Boston Studies in the Philosophy of Science 282

Kathryn S. Plaisance
Thomas A.C. Reydon *Editors*

Philosophy of Behavioral Biology

Springer

Philosophy of Behavioral Biology

BOSTON STUDIES IN THE PHILOSOPHY OF SCIENCE

VOLUME 282

For further volumes:
http://www.springer.com/series/5710

Kathryn S. Plaisance • Thomas A.C. Reydon
Editors

Philosophy of Behavioral Biology

Springer

Editors
Kathryn S. Plaisance
Centre for Knowledge Integration
University of Waterloo
200 University Ave W
Waterloo, ON N2L 3G1
Canada
kplaisan@uwaterloo.ca

Thomas A.C. Reydon
Institut für Philosophie & Center for
Philosophy and Ethics of Science (ZEWW)
Leibniz Universität Hannover
Im Moore 21
30167 Hannover
Germany
reydon@ww.uni-hannover.de

ISSN 0068-0346
ISBN 978-94-007-1950-7 e-ISBN 978-94-007-1951-4
DOI 10.1007/978-94-007-1951-4
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2011937557

Contents

Part I Introduction

1 The Philosophy of Behavioral Biology 3
Kathryn S. Plaisance and Thomas A.C. Reydon

2 Knowledge for What? Monist, Pluralist, Pragmatist Approaches to the Sciences of Behavior 25
Helen Longino

Part II Genetic Explanations of Behavior

3 Genome Wide Association Studies of Behavior are Social Science 43
Eric Turkheimer

4 Genetic Traits and Causal Explanation 65
Robert Northcott

Part III Developmental Explanations of Behavior

5 From Cell-Surface Receptors to Higher Learning: A Whole World of Experience 85
Karola Stotz and Colin Allen

6 Re-Conceiving Nonhuman Animal Knowledge Through Contemporary Primate Cognitive Studies 125
Andrew Fenton

Part IV Evolutionary Explanations of Behavior

7 Evolving the Future: Sketching a Science of Intentional Change 149
David Sloan Wilson

8 Human Artistic Behaviour: Adaptation, Byproduct, or Cultural Group Selection? 167
Johan De Smedt and Helen De Cruz

9 Sensory Exploitation: Underestimated in the Evolution of Art As Once in Sexual Selection Theory? 189
Jan Verpooten and Mark Nelissen

10 Heuristic Evolutionary Psychology 217
Armin W. Schulz

11 Evolutionary Psychology and the Problem of Neural Plasticity 235
Chuck Ward

12 Free Will, Compatibilism, and the Human Nature Wars: Should We Be Worried? 255
Brian Garvey

13 Altruistic Emotional Motivation: An Argument in Favour of Psychological Altruism 275
Christine Clavien

14 The Neurobiology of Altruistic Punishment: A Moral Assessment of its Social Utility 297
Rebekka A. Klein

Part V Neurobiological Explanations of Behavior

15 Behavioral Traits, the Intentional Stance, and Biological Functions: What Neuroscience Explains 317
Marcel Weber

16 From Reactive to Endogenously Active Dynamical Conceptions of the Brain 329
Adele Abrahamsen and William Bechtel

Index 367

Contributors

Adele Abrahamsen Project Scientist, Center for Research in Language, University of California, San Diego, La Jolla (CA), USA, adele@crl.ucsd.edu

Colin Allen Provost Professor of History and Philosophy of Science, and Cognitive Science Program, Indiana University, Bloomington (IN), USA, colallen@indiana.edu

William Bechtel Professor of Philosophy, Department of Philosophy, and faculty member, Center for Chronobiology and Interdisciplinary Program in Cognitive Science, University of California, San Diego, La Jolla (CA), USA, bill@mechanism.ucsd.edu

Christine Clavien Junior Postdoctoral Lecturer, Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland, christine.clavien@unil.ch

Helen De Cruz Postdoctoral Research Fellow, Centre for Logic and Analytical Philosophy, Research Foundation Flanders (FWO), Katholieke Universiteit Leuven, Leuven, Belgium, and Templeton Research Fellow, Somerville College, University of Oxford, Oxford, United Kingdom, helen.decruz@hiw.kuleuven.be

