Evolving Ethics: The New Science of Good and Evil

Steven Mascaro
Kevin B. Korb
Ann E. Nicholson
Owen Woodberry

Published in the UK by Imprint Academic
PO Box 200, Exeter EX5 5YX, UK

Published in the USA by Imprint Academic
Philosophy Documentation Center
PO Box 7147, Charlottesville, VA 22906-7147, USA

ISBN 9 781845 402068

A CIP catalogue record for this book is available from the
British Library and US Library of Congress

To Charles Darwin, Alan Turing,
and their progeny

About the Authors

Steven Mascaro, Ph.D., earned his doctorate in Computer Science at Monash University (2009) with the thesis *Abortion, Rape and Suicide*. He is a private consultant working with Bayesian network technology and web applications and continuing evolutionary and ethical simulation research as time permits.

Kevin B. Korb, Ph.D., earned his doctorate in the philosophy of science at Indiana University (1992) working on the philosophical foundations for the automation of Bayesian reasoning. Since then he has lectured at Monash University in Computer Science, combining his interests in philosophy of science and artificial intelligence in work on understanding and automating inductive inference, the use of MML in learning causal theories, artificial evolution of cognitive and social behavior and modeling Bayesian and human reasoning in the automation of argumentation.

Ann E. Nicholson, D.Phil., did her undergraduate computer science studies at the University of Melbourne and her doctorate in the robotics research group at Oxford University (1992). She spent two years at Brown University as a post-doctoral research fellow before taking up a lecturing position at Monash University in Computer Science. In addition to her interest in ALife simulations for investigating evolutionary ethics, her research spans many areas of Artificial Intelligence, including probabilistic reasoning, Bayesian networks, planning, user modeling and knowledge engineering.

Owen Woodberry is completing his Ph.D. at Monash University in Computer Science, exploring the use of evolutionary simulation to shed light

on evolution theory, focusing on the units of evolutionary selection and the evolution of aging. He also has interests in Artificial Intelligence, Knowledge Engineering Bayesian Networks, Environmental Science and Teaching. In addition to his academic duties, he works as a consultant for the company Bayesian Intelligence, which specializes in Bayesian Networks.

Contents

List of Figures

List of Tables

Preface

The seeds of this book were planted at the turn of the millenium, when a computer science student curious about the philosophical potential of computers encountered two like-minded lecturers. With their encouragement, this student embarked on a thesis exploring the potential of evolutionary and ethical simulation. These three were joined a few short years later by another student, equally curious about what simulations could say about foundational issues in evolutionary theory. After a considerable virtual journey led to the first author's successful PhD, Mark Bedau suggested the thesis might serve as the basis for an interesting book. For this suggestion, all four authors are very grateful. This book is the culmination of that suggestion, bringing together various interrelated strands of research pursued collectively by the authors. It combines the major part of the work of two PhDs, but blends and leavens the work leading, we hope, to a nourishing result.

Early on, philosophy described all of our attempts to advance the state of human knowledge. Physics commingled with biology, medicine, religion, politics and logic as well as matters now traditionally of philosophy, such as metaphysics, ethics, epistemology and esthetics. While Plato and others drew a boundary around natural philosophy, the distinction had never been methodological. After long ages, this began to change with some natural philosophers of the Renaissance striking out into new territory, enchanted with the methods of experiment. Physics left first, followed by others such as medicine and economics and, most recently, psychology. Other fields, for which experiment seemed impractical or pointless, remained — ethics among them. Our hope here is, first, to show that the experimental method is of as much use to ethics as it is physics, but, second, and more importantly, to show that simulation can act as a bridge between the analytical tradition of philosophy and the experimental tradition of science.

A question sometimes asked of agent-based modelers is, Why not use

game theory rather than simulation? The question seems motivated by a persistent belief that simulation is the inferior option — that we drag it out only for pragmatic reasons, but would happily return into the arms of game theory if at all possible. The closed-form equations that game theory produces are simpler, more convenient and more certain than what we learn from simulations, so perhaps it is true that simulation should be a last resort. But, if so, it is a "last resort" with a vastly wider domain of applicability, as demonstrated clearly by the simulations in this book. To understate the matter, all living creatures — humans, animals, plants — are heterogeneous, whether across species, within species or even with a single kin group. Game theory does not even try to capture this heterogeneity — and if it tried, the result would surely end up indistinguishable from agent-based simulation.

Simulation research is still quite young and the practical limits of computing — power, memory and software methods — will decide how well our programs model the physical systems of interest to us for many years to come. The simulations in this book certainly reflect this, trading off detail for practical performance. We simulate as we are able; but it is clear that the most important, insightful and even groundbreaking simulations are ahead of us, and not behind us.

This work is aimed at an assortment of readers: philosophers, evolutionary biologists, economists, sociologists, psychologists, computer scientists, simulationists and the generally curious. The book is written to allow readers with different backgrounds and interests to dive in where they wish. A glossary at the back (with first occurrences of entries bolded in the text) may also help with this. Most will want to begin with the discussion of Chapter 1. Anyone wanting a review of ethical systems, and especially of utilitarianism, should look to Chapter 2; common arguments against utilitarianism are given in that chapter, along with our rebuttals. Chapter 3 argues the case that there is little difference between computer simulations and experiments, either methodologically or epistemologically. Chapters 4 through 6 then present our experimental work. A brief history of Artificial Life (ALife) is drawn at the beginning of Chapter 4, followed by a description of the common design elements of our otherwise variegated simulations. Chapter 5 applies evolutionary artificial life simulation techniques to the investigation of some fundamental issues of evolutionary theory, including the levels of selection debate. Our analysis in this chapter highlights the relation between group and kin selection, usually held to be antagonistic, but which we find to be supportive. Chapter 6 turns to a small selection

of ethical and unethical behaviors. While investigations of cooperation and altruism are very common, as we discuss, we also explore two behaviours that have not been deeply explored via simulation: rape and abortion.

Many of the curlier details of the simulations have been glossed over for this presentation. The reader wanting more information is encouraged to explore our papers and technical reports on the topic, which are available from our website:

`http://www.csse.monash.edu.au/evethics`

Also available from the website is a simplified demonstration simulation in Netlogo for each of the major simulations in the book. We hope that you will be inspired to replicate, critique and expand the scope of these simulations, and, ultimately, to contribute to the growing use of simulation and computers in philosophy, science and ethics.

Acknowledgments We thank Alan Dorin for the cover image of our home world. Some of the ideas appearing here were tested at the Center for Logic and Philosophy of Science, Tilburg University, the Institut d'Histoire et de Philosophie des Sciences et des Techniques, University of Paris, and the 4th Australian Conference on Artificial Life, Melbourne, 2009; we are grateful for those opportunities. The second author is grateful to Volker Grimm and Steven Railsback for the chance to participate in their 2008 Summer Individual-based Modeling School in Bad Schandau. Those who assisted with reviewing, discovering errors or providing other useful comments include Mark Bedau, John Bigelow, Nick Bostrom, Allie Ford, Roman Frigg, Stephan Hartmann, Erik Nyberg, Julian Reiss, Geoff Webb. We thank our editor, Anthony Freeman, for his patience.

Chapter 1

A Science of Ethics

Throughout history philosophers have studied and debated ethical questions without the help of real-world experiments. While ethical experiments could answer many important questions, most such experiments would themselves clearly be unethical. Some empirical assistance to ethical theorizing has been found in the recent past. Since Charles Darwin, many have found the story of the evolution of cooperative and social behavior so compelling that they have claimed to find justifications for ethical behavior within that evolutionary history. This is the program of **evolutionary ethics**,[1] advocated by Julian Huxley (1927), E. O. Wilson (1978) and many others. More empirical assistance comes from **evolutionary psychology**, which attempts to apply the concepts of evolutionary biology, and the circumstances of evolution's activity, to solving problems about current social behavior. The direct application of the facts of evolution to justifying ethical norms, however, can only get anywhere by way of the **naturalistic fallacy** of inferring ought from is. If we are to learn anything about ethics from evolution, we need a less direct route.

The very first substantial application of computer simulation was John von Neumann's simulation of nuclear reactions for the design of the hydrogen bomb, using the very first computer, the ENIAC (Goldstine, 1993). Since then, every scientific discipline — from Astronomy to Zoology — has adopted computer simulation techniques to explore beyond the limits imposed by time, money, and social and ethical constraints in a new era of scientific experimentation (see, e.g., Humphreys, 2004, Racynski and Bargiela, 2007, Frigg and Reiss, 2009). We argue that these new experimen-

[1] Boldface for phrases or their cognates in ordinary text indicates a corresponding entry in the glossary.

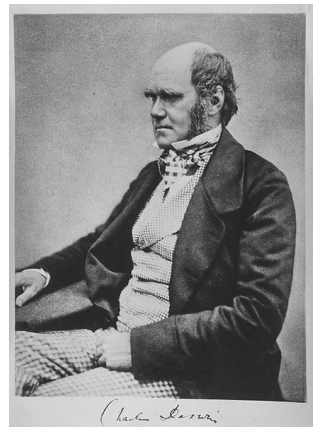


Figure 1.1: Charles Darwin.

talists are on the right track — that these computer simulation experiments have the same epistemological standing as traditional physical experiments.

In addition to experimental simulation launching new approaches to the study of the traditional experimental sciences, computer simulation has overtaken sciences which have been heavily dependent upon non-experimental techniques, such as economics, epidemiology, and sociology, offering experimental options from the 1990s. These social simulationists have drawn upon research on complex systems and the evolution of social behavior (e.g., Axelrod, 1984) to find new and fruitful applications of simulation.

We extend the application of computer simulation further. In particular, we draw together all of the above ideas into a new experimental ethics. The simulation of ethical behavior allows us to examine that behavior in ways never previously available. We can systematically alter the conditions within which ethical actions occur, the available behaviors themselves, and their impact on societies. By evolving these behaviors we can establish the evolutionary scenarios which do, and do not, support their establishment — not just cooperative behaviors, but altruistic behaviors, and selfish and other unethical behaviors as well. We are thus able to map out the evolutionary limits and possibilities for ethical action. By examining the distribution of **utilities** in the population we can also consider the moral status of contentious behaviors from a **consequentialist** perspective. All of this opens up an entirely new approach to the scientific study of ethics.

The potential for philosophical studies goes beyond ethics. There is no reason our computer simulation methods cannot be applied also to, for

example, social epistemology and the philosophy of scientific method, as has already been done in preliminary forms (e.g., Hegselmann and Krause, 2006 and Gooding and Addis, 2008). We look forward to these evolving. And there is no reason questions of traditional epistemology cannot be attacked similarly, looking at the evolutionary conditions for different criteria of belief and justification to make sense. Indeed, we know of no reason why there should not be an experimental assault upon many of the traditionally a priori enterprises, investigating mind and matter, meaning and language, social and individual decision-making.

