on a current practice in order to discover an alternative practice opportunity. The detractor, however, may object that this assertion is problematic when an agent endeavours to less than maximise on the practice of rational complacency. The problem, claims the detractor, is that by less than maximising on rationally complacent behaviour, an agent is simultaneously behaving strategically and complacently.

The problem under consideration here will be recognised by those who are familiar with Newcomb's problem. Newcomb's problem is characterised by a situation in which a rational agent is given the choice between 1) taking the first of two boxes which contains one million dollars and 2) taking both boxes, the second of which contains one thousand dollars, making the total reward one-million-and-one-thousand dollars. The choice is difficult in the following manner. Prior to entering the room in which the two boxes are placed, the rational agent is subjected to a scanning device that predicts whether the agent will choose one box or two. The device is 99.9% percent accurate. If the device predicts that the agent will choose *only the first box*, both boxes will be filled with their respective cash rewards. If the device predicts that the agent will choose *both boxes*, only the second box will hold its reward of one thousand dollars.

The problem for the agent is whether to act according to the *principle of maximising expected utility* or the *principle of dominance*. The principle of maximising expected utility directs the agent to choose only the first box according to the following. The predictor would expect the agent, given an option between a 99.9% chance of receiving a million dollars and a 0.1% chance of receiving one-million-and-one-thousand dollars, to choose the former. As a result, the respective rewards are placed in both boxes. According to the principle of dominance, however, the agent should take both boxes. The dominance principle asserts that, if the states determining the outcomes of options are causally independent, and one option is better than another, it is rational to choose the better of the two options. Particular to this example, since one's choice does not change the decision made by the predicting device, and thus has no causal impact on the contents of the first box, and choosing both boxes gets you an additional one thousand dollars, you should take both boxes.

According to the theory of strategy as methodological critique, the rationally complacent agent will, on the principle of maximising expected utility, adopt the practice of choosing only the first box. But for the strategic agent things are not so simple. The strategic agent's chances taking a million dollars are initially, as with the rationally complacent agent, 99.9%. But the strategic agent engages in reconnaissance to less than maximise on total expected utility, and upon doing so entertains the principle of dominance. The deliberation is likely to proceed as follows.

1) The principle of maximising expected utility says that taking only the first box delivers a 99.9% chance of receiving one million dollars.
2) But the principle of dominance says that there is a 99.9% chance of receiving one-million-and-one-thousand dollars by taking both boxes.
3) But there is a 99.9% chance that the device will predict that I will do 2), in which case there will only be one thousand dollars in the second box.
4) Therefore, I should take only the first box, which is point 1). It would seem, then, that
5) it is strategic to behave complacently.

The problem, then, is that an attempt to model the practice of rational complacency causally leads to a recursive argument.

Again, those familiar with Newcomb's problem will recognise claim 5) to be similar to the position asserted by Robert Nozick in *A Theory of Rationality*. Nozick argues that in the case of functional agents, as our agents are, behaviour is explained causally. As a result, the principle of dominance must apply. In the case of Newcomb's problem, then, a rational agent should take both boxes. To assert that an agent should take 'one box' in a Newcomb's problem, says Nozick, is to invite *evidential* reasoning where the evidence supporting the 'one box' choice can result only from empirical observations about how well a 'one box' solution works in practical situations. In light of Newcomb's problem and Nozick's subsequent insights, the theory of strategy as methodological critique is weakened: but not seriously so, for the following reason. The explanatory force of a causal model is not dependent on the extent to which the model explains its algorithmic constituents. Let me explain this claim.

The theory of strategy as methodological critique is a causal explanation modelling the relationship among practice constituents according to the algorithms A_{RC} for the rationally complacent algorithm and A_S for the strategic algorithm. These two algorithms relate the independent practice constituents P_in and the dependent practice constituents P_dn. For a model to have explanatory efficacy, its algorithm must logically relate the practice constituents P_in and P_dn.

The detractor, however, is asking that the algorithm explaining strategic behaviour logically relate the independent and dependent constituents of the practice of rational complacency. It is not necessary, however, that the model logically relates the *practice constituents of the algorithm for strategic behaviour*, P_{A_S}, with P_dn, the model's dependent practice constituents. Take Newton's theory of gravity for example, a causal model mathematically explaining the relationship among its independent practice constituents G, M_1, M_2, R and its dependent practice constituents F. The algorithm directs one to multiply G, M_1, and M_2, and then divide by R squared. The practice constituents of the algorithm are multiplication and division. Yet we certainly do not expect the algorithm explaining the theory of gravity to also explain the practices of mathematics and division. Asking one to relate the constituents of the algorithm for strategic behaviour with the constituents of the practice of rational complacency invites an analogous expectation.

Practices are defined interactively. That is to say, the constituents of the situation in which the agents are interacting define the practice. Agents themselves are characterised by a function, or algorithm, so posited as to relate behavioural inputs and outputs. While it is true that 1) the practice constituents and 2) the agent's algorithmic constituents may be one in the same thing, it is not necessary that the algorithm relating functional behaviour logically relate practice constituent and algorithmic practice constituents.

In the worry we have been discussing, the algorithmic practice constituents of strategic behaviour relate to the practice constituents of rational complacency recursively. But the theory of strategy as methodological critique does not purport to relate logically the algorithmic practice constituents of strategic behaviour with the practice constituents of rational complacency. As we have seen, the explanatory force of the theory of strategy does not turn on such a relation.

Segue

Methodological Critique: A Recap

Claim 5) is thus far based on both a reductive approach and intuitive probing. At this time I will subject my proposal for a transcendental concept of strategic behaviour to evidential and explanatory scrutiny. First, let me retrace the path that has led us to this proposal.

The notion of strategy plays a considerable role in game theory, in the military and political sciences, and in the practical affairs of business. And while ample definitions are offered in each of these areas of study, no single definition is appropriate in all cases. Furthermore, the term strategy seems to be used interchangeably with the term rationality.