Johan De Smedt Research Fellow, Department of Philosophy and Ethics, Ghent University, Ghent, Belgium, and visiting research scholar, Uehiro Centre for Practical Ethics, University of Oxford, Oxford, United Kingdom, johan.desmedt@ugent.be

Andrew Fenton Assistant Professor, Department of Philosophy, California State University - Fresno, Fresno (CA), USA, andrew.fenton@gmail.com

Brian Garvey Lecturer in Philosophy, Department of Politics, Philosophy and Religion, Lancaster University, Lancaster, United Kingdom, b.garvey@lancaster.ac.uk

Rebekka A. Klein Dilthey-Fellow granted by Volkswagen Foundation, Institute for Systematic Theology, University of Halle-Wittenberg, Halle/S., Germany, kleinrebekka@hotmail.com

Helen Longino Clarence Irving Lewis Professor of Philosophy, Department of Philosophy, Stanford University, Stanford (CA), USA, hlongino@stanford.edu

Mark Nelissen Professor of Behavioral Biology, Department of Biology, University of Antwerp, Antwerpen, Belgium, mark.nelissen@ua.ac.be

Robert Northcott Lecturer, Department of Philosophy, Birkbeck College, London, United Kingdom, r.northcott@bbk.ac.uk

Kathryn S. Plaisance Assistant Professor, Centre for Knowledge Integration, cross-appointed to the Department of Philosophy, University of Waterloo, Waterloo (ON), Canada, kplaisan@uwaterloo.ca

Thomas A.C. Reydon Junior Professor of Philosophy of Biology, Institute of Philosophy & Center for Philosophy and Ethics of Science (ZEWW), Leibniz Universität Hannover, Hannover, Germany, reydon@ww.uni-hannover.de

Armin W. Schulz Lecturer, Department of Philosophy, Logic, and Scientific Method, London School of Economics and Political Science, London, United Kingdom, a.w.schulz@lse.ac.uk

Karola Stotz Australian Research Fellow, Department of Philosophy, University of Sydney, Sydney, Australia, karola.stotz@sydney.edu.au

Eric Turkheimer Professor, Department of Psychology, University of Virginia, Charlottesville (VA), USA, ent3c@virginia.edu

Jan Verpooten PhD Research Fellow, Konrad Lorenz Institute for Evolution and Cognition Research, Altenberg, Austria and Department of Biology, University of Antwerp, Antwerpen, Belgium, jan.verpooten@ua.ac.be

Chuck Ward Associate Professor, Department of Philosophy, Millersville University, Millersville (PA), USA, chuck.ward@millersville.edu

Marcel Weber Professor, Department of Philosophy, University of Geneva, Geneva, Switzerland, marcel.weber@unige.ch

David Sloan Wilson SUNY Distinguished Professor, Departments of Biology and Anthropology, Binghamton University, Binghamton (NY), USA, dwilson@binghamton.edu

Part I
Introduction

Chapter 1
The Philosophy of Behavioral Biology

Kathryn S. Plaisance and Thomas A.C. Reydon

1.1 Introduction

1.1.1 Background

This volume offers a broad overview of central issues in the philosophy of behavioral biology, addressing philosophical issues that arise from the most recent scientific findings in biological research on behavior. It thus exemplifies an approach to philosophy of science that is scientifically informed as well as interdisciplinary. Accordingly, it includes chapters by professional philosophers and philosophers of science, as well as practicing scientists.

The volume originates from the conference, "Biological Explanations of Behavior: Philosophical Perspectives", held in Hannover, Germany, in June 2008. Participants in this conference represented the fields of behavioral genetics, evolutionary biology, cognitive science, philosophy of biology, philosophy of science, and communication studies. Conference presentations were organized into three main themes: explanations in behavioral genetics, developmental explanations of behavior, and the evolution of behavior. The book largely mirrors this organization, in addition to representing another theme in the philosophy of behavioral biology, namely neurobiological explanations of behavior. In what follows, we sketch out an

K.S. Plaisance (✉)
Centre for Knowledge Integration & Department of Philosophy,
200 University Ave West, University of Waterloo,
Waterloo, ON N2L 3G1, Canada
e-mail: kplaisan@uwaterloo.ca

T.A.C. Reydon
Institute of Philosophy & Center for Philosophy and Ethics of Science (ZEWW),
Leibniz Universität Hannover, Im Moore 21, 30167 Hannover, Germany
e-mail: reydon@ww.uni-hannover.de

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*,
Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_1,

overview of the book, both by describing some of the major themes and philosophical context, as well as providing detailed summaries of each of the chapters.