While we have a wider agenda of advancing computer simulation generally, our aims specific to this book are to introduce a new experimental science — the evolutionary **social simulation** of ethics, to justify its use as an empirical method, and to describe and illustrate its advantages. Making sense of our evolutionary simulations requires us also to make sense of some basic issues about evolution, especially why evolution works. If evolution depended solely upon natural selection at the individual level, as it would on a traditional interpretation, we would be at a loss to explain our simulation results. More sophisticated interpretations of evolution are more accommodating. So, in this introduction we shall review the basic elements of our evolutionary stories: ethics; theories of evolution by natural selection; computer simulation and especially **artificial life (ALife)** simulation; and how these are put together in the ALife simulation of the evolution of ethical behavior.

1.1 Ethics

Ethics is the study of the "shoulds" or "oughts" of human behavior. The terms 'morality' and 'descriptive ethics' are used to describe the study of a group's principles of behavior (Singer, 1994). That is, **descriptive ethics** is the study of what people in fact *believe* ought to be done. We could as justifiably locate its study within anthropology as within philosophy. By contrast, we may be interested not so much in *beliefs* about what is right, but in *what is right*. That is, we may be interested in **normative ethics** or moral philosophy (or simply ethics): the systematic study of *how we ought to act*. Normative ethics is usually investigated from within a particular ethical system, such as **virtue ethics**, **deontological ethics** and **utilitarianism**, to name the major players (which we will discuss in the next chapter). **Metaethics** attempts to arbitrate between these systems, aiming to supply a theory of ethical study which might give us reasons for or against the

various ethical systems.

The relation between descriptive and normative ethics has been, and remains, fraught. A naturalistic ethics, as described by G.E. Moore (1903), holds that good is defined by reference to natural objects — that is, it defines *ought* in terms of *is*, effectively identifying normative with descriptive ethics. However, since Hume's (1739) argument that ought-statements cannot be derived from is-statements, philosophers have tended to steer clear of naturalistic normative theories. Moore dubbed the inference of *ought* from *is* the 'naturalistic fallacy' and accused naturalistic ethics of committing it. The relation between descriptive and normative ethics, however, is not settled by any of this. If we cannot infer the one from the other, we can certainly *inform* the one by means of the other.

There are at least two grounds for an informative relation between the descriptive and the normative. First, it is an undisputed principle of ethics that we cannot be obligated to do what we are incapable of doing. And what we are capable of doing is contingent upon many things: our physical natures, our environment and our cultures and beliefs. So, an investigation of these many features descriptive of us and our surrounds is actually essential for a proper understanding of the normative. Similar considerations have given rise to much of the interest in recent decades in "naturalized" epistemology and "naturalized" philosophy of science, both of which aim to identify normative standards for acquiring knowledge (e.g., Quine, 1969, Giere, 1985). Likewise, evolutionary ethics seeks to base ethics somehow on our evolutionary heritage (e.g., Huxley, 1927), and, while evolutionary ethics has generally been accused of committing the naturalistic fallacy, a viable alternative is to use it to help normative studies determine the boundaries of evolutionarily possible ethical behavior.

The second support for a relation between the descriptive and the normative is the metaethical process of **reflective equilibrium** attributed to Nelson Goodman (1956) and John Rawls (1972), but already to be found in a nascent form in Aristotle's *Nichomachean Ethics*. Reflective equilibrium, in brief, treats a normative theory analogously to scientific theories, with the direct empirical evidence being a core set of normative judgments common to a population. The normative theory should provide potential explanations of its evidence, our intuitions, which have the usual virtues of scientific explanations, such as being unifying (consilient), generalizing, simple, and compatible with related scientific theories. The normative theory is thus required to get considered judgments of right and wrong the same way as those judgments in the population which are (nearly) universal, but is free

to carve up the remaining, unclear judgments as it will ("spoils to the victor", as Lewis, 1986, p. 203, puts it). The freedom to theorize, however, is far more constrained than that, since, as with any scientific theory, it is also obliged to be consistent with the full range of accepted scientific theories today. An ethical theory that is in reflective equilibrium today must do justice to our evolutionary ancestry, our best theories of cognition and social behavior, and much else besides.[2] These kinds of considerations are responsible in part for recent interest in a kind of experimental moral philosophy, which has aimed at identifying core moral intuitions in ways more systematic than traditional philosophy has done — that is, by substituting empirical inquiry into moral judgments for armchair speculation by philosophers (e.g., Nichols and Knobe, 2008). We applaud those efforts, and we offer another way to get out of the armchair to do empirical ethics, namely by moving over to the computer.

In order for computer simulation studies to be informative about ethics, we must adopt a point of view which allows us to measure the outcomes. Utilities are the natural currency for measuring ethical outcomes. Utilities also support a very natural ethical system, namely utilitarianism, the thesis that *what action is best collectively* is *what action is best*. Utilitarianism is, in fact, the only ethical system which *allows* us to measure the outcomes of computer simulations and judge them as better or worse.[3] And so we adopt utilitarianism both on the grounds that it is unavoidable in studies like these and because it provides a plausible candidate system for achieving the kind of reflective equilibrium theory we mentioned above. Beyond that, however, utilities provide a general-purpose apparatus for measuring the impact of actions on both individuals and societies, and they allow us to investigate the evolution of utility with potential for informing us about the evolution of action and agency (Chapter **??**).

[2]This approach is also related to Quine's *Web of Belief* (Quine and Ullian, 1978). Much more might be said about the reflective equilibrium method. Certainly some qualifications are required. For example, the universality of opinion for the "base" cases depends upon the population of interest and for the method to have any value this must at least exclude manifestly deviant people, such as the seriously mentally ill. Furthermore, the base cases, even with an agreed population and universal agreement, cannot be treated as inviolate, any more than Karl Popper's "basic sentences" in observational science were inviolate, which is already suggested by our reference to Quine. But examination of these subtleties would take us far afield. Here we are satisfied to accept *something like* what we've outlined as a promising, indeed the best available, approach to generating a normative ethical theory.

[3]According to Stephan Hartmann, this was a point made some time ago by Patrick Suppes as well, in conversation.

1.2 Evolution

Evolution theory has become a keystone theory in science, supporting much of what we understand about life from the lowest levels of molecular biology to abstractions about human behavior, such as **altruism**, love and purpose. Evolution is also fundamental to our uses of simulation to investigate questions about biology and behavior. In order to understand this usage, we first introduce the *received view* of evolution — the standard interpretation of the "neosynthesis" of Mendelian genetics and Darwinian theory promulgated in the mid-20th century by, for example, Huxley (1942), Dobzhansky (1951) and Mayr (1976) — and its close relative, the "**gene**'s eye view" of evolution, largely based on Hamilton (1964) and developed and popularized by Williams (1966) and Dawkins (1976).

1.2.1 The Received View

Evolution has three necessary ingredients, which, when combined properly in a reproductive population are also jointly sufficient — i.e., they inevitably get evolution going:

Necessitata of evolution:

1. ***Heritability.*** Phenotypic traits must have a tendency to be passed on to the next generation.

2. *Selection.* Some of those phenotypic traits must have *different* tendencies to be passed on to the next generation.

3. *Variation.* The traits passed on to subsequent generations must not always be perfect copies; they must vary.

The inevitability of evolution is not a logical inevitability. Evolution is a stochastic process, subject to a great deal of randomness, for example, in mutations and other genetic modifications during reproduction and also in accidents that suddenly remove individuals from the evolutionary process through death. However, evolution is a practical inevitability: out of, say, one hundred thousand simulations of evolution incorporating the above three features in well understood ways — ways we describe in this book — we would typically find zero of them showing no evolutionary process. Creationists might regard the results of computer simulation to be no more than a "proof of concept", demonstrating only that evolution in biology is

possible. However, the manifest applicability of the three necessitata to real biology leaves no doubt as to the reality of biological evolution.

To be sure, in a full and proper account of evolution much that is implicit in our three necessary conditions would need to be enlarged upon. For an example, the variation of condition 3 must lie within some intermediate range: too little added variation can result in selective pressures turning an initially varied population into virtual clones of each other; too much variation will kill everything off. For another example, the abiotic environment in which evolution unfolds must be neither too chaotic nor too uniform. A completely unchanging and uniform environment will not provide differential selection pressures to push evolution along. A radically changing environment, on the other hand, will not allow selected traits to be adaptive in successive environments, making cumulative **adaptations** impossible. Even more radically changing environments won't allow traits to be inherited, because they won't support continued life at all. We cannot give a full accounting here of the conditions for evolution or their interpretation; instead, we refer the reader to the neosynthetic texts cited above or to the excellent introduction to the philosophy of biology, *Sex and Death*, by Sterelny and Griffiths (1999).

"Survival of the fittest" is a slogan introduced by Herbert Spencer, meant to sum up the import of Darwin's (1859) theory (Spencer, 1864). If **fitness** is understood as 'being selected for', then, as many have pointed out, the slogan degenerates into a boring tautology. However, a more plausible and useful interpretation of **fitness** is as the number of expected descendants of an organism. Expectations are not always fulfilled, so no tautology remains in Spencer's slogan. Regardless, the slogan is suspect. It has been used to emphasize a very narrow suite of ideas about fitness, ideas of strength, speed and the physical fight for survival. No doubt, during the many millions of years of evolution, especially the recent ones, fitness has often been enhanced by improvements in communication, **cooperation**, symbiotic coadaptations, immune systems and also plumage, for those of us who must dress for success. Bloody fights to the death are relatively less common.

Computing an expected value requires having a probability distribution, and, by introducing probabilities of descendants into fitness, we allow for the tendencies (and uncertainties) of selection pressures to manifest themselves. Traits which raise the probability of reproduction will have higher fitness than those which do not. *Survival* is not actually of the essence here — a praying mantis that is likely to be eaten immediately after copulation

may well have high fitness — so that is another way in which Spencer's slogan was misguided. And a final, important point about fitness is that what matters is *relative* or *differential* fitness and not absolute fitness. A spider with one million expected descendants may well be extremely unfit; it will be unfit if its conspecific spiders have ten million expected descendants. In generations to come, the traits of these latter spiders will completely swamp those that are making the first spider unfit. It is only the ratios of fitness *within* a **species** that determines which traits will win out.

In the received view the carrier of heritable traits is the genotype, chromosomal DNA. Differential selection pressure at the phenotypic level indirectly puts differential pressure on the genotypic level, with the result that the gene pool alters over time, leading generally to better adapted organisms.

This much is common to modern interpretations of evolutionary biology, even though it has become clear that the received view is too narrow. (For an important example, epigenetic inheritances outside of nuclear DNA are by now well established; e.g., Jablonka and Szathmáry, 1995.) The received view also has some commitments particular to it. Of special interest here is the idea that selection is **individual selection** — that individual organisms are the **units of selection**. Every modern theory of evolution has it that selection pressures originate with the interaction of the phenotypic traits of individuals and their environments, but it doesn't necessarily follow that the best description of what is being selected for and against *is* the individual. In fact, that commitment by the received view leaves it at a considerable disadvantage to its relative, the "selfish gene" view of Dawkins and others. Individual selection cannot explain the evolution of altruism.