In order to unearth a distinction between rationality and strategy a thought experiment was offered in which agents were interacting under the minimal social conditions. The results of the thought experiment were threefold. First, rational behaviour is either complacent or non-complacent. Second, a condition in which agents behaved rationally and non-complacently was identified. What makes non-complacent rationality distinct from complacent rationality is that the former is characterised by, at minimum introspective authoromorphic reconnaissance, and beyond, reconnaissance that less than maximises the utility of a current practise in order to discover an alternative practice of higher utility. While deliberation plays a role in both types of rationality, rational complacency lacks the cognitive virtues required to discover alternative practices and instead employs deliberation for the purpose of retro-justification. Third, strategic behaviour was identified as the use of non-complacent cognitive virtues in accessing positive alternative practice opportunities. What immediately follows from these three results is that strategic behaviour is characterised by non-complacent rationality. In the context of the cognitive virtues of rational agents, therefore, rational behaviour is either complacent or strategic.

As noted, this proposal arose from both intuitive probing and logical reduction and should be taken, at this point, as such: a proposal. What stands between the point of proposal and the point of assertion is evidential and explanatory justification of the proposal. Having said that, next I will provide the justificatory evidence supporting the transcendental concept of strategy in fields in which the term strategy is taken to play dominant roles.

Part III

Evidential and Explanatory Justification

Chapter 7

Strategy in the Military and Political Sciences

The intuitive notion of strategy is historically rooted in the military and political sciences. It is certainly possible to chronicle what are taken to be the important advances in 'military strategy'. But while the evolution of military strategy is itself interesting, what is important to this investigation is establishing a correlation between military strategy and methodological critique. Correlating the two will be accomplished by a review of significant advances in strategic thinking within the rubric of the military and political sciences. In each case, the advancement being considered will be analysed for evidence of methodological analysis. As for what will count as methodological analysis, recall, as per the preceding section, that methodological critique is evidenced by non-complacent rationality delivering positive alternative practice opportunities.

Following Kenneth Booth's classifications for the development of strategic thinking in the military arena, we can organise military strategic analysis according to the following:

1. Pre-Napoleonic strategic analysis,
2. Strategy in the era of nationalism and industrialism,
3. The generation of World Wars (1914-1945) and the nuclear strategists.[89]

Each of these classifications will be explored in search of evidence that correlates strategic military thinking with methodological critique.

7.1 pre-Napoleonic retro-justification

For the most part, the pre-Napoleonic method of strategic analysis was a matter of retro-justifying non-deliberative practices. New practices were usually established as a matter of happenstance. The battle of Agincourt is perhaps the best example retro-justification. In 1415, Henry V of England, claiming title to French land, had seized the port of Harfleur en route to Calais. According to most accounts, the 6,000 strong English force, primarily longbowman, encountered a French force of 25,000 men. The English initially sought a truce, which the French rejected. The English retired to wooded cover but were forced to meet the French in a narrow opening among the woods. The French attacked with their cavalry. The wet weather and mud slowed the French cavalry such that the English longbowman were able to halt the French cavalry and inflict casualties upon a full one-fifth of the French force before the French withdrew.

[89] Ken Booth, *Contemporary Strategy: Theories and Policies*, 1975, pp22-41.

After reflecting upon the manner in which the battle had unfolded, the English force, under Henry V, formalised the practice of employing longbowman under conditions which favoured a similar outcome to that of Agincourt. As with the Battle of Agincourt, most pre-Napoleonic military practices were formalised after reflection on what had serendipitously transpired during a battle. In formalising the practices that delivered positive results in battle, military strategists attached a positive alternative practice opportunity to the practice formalised. Methodological critique as retro-justification is prevalent throughout pre-Napoleonic warfare. The Egyptians employed the mace as a weapon of the foot soldier until helmets were introduced and the use of the sword came into practice. The mobility of the chariot and later the concept of a cavalry redefined the role of the phalanx.

Along with the retro-justification of a practice, there are certain pre-Napoleonic narrators who took the bold step to link overtly warfare with political ends. This is an important step since – prior to Napoleon – the role of a soldier, and often the commander, was couched in religious terms. Going into battle meant putting one's life in the hands of the gods: for example, the Viking who died in battle took his place in Valhalla.

The reduction of warfare to a political agenda, rather than a theological one, is thought to have begun with Thucydides (c.460-c.400bce) and is taken to have fully developed with Machiavelli. There are certainly others who contributed to the link between warfare and politics;[90] however, it is only Thucydides and Machiavelli on which I will focus here.

Thucydides' accounts of the Peloponnesian Wars are often taken as the first historical record of a strategic analysis. After his discharge from actual fighting,[91] Thucydides interviewed both Spartan and Athenian soldiers, and thus concentrated on providing an unbiased account of how the war was unfolding. But unlike other historians of his time (such as Herodotus) Thucydides was focused on the political implications of the conflict.

> While Herodotus presented a widely elaborated, leisurely story, which paid attention to many aspects of like and dealt with each one as a rounded whole, Thucydides concentrated his vision upon political and military events and divided his treatment rigidly by campaigning seasons. His aim was not primarily to chronicle these events but for more so to illuminate the forces at work so as to aid leaders.[92]

It is possible to view Thucydides' analysis as a retro-justification of the practices that were successful in warfare and the practices that were not. Thucydides himself claims that his historical account is given to serve as a practical guide for "like events which may be expected to happen hereafter in the order of human things".[93]

---

[90] Sun Tzu, for example.

[91] Thucydides, at the time an Athenian general, was unable to secure Amphipolis (in Thrace) which was besieged by the Spartans, and was thus discharged.

[92] Chester G. Starr, *A History of the Ancient World*, Oxford University Press, 1991, p351.

[93] Ibid.

Perhaps Niccoló Machiavelli's 'The Duties Of A Prince With Regard To The Militia' in *The Prince* best exemplifies a retro-justified link between military action and political ends.[94] Machiavelli's methodological critique did not influence his contemporaries or the methods of sixteenth century warfare. Still, "the military innovators of the time were pleased to find a work in which aspects of their practice were explained and justified".[95] What Machiavelli did was nothing short of taking to a logical conclusion the conceptual link between state and militia. In Machiavelli's letter to the Medici, Machiavelli articulates what he takes to be the link between head of state and the military: Says Machiavelli:

> [a] prince should ... have no other aim or thought, nor take up any other thing for his study, but war and its organisation and discipline, for that is the only art that is necessary to one who commands, and it is of such virtue that it not only maintains those who are born princes, but often enables men of private fortune to attain that rank.[96]

The prince to which Machiavelli refers is a political leader, not a religious one. In order to fulfil the role of prince, a political leader must place primary emphasis on his ability to organise and conduct warfare.