1.1.2 Motivation & Content

One of the major motivations for this volume, and the conference that preceded it, was the observation that there were many philosophically interesting and fruitful research questions about the nature of behavior that did not fall neatly within one area, such as philosophy of biology or philosophy of psychology.[1] Thus, one must often take an interdisciplinary approach when considering scientific explanations of behavior, drawing from biology, psychology, cognitive science, anthropology, etc., and from disparate areas from within each of these disciplines. For example, with respect to psychology alone, many papers in this volume make use of and analyze research in behavioral genetics, socialization research, evolutionary psychology, and neuropsychology, to name a few. Part of our aim in this volume (and in the conference that preceded it) is to map out the philosophical domain where these different areas of work intersect and identify what might be considered the philosophy of behavioral biology.

Furthermore, just as the philosophy of behavioral biology draws on many areas of research, it also looks at a variety of behaviors in many types of organisms. With respect to humans, specific traits are considered (e.g., intelligence, personality, and schizophrenia), as well as more general behaviors such as artistic behavior or phenomena like free will and altruism. With respect to animals, scientific explanations of the development of means of communication and intentional behavior are examined. Appropriately, then, the papers in this volume reflect work from philosophers working in a wide variety of subfields – many of whom tend to take interdisciplinary approaches, and in some cases conduct their own scientific research – as well as practicing scientists (most notably evolutionary biologist David Sloan Wilson and behavioral geneticist Eric Turkheimer).[2]

1.1.3 Audience

As a result of the interdisciplinary nature of this book, we think it will be of interest to a broad audience consisting of philosophers of science, philosophers of biology, philosophers of psychology, theoretical biologists, evolutionary psychologists,

[1] Interestingly, Karola Stotz and Colin Allen make a similar point about scientific disciplines in their paper; to address this, they "promote a biologically-informed psychology and a psychologically-informed biology."

[2] In addition to including scholars from a wide range of disciplines, this volume also displays a great deal of diversity in terms of gender, nationality, and academic rank (including chapters by graduate students, postdoctoral fellows, and assistant, associate and full professors).

behavioral geneticists, cognitive scientists, and behavioral biologists more generally. In addition, some of the chapters may be of interest to those working in other areas of science or philosophy. For example, Eric Turkheimer's chapter addresses limitations of research in molecular genetics; Christine Clavien's and Rebekka Klein's discussions of altruism connect to important questions in moral philosophy; and Brian Garvey's chapter on the question of free will addresses central issues in moral philosophy as well as philosophy of mind.

For philosophers of science, including philosophers of biology and psychology, the specific papers included in this volume will be central to much of the research that focuses on philosophical issues in biological explanations of behavior, as well as more general philosophical issues such as causation and explanation. Theoretical and philosophically-minded biologists and psychologists will also find interesting and relevant work that examines the concepts, methods, and inferential reasoning of scientific research in those fields. Furthermore, some of the papers in this volume explicitly address important methodological implications of research in behavioral biology that will be of use to practicing scientists. For example, Eric Turkheimer's chapter on Genome Wide Association Studies (GWAS) could affect how research in behavioral genetics is done, and Adele Abrahamsen and Bill Bechtel's argument that the brain should be thought of as an endogenously active mechanism might lead to a new approach in neuroscience.

While the papers herein will certainly be of use to scholars in terms of their research, this volume is also intended to serve as a useful resource for teaching higher-level courses, graduate seminars, and reading groups. It offers both an overview of the issues in philosophy of behavioral biology, as well as examples of current controversies in specific fields.