1.2.2 The Gene's View

Biological altruism is any act which damages the individual's fitness to the benefit of that of others. There are many examples in nature. The actions of sentinels, in howler monkeys for example, typically put them at greater risk of predation while improving the chances of their peers to escape unharmed (Wilson, 2005). A somewhat strained attempt to explain away sentinel behavior might put it down to **reciprocal altruism**, if a round-robin or random schedule for acting as sentinel is applied. And reciprocal altruism is clearly not altruism at all, but a simple exchange of services conducted over time. But there are also more extreme forms of altruism offering no possibility of reciprocity, such as matriphagy in the *Chiracanthium Japonicum* spider, with the mother spider giving her life to foster her offspring's development

(Toyama, 2001). Altruistic behavior occurs across a wide range of species and clearly has some kind of adaptive value, but what kind? It is clearly not adaptive for the *individual*, since the individual fitness is what is being sacrificed. If individual fitness exhausts the fitness story, then altruistic behaviors which reduce it must over time simply be erased from the gene pool, so altruism can only ever evolve away. For the received view, biological altruism is just an anomaly.

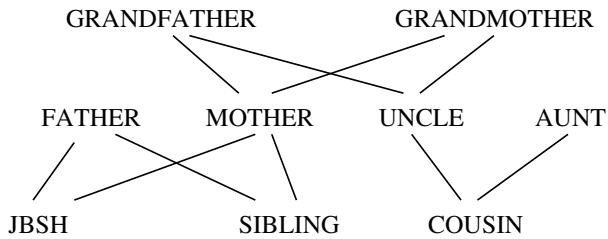
Hamilton's (1964) **kin selection** theory, expanding fitness to **inclusive fitness**, provides a clear and persuasive explanation of the evolution of altruism. Inclusive fitness adds together the individual fitness effects (f_i) of an **allele** (a), weighted by Wright's (1922) coefficient of relatedness to the allele's owner (r_i):

$$F(a) = \sum_i r_i f_i$$

Individual fitness requires i to range over the actor alone, when, of course, $r_i = 1$. Inclusive fitness allows i to range over a whole population, with r_i reflecting the probability of alleles being shared. Thus, inclusive fitness measures the expected number of copies of a gene within the gene pool over future generations. So, genes with positive inclusive fitness may be expected to spread and become established in a population, and so also altruistic actions, insofar as they are genetically predisposed. As J.B.S. Haldane famously quipped, "Would I lay down my life to save my brother? No, but I would to save two brothers or eight cousins" (see Figure 1.2).[4] In Chapter **??** we will illustrate the evolution of altruistic suicide through kin selection.

Richard Dawkins (1976) has taken this idea of fitness for genes (alleles) and built up an entire *Weltanschauung* out of it, the world as seen by a "selfish gene". From the gene's perspective, organisms are simply means to an end, that end being to replicate oneself. Genes are in control, and bodies are like vehicles that they drive around. Of course, it isn't really about

[4]If we were to take this story exactly literally, apparently we would end up being more closely related to monkeys than our siblings, since reportedly we share 98.8% of our DNA sequence with the former (Chen et al., 2006) and supposedly we share 50% with the latter! The explanation, of course, is that the coefficient of relatedness is the probability of the two alleles being direct copies of an allele of a common ancestor in the system portrayed (i.e., with "root" nodes (e.g., AUNT) counterfactually assumed to have no common ancestry with JBSH, etc.). Since we share 98.8% of our alleles with monkeys, and 99.9% with randomly selected humans, most alleles will turn out to be the same between two humans (primates) regardless of the coefficient of relatedness between them. Nevertheless, these coefficients (at least when the system is somewhat expanded) are exactly what are needed to compute the tendencies of fitter alleles to replace less fit alleles in population genetics.



Each link weight represents 1/2 of the genes being shared
Relatedness = sum of products of link weights over all paths

$$R(\text{JBSH, SIBLING}) = (1/2)^2 + (1/2)^2 = 1/2$$

$$R(\text{JBSH, COUSIN}) = (1/2)^4 + (1/2)^4 = 1/8$$

Figure 1.2: Coefficients of relatedness.

selfish genes fighting other selfish genes for the opportunity to replicate; organisms are built by a grand assemblage of genes who all must cooperate to get the organism developed properly and so for any of them to be replicated. Other, perhaps less sensational, versions of gene selectionism are advocated by Williams (1966) and Hull (1988).[5]

Aside from its amusing new imagery, Dawkins' interpretation does not appear to be very far removed from the received view itself. It importantly locates fitness at the level of genes, which, as we've pointed out, can only be advantageous, allowing for an explanation of the evolution of altruism without losing anything, since inclusive fitness subsumes individual fitness. What selfish gene theory does not do is imply genetic determinism: although it emphasizes the role of genes in determining development, neither Dawkins nor any other advocate would claim that the genetic endowment simply overrides the developmental environment; genes clearly work within some (benign) environment to guide development. In this, Dawkins' and the received view are one. They both have been blind to non-genetic in-

[5]Richard Dawkins and the philosopher Daniel Dennett found this point of view so gripping that they have been driven to extend it to *ideas*, or *memes* as they would have it: memes drive humans around and make them put together compelling arguments, stories and songs so as to reproduce themselves in other people's heads (Dawkins, 1976, Dennett, 1995). It is clearly correct that ideas, beliefs, etc., as realized in human culture, possess the three necessary conditions to evolve: they are passed on from generation to generation (heritability); some are preferred to others (selection); and they are altered while being passed on (variation). Nevertheless, the added explanatory value of attributing the spread of an idea to positive selection pressure rather than an old-fashion appeal to, say, its persuasive power seems to be small, so perhaps memes about memes won't replicate much farther.

heritances, such as epigenetic inheritances (Jablonka and Szathmáry, 1995) and ecological inheritances (Odling-Smee et al., 2003).[6]

And both views have been hostile toward the possibility of *groups* of organisms being somehow a focus for selection pressures.[7] However, **group selection**, the differential ability of *groups of organisms*, whether tribes, communities or species, to reproduce themselves, is another potential explanation of the evolution of biological altruism. Ignoring kin selection pressures for the moment, we can agree with Maynard Smith (1976) that the individual fitness of altruists will be lower than that of selfish organisms, so that altruists will tend to die out in groups, and yet groups with altruists may well have greater expected longevities and so have greater opportunities to found new groups than entirely selfish groups. It follows that, under the right circumstances, the group selection pressure for altruists will outweigh the individual selection pressure against altruists, and so altruism will be evolutionarily stable. Maynard Smith (1976) himself, and most biologists, thought that the required circumstances are unlikely to be realized in nature. We will argue that they are mistaken and, in particular, that kin selection will often bring those very circumstances about. We will illustrate the operation of group selection, and its interaction with simultaneous kin selection, in Chapter **??**.

The idea that gene selection and group selection are compatible and, indeed, simultaneously active has become respectable in recent decades (in **multilevel selection** theory; e.g., Wilson, 1997). The competition for explanatory exclusivity — between genes and groups, and between DNA and epigenetic inheritance — is an ill-conceived struggle; by dropping claims to exclusivity, all of these become compatible with each other.

1.3 Simulation

Our technique for experimental investigation here is computer simulation. So, what are computer simulations? The PC game "The Sims" is a well known example: it simulates the life and times of various characters who worry about getting jobs and cleaning toilets. Aircraft and naval piloting simulators simulate conditions involved in normal and abnormal maneuvers

[6]Contrary to the extremism of Fodor and Piattelli-Palmarini (2010), while the received view has ignored these sources of heritability, there is nothing incompatible with the main ideas of the received view and these additional sources of inheritance.

[7]We note that W. D. Hamilton has not shared this hostility; in fact, Hamilton (1975) gives an excellent analysis of the conditions for group selection.

of aircraft and ships. And "Second Life" simulates a large range of human and non-human activities. However, none of these are simulations in the sense we mean here: in all of them a human user plays an essential and central role, which is not to the point in simulation science.

Simulation science is about expanding our scientific knowledge, rather than entertainment or instruction. The simulations of interest to us here are those in which the *entire* simulation occurs within a computer, as a **computer process**, imitating physical or social processes in the wider world, whether chemical, astrophysical, evolutionary or ethical. These scientific simulations are commonly reported by philosophers of simulation to be nothing more than substitutes for analytic methods of solving integrations and partial differential equations, when the latter turn out to be too hard for us to do (e.g., Humphreys, 1991, Winsberg, 2001, Frigg and Reiss, 2009). In fact, however, simulations go far beyond that numerical computing role (Korb and Mascaro, 2009). Given a simulation model instantiating some theory, we can, of course, determine what the theory predicts for the future, by setting the simulation parameters to values representing the present and running it. This substitutes for deductions of predictions when deductions are hard. But we can also do many other things. We can be, and often are, surprised by what our simulations turn up: no one can envision the full range of consequences of a theory for any situation, and so no one can make predictions for all of them. In other words, beyond predicting the future, we can *explore* the future, insofar as our guiding theory is the right one. And, whether or not our theory is correct, we can explore the theory-space by altering the structure and dynamics of the simulation and observing how things would then unfold. We can also perform sensitivity analysis to determine how responsive effect variables are to initial conditions or to dynamic parameters. Numerical computation is such a narrow part of the life of a simulationist one must wonder how much actual experience these philosophers have had with computer simulation! In addition to all of those considerations, artificial life simulation in particular doesn't even begin to fit the numerical computation model, because it doesn't even begin with equations to be solved.

1.3.1 Artificial Life Simulation

Artificial life (ALife) is the imitation of living processes using computer simulation (or other technology). The term 'Artificial Life', due to Chris Langton, is just two decades old, but the idea is much older. Arguably, the first serious ALife research began in 1911, when Leduc experimented

with colloidal solutions to emulate metabolic functions, mitosis, and other activities associated with life (Keller, 2003). Later, von Neumann (1951), inspired by Stanislaw Ulam, invented **cellular automata** and designed the first (somewhat complicated) self-replicating cellular automaton. Self-replicating cellular automata were later made famous — and much simpler — using Conway's "Game of Life" (Berlekamp et al., 1982).

In addition to individuals, many fields have also contributed to ALife. Cybernetics passed on the study of self-organization and complex systems (Wiener, 1948). Evolutionary **algorithms** were developed as search and optimization techniques from the 1950s and 60s (Selfridge, 1959, Holland, 1975, Fogel et al., 1966, Schwefel, 1981), leading to a highly active community of evolutionary ALife researchers, including us. Finally, artificial intelligence contributed (and continues to contribute) many systems of interest in ALife, such as **Bayesian** and neural networks, **decision trees**, planning systems and more.

While dubbing the field 'Artificial Life', Langton (1989) asserted that whereas biology focuses on life as it is, ALife aims to increase our understanding of life both as it is and as it could be. That definition leaves the domain of ALife not so much open-ended as unclear. No one knows what life could or could not be. But actually, ALife is not so much a *domain* of study as it is a *methodology*, a set of techniques. Indeed, most of the researchers actively using these techniques don't call themselves ALife researchers at all: in ecology they call themselves **Individual-Based Modelers (IBMers)** (Grimm and Railsback, 2005) and in the social sciences they call themselves **Agent-Based Modelers (ABMers)** (Epstein and Axtell, 1996). Quite likely, this coyness has to do with disciplinary obligations: adopting the term "ALife" would make a social scientist or ecologist sound too much like a computer scientist. Regardless, IBMers and ABMers are all applying ALife simulation to investigate problems within their various sciences. Despite the wide adoption of ALife methods, the full range of potential applications has only begun to be explored; our exploration of ethics strikes out in a new direction.