The evidence I have presented here suggests that formal changes in pre-Napoleonic thinking were limited to the retro-justification of a practice. It is worth noting that there are examples where methodological critique was performed prior to adopting a practice. Hannibal, for example, has been cited as one whose behaviour would lead his opponents to believe that he was engaged in one practice when he had, in fact, adopted another that delivered a higher total expected utility. Says Field-Marshall Montgomery of Alamein, "[Hannibal] was a master of psychology...in his ability to mislead and mystify his opponents".[97] Still, in pre-Napoleonic times, retro-justification is the predominant means of adopting new practices.

7.2 – the era of nationalism and industrialism

The use of psychology as a means of predicting an opponent's behaviour certainly occurred in pre-Napoleonic times. Yet, as a means of accessing a positive alternative practice opportunity, interest in psychology became pervasive during the era of nationalism and

---

[94] Niccoló Machiavelli, *The Prince*, The Modern Library, Inc., 1940, pp53-55.

[95] Peter Paret, *Makers of Modern Strategy from Machiavelli to the Nuclear Age*, Princeton University Press, 1986, p28.

[96] Op.Cit., p53.

[97] Field-Marshall Viscount Montgomery of Alamein, *A History of Warfare*, The World Publishing Co., 1968, p97. Hannibal's manoeuvre at Cassino provides an example. The Roman forces planned to pin Hannibal down in a mountain pass near Cassino. Expecting this, a small contingent from Hannibal's army attached torches to the horns of a heard of cattle and drove them toward the pass. Roman reconnaissance, during the night, observed the mass of lights moving toward the pass and reported that the Carthaginian forces had mobilised. The Romans executed their plan to trap Hannibal in the pass only to find they were defending the pass from torch-laden cattle. Meanwhile, Hannibal led his army through an alternative pass to safety.

industrialism. Carl von Clausewitz provides the archetypal description of strategic behaviour for this era.

Prior to Machiavelli, war had often been justified according to the divine rights of feudal lords. In contrast, the fallout from the intellectual advancements of the Enlightenment resulted in an approach to warfare that attempted to model interacting agents mathematically. Von Clausewitz vehemently rejected both of these approaches as a means of analysing warfare. Instead von Clausewitz insisted that a theory of war must be built from the basic premise that war was the intellectual use of, or the threat to use, violence. This intellectual approach, according to von Clausewitz, requires that the emotional distress and psychological brutality of war are treated as facts, or as tools, in determining the best course of action to secure victory and battle:

> The maximum use of force is in no way incompatible with the simultaneous use of the intellect. If one side uses force without compunction, undeterred by the bloodshed it involves, while the other side refrains, the first will gain the upper hand...it would be futile – even wrong – to try to shut one's eyes to what war really is from sheer distress at its brutality.[98]

Von Clausewitz prescribed practices designed to bend the enemy to one's own will. On this view, a rational approach does not involve putting one's destiny in the hands of a god. Instead, practices are established by predicting the enemy's behaviour on the basis of physical conditions and psychological analysis. Reconnaissance is required to determine an opponent's beliefs and desires. At the same time, opponents will likely endeavour to conceal their beliefs and desires leaving gaps in the information available: gaps, however, that can be filled in authoromorphically.

From the time of Napoleon to the end of the Second World War, the role of authoromorphism as a means of predicting an opponent's behaviour increased dramatically. The success of authoromorphism led to a series of advancements in methods of reconnaissance. These methods in turn delivered substantive positive marginal total expected utilities. Napoleon is taken as a model for using psychology in reconnaissance. Says Field-Marshall Montgomery of Alamein:

> [Napoleon] formulated his plans on the basis of information supplied by his staff...Information was kept up to date and immediately accessible on every relevant subject. Minute research preceded the organization of a campaign...He did as much as possible in advance to determine the course of the battle.[99]

The Battle of Austerlitz is often told in a manner that marks Napoleon's ability to organise the circumstances of battle in advance such that both sides engaged in practices that

---

[98] Carl von Clausewitz, "What Is War?", taken from Peter Paret, *Clausewitz And The State*, Princeton Press, 1985, p383.

[99] Field-Marshall Viscount Montgomery of Alamein, *A History of Warfare*, The World Publishing Co., 1968, pp345-346.

maximised Napoleon's total expected utility. At Austerlitz, the circumstances seemed to heavily favour the allied forces. The Allies, primarily Russians, greatly outnumbered the French, some 90,000 to 20,000. Further, Napoleon's field supplies had not been renewed in some time. Napoleon was aware that the numbers favoured the Allies. In addition, he reinforced what appeared to be his unfavourable conditions by refusing to allow his own army to redress thus giving his army the appearance of being ragged and badly beaten. At the same time, Napoleon reinforced the working condition of his men's weapons, issuing new rifles where appropriate.

The Russian Commander, Tsar Alexander I, was convinced that Napoleon was leading an unprepared and tired force that, with the Morovian Mountains situated to his north, would be trapped by flanking the French force from the south. Alexander did have some concerns with this plan. The move south would place the Littawa River and Lakes at the Russians' backs as the army swept south before turning to strike north. Alexander, however, dismissed this concern since the numbers so heavily favoured the Russians. Napoleon's own reconnaissance confirmed Alexander's plan. What's more, Napoleon suspected Alexander of rational complacency believing that Alexander expected Napoleon to retreat should the Russians flank from the south. Napoleon, on this assessment, positioned his main force at what would be the narrowest span of the flanking allied army and waited.

Given the knowledge of the situation, both the Russian and the French commanders acted rationally. Yet the battle unfolded as if Napoleon had simply willed it. The Allies moved south preparing to flank north. Napoleon held off the advancing army at the southern extremity until the Allies' main force was stretched across the shore of the lakes. Napoleon then attacked. The Russian lines, stretched thin, and with their backs against the water, were unable to hold off the French. The Russians broke ranks, and the soldiers retreated to their rear across the ice on the frozen lakes which, once the Russians were upon it, Napoleon smashed with his canons.

This Battle of Austerlitz provides a vivid example of how an opponent's rational complacency can be leveraged to maximise one's own utility. Throughout the age of nationalism and industrialism, the use of both reconnaissance and authoromorphism dominated military strategy. Following von Clausewitz, the psychological aspects of warfare became interwoven in predicting an enemy's behaviour, and thus total expected utility. With an increased use of reconnaissance came a reliance on authoromorphism to fill in the gaps in incomplete information.