1.1.4 Structure

This volume is divided into an introductory part, as well as four parts that focus on different approaches to scientific explanations of behavior: genetic, developmental, evolutionary, and neurobiological. The introductory part (Part I) includes this chapter as well as Helen Longino's paper, "Knowledge for What? Monist, Pluralist, Pragmatist Approaches to the Sciences of Behavior", which is based on her keynote address for the 2008 conference from which this book originated, and which examines and compares various approaches to the scientific study of behavior. Part II includes two papers on genetic explanations of behavior, one written by a behavioral geneticist and the other a philosopher of science, both of which focus on human behavioral traits. Part III consists of two chapters on developmental explanations of behavior, with an emphasis on non-human animal learning. Part IV is the largest section, comprising eight papers on the evolution of behavior and addressing a variety of themes such as artistic behavior, research in evolutionary psychology, and altruism. Finally, in part V, neurobiological explanations of behavior are discussed.

There are a couple of things worth noting about this structure. First, there is the obvious unevenness of the sections, with the section on evolutionary explanations

of behavior dominating the various approaches, so much so that it comprises half the volume. This indicates to us a tendency in philosophy of behavioral biology to focus on evolutionary explanations over and above genetic, developmental, or neurobiological accounts.[3] Also, this particular emphasis is consistent with what we find in philosophy of biology more generally, where much of the work in that area has focused on conceptual issues in evolutionary biology.

Second, while we have been able to group the papers into these various sections in a relatively straightforward manner (by considering the approach that is the main focus of the paper), it is not the case that each paper only addresses scientific explanations that fall within that theme; rather, many of them draw on multiple approaches. For example, the paper by Karola Stotz and Colin Allen argues for a developmental approach to studying animal behavior, but in doing so they acknowledge the various roles of genetic and other kinds of factors in learning and development, arguing that these factors cannot be separated. In addition, Rebekka Klein's paper examines evolutionary explanations of altruistic behavior, though it also draws on research from neurobiology, thus providing a bridge between the two sections. As a final – and perhaps the most notable – example, Helen Longino explicitly analyzes and compares several approaches to studying behavior, including both single-factor approaches (namely those that look at the role of one type of influence, such as genetics) as well as integrative approaches that address how interactions between genetic and environmental factors influence human behavior.

1.2 Summaries of the Chapters

1.2.1 Part I: Introduction

The chapter by Helen Longino examines a variety of approaches to the scientific study of human behavior, arguing that while these approaches may at first seem to be in conflict with one another and thus amenable to comparison, each is in fact partial, focusing only on a subset of causal factors. Longino focuses on those approaches that seek to provide proximate explanations – as opposed to ultimate ones, in Ernst Mayr's terms – thus disregarding evolutionary accounts of behavior in her analysis. Instead, she focuses on single-factor approaches such as behavioral genetics, neurophysiology, and social-environmental research, as well as integrative approaches represented by developmental systems theory (DST) and gene-environment

[3] Interestingly, Helen Longino addresses a variety of scientific approaches to studying human behavior, including behavioral genetic, developmental, and neuroscientific approaches, but intentionally *excludes* evolutionary accounts in her analysis.

interaction (GxE) accounts. Given the partial nature of each approach (or their methodological difficulties, in the case of DST), Longino argues that the best way to view these different approaches is not to compare them to see which produces the correct account, but rather to take a pluralist stance. Furthermore, this pluralist stance ought to be supplemented with pragmatism, whereby one considers what kinds of questions a particular approach actually addresses, in order to make use of the knowledge that results from each of them.

Longino goes on to provide an overview of each approach, as well as some of the more important assumptions underlying them. First is behavioral genetics, which includes both classical, quantitative behavioral genetics (as seen, for example, in twin and adoption studies), as well as molecular behavioral genetics. Both of these research programs aim to identify genetic contributions to behavior, with the former estimating correlations between genetic and phenotypic variation (i.e., heritability) and the other searching for specific genes associated with the trait in question.[4] The second approach Longino discusses is neurophysiology/neuroanatomy, which aims to identify the role of neural structures and processes in behavior. This includes, for instance, studies seeking to find associations between neurotransmitters like serotonin and behavioral traits like depression and aggression. Third, Longino discusses what she calls social/environmental approaches, which aim to understand the role of environmental factors in behavior, including both macro-level variables (e.g., social class and race), as well as micro-level variables (such as family, peers, or media exposure).