Rather than beginning with equations to analyse, ALife researchers typically begin with complex behaviors to explain. For example, how do trout make the tradeoff between occupying rich feeding grounds and avoiding predation risk (Railsback et al., 1999)? Or, what is the relation between average household size and the spread of influenza in southern California (Stroud et al., 2007)? Or, again, why do the different sexes most commonly invest different amounts of time, matter and energy in their offspring

(Chapter **??**)? In order to answer these kinds of questions, ALife modelers attempt to build simulations which show these behaviors as **emergent properties** of the system. A property is *not* emergent if it is explicitly programmed into the simulation. The underlying concern here is epistemological: we cannot learn anything from our simulations if our results have been cooked. For example, suppose we are interested in investigating the **Lotka-Volterra model** of predator-prey populations, wanting to uncover conditions under which it is, and is not, realized (Volterra, 1931). It will obviously be pointless to simply program two variables, `predator-numbers` and `prey-numbers`, and assign them the values prescribed by the Lotka-Volterra equations. The only thing we could possibly learn is whether we know how to write such a **computer program**. For an ALife model, what we shall do instead is design a program which represents some geographical region, some prey animals and some predator animals, food for the prey, and some attack and defense options for the animals. The level of detail with which we model the geography, flora and fauna will depend upon the problems of interest to us, and especially whether we are interested in very generic population dynamics or those specific to some species or habitat. However, variables reporting the population levels will have to be epiphenomenal: they must have no causal role in the simulation at all, but instead be restricted in use to generating summary statistics for output. Whether the Lotka-Volterra equations are satisfied or not will then depend upon the dynamics and conditions under which the simulation runs; it can only *emerge* as a long-range result of what we have programmed.

Many have held properties to be emergent only if they are *surprising* to the researcher. While many emergent properties certainly are surprising, this is no necessary condition. Many other emergent properties are fully expected to show up: if we have programmed our simulation right, and set initial conditions within some reasonable range of possibilities, it would simply be amazing if the Lotka-Volterra population cycles did *not* show up! Rather than grounding emergence in a subjective emotional response, a better approach is to think of properties and behaviors of the simulation as existing in layers. At the bottom layer is the program, the computer instructions we have explicitly coded. At one or more levels above that are regular properties and relations between states of the running process. They are *above* the program at least in the sense of being **supervenient** upon the program: the program realizes these properties, however, any number of very different programs could do the same (hence, **multiple realizability** is characteristic of supervenience). This account is modeled on supervenience

theory in the philosophy of mind, which avoids the troubles of **reductionism** — the attempt to find bridging laws *identifying* mind and brain — by pointing out that mental properties are multiply realizable (Kim, 1993).

This yields *bottom-up computer simulation*, or "BUCS" for short (Goldspink, 2002). Epstein and Axtell (1996) suggest that BUCS is particularly valuable because it forces the researcher to look for simple explanations (in the form of simulations) for complex systems. Aside from the above epistemological concerns, another motivation is the perceived failure of top-down methods in artificial intelligence to produce a general intelligence. The hope has been that BUCS, in contrast, would be able to recreate the wide variety of living processes that exist around us, and to create new ones that do not. Whether it has succeeded is debatable (cf. Bedau, 2006); however, it has been an unquestionably fruitful approach to research.

BUCS weds magnificently with evolution: the detailed outcomes of evolution are notoriously unpredictable, indeed emergent; those which appear and reappear regularly in our evolutionary simulations become candidates for interesting emergent properties. We shall consider further how to judge the status of emergent properties, and more generally the epistemological status of simulations, in Chapter **??**.

1.4 Evolving Ethical Behavior

In putting evolution and ALife together we make possible the evolution of emergent, complex ethical behavior. Of course, this is nothing new, since Evolution has already done the same Herself. The advantage of repeating the evolution of ethical behavior is that we can study the process, rather than just be its outcomes.

The study of behavior from an evolutionary perspective has been developing rapidly, most prominently in evolutionary psychology. Evolutionary psychology is based on three main principles: that behavior has an evolutionary explanation; that behavioral traits arose in an **evolutionary environment of adaptation (EEA)**, which often differs from present day environments; and that minds are modular. Modular minds can be separated into components that have evolved partially independently of each other, much as a genome that can be separated into genes (Fodor, 1983).

Evolutionary psychologists have been investigating a wide range of ethical behaviors, including those that we investigate here. The ethical studies of evolutionary psychologists, however, are limited to observational measurements and theoretical reasoning. Thus, for example, de Catanzaro

(1995) analyses suicide notes for signs that the act is aimed at benefiting kin, in accord with kin selection theory. Thornhill and Palmer (2000) propose evolutionary hypotheses (the adaptive and **by-product** hypotheses) for the existence of rape and assess their hypotheses based on the available victim data, which is notorious for its unreliability. And Lycett and Dunbar (1999) look at abortion data for single and married women of various ages, concluding that evolution may have shaped women's decision-making about abortion. All of these investigations proceed within the observational and deductive realm; evolutionary ALife simulation permits the addition of experimentation.

1.4.1 The Iterated Prisoner's Dilemma

Perhaps the most popular and commonly simulated model for ethical behavior is the Iterated **Prisoner's Dilemma** (IPD), from the game-theoretic work of Dresher and Flood (Dresher, 1961). The IPD simulations were the first simulations of significant interest to biology, and particularly to evolutionary psychology.

In the basic (non-iterated) Prisoner's Dilemma two prisoners are separated and each given two options: to rat out the other prisoner (defect), or to keep mum (cooperate). We are to assume that the prisoners have no means of communication and do not have pre-existing, binding commitments to each other. The payoffs to each prisoner of choosing cooperation or defection will depend on what the other prisoner chooses. In the original story these payoffs are reductions in prison terms (so '2' represents 2 years off), but more generally these should represent the **agent**'s utilities, so that, for example, any psychological or social disvalue in ratting someone out is already accounted for. Table 1.1 shows one possible matrix of payoffs for two prisoners, Alice and Bob.[8] The first value in each cell's pair indicates the payoff to Alice given the choices taken by both prisoners, and the second value indicates the payoff to Bob. We can see that Alice should defect if Bob cooperates, because it would pay her more (2 instead of 1). Furthermore, we can see that Alice should still defect if Bob instead chooses to defect, because it again would pay her more (0 instead of -1). Thus, defection is the **dominant strategy** for Alice, and by a symmetrical argument for Bob as well — that is, defection is *always* preferred regardless of the

[8] Note that there may be quite different payoff matrices for this situation, some of which will lead to different conclusions.

other prisoner's choice.[9]

		Bob	
		Cooperate	Defect
Alice	Cooperate	(1, 1)	(-1, 2)
	Defect	(2, -1)	(0, 0)

Payoffs: (Alice, Bob)

Table 1.1: The prisoner's dilemma. Rows represent Alice's choices, columns Bob's choices. Values are: (Alice's payoff, Bob's payoff).

Paradoxically, despite the domination argument, double defection yields a smaller payoff to each than double cooperation; but, since it is the dominant strategy (and they can't communicate with each other), both players will defect regardless. However, it is possible that players will choose a different strategy under the *iterated* version of the game, when they must make the same kind of choice, say, 100 times in succession. The results of prior rounds will be informative about their opponents, and so they offer a form of communication. And indeed, when Axelrod and Hamilton (1981) hosted a computer tournament in which submissions were subjected to a round-robin of IPDs with other submissions, the winning strategy was not `always-defect`, but `tit-for-tat`. The `tit-for-tat` strategy involves cooperating on the first turn and then reciprocating whatever choice the opponent made last, cooperating when the other player does, punishing the other player when not. It seems that in a fairly wide variety of environments (types of opponents) `tit-for-tat` does best in accumulating utilities.

Axelrod (1984) went on to confirm this result, evolving the `tit-for-tat` strategy in a **genetic algorithm**, in a process he called the "evolution of cooperation". In the right circumstances not only does `tit-for-tat` evolve, but it can be an **evolutionarily stable strategy** (ESS). For example, a population of suckers (`always-cooperate`) can be invaded, and eliminated, by defectors, but a population of tit-for-tatters cannot.

This is not yet the evolution of very interesting ethical behavior: the goal of the evolved "organisms" is not, for example, altruistic, but strictly selfish, maximizing their *own* utilities over time. The fact that `tit-for-tat` helps other tit-for-tatters maximize their utilities as well is an incidental by-product, so calling this behavior *cooperation* is already rather optimistic —

[9] This also implies that defection by both is the only **Nash equilibrium** for this game.

it is based more on the tags chosen for the actions than on the meaning of 'cooperation'! Attempts in these terms to explain genuine cooperation — where the activating goal is mutual — are much like attempts to explain altruism in terms of Trivers' (1971) concept of reciprocal altruism: they are aimed at explaining them *away*. However, we shall argue that altruism (and cooperation) are real enough, and that they can be explained in evolutionary terms. And we will back these claims with our simulations in Chapter **??**.

1.5 Experimental Philosophy

Experimental philosophy has suddenly become a big business. There are experimentalist philosophers investigating, amongst many other fields, epistemology (Bishop and Trout, 2005), philosophy of language (Machery et al., 2004), philosophy of science (Stotz and Griffiths, 2004) and also ethics (Knobe and Doris, 2008).[10] Part of what has stimulated this is a long-developing revolt against intuitive analytic philosophy, in particular the idea that philosophers have privileged access to concepts and ideas through their intuitions, so that the first step in analytic philosophy — getting clear about the concepts involved in some philosophical problem — can be done in the comfort of the philosophical armchair, running little thought experiments. (Of course, the succeeding steps can be done in the armchair as well, since they are deductive — indeed, as the joke goes, without even the need for a wastepaper basket.) Dennett (1991) denounced philosophical thought experiments which pretend to offer insights about circumstances with which we are wholly unfamiliar, for example, "brains" that replace neurons with humans and neural firings with instructions on paper. The experimental revolutionists go farther, demanding that the intuitions of philosophers be dethroned entirely and replaced by the intuitions of the masses. This shouldn't be a surprising step. We might, in fact, wonder why it has taken so long to arrive. Introspective psychology, for example, which supposed that psychologists have accurate insights into the workings of their own minds, died out in the 1920s, under joint pressure from the Behaviorists and the Freudians. Philosophers, no doubt, have no greater claims to special insight than psychologists. And analytic philosophers on the whole are a very unusual class of people, self-selected, well-educated and (once they get tenure[11]) well looked after. It's hardly likely that the ideas of ordinary discourse and folk psychology, the meat and potatoes of human ideation, are going to be

[10] For a collection of recent experimental philosophy, see Nichols and Knobe (2008).

[11] Excepting Australia, where tenure has been denatured.

properly served up by the likes of them!

So, experimental philosophers have the common aim of substituting empirical evidence about human ideation for unreliable, biased philosophical intuitions about our ideas. Many also take reflective equilibrium seriously as a useful approach towards explaining whatever intuitions one finds in field work. We share these interests and aims, however, we wish the experimentalists would go a little farther than they have towards *experimentalism*. In particular, most "experimentalists" are fully satisfied with opinion surveys about ideas, substituting some more representative sample of intuitions for the unrepresentative sample of one armchair occupant. But a sample survey is hardly the same as a controlled experiment.