7.3 – the world wars and the nuclear strategists

Authoromorphism took on entirely different complexity in the wake of the Industrial Revolution. While the acquisition of knowledge through traditional means such as espionage continued to be practised the industrial revolution brought with it an increased interest in scientific knowledge and technological advancements. Energy was invested into the development of weapons of war which themselves were intended to render the opponent

into a state of psychological resignation. One might point to the use of automatic rifles, tanks, or aeroplanes as examples of technological advancements in weaponry. However, Japan's surrender following the atomic bombs dropped on Hiroshima and Nagasaki is clear example of bending an enemy's will to one's own through the use of technologically advanced weaponry.

The modern view that one's armed forces should be both mechanised and technologically advanced was developed by Colonel B.H. Liddell Hart and introduced in what he called the *New Model Army*. So important was technology to military strategy, claimed Liddell Hart, that military technological development should form the basis for national defence policies.[100] Liddell Hart supported his view with two tactical platforms particular to his time. On the first, the tank played the role of a modern cavalry thus reducing reliance on roads and railways for mobilisation. On the second, the role of the air force was elevated from primarily reconnaissance to that of an offensive weapon in co-ordinated air-land attacks. The German Blitzkrieg is an example that characterises the Liddell Hart theory for mechanised warfare.[101] Aside from the tactical advancements of his time, Liddell Hart advocated the *a priori* identification of alternative practices of war. The criterion for an effective alternative practice was that it must deliver "a decisive blow against the Achilles' heel of the enemy army, the communication and command centres which form its nerve centre".[102] Upon identifying the practice that would deliver such a blow, governmental treasury support was required to fund the technological advancements required to actualise whatever practice was taken to meet the criterion.

From 1919 to 1945 technological advancements became the primary means of achieving practices of higher and higher total expected utility. But the advent of the atomic bomb and the subsequent development of nuclear weapons in the years that followed the Second World War issued in still another complex to military strategy. In the nuclear age, the threat of destruction was itself sufficient to bend the enemy's will to one's own. Consider the threat of mutually assured destruction (MAD). Those advocating MAD hold that for two opponents, both having the capacity to inflict a level of damage on the other that the other would take as prohibitive, the threat of damage is itself sufficient to ensure that neither agent will strike the other.[103] Interestingly, by arming oneself to the point of mutually assured destruction, one is engaged in the no-lose-stay loop where to stay is not to strike first. While

---

[100] Brian Bond and Martin Alexander, "The Doctrines of Limited Liability and Mobile Defense", found in Peter Paret, *Makers of Modern Strategy*, Princeton University Press, 1986, p600.

[101] See David McIsaac, "Voices from the Central Blue: The Air Power Theorists" found in Peter Paret, *Makers of Modern Strategy*, Princeton University Press, 1986, p626.

[102] B.H. Liddell Hart, *The Future of War*, New York Press, 1925, pp79-85.

[103] See Paul Viminitz, "Nuclear Warfare" in the *Encyclopaedia of Applied Ethics*, Volume 3, Academic Press, 1988, p358.

it is arguable that MAD is an effective means of maintaining the no-lose-stay loop, the risk of someone finding a positive total expected utility in striking first is worrisome.

The technology of the nuclear age makes defence against nuclear destruction difficult. What is more, nuclear destruction makes testing nuclear deterrence unreasonable. As a result, strategies for deterrence can only be developed and analysed authoromorphically. Nuclear warfare strategists must develop their plans on the ascription of beliefs and desires to their opponents. This is especially so since an opponent is likely to do their best to conceal certain beliefs and desires from you. By removing empirical testing, deterrence strategies rely heavily on rational choice theory, including game theory. The history of military strategy is an interesting transition from the rationally complacent, empirically based, retro-justification of the pre-Napoleonics to the rational decision theories of nuclear deterrence strategists and the reliance on game theoretic analysis to determine practices with positive alternative practice opportunities.

7.4 strategies for game theoreticians

Does the transcendental concept of strategy have explanatory force for the game theoretician? This question is answered by understanding the role of rational complacency under game theoretic conditions. In this section I will show that: 1) Leader and Prisoner's Dilemma allow the interacting agents to behave complacently; and 2) the *ceteris paribus* conditions in Battle of the Sexes rule out rationally complacent behaviour; and 3) Chicken can be interpreted either way. Furthermore, I will show that the rational complacency supports the claim that it is possible to think one's way out of a prisoner's dilemma.

In Leader, rationally complacent behaviour is overtly available to both agents in the C,C outcome. Recall that in Leader the C,C outcome is that where neither agent goes first. Both players wait for the other to turn and each is satisfied with sitting indefinitely. As with the agents in the minimal social situation, the practice of waiting has no immediate risk. At the same time, the practice does not advance the player toward a higher total expected utility and will eventually result in a negative marginal expected utility as both agents' fuel reserves dwindle. Thus, in Leader, the complacent choice C,C is rational but not strategic.

Similarly in Prisoner's Dilemma, a rationally complacent choice is available in the 'keeping the faith' choice albeit accounted for under the *ceteris paribus* conditions. Recall that players can either 'rat out' the other agent or 'keep the faith'. Neither prisoner can choose to clam up, say nothing, and allow circumstantial evidence to decide their fate. However, the jail time associated with 'keeping the faith' is indistinguishable from the jail time associated saying nothing. In fact, the 'say nothing' condition (inherent in the 'keep the faith' choice) reinforces that 'keeping the faith' is a rationally complacent behaviour since for both players ratting out the other delivers a higher total expected utility than keeping the faith. Thus, in Prisoner's Dilemma, to 'keep the faith' is rational and can be retro-justified as such, but to 'rat out' the opponent is strategic.

There is another point to be made here. Recall from chapter 4 the Prisoner's Dilemma preference matrix reproduced in Figure 26.