Longino points out that disagreements among researchers using different approaches are not about *which* of these factors (genetic, neurophysiological, or environmental) play a role in human behavior, as they'd agree that all of them are important. Rather, they disagree about which kinds of factor are the *most* important, what methods ought to be used to estimate their relative importance, and how various interactions (e.g., between particular genotypes and environments) ought to be accounted for. As Longino puts it, "The debates, then, are less about ontology than about methodology: given that all the factors identified in the various approaches play some role, which approach is likely to be most informative about the etiology of behavior?" As she goes on to show, each approach offers something different.

The assumptions underlying each approach illustrate their partial nature. First, all assume that the behavioral traits being studied are well defined (an assumption that Longino has elsewhere critiqued at length, as she notes in her paper). Second, each of the single-factor approaches assumes that one can legitimately separate the various causes underlying human behavior.[5] Third, and most important for the

[4] Longino mentions Genome Wide Association Studies (GWAS) as an example of a popular method in molecular behavioral genetics. Interestingly, in chapter 3, Turkheimer specifically addresses recent failures of GWAS and discusses why he thinks it is not likely to be successful.

[5] This is an assumption with which Stotz and Allen disagree (see chapter 5).

purposes of taking a pluralist stance, each approach focuses on a limited range of possible causes, largely ignoring certain causal factors. As Longino puts it, "Each approach effectively situates itself in a different causal universe, making comparative assessment impossible."

One might wonder, then, whether the integrative approaches fare any better. According to Longino, while they might at first seem more promising, they encounter serious difficulties. For DST, which considers the developmental system of both the organism and its environment, separation of causes is not possible – the causal space or universe in which DST researchers work is comprehensive. As Longino points out, this is in some sense the correct picture, given the complexity of organisms (especially humans) and the development of behavioral traits. Unfortunately, with respect to human behavior at least, DST is methodologically untenable as most studies of human behavior are non-experimental for obvious practical and ethical reasons. Another, more restricted yet methodologically tractable, approach is the GxExN approach (also referred to as 'gene-environment interaction') introduced by Avshalom Caspi and Terri Moffitt. One of the main questions addressed here is, how do genes and environments interact to affect a particular neurological substrate so as to bring about a particular psychiatric disorder, such as schizophrenia or depression? For a few behavioral traits, Caspi and Moffitt have found that individuals who have *both* a particular genotype *and* have experienced a particular environmental input (such as childhood abuse) are more likely to exhibit the trait in question than those who are subject to only one type of factor. While this approach holds promise, it is also limited to disorders rather than behavioral traits subject to normal variation (such as intelligence or personality), and its findings have proven difficult to replicate.

Based on her analysis of these various approaches, Longino concludes that pluralism is the best stance to take with regard to the study of human behavior. The alternative is monism, which holds that there is only one correct account and that it is possible to figure out which one that is. However, Longino stipulates that it may be possible for many accounts to be correct, as each approach is restricted with respect to the types of explanatory factors that it can invoke, and thus one cannot legitimately compare the various accounts to one another. As she puts it, "The pluralist will propose that our task as philosophers is not to participate in debates about which of these approaches is the correct one, but to understand and help to articulate their scope, their evidential requirements, and their limitations." Longino supplements this pluralist perspective with pragmatism, arguing that the approaches discussed above should be evaluated in light of the practical goals at hand: "Pragmatism, as a second-order sorting procedure, recommends that we evaluate theories and models with respect to the specific questions they set out to answer and the kinds of intervention in the world the answers make possible." Thus, which approach we look to for answers depends on the questions we're asking and the kinds of interventions or policies that we're seeking.