For example, Marc Hauser (2006) conducted an on-line survey confirming armchair intuitions about the morality of killing and allowing killing. Around 90% of respondents thought it permissible to divert a tram headed for five innocent victims onto a track with one innocent victim; on the other hand, 90% also thought it impermissible to make the apparently same kind of utilitarian decision when it involved shoving an innocent fat man in front of the tram, with his death stopping the tram ahead of the five innocent victims. The best explanation for such intuitions seems to involve a contrast between allowing someone to die and being an active agent in his death. This intuition, if real, poses a *prima facie* problem for utilitarians. But the very first thing we should do in response to this is to question the reality of the intuition; there is no experimental evidence supporting its reality. Sample surveys are one thing and experiments another. To put this more pointedly: people like to think well of themselves. They spend much of their time and energy building up a world view and a view of themselves, and the latter is almost invariably positive. You may think Adolf Hitler was the epitome of evil, but it is clear that Hitler himself thought no such thing. This (extremely strong) tendency is revealed in asking people to remember long-past incidents occurring within the presence of others: no two stories are alike, and they often differ in predictable ways, those ways favorable to the story teller (as in Kurosawa's film *Rashomon*). In short, asking someone to imagine what they would do in circumstances C almost invariably results in that person reporting what she or he thinks sounds best. Putting that person *into circumstances C* will often yield entirely different results. Experimental philosophy could, apparently, learn a lot from recent experimental economics and, in particular, from adopting genuinely experimental methods.[12]

[12]We will grant that the proposed experiment introduces additional complexity, such

By the way, this criticism is not merely an evasion by some utilitarians of an unpleasant counterexample. The methodological defects of experimental philosophy need to be addressed by all, regardless of the subdiscipline or point of view. And, we shall defend utilitarianism in any case, even against possible counterexamples of the above type, in Chapter **??**.

1.5.1 Experimental Simulation

One of the major reasons why simulation methods have gained such prominence across the sciences is that they allow easy, and easily controlled, access to experimental techniques. It is far easier to manipulate variables within a simulation than within the process being simulated. Very often, there simply is no possibility of manipulating anything in the worldly process. Ignoring our simulations, then, the only kind of empirical evidence achievable is observation. So, for example, until the rise of computer simulation, empirical astronomy was almost purely an observational affair. Now, however, simulations of the origins of solar systems and the deaths of stars are commonplace.

What can be learned from observation is substantial. What can be learned from experimental intervention is far more substantial. This is quite intuitive. Observed associations between types of events commonly lead us to hypothesize a causal connection between them: smoking and lung cancer, CO_2 and global warming, asteroidal impacts and mass extinctions. In each of these cases, and many more, the observed associations, while supporting the hypothesis of causality, failed to settle the causal question — there were (or are) seemingly endless debates with skeptics. Experimental interventions, of course, could settle the causal questions. Interventions on humans smoking, the earth's atmosphere and asteroids may not be practical options, but they are options in principle. Indeed, the smoking question has been settled, with experimental interventions in test animals helping to rule out skeptical doubts, such as Fisher's (1957) suggestion of a common genetic cause for smoking and cancer. That asteroids have caused at least some of the mass extinctions has been thoroughly confirmed by a vast array of observational data, leaving no reasonable alternative explanation for

---

as questions about the moral courage of participants. However, our point is in part that survey samples are not as simple as they seem and are not clear and direct measures of the conceptual structure of the subjects. In particular, they are subject to systematic biases. Therefore, alternative measurements, and especially measurements that probe beyond the subjects' models of themselves, are going to be useful, whether or not they are also difficult to use. There are no simple answers to difficult philosophical problems.

the K-T boundary extinction of the dinosaurs and half of all other species (Ward, 1995). So, that's a win for observation, but it came at the cost of massive amounts of observational and theoretical research over a decade (the 1980s). The CO_2 cause of global warming is also, belatedly, coming to public consensus, well after computer simulation experiments put the issue beyond doubt for specialists. The theoretical and simulated causal connection between rises in CO_2 levels and terrestrial temperatures is far more compelling than the associational evidence alone of a past correlation between them (Randall et al., 2007).

Observations by themselves cannot reveal the difference between a correlation of effects of a common cause, e.g., genes causing both smoking and cancer, and a correlation arising from causation, e.g., smoking causing cancer. This is a problem of underdetermination, of multiple viable hypotheses remaining after the evidence comes in. Evidence from experimental interventions can eliminate far more hypotheses than can observational evidence. Indeed, in ideal circumstances, interventional evidence can uniquely identify the true hypothesis, simply eliminating the underdetermination problem (Korb and Nyberg, 2006). Experimental simulation opens up the possibility of gathering interventional evidence when otherwise only observational evidence would be available, whether because of physical, social or ethical limits upon those interventions.

So, experiments are a major part of what simulationists do. Making sense of experimentation and experimental methods, and especially how and why we can learn from them, will be an important topic in this book. This is necessitated by a frequently skeptical reaction to the possibility of learning *anything* from computer experimentation. In the various sciences in which computer simulation plays an important role (that is to say, in all sciences) the same skeptical concerns can be raised. They, however, no longer play a serious role in discussions in experimental physics, cosmology, chemistry, chemical engineering, etc. On the other hand, they do continue to be pressed for ALife simulations which attempt to inform us about the real world — ecological IBMs, social scientific ABMs and evolutionary simulations. What we will find in Chapter **??** is that there is no interesting epistemological difference between ALife simulations and the other simulation sciences: it is incoherent to accept the experimental verdicts of astrophysical simulations and reject those of ALife simulations, so long as certain preconditions are satisfied. Perhaps more surprisingly, we shall find that there is no interesting epistemological difference between simulated experiments in any of the sciences and real-world experiments in those sci-

ences. The calls of some philosophers of simulation (e.g., Winsberg, 2003) for a radically new epistemology to fit the radically new method of simulation will be firmly rejected.

1.5.2 Experimental Ethics

What is true of simulation across the sciences is true of simulated ethics, only more so. The kinds of experiments we perform with the virtual beings of Chapters **??** and **??** are not the kinds of things readers will want to try at home, outside of their computers! To be sure, most of them would be impossible, but many of the possible ones would be abhorrent. However, by conducting genuine experiments about rape, abortion, suicide and so on we can discover much of interest. We can discover environments where such behaviors have, and do not have, adaptive value. We can learn something about how and why they are adaptive, when they are — and vice versa. And, under the hypothesis of utilitarianism, we can learn about the morality of the behaviors under different evolutionary scenarios. There are, of course, substantial limits to what we can learn from the simulations we have performed. Some practical limitations arise from the fact that our work here is largely preliminary. This is a new field, and we are, at best, a scouting party. For an important example, our simulations here are fairly generic: we have not attempted to simulate any particular species, but rather the behavior of some very large class of species. We have aimed to cast light on general questions about the evolution of altruism and selfishness, **parental investments** and neglect. Inferences about any *particular* species will have to be qualified by considerations about their particular circumstances. We hope that others will pick up where we have left off and start simulating such particular circumstances. There is only knowledge to be gained.

1.6 Conclusion

We endorse a more scientific philosophy, one that sees little or no difference between the conjectures and theorizing of biologists and those of philosophers of biology, beyond disciplinary conventions and habits. Every science has its philosophy: physics and the philosophy of physics; AI and the philosophy of AI; psychology and philosophical psychology. Applied ethics has ethics, and ethics has metaethics. And every theory has empirical consequences — or else, as Karl Popper pointed out, we have a worthless theory. So, one of our goals is to liberate philosophy, to get it out

of doors, whether literally, as in experimental philosophy, or virtually, as in our experimental ethics.

This is not Scientism, the worship of all things pronounced Scientific. There is plenty that is wrong in science, just as there is in philosophy. But we think a genuine union of science and philosophy will do far more good than harm. To be sure, some of the worst philosophy has been done by scientists waxing philosophical. And some of the worst science already has been done by philosophers waxing empirical. The disciplinary divides probably cut deepest in methodological practices, and so mistakes are most likely when researchers try out new methods for new kinds of problems. But the presence or probability of error is no reason to abandon new methods; it is rather an opportunity to learn from them.

In conclusion, we hope this trek through artificial evolution and ethics will be both enlightening and entertaining. To enhance both aspects, we make available a variety of simulations illustrating our experiments at

`http://www.csse.monash.edu.au/evethics`

These are written in NetLogo, a user-friendly computer simulation language also available on the net at

`http://ccl.northwestern.edu/netlogo`

Reading the rest of this book. The next three chapters can be thought of as introducing the experiments of two chapters following them, which lie at the conceptual center of our work. Chapter **??** presents our experimental investigations of some of the key ideas in evolution theory, including levels of selection (gene, individual, group and species) and the evolution of altruism. Chapter **??** reports on the more ethically oriented experimental simulations. Preparatory to them, Chapter **??** describes the common structure and operation of most of our simulations, and so is a necessary prerequisite to understanding them. The next two chapters, on the other hand, can stand on their own. Chapter **??** introduces ethics and evolutionary psychology and defends our preferred ethical system, utilitarianism, against some misunderstandings and some counterarguments. While our experimental work collects and reports on utilitarian statistics, the defense of their use may be skipped by the uninterested. Chapter **??** defends the view that we can learn about the real world from the virtual world; those who need no persuasion, and who feel no urge for a dose of epistemology, may easily pass over it.

Bibliography

Aristotle (325BC/1998). *The Nicomachean Ethics*. Oxford: Oxford Univ.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Axelrod, R. and W. D. Hamilton (1981). The evolution of cooperation. *Science 211*, 1390–1396.

Batterman, R. (2007). Intertheory relations in physics. *Stanford encyclopedia of philosophy*. http://plato.stanford.edu/entires/physics-interrelate/.

Bedau, M. (2006). The evolution of complexity. *Symposium on the Making up of Organisms*, Ecole Normale Suprieure.

Berlekamp, E., J. Conway, and R. Guy (1982). *Winning Ways for Your Mathematical Plays*, Volume 2. New York: Academic Press.

Bishop, M. and J. D. Trout (2005). *Epistemology and the Psychology of Human Judgment*. Oxford: Oxford University Press.

Chen, W.-H., X.-X. Wang, W. Lin, X.-W. He, Z.-Q. Wu, Y. Lin, S.-N. Hu, and X.-N. Wang (2006). Analysis of 10,000 ESTs from lymphocytes of the cynomolgus monkey to improve our understanding of its immune system. *BMC Genomics 7*.

Darwin, C. (1988/1859). *On the Origin of Species*. Washington Square, NY: New York University Press.

Dawkins, R. (1976). *The Selfish Gene*. Oxford: Oxford University Press.

de Catanzaro, D. (1995). Reproductive status, family interactions, and suicidal ideation: Surveys of the general public and high-risk groups. *Ethology & Sociobiology 16*, 385–394.

Dennett, D. (1991). *Consciousness Explained*. Little, Brown & Co.

Dennett, D. (1995). *Darwin's Dangerous Idea*. New York: Simon and Schuster.

Dobzhansky, T. (1951). *Genetics and the Origin of Species*. New York: Columbia University.