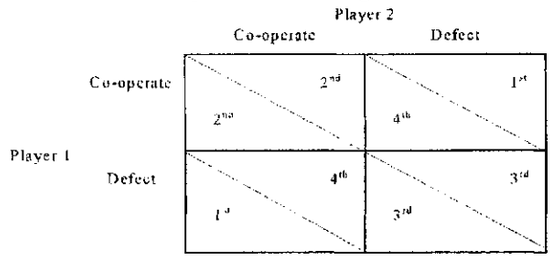

Figure 26: Preference Matrix for Prisoner's Dilemma

Recall also that defection in a Prisoner's Dilemma is the dominant strategy. The description provided above captures defection as strategic since to defect is the dominant strategy and is more rational than rationally complacent co-operation. Suppose, however, that we stipulate that defection is a rationally complacent disposition and co-operation is strategic. For the defecting agent, defecting is their 1st choice since it maximises utility regardless of the outcome. Therefore, for both agents, defecting is 1st in both the case where the other agent co-operates and where the other defects (as shown in Figure 27). The upshot is that for a rational agent in a Prisoner's Dilemma, co-operation can never be strategic.

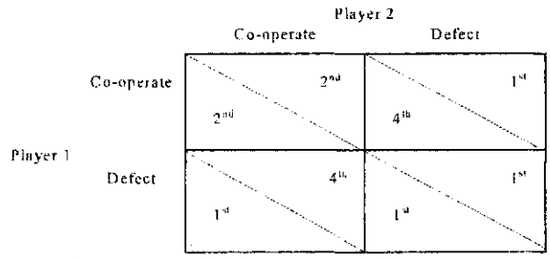

Figure 27: Alternative Preference Matrix for Prisoner's Dilemma.

In Battle of the Sexes the game is so framed that both players are given a choice between going to a romantic movie and going to a boxing match. The rationally complacent practice of staying home is not at all present as an option. The choices are presented in a way that the *ceteris paribus* condition requires that each player *must* adopt a practice other than 'do nothing'. Presumably, then, all the choices available in the Battle of the Sexes game deliver a positive alternative practice opportunity over rational complacency (which is not an option).[104] Thus, all choices in Battle of the Sexes are strategic.

Finally, in Chicken, the ruling out of a rationally complacent choice may be interpreted in two ways. In the first, rationally complacent behaviour is ruled out by the *ceteris paribus* condition that both players do not simply sit in their cars without depressing the

[104] The practice of rational complacency is accounted for by the *ceteris paribus* condition where all things being equal includes the condition that both players must adopt a practice other than rational complacency.

accelerator pedal or managing the steering wheel. In this scenario, as with the two described above, all choices are strategic. The second, and more interesting, interpretation is characterised by one player driving head on toward the other player and simply not recognising the risk associated with the impending crash. But the game of chicken is usually framed in a way that a defecting agent is keen to make the other agent aware that she is defecting.

It may, in fact, be strategic for an agent to dismiss the impending danger in a way that the other is made aware of the dismissal. By dismissing the danger, the first agent is likely to appear irrational to the second. By the first rendering itself irrational, it is entirely up to the second agent to choose between crashing head-on and swerving. Whether the first agent – in dismissing the impending crash – is doing so strategically or doing so as a result of rational complacency is known (or not known) only by the dismissing agent. Third party speculation of what the agent knows is unhelpful since functionally the behaviour associated with knowing or not knowing is indistinguishable.

Ceteris paribus seems to rule out a choice where the defecting agent is unaware of the danger of a crash and is self-satisfied with its current practice, either happily sitting (in the car) in the middle of the road or driving down the road unaware of the impending crash. But it is only the agent who defects in a game of Chicken that can know if the choice is made strategically or not since it would be strategic to claim rational complacency.

The transcendental notion of strategy as methodological critique does provide explanatory force within the context of game theoretic modelling. Where the explanatory force may not be overtly evident are in the games of Battle of the Sexes and Prisoner's Dilemma, where the game is so designed that the practice of rational complacency is ruled out by the *ceteris paribus* conditions. Having said that, it is important to note that there exists a problem in the game of Chicken in determining whether or not the defect choice is strategic or not. This problem, however, is one that is best solved within the theory of mind rather than taken as a problem with the transcendental definition of strategy.

7.5 empirical evidence from the world of business

There exist compelling evidential and explanatory support for the transcendental concept of strategy in military history and in game theory. The question remains as to how well the transcendental concept of strategy explains Mintzberg, Ahlstrand and Lampel's description–definition dilemma introduced in Chapter 3. Recall:

> [a]sk someone to define a strategy and you will likely be told that a strategy is a plan, or something equivalent – a direction, a guide or course of action into the future, a path to get from here to there. Ask [the same] person to describe the strategy that his or her own organisation or that of a competitor actually pursued over the past five years – not what they intended to do but what they really did. You will find that most people are perfectly happy to

answer the question, oblivious to the fact that doing so differs from their very own definition of the term.[105]

First let us consider Mintzberg and team's notion of a strategy as a plan. On the transcendental concept, strategic behaviour is that which, through deliberation and reconnaissance, identifies positive alternative practice opportunities. A strategy, then, is the formal articulation of deliberation and reconnaissance. This formal article includes the process for deliberation, the method of reconnaissance, the identification of the preferred practice, and the change in physical conditions required in order to realise that practice. The plan may be subdivided in such a way that outlines a series of intermediary practices necessary to engage the final practice. In accordance with Mintzberg et al., a strategy is just such a plan.

Let us turn our attention to the descriptive responses concerning the strategies pursued over the past five years. Here we have rational agents, only moments after calling a strategy a forward-looking plan, describing a strategy in terms of what actually happened over the past five years with no comparison to what may have been planned five years ago. But recall that agents often adopt a position of rational complacency. A rationally complacent agent would be self-satisfied with its current practice; the agent would be unwilling to expend energy in evaluating what was planned five years ago and what actually occurred in order to understand errors in planning. By committing the fallacy of planning, the agent rationally – albeit complacently – focuses on the constituents of the current condition. Thus, the rationally complacent agent answers the question of what strategy was followed with a retro-justification of the current practice including a historical account of relevant deliberations, reconnaissance, and changes in physical conditions. Since the cognitive disposition from which the agent derives this description is an internally focused self-satisfaction, there is no need for the agent to compare between the plan from five years ago and the actual states of affairs that unfolded. The agent may not even be aware that such a comparison is possible.