Dresher, M. (1961). *The Mathematics of Games of Strategy: Theory and Applications*. Englewood Cliffs, NJ: Prentice-Hall.

Epstein, J. M. and R. Axtell (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. Cambridge: MIT Press.

Fisher, R. (1957, 3 August). Letter. *British Medical Journal*, 297–8.

Fodor, J. and J. Piattelli-Palmarini (2010). *What Darwin Got Wrong*. London: Profile Books.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

Fogel, L. J., A. J. Owens, and M. J. Walsh (1966). *Artificial Intelligence through Simulated Evolution*. New York: Wiley.

Frigg, R. and J. Reiss (2009). The philosophy of simulation: hot new issues or same old stew? *Synthese 169*, 593–613.

Giere, R. N. (1985). Philosophy of science naturalized. *Philosophy of Science 52*, 331–356.

Goldspink, C. (2002). Methodological implications of complex systems approaches to sociality: Simulation as a foundation for knowledge. *Journal of Artificial Societies and Social Simulation 5*(1).

Goldstine, H. H. (1993). *The Computer from Pascal to von Neumann*. Princeton: Princeton University.

Gooding, D. C. and T. R. Addis (2008). Modelling experiments as mediating models. *Foundations of Science 13*, 17–35.

Goodman, N. (1956). *Fact, Fiction, and Forecast*. Indianapolis: Bobbs-Merrill.

Grimm, V. and S. Railsback (2005). *Individual-based Modelling and Ecology*. Princeton: Princeton University Press.

Hamilton, W. (1964). The genetical evolution of social behavior I & II. *Journal of Theoretical Biology 7*, 1–16 & 17–52.

Hamilton, W. (1975). Innate social aptitudes of man: An approach from evolutionary genetics. In R. Fox (Ed.), *Biosocial Anthropology*, New York, pp. 133–155. John Wiley and Sons.

Hauser, M. (2006). *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. New York: Harper Collins.

Hegselmann, R. and U. Krause (2006). Truth and cognitive division of labour. *Journal of Artificial Societies and Social Simulation 9*. `http://jasss.soc.surrey.ac.uk`.

Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor: University of Michigan Press.

Hull, D. (1988). *Science as a Process*. Chicago: Chicago University Press.

Hume, D. (1739). *A Treatise of Human Nature*. London: Penguin (1984). Edited by E. C. Mossner.

Humphreys, P. (1991). Computer simulations. In *Philosophy of Science Association 1990*, Volume 2, pp. 497–506.

Humphreys, P. (2004). *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford: Oxford University Press.

Huxley, J. S. (1927). *Religion without Revelation*. London: Ernest Benn.

Huxley, J. S. (1942). *Evolution, the Modern Synthesis*. London: Allen and Unwin.

Jablonka, E. and E. Szathmáry (1995). The evolution of information storage and heredity. *Trends in Ecology and Evolution 10*, 206–211.

Keller, E. F. (2003). *Making Sense of Life: Explaining Biological Development with Models, Metaphors, and Machines*. Cambridge, MA: Harvard University Press.

Kim, J. (1993). *Supervenience and Mind: Selected Philosophical Essays*. Cambridge: Cambridge University Press.

Knobe, J. and J. Doris (2008). Strawsonian variations: Folk morality and the search for a unified theory. In J. D. et al. (Ed.), *Handbook of Moral Psychology*. Oxford: Oxford University Press.

Korb, K. B. and S. Mascaro (2009). The philosophy of computer simulation. In C. Glymour, W. Wei, and D. Westerstahl (Eds.), *Logic, Methodology and Philosophy of Science: Proceedings of the Thirteenth International Congress*, pp. 306–325. College Publications.

Korb, K. B. and E. Nyberg (2006). The power of intervention. *Minds and Machines 16*, 289–302.

Langton, C. (Ed.) (1989). *Artificial Life: The Proceedings of an Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems*, Redwood City, CA. Addision-Wesley.

Lewis, D. (1986). *Philosophical Papers*, Volume II. Oxford: Oxford University Press.

Lycett, J. E. and R. Dunbar (1999). Abortion rates reflect optimisation of parental investment strategies. *Proceedings of the Royal Society, B.: Biological Sciences 266*(1436), 2355–2358.

Machery, E., R. Mallon, S. Nichols, and S. P. Stich (2004). Semantics, cross-cultural style. *Cognition 92*, 1–12.

Maynard Smith, J. (1976). Group selection. *Quarterly Review of Biology 51*, 277–283.

Mayr, E. (1976). *Evolution and the Diversity of Life*. Cambridge, MA: Harvard University.

Moore, G. E. (1903). *Principia Ethica*. Cambridge: Cambridge University Press.

Nichols, S. and J. M. Knobe (Eds.) (2008). *Experimental Philosophy*. Oxford: Oxford University.

Odling-Smee, F. J., N. Laland, and M. W. Feldman (2003). *Niche Construction*. Princeton: Princeton University.

Quine, W. and J. Ullian (1978). *The Web of Belief*. New York: Random House.

Quine, W. V. O. (1969). Epistemology naturalized. In *Ontological Relativity and Other Essays*. Columbia University.

Racynski, S. and A. Bargiela (2007). *Modeling and Simulation: Computer Science of Illusion*. Research Studies Press. `http://portal.acm.org/`.

Railsback, S. F., R. H. Lambersion, B. C. Harvey, and W. E. Duffy (1999). Movement rules for individual-based models of stream fish. *Ecological Modelling 123*, 73–89.

Randall, D. A., R. A. Wood, S. Bony, R. Colman, T. Fichefet, J. Fyfe, V. Kattsov, A. Pitman, J. Shukla, J. Srinivasan, R. J. Stouffer, A. Sumi, and K. E. Taylor (2007). Climate models and their evaluation. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller (Eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Rawls, J. (1972). *A Theory of Justice*. Oxford: Oxford University Press.

Schwefel, H.-P. (1981). *Numerical Optimization of Computer Models*. New York: Wiley.

Selfridge, O. G. (1959). Pandemonium: A paradigm for learning. In *Proceedings of the Symposium on the Mechanization of Thought Processes*. London: Her Majesty's Stationery Office.

Singer, P. (1994). Introduction to 'Ethics'. In P. Singer (Ed.), *Ethics*. Oxford: Oxford University Press.

Spencer, H. (1864). *Principles of Biology*. D. Appleton.

Sterelny, K. and P. E. Griffiths (1999). *Sex and Death: An Introduction to the Philosophy of Biology*. Chicago: University of Chicago.

Stotz, K. and P. E. Griffiths (2004). Genes: Philosophical analyses put to the test. *History and Philosophy of the Life Sciences 26*, 5–28.

Stroud, P., S. Del Valle, S. Sydoriak, J. Riese, and S. Mniszewski (2007). Spatial dynamics of pandemic influenza in a massive artificial society. *Journal of Artificial Societies and Social Simulation 10*. `http://jasss.soc.surrey.ac.uk`.

Thornhill, R. and C. T. Palmer (2000). *A Natural History of Rape*. London: MIT Press.

Toyama, M. (2001). Adaptive advantages of matriphagy in the foliage spider, *Chiracanthium Japonicum*. *Journal of Ethology 19*, 69–74.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology 46*, 35–57.

Volterra, V. (1931). Variations and fluctuations of the number of individuals in animal species living together. In R. Chapman (Ed.), *Animal Ecology*.

von Neumann, J. (1951). The general and logical theory of automata. In A. H. Taub (Ed.), *John von Neumann: Collected works.*, pp. 288–328. New York: Pergamon Press.

Ward, P. (1995). *The End of Evolution: Dinosaurs, Mass Extinction and Biodiversity*. London: Weidenfeld and Nicholson.

Wiener, N. (1948). *Cybernetics: Or, Control and Communication in the Animal and Machine*. Cambridge, MA: MIT Press.

Williams, G. C. (1966). *Adaptation and Natural Selection*. Princeton, NJ: Princeton University Press.

Wilson, D. (1997). Introduction: Multilevel selection theory comes of age. *American Naturalist 150, Supplement*, S1–S4. This introduces a special issue on multilevel selection.

Wilson, D. (2005). Human groups as adaptive units: toward a permanent consensus. In P. Carruthers, S. Laurence, and S. Stich (Eds.), *The Innate Mind: Volume 2, Culture and Cognition*, pp. 78–90. Oxford University Press.

Wilson, E. (1978). *On Human Nature*. Cambridge: Harvard University.

Winsberg, E. (2001). Simulations, models, and theories: Complex physical systems and their representations. In *Proceedings of the 2000 Biennial Meetings of the Philosophy of Science Association (Supplement to Philosophy of Science Vol. 68 No. 3)*, pp. S442–S454.

Winsberg, E. (2003). Simulated experiments: methodology for a virtual world. *Philosophy of Science 70*(1), 105–125.

Wright, S. (1922). Coefficients of inbreeding and relationship. *American Naturalist 56*, 330–338.

Glossary

Adaptation Adaptation either refers to an evolutionary process or to a trait resulting from such a process. The adaptive process fixes traits in a population through the action of natural selection. Adaptive traits are structures, behaviors or strategies that increase fitness and so are likely to be fixed in the population by natural selection.

Agent There are two senses of 'agent' employed in the literature that we refer to. (1) An agent is a behaving system with intentionality — i.e., with beliefs, desires and purposes — and so capable of moral responsibility. (2) An agent is a behaving system — i.e., a system that has some dynamics associated with it. Pieces of code interpreted in **computer processes** which have little or no cognitive ability are often called "agents" by their designers, in much the same hopeful way they call their tiny representational abilities "knowledge".

Agent-Based Modeling (ABM) The study of social systems using computer simulation of individuals within an environment; a form of **ALife** simulation.

Aging Aging is the general deterioration of an organism via internal causes, leading to its eventual death.

Algorithm An algorithm is a type of procedure for implementing a function from a range of possible inputs to particular outputs. The three defining features of an algorithm are: (1) definiteness (its steps must be "primitive" and well understood, as in the steps of a Turing machine); (2) finiteness (it must stop); (3) functionality (its output must always be the same given the same input). An algorithm may be implemented by any number of distinct programs (Turing machines).

Stochastic algorithms incorporate the appearance of indeterminism, by using pseudo-random number generators to sample from prob-

ability distributions. This introduces variation into some aspect of the performance of the program, for example in mutations during artificial reproduction. Nevertheless, these are algorithms, since the pseudo-random number generators are deterministic.

Allele An alternative form of a DNA sequence at some **locus** in the chromosome.

Allopatric Speciation Speciation that occurs after a population has divided into isolated, geographical groups.

Altruism An altruistic act is one which harms the acting agent while benefitting others. The actor incurs negative utility while others derive positive utility. See also **Biological Altruism**.

Artificial Life (ALife) The study of the basic processes of life, real or possible, using computer simulation or other technology.

Artificial Neural Network (ANN) An artificial neural network is a model of neural processing in the brain. An ANN is composed of interconnected neurons, where each neuron is a function taking inputs either externally or from other neurons, and yielding an output. Commonly, the function involves a weighted sum of the inputs which is then passed through a continuous equivalent of a threshold function such as the logistic function (called the neuron's activation function) to yield the output.

Bayesianism, Bayesian A Bayesian takes probability functions to be the (or a) key means to represent and reason about uncertainty. Bayesians use probability to represent subjective degrees of belief in propositions and generally advocate conditionalization — adopting the conditional probability function based upon the evidence acquired — as a useful method of modeling evidential learning. This approach can be applied to any domain in which probabilities (or degrees of belief) arise, such as in the expected value calculations of decision theory, scientific theory **confirmation** and causal modeling.

Bayesian Network (BN) A Bayesian network is a graphical means of representing probability distributions and associated software for performing evidential updating. The graphs are sets of variables (or nodes) connected by directed arcs without cycles (i.e., directed acyclic graphs).

A directed arc represents a conditional probabilistic dependency between a parent variable (the arc source) and the child variable (the arc destination), such that the state of a child depends on the state of all its parents.

Biological Altruism An **allele** exhibits biological altruism if and only if its presence reduces the individual fitness of its owner while increasing the individual fitnesses of others. Phenotypic traits (such as behaviors) may analogously be considered altruistic if they do the same. See **Altruism**.

By-product (Piggyback Trait) A by-product is a trait that is evolutionarily neutral or harmful but has survived the evolutionary process because of its necessary connection with other, adaptative, traits. See **Adaptation**.

Cellular Automaton A cellular automaton is a grid of cells (of any finite number of dimensions, but frequently two dimensions) in which each cell can take on a finite number of states (often just the two states "on" and "off"). The state of each cell changes over (discrete) time and is a function of its own state and that of neighbouring cells from the previous time step. The most famous cellular automaton is The Game of Life by John Conway.

Co-evolution Co-evolution describes scenarios in which two evolving species inhabit each others' environments and affect each others' evolutionary histories. Commonly discussed co-evolutionary scenarios include predators and prey, hosts and parasites, and flowers and pollinators.

Computer Program A program is a sequence of instructions for a (virtual) machine. The beginning and end of a program are its first and last sentences.

Computer Process A computer process is a process run by a (virtual) machine which interprets a **computer program**, generating outputs given some inputs. Like all processes, a computer process has a temporal beginning and end; these are usually not directly related to the beginning and end of the program being interpreted.

Confirmation, Bayesian Confirmation Theory In the philosophy of science confirmation refers to cases in which evidence is discovered that

increases the probability of a given hypothesis or theory. The opposite is disconfirmation, in which the probability of the hypothesis or theory is decreased by the new evidence.

Consequentialism Consequentialism holds that the effects or consequences of an act must be taken into account when determining the ethical value of an act. See **Deontological Ethics**.

Cooperation Cooperation is coordinated activity aimed at achieving a common goal. The common goal often provides mutual benefits, but this is not required. Since participants must be capable of possessing goals, they must also possess agency. See **Agent** (1).

Cultural Evolution In the context of evolutionary biology, cultural evolution is the change in culture over time due to processes analogous with those found in biological evolution: namely, **heritability**, selection and variation. Cultural evolution substitutes elements of culture, such as ideas, theories and customs, learning and mistakes and creativity, for the corresponding elements of biological evolution (such as genes, reproduction and mutation). In most such models, the distinction between genotype and phenotype is either absent or ill-defined.

Decision Theory Decision theory examines how choices are or can be made under uncertainty. At the heart of decision theory is the idea of the expected value of a choice (a sum of all the possible outcomes, good, bad or neutral, weighted by the probability of each outcome), allowing choices to be placed on an interval scale. The value of outcomes is typically assessed in accordance with utility theory. See **Utility Theory**.

Decision Tree (Classification Tree) A decision tree is a classifier function that operates by recursively classifying the input into increasingly finer-grained classes. It is represented by a tree (typically visualized upside-down), with branches, leaves (or output) and a single root. Branches split the input data into two or more classes; once a branch identifies which class the input falls into, it passes execution to either the next branch for that class or, as a stopping condition, the leaf associated with that class. See **Production Rule**.

Deme A deme is a locally isolated and interbreeding sub-group of a population.

Deontological Ethics Deontological ethics holds that goodness inheres in either acts, duties, rules or rights. See **Consequentiaism**.

Descriptive Ethics The study of what people believe about how we ought to behave. See **Normative Ethics**.

Dominant Strategy A strategy (or action) dominates another in game theory if and only if the **utility** of each possible state on the first strategy is greater than that for the same state on the second.

Egoism Ethical egoism (or just 'egoism') is the belief that people ought to do what is in their own self-interest. See **Hedonism**.

Emergence, Emergent Property Emergent properties of a system A are its higher-level properties that cannot be defined in terms of the **supervenience** base B (and its properties) which implement it. See **Supervenience**.

Environment of Evolutionary Adaptation (EEA) The circumstances in which a characteristic, or set of characteristics, evolved. In particular, this is used to refer to the prehistoric circumstances in which the human mind evolved, contrasting them with the historic circumstances in which the human mind is currently operating.

Evolutionary Algorithm An evolutionary algorithm uses the operators of evolution (selection, reproduction and mutation) to successively modify or (in optimization problems) improve the current population of entities or solutions. Examples of evolutionary algorithms include evolution strategies, evolutionary programming, genetic programming and genetic algorithms. The main difference between these algorithms is what representations are used in the genomes or members of the population — whether bit strings, numbers, graphs, vectors, state machines or programs. See **Genetic Algorithm**.

Evolutionary Ethics Evolutionary ethics is the attempt to found ethical norms on evolutionary history.

Evolutionary Psychology Evolutionary psychology is the study of animal behavior from an evolutionary perspective. More particularly, it is a school of thought in such studies characterized by the beliefs that most behavior has an evolutionary explanation, that behaviors have arisen in "evolutionary environments of adaptation" (EEA), and that

cognitive functions have often evolved in semi-independent modules. See **Sociobiology**.

Evolutionarily Stable Strategy (ESS) An evolutionarily stable strategy is a behavior determined (partially or fully) by genetics such that, if adopted by all members of a population, no alternative strategy can invade and replace the ESS. In other words, under the circumstances (where most of the population has adopted the ESS), the ESS is fitter than its alternatives and so resists invasion. The behaviors may be pure or **mixed strategies**.

Filial Infanticide The killing of one's own offspring shortly after birth, while the offspring is still a dependant.

Fisher's Reproductive Value The reproductive value of an organism of a given age reflects the expected future number of offspring of the organism. Reproductive value is calculated by summing (from the organism's current age onwards) the probability of reaching a given age multiplied by the average number of offspring produced by an individual at that age and dividing the sum by the average population fitness.

Fitness Individual fitness is often defined as the expected number of offspring reaching maturity. We prefer to define it as the expected number of descendants, which accommodates both uncertainty about the future (through probabilistic expectation) and issues about the viability and fertility of offspring. For practical measurements, we often substitute descendants over two generations. See **Inclusive Fitness**.

Gene A sequence of DNA which codes for some protein(s).

Genetic Algorithm (GA) A genetic algorithm is a type of optimizing evolutionary algorithm that operates on a population of chromosomes (traditionally, raw bit strings) without any environment. The GA handles the reproduction, mutation and selection of these genomes. Selection for reproduction is directed by an artificial fitness function ("objective function") and occurs between discrete, successive generations. Evolutionary ALife simulations, by contrast, embed genomes within interacting agents that exist in a wider environment, where fitness is determined (as in biology) by agents' abilities to survive and reproduce.

Group Selection Group selection is selective pressure arising from the differential ability of groups to propagate themselves by establishing new groups, whether from greater rates of founding colonies, greater group longevity, or both. Given a correlation between this kind of group fitness with allele frequencies across groups, allele representations within the total population will tend to correspond to group selection pressure.

Hamming Distance Given two bit strings of equal length, the Hamming distance is the number of locations at which the two strings differ.

Hedonism Hedonism is the belief that pleasure is the only intrinsic good. Ethical hedonism (often just 'hedonism') merges egoism with hedonism yielding the belief that people ought to do what maximizes their own pleasure and minimizes their own pain. See **Egoism**.

Heritability Given a probability distribution over environments, the heritability of a trait is the amount of its variance which is explained genetically. (For standardized variables, this will be equal to one minus the amount of its variance explained by environmental variation.)

Homomorphism A homomorphism from system A to system B is a mapping of objects, functions and relations from system A onto system B such that all relations (and functions) within system A are preserved under the mapping in system B.

Inclusive Fitness Inclusive fitness measures the total fitness effects of an allele over a population. See **Kin Selection**.

Individual-Based Modeling (IBM) The study of biological systems using computer simulation of individuals within an environment; a form of **ALife** simulation.

Individual Selection Individual selection is differential natural selection pressure operating upon individual organisms through their different individual fitnesses. See **Kin Selection** and **Group Selection**.

Induced Abortion Induced abortion is the termination of a pregnancy by choice. See **Spontaneous Abortion**.

Isomorphism An isomorphism between system A and system B is a homomorphic mapping from system A to system B such that its inverse

is a homomorphic mapping from system B to system A. See **Homomorphism**.

Kin Selection Kin selection identifies the impact of a phenotypic trait upon the fitness of the kin of the trait's bearer as the relevant factor for determining the spread of alleles coding for that trait. See **Inclusive Fitness**.

Levels of Selection Levels of selection refers to the type of selection pressure that may be active in an evolutionary system. The levels generally recognized as subject to significant selection are the gene, the cell, the individual, the group (**deme**) and the **species**. See **Multilevel Selection**.

Locus A location within DNA that repeatedly plays host to the same gene (and, so, some set of alleles).

Lotka-Volterra Model The Lotka-Volterra model is a simple model of predator-prey interactions described by a pair of differential equations. Each differential equation describes the rate of change in the numbers of either predators or prey given the current numbers of predators and prey, growth rates, encounters between predators and prey and death rates.

Mental Module A mental module is a substructure within the mind that has an evolved function (such as a module for language or facial recognition). In contrast to phrenology, mental modules need not (and are not generally expected) to correspond one to one with a physical substructure in the mind. See **Evolutionary Psychology**.

Metaethics The attempt to provide a theory of ethical study that allows us to choose between ethical systems.

Multilevel Selection Multilevel selection is selection pressure that acts simultaneously at multiple levels or, equivalently, on multiple biological units. The units generally recognized as subject to significant selection are the gene, the cell, the individual, the group (**deme**) and the **species**. It can be contrasted with the traditional (modern synthesis) view, in which genes are considered the only unit of consequence. See **Levels of Selection**.

Multiple Realizability See **Supervenience**.

Mixed Strategy A set of behaviors alternative to each other which are selected according to some probability distribution.

Naive Bayes A naive Bayes model is a simple Bayesian network used for (probabilistic) classification in which a single class node is the lone parent to all other nodes. The child nodes are independent of each other, given the class, and represent the attributes of some entity. See **Bayesian network**.

Nash Equilibrium A Nash equilibrium in game theory is a set of strategies such that no player can increase its utility by switching to an alternative strategy when the other players do not change strategy. See **Dominant Strategy**.

Naturalistic Fallacy Moore defined the naturalistic fallacy as the error of inferring an object's goodness from its natural properties. The term now generally refers to any attempt to directly derive 'ought' from 'is'.