Mintzberg, Ahlstrand and Lampel have suggested that strategy formation is best mapped according to one's view of the world and the internal processes one calls upon to interact with the world. The internal processes available for interacting fall along the deliberative continuum described in Chapter 5. Recall that ideal panoramic rationality set is one extreme of the deliberative continuum and rational bankruptcy the other.

---

[105] Henry Mintzberg, Bruce Ahlstrand, Joseph Lampel, *Strategy Safari A Guided Tour Through The Wilds Of Strategic Management*, The Free Press, 1998, p180.

According to Mintzberg, Ahlstrand and Lampel, if one views the world as comprehensible and controllable, and calls upon internal processes that utilise idealised panoramic rationality, then one is likely to focus on scenario planning and game theoretic analysis to access an alternative practice. If, however, one views the world as comprehensible and controllable and calls upon internal processes that utilise heuristic algorithms for decision-making, then one is more likely to focus on identifying positive alternative practice opportunities rather than focusing on the method by which to access an alternative practice. On the other hand, one may view the world as chaotic and unpredictable. If one views the world as chaotic and unpredictable, and calls upon internal processes that utilise heuristic algorithms for decision-making, then one is more likely to focus on distinguishing meta-practices from alternative practices rather than being too quick to switch when one's current practice happens to deliver a negative marginal expected utility.[196] If, however, one views the world as chaotic and unpredictable, and calls upon internal processes that utilise idealised panoramic rationality, then one is likely to focus on the distributional data of a given practice in order to avoid the fallacy of planning.

Consider the following example in which strategic behaviour overcomes the fallacy of planning. Suppose an agent makes, ships, and sells widgets. Further suppose that each month the agent forecasts the profit (in utility) she expects to gain by making, shipping and selling widgets. Suppose also that in each of the months when profits are less than expected the agent compares what she forecasted to what she actually made and asks the question 'why is the actual profit different than my forecast?'. The forecast and actual data is shown in Figure 28. Also shown in Figure 28 are the answers to the question – 'why is the actual different than forecast?' – asked in the months where profits are less than forecast.

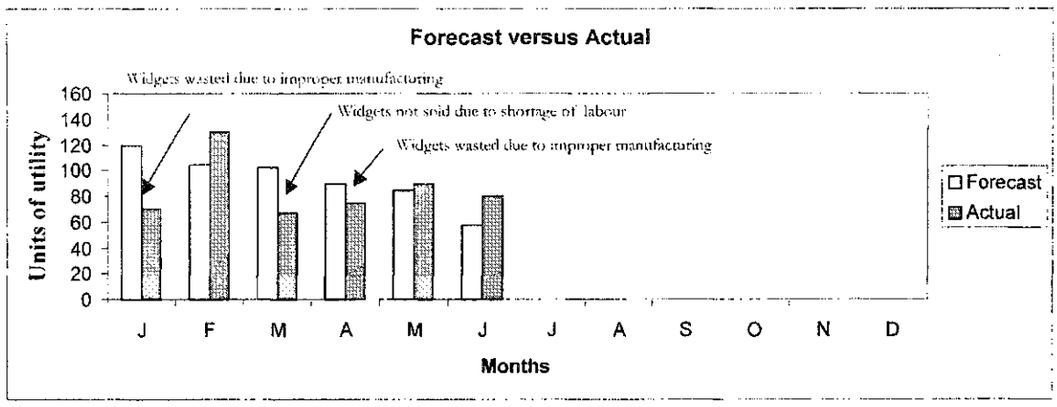

Figure 28. Forecast versus Actual data.

[196] The restraint from switching is primarily driven by bounded rationality.

First, let us consider how this agent might be considered rationally complacent. For the rationally complacent agent, the forecasted profits are the maximum expected utility for the current practice used to make, ship and sell widgets. On the months where the actual profits exceed forecast, the complacent agent is satisfied that the current practice did, in fact, maximise on expected utility. That the practice delivered greater than what was expected is not significant. It may be the case that the agent is happily surprised that actual profits exceeded the maximum, nonetheless, this does not change the complacent agent's view that what is forecasted is that which maximises profits on the current practice.

On the months where profits fall below forecast, the rationally complacent agent may be compelled to ask why the actual is less than forecast. In Figure 28, the answers to this question are shown. Notice that the answer may be that the widgets had to be thrown away due to the fact that they were improperly manufactured. In the same way that an agent explains a child's death in theological terms, or the car driving agent who sees running out of gas as an anomaly, the rationally complacent agent in this case views improper manufacture of widgets as an anomaly and will adjust the practice curve to match whatever utile the practice delivers.

For the rationally complacent agent, the difference between forecast and actual is a retro-justification for continuing to use the same practice. Notice that the rationally complacent agent presupposes the practice she is using will maximise her utility. This presupposition is evident in the question 'why is the actual different than the forecast?'. Notice that the question presupposes that the world 'ought' to have turned out like the agent had forecasted it would. For the rationally complacent agent, that the world did not turn out as it 'ought' to have had is not a reflection of the forecaster's ability but instead a result of some strange anomaly that occurred as events transpired. The rationally complacent agent's view of how the world 'ought' to turn out will not be altered by how events actually unfold.

The strategic agent, on the other hand, differs form the rationally complacent agent in at least three ways. In the first case, the strategic agent is similar to the rationally complacent agent in that the strategic agent will ask 'why *is* the actual different than the forecast?'. The strategic agent differs, however, in that she will ask this question even when actual exceeds forecast.

In the second case, the strategic agent overcomes the fallacy of planning. Recall that, according to Kahneman and Tversky, fallacies of planning are a consequence of neglecting distributional data and adopting as an internal approach to prediction, in which one focuses on the constituents of the specific problem rather than on the distribution of outcomes in similar cases. In the example above, the rationally complacent agent neglected considering that in both January and April, profits did not meet forecast due to widgets being improperly manufactured. The strategic agent, on the other hand, accounts for this distribution data. Using reconnaissance, the strategic agent can establish trends in the data collected and take steps to address similar problematic outcomes occurring in similar cases. In the example described here, the strategic agent may implement a change in practice to minimise improper

manufacturing as a cause of variability between forecast and actual. In this case, the strategic agent might institute Statistical Process Control procedures, which use ongoing (say hourly) data collection and trending analysis to allow the agent to take corrective action in the manufacturing of the widgets before the widgets are improperly made and thrown away. So, in this example, the strategic agent is able to overcome the fallacy of planning by analysing distributional data using Statistical Process Control techniques

The third difference between the rationally complacent agent and the strategic agent is perhaps the most important. As with the rationally complacent agent, the strategic agent asks, in any given month, why the actual profits are different than the forecasted profits. Additionally, however, the strategic agent asks the reverse: 'why is the forecast different than the actual?'. This distinction is an important one. Recall that the rationally complacent agent was satisfied with presupposing that the world 'ought' to have turn out as the agent forecasts it will; even on retrospection, the rationally complacent agent's view of how the world 'ought' to turn out will not be altered by how events actually unfolded. The strategic agent, however, frames the question in a way that makes no normative claims about how the world 'ought' to unfold. For the strategic agent, the difference between forecast and actual can be explained by the accuracy of the forecasting methodology itself.

Again, let us refer back to Kahneman and Tversky's experiments highlighting the fallacy of planning. Recall the investment reports that described both the favourability of the investment and the reliability of the reports. Rationally complacent agents neglected the reliability information and based their investment on the favourability of the report itself. The strategic agent, however, is concerned with the reliability of the forecast. By asking 'why is the forecast different than the actual?', the strategic agent focuses on reconnaissance aimed at uncovering the distributional data of outcomes in similar cases.

The detractor might counter that while the strategic agent may be questioning the reliability of the reports, the questioning is nonetheless reactive and, similar to the rationally complacent agent whose test determines whether to throw widgets away or not, the strategic agent can only react to the reliability information. Recall however, that the strategic agent was able to overcome reactive testing techniques by using Statistical Process Control techniques to take corrective action in manufacturing the widgets before the widgets were improperly made and thrown away. Similarly, the strategic agent can collect data on the difference between forecasted and actual monthly profits, analyse this data, use trend analysis to predict when variability in the forecasting process is likely to produce forecasts with variances outside of acceptable limits, and take pre-emptive action to minimise the occurrence of unacceptable variability in forecasts.

The conclusion here is that the use of Statistical Process Control can improve forecasts when applied to the answer to the question 'why is the forecast different than the actual?'. There is valid reasoning to support this conclusion. There is also strong empirical evidence. This technique has been used to forecast the monthly capital expenditures over a

five-year timeframe. The use of questioning and trend analysis delivered a threefold reduction in the variability between actual dollars spent to forecasted expenditure.[107]

The transcendental notion of strategy as methodological critique does provide explanatory force within the arena of practical affairs and business. In the arena of practical affairs, rational complacency is less distinct from strategic behaviour. Still, a methodological critique solves the pervasive definition-description dilemma; the use of total expected utilities of practices collapses the short-term/long-term distinction thereby eliminating any sorites problems; and perhaps most importantly, the depth-logic supporting a methodological critique reinforces the foundation upon which prescriptions for strategy formation are built.

It is interesting to note that business authors are careful in how they present the notion of rational complacency to their audiences. Authors tend to be reluctant to discuss rational complacency as such. Instead rationally complacent behaviour is couched in terms of opportunities, or *watch-outs*, when one is following a certain prescribed method for strategy formation. The delicacy with which the notion of rational complacency is handled accounts, to some extent, for the ambiguity in defining a business strategy. What is clear, however, is that a methodological critique is essential to forming, identifying and adopting successful business practices.

[107] This technique was tested empirically during my tenure as Manager for Engineering and Technical Services, for a packaged consumer goods company. It is also worth noting that it is my experience that agents are reluctant to apply Statistical Process Control techniques to human processes such as forecasting since SPC has traditionally been used to control manufacturing process equipment.

Concluding Remarks

What I have left to offer in this thesis are concluding remarks concerning the agents we have been studying. Recall that the agents under investigation are based on Rawls' and Danielson's rational agents. Of the two, Danielson relies more heavily on a distinction between rational and strategic than does Rawls. As a result, my concluding remarks address Danielson's distinction alone. Recall that Danielson identified four dispositions of interactive rational agents: 1) unconditional straightforward maximising, 2) unconditional co-operation, 3) conditional co-operation, and 4) reciprocal co-operation. Danielson's *rational thesis* distinguishes the co-operative dispositions as *a more rational means* to a rational end than a straightforward maximiser disposition. And, lastly recall that Danielson does not call on the adjective 'strategic' to single out one disposition over another.

For Danielson, each disposition is itself a strategy. Notice, however, that agents unconditionally disposed to either co-operate or defect do not change their practices. For example, regardless of what disposition an unconditional defector encounters, as indicated by its appellation, the defector will do so unconditionally. It is not only possible but also reasonable to describe the unconditional defector and the unconditional co-operator as rationally complacent since neither disposition includes engaging in reconnaissance in order to identify alternative dispositions.

The conditional and reciprocal co-operators, on the other hand, are disposed to choose among alternative dispositions: conditional co-operators do so depending on the disposition of the other agent; reciprocal co-operators do so depending on a previous interaction with the other agent. These agents engage in reconnaissance and adopt alternative practices when doing so delivers positive alternative practice opportunities. It seems reasonable, then, that agents disposed to co-operate conditionally or reciprocally should be distinguished as *strategic agents*.

The transcendental concept of a strategic agent asserts that, in agreement with both Danielson's rationality thesis and our intuitions about strategic behaviour, those agents disposed to reciprocal and conditional co-operation are *more rational* than rationally complacent agents. The depth-logic explored in this thesis gives us good reason to distinguish the adjectives 'strategic' and 'rational'. Using 'strategic' intuitively implies that the predicated agent is rational but not complacently so. The strategic agent is one that expends energy on reconnaissance and deliberation so as to less than maximise the total expected

utility of a current practice in order to discover an alternative practice offering a positive alternative practice opportunity.

The transcendental definition is an unambiguous and practical description of a strategic agent's behaviour appropriate in areas beyond that in which the adjective prolifically appears: for those wishing to ornament the vernacular, the adjective 'strategic' captures all the romance associated with throwing off the chains of rational complacency; for those requiring precision and rigour, the strategic agent is one that expends energy on methodological critique; and for the rest, that their intuitions are confirmed will very likely go unnoticed.

Bibliography

Bermúdez, José Luis. *The Paradox of Self-Consciousness.* Cambridge: The MIT Press, 1998.