Normative Ethics (Also, just "Ethics".) The study of how we ought to behave. See **Descriptive Ethics**.

Parapatric Speciation Speciation that occurs after a small group partially splinters from a larger group into a new, adjacent but not isolated geographical niche. See **Peripatric Speciation**.

Parental Investment Parental investment is any investment in an offspring that boosts that offspring's chance of survival but comes at the cost of investing in other offspring. It may refer to either material or behavioral support. See **Reproductive Strategy**.

Peripatric Speciation Speciation that occurs after a small group splinters from a larger group into an isolated geographical niche. See **Parapatric Speciation**.

Phyletic Gradualism Phyletic gradualism refers to evolutionary histories in which all evolutionary change occurs gradually (with no sudden jumps in phenotype space), including evolutionary change that gives rise to new species.

Pleiotropy Pleiotropy describes the genetic effect in which a single gene gives rise to multiple phenotypic traits. See **Polygenetic**.

Polygenetic A phenotypic trait is polygenetic if it is caused or influenced by multiple genes. See **Pleiotropy**.

Pop Sociobiology Pop Sociobiology is a term used by critics to describe sociobiology that relies heavily on just-so stories to explain the origins of modern human behavior. Fierce criticism of pop sociobiology lead to development of the methodologically more rigorous field of evolutionary psychology. See **Sociobiology** and **Evolutionary Psychology**.

Pre-adaptation A pre-adaptation is a trait that has evolved as an adaptation to one situation, but is subsequently put to a different use in an evolutionarily novel situation. See **Adaptation**.

Price Equation The Price equation is a generalization of kin and group selection models (and of selection models in general) which allows us to separate the selection effects acting within groups from those acting between groups. The equation is:

$$\bar{w}\Delta\bar{z} = \mathrm{Cov}(w_i, z_i) + \mathrm{E}(w_i \Delta z_i) \tag{1.1}$$

On the right hand side, z is the character of interest (e.g., height, eye color or altruistic disposition) assumed to be representable by a real number, i identifies a subgroup of the population that shares the same value for z, z_i is the shared value itself (e.g., tall, blue or selfish), Δz_i is the change in this character from generation to generation, and w_i is the average absolute fitness of the subgroup i with trait z_i. On the left hand side, $\bar{w}$ is the average absolute fitness across the population overall and $\Delta\bar{z}$ is the average change from generation to generation in the character z over the population overall. By dividing through by $\bar{w}$, evolutionary change can be explicitly phrased in terms of relative fitness (i.e. $w_i/\bar{w}$).

The covariance term represents how fitness (w_i) varies with the value of the character (z_i) — if this term is positive, larger z values lead to higher fitness; if negative, smaller z values lead to higher fitness. The expectation term describes the fidelity or bias with which traits are transmitted to offspring. The terms can also be adapted to refer to groups containing altruists rather than individuals, in which case the covariance models the contribution of altruists to group fitness, while the expectation term models the loss due to the in-group loss of altruists. See **Group Selection**, **Kin Selection** and **Individual Selection**.

Production Rule A production rule is a condition-action (or if-then) rule that links an observation of the world (the condition) with an action to perform (the action). In agents, a set of production rules is typically used to link sensory data with motor function. See **Decision Tree**.

Prisoner's Dilemma The Prisoner's Dilemma is a game (in the game-theoretic sense) based on the hypothetical case of two prisoners, collaborators in some crime, who are separated and questioned by police. Each prisoner has two choices: to inform ("defect") or to stay silent ("cooperate"). If both stay silent, the prisoners receive a modest prison term; if both inform, they receive longer terms. However, if one informs while the other stays silent, the informant receives the minimum prison term while the other receives the maximum. The dominant strategy is to inform, since informing is rewarded with the shorter prison term regardless of what choice the other prisoner makes. However, iterating the game can lead to different conclusions. See **Cooperation** and **Stag Hunt**.

Pro-choice The position that a woman should have control over her own body during pregnancy and thus that she can choose to abort. See **Pro-life**.

Pro-life The position that human life begins at or just after conception and that abortion should be entirely prohibited or permitted only under extreme circumstances. See **Pro-choice**.

Punctuated Equilibrium Punctuated equilibrium refers to evolutionary histories in which long periods of morphological and behavioral stasis (or near stasis) are punctuated by short periods of rapid evolutionary change. Such punctuations often result in the appearance of new species. See **Phyletic Gradualism**.

Reciprocal Altruism Reciprocal altruism is the exchange of altruistic acts between individuals over time such that both individuals enjoy a net benefit. See **Cooperation** and **Altruism**.

Reduction A reduction of system A to system B is the provision of necessary and sufficient conditions for the properties and relations of A in terms of a different system B (the reduction base). In other words, there is an **isomorphism** between the two systems. (Cf. Batterman, 2007, on reductive bridge laws establishing synthetic type identities.) System B is typically taken as metaphysically more fundamental.

Examples include supposed reductions of biology to chemistry, of chemistry to physics, and of mental states to neurochemical states. See **Supervenience**.

Reflective Equilibrium Reflective equilibrium is a process of achieving a state of coherence between a core set of judgments about some domain, a theory about how those judgments should be made and wider theories relevant to the domain (e.g., including theories of human judgment).

Reinforcement Learning Reinforcement learning is a set of machine learning techniques for learning in a stochastic environment. An agent performs a sequence of actions affecting the environment. When the agent receives some (positive or negative) reward from the environment, it applies some algorithm to decide how much of the reward to attribute to the different actions leading to that algorithm (in a credit assignment process). The overall goal is to activate actions leading to positive rewards more often and suppress those leading to negative rewards.

Reproductive Strategy A reproductive strategy is the approach an organism takes to maximizing its genetic contribution to future generations by optimizing the division of parental investments amongst its expected offspring. When the environment is harsh and unpredictable, organisms will produce many offspring (since each will have little chance of surviving the environment) but invest little in each. When the environment is safe and predictable, organisms will predominantly compete with each other, and therefore invest a large (competitive) amount in each, and necessarily produce fewer offspring.

Sexual Dimorphism A species exhibits sexual dimorphism when its two sexes differ in morphology or behavior.

Sexual Selection Sexual selection refers to the competitive processes that occur within one sex for access to desirable members of the other sex. Typically, these processes are broken down into inter-male agression versus inter-female mate choice, but these sexual roles may often be reversed.

Simulation A (computer) simulation is a (computer) process that mimics features of a target physical process, such that a common dynamical theory is capable of describing both the simulation and its target

process. For practical and theoretical reasons, a simulation is strictly simpler than its target, which entails a homomorphism from the target process to the simulation and the lack of an isomorphism. See **Homomorphism** and **Isomorphism**.

Social Simulation The field of social simulation employs simulations that contain societies of interacting agents, typically implemented in a bottom-up fashion, to explore social and societal phenomena. See **ALife**.

Sociobiology Sociobiology is a field that attempts to integrate a range of fields that study social behavior within both biology and sociology, including ethology, anthropology and behavioral economics. Sociobiology's approach to the study of social behavior is firmly rooted in evolution theory, and evolutionary psychology can be considered both a taxonomic and historical offshoot. See **Pop Sociobiology** and **Evolutionary Psychology**.

Species A species is commonly defined as a group of organisms that are capable of interbreeding and producing viable offspring. Especially for asexual species, alternative definitions are used in which similarity in genotype or phenotype is central.

Species Selection Selection which operates at the level of the species, in the form of extinctions and the propagation of new species. See **Multilevel Selection**.

Spontaneous Abortion Spontaneous abortion is the termination of a pregnancy that occurs via internal (non-intentional) causes. See **Induced Abortion**.

Stag Hunt The Stag Hunt is a game (in the game-theoretic sense) based on a scenario described in Rousseau's *The Social Contract*: two individuals on a hunt may choose either to hunt stag, which can only be successful if both cooperate, or to hunt hare, which can be successfully done alone. A stag yields more than twice the food of a hare, which implies that the game contains two Nash equilibria: either both cooperate to hunt stag (which yields a greater payoff) or both hunt hare (which involves less risk). See **Cooperation** and **Prisoner's Dilemma**.

Supervenience A system A is supervenient upon system B (the supervenience base) if and only if (a) B realizes (implements, instantiates) A; and (b) B is a member of a wider class of systems $\mathscr{B}$ any one of which could realize A. A is, therefore, said to be multiply realizable. In other words, there is a **homomorphism** from B to A.

(NB: Some people prefer to allow $\mathscr{B}$ to be a singleton set. This, however, loses the key characteristic of multiple realizability and conflates supervenience with **reduction**.)

Sympatric Speciation Speciation that occurs within a population in a single geographical area. Such speciation is considered rare, but may result if the population begins to exhibit polymorphic types that either cannot interbreed or find it difficult. See **Allopatric Speciation**.

Three Laws of Robotics The Three Laws of Robotics from Isaac Asimov's robot stories describe a set of "ethical" rules for robots to follow. With the addition of the later zeroeth law, the rules are as follows:

0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.
1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey any orders given to it by human beings.
3. A robot must protect its own existence.

Token An individual instance of some type of thing or event. Example: a dollar bill.

Type A collection of individual things or events characterizable by a set of properties held by all of those individuals. Example: a dollar.

Units of Selection See **Levels of Selection**.

Universal Computation, Universal Turing Machine (UTM) Devised by Alan Turing, a Turing Machine is a machine that can read and write symbols one at a time on an unbounded tape according to a table of rules. A Turing Machine is capable of performing any computation if given the right rules and symbol inputs. A Universal Turing Machine is capable of performing any computation (including the simulation of another Turing Machine) by changing the symbols on the tape alone.

Universalizable In ethics, a principle is universalizable if it can be adopted by everyone without logical inconsistency or absurd consequences. We appeal to universalizability when we ask, What if everyone did that?

Universal In evolutionary psychology, a behavior is universal if it is present in every member of the species (typically humans) or is pervasive within every group or culture, setting aside pathological cases.

Utilitarianism Utilitarianism is an ethical system which states that we should act so as to maximize the sum of expected utilities across a population. See **Utility** and **Consequentialism**.

Utility Utilities are theoretical entities used to explain agents' behavior under an assumption of Bayesian rationality — i.e., that agents aim to maximize their expected (probability-weighted) utilities. Utility functions map pairs of states of the world and actions to real numbers. A single unit of utility is sometimes called a "utile". Informally, utilities report the pleasantness or unpleasantness of the situation that an agent finds itself in, where "pleasant" is understood in a wide sense, incorporating any sensation that might have intrinsic value to the agent (e.g., the satisfaction of solving a problem would count as pleasant).

Utility Theory Utility theory is based on the principle that the set of preferences used in expected value calculations can be modeled by a cardinal (and potentially ratio-scale measurable) utility function. See **Utility** and **Decision Theory**.

Validation In simuation research validation refers to establishing (or testing) whether the simulation model corresponds to the targeted physical process. This corresponds (somewhat confusingly) to what Logical Positivists called **verification** and what in the philosophy of science generally is called confirmation. See **Verification** and **Confirmation**.

Verification In simulation research verification refers to establishing (or testing) whether the simulation model correctly implements the theory being investigated, including determining whether or not it is bug free. This usage is in contrast to that within the philosophy of science. See **Validation**.

Index