Beaufre, André. *An introduction to strategy : with particular reference to problems of defense, politics, economics, and diplomacy in the nuclear age*; translated by R. H. Barry ; with a pref. by B. H. Liddell Hart. New York: Praeger, [1965]

Baylis, John; Booth, Ken; Garnett, John; Williams, Phil. *Contemporary Strategy: Theories and Policies.* London: Croom Helm, 1975.

Coleman, Andrew M. *Game theory and its applications in the social and biological sciences*, 2nd ed. Oxford [England]: Butterworth-Heinemann, c1995.

Crumley, Jack S. *Problems of Mind.* London: Mayfield Publishing, 2000.

Danielson, Peter. *Artificial Morality: Virtual Robots for Virtual Games.* London: Routledge, 1992.

Danielson, Peter. *Modeling Rationality, Morality and Evolution.* Toronto: Oxford University Press, 1998.

Darwin, Charles. *On The Origin Of Species.* New York: The Modern Library, 1936.

Dennett, Daniel. *The Intentional Stance.* Cambridge, [Mass.]: MIT Press, 1987.

Dennett, Daniel. *Consciousness Explained.* Boston: Little Brown, 1991.

Dorf, Richard C. *Modern Control Systems.* Don Mills [ON]: Addison-Wesley Publishing Company, 1986, p4.

Earle, Edward Mead. *Makers of modern strategy; military thought from Machiavelli to Hitler.* Princeton [N.J.]: Princeton University Press, [1960, c1941]

Fraser, N.M.; Hipel, K.W. *Conflict Analysis, Models and Resolutions.* New York: North-Holland Press, 1984.

Greene, Brian. *The Elegant Universe.* New York: Vintage Books, 1999.

Harman, Gilbert. *The Nature of Morality.* New York: Oxford University Press, 1977.

Hobbes, Thomas. *Leviathan, or the matter forme and power of a commonwealth, ecclesiasticall or civill.* London: Cambridge University Press, 1935.

Hume, David. *Enquiries concerning the human understanding and concerning the principles of morals.* Second edition. London: Oxford University Press, 1902.

Kahneman, Daniel; Slovic, Paul; Tversky, Amos. *Judgement under uncertainty: Heuristics and biases.* New York: Cambridge University Press, 1992.

Kelley, Harold H.; Thibaut, John W. *The social psychology of groups.* New York: Wiley, 1959.

King, J. Charles; McGilvray, James A. *Social and Political Philosophy.* New York: McGraw-Hill, 1973.

Lehrer, Keith; Wagner, Carl. *Rational Consensus in Science and Society.* Boston: D. Reidel Publishing, 1981.

Liddell Hart, B.H. *The Future of War.* New York: New York Press, 1925.

Lipsey, Richard G.; Purvis, Douglas D.; Sparks, Gordon R.; Steiner, Peter O. *Economics,* Fourth Edition. New York: Harper & Row Publishers, 1982.

Machiavelli, Nicolo. *The Prince & The Discources.* With introduction by Max Lerner. New York: The Modern Library, 1940.

Mill, John Stuart. *Utilitarianism;* edited, with an introduction, by George Sher. Indianapolis: Hackett Publishing, 1979.

Mintzberg, Henry; Ahlstrand, Bruce; Lampel, Joseph. *Strategy safari : a guided tour through the wilds of strategic management.* Toronto: Free Press, 1998.

Mithen, Steven. *The Prehistory of the Mind.* London: Thames and Hudson, 1996.

Montgomery of Alamein, Bernard Law Montgomery, Viscount. *A History of Warfare.* Cleveland: World Publishing,1968.

Morgenstern, Oskar; Von Neumann, John. *Theory of games and economic behavior.* Princeton: Princeton University Press, 1944.

Nesbitt, Richard & Ross, Lee. *Human Inference: Strategies and Shortcomings of Social Judgement.* Englewood Cliffs, [N.J.]: Prentice-Hall Inc, 1980.

Paret, Peter. *Clausewitz and the State.* New Jersey: Princeton University Press, 1976.

Paret, Peter; Craig, Gordon A.; Gilbert, Felix *Makers of modern strategy: from Machiavelli to the nuclear age.* Princeton[N.J.]: Princeton University Press, 1986.

Pascal, Blaise. *Pensées;* translated with an introduction by A.J. Krailsheimer. London: Penguin Books, 1995.

Poff, Deborah C.; Waluchow, Wilfred J. *Business Ethics in Canada.* Scarborough: Prentice-Hall Canada Inc., 1991.

Pojman, Louis P. *The Theory of Knowledge.* Belmont [Ca.]: Wadsworth Publishing, 1998.

Quine, Willard von Orman. *Word and Object.* Cambridge: Technology Press of the Massachusetts Institute of Technology, 1960.

Rapoport, A.; Guyer, M.; Gordon, D. *Two-person game theory; the essential ideas.* Ann Arbor: University of Michigan Press, 1966.

Rapoport, A.; Guyer, M.; Gordon, D. *The 2x2 Game.,* Ann Arbor: U. Michigan Press, 1976.

Rawls, John. *A Theory of Justice.* Cambridge [Mass.]: Belknap Press of Harvard University Press, 1971.

Rescher, Nicholas. *Induction: an essay on the justification of inductive reasoning.* Pittsburgh [Pa.]: University of Pittsburgh Press, 1980.

Rousseau, Jean-Jacques. *Rousseau's Social Contract*; translated and edited by H.M. Tozer. New York: Charles Scribner's Sons, 1895.

Russell, Bertrand. *The Conquest Of Happiness*. New York: Liveright Publishing, 1930.

Sidowski, Wyckoff, and Tabory. "The Influence of Punishment and Reinforcement in a Minimal Social Situation", *The Journal of Abnormal and Social Psychology*, 52, pp 115-119.

Sidowski, "Reward and Punishment in Minimal Social Situation", *Journal of Experimental Psychology 54*, pp 318-326.

Simon, Herbert A. *Administrative Behaviour*. New York: The Free Press, 1945.

Starr, Chester G. *A History of The Ancient World*. New York: Oxford University Press, 1991.

Thorndike, E.L. *Animal Intelligence*. New York: MacMillan, 1911.

Von Clauswitz, Carl. *On War*; edited and translated by Michael Howard and Peter Paret. Princeton [N.J.]: Princeton University Press, 1976.