1.2.2 Part II: Genetic Explanations of Behavior

The two chapters in this part of the volume examine behavioral genetic explanations of behavior, with a focus on genetic explanations of human traits. While behavioral geneticists have documented high correlations between genetic and phenotypic variance – i.e., high heritability estimates – for a number of traits like intelligence, extraversion, schizophrenia, and height, the authors point out that high heritability does not indicate how many genes are involved (Turkheimer), nor even that it makes sense to label a trait 'genetic' (Northcott). Turkheimer expands on the former point by examining the failed attempts to locate specific genes underlying heritability estimates, while Northcott draws on philosophical theories of causation to show how and why explanatory context matters in terms of whether we label a trait 'genetic'.

In chapter 3, Turkheimer notes that recent attempts to locate particular genes through Genome Wide Association Studies (GWAS) of height (a highly heritable trait) have not been very successful: only a few genetic variants have been identified, and, taken together, they only account for only about 5% of the total variation. Turkheimer explains this failure – a failure he predicted as part of his "gloomy prospect" – by demonstrating similarities between GWAS and social science and explaining why social scientists are unable to provide general causal explanations of human behavior.

Heritability estimates reflect *associations* between genetic variation and variation in a particular trait for a particular population; however, what we really want to know is whether these associations reflect an underlying *causal* process – hence the search for specific genes. As Turkheimer points out, there has been some success in identifying associations between behavioral traits (e.g., schizophrenia) and genetic variants. However, those associations have been numerous (on the order of half a million in the case of height), small (accounting for a tiny percentage of the variance), and difficult to replicate. Overall, then, they haven't added up to a causal explanation of the trait in question. More fundamentally, though, researchers have had a difficult time sorting out which of these associations are actually causal to begin with. The method used to do such sorting is null hypothesis significance testing (NHST). As Turkheimer explains, however, finding a statistically significant correlation does not guarantee a causal relationship due to the phenomenon of population stratification, where the gene variant associated with the trait in question is also associated with an environmental factor that is the actual cause (chopstick use is the classic example of population stratification). Thus, in such cases, the correlation between the genes and the behavior is a spurious one. Unfortunately, Turkheimer concludes, "NHST has not succeeded in discriminating actual causal processes from spurious correlations and non-causal associations."

Turkheimer goes on to identify an interesting analogy to GWAS in social science: the Environment Wide Association Study, or EWAS, where researchers have tried to identify the specific environmental factors underlying a behavioral trait (such as juvenile delinquency). As with GWAS, attempts to identify such specific factors

have failed in the case of the environment, despite the use of a variety of statistical methods, which Turkheimer documents in detail: multiple regression (in some contexts referred to as Analysis of Covariance, or ANCOVA), Principle Component Analysis (PCA), instrumental variable regression, and propensity score analysis. As Turkheimer shows, every one of these methods is flawed in that they rely on correlations obtained from non-experimental methods, and thus cannot guarantee that any of the identified associations between a genetic or environmental factor on the one hand, and the behavioral trait on the other, indicate a causal relationship.

Despite the problems with traditional social science methods, there are some methods that are able to address the problem of population stratification, namely *quasi*-experimental designs. In behavioral genetics, these are known as within-family designs, and include twin and adoption studies. For example, by comparing the behavioral traits of pairs of monozygotic (identical) twins reared together, behavioral geneticists can obtain estimates of the nonshared environmental variance component – a measure of phenotypic differences that cannot be attributed to having different genotypes or being reared in different home environments. Interestingly, for many traits that have been studied, estimates of nonshared environmental variance have often hovered around 50%, and behavioral geneticists have tried to account for this by identifying specific environmental factors – such as differences in parental treatment, non-overlapping peer groups, unshared experiences, etc. – underlying the variation. However, like the search for specific genes using GWAS, these studies have also largely failed: many associations were identified, but none accounted for more than 2–3% of the variance component.

Turkheimer concludes, then, that GWAS is a social science as it is characterized by the following features: "1) There are a large number of potential causes, individually small in their effects. 2) The causes are non-independent and non-additive. 3) Randomized experimentation is not possible." The problem, Turkheimer goes on to explain, is not that there are many small causal factors (this is true for other areas of biology), but that the effects are interactive and thus nearly impossible to tease apart, which is made even more difficult by the fact that randomized experimentation is neither ethical nor feasible. "The problem lies in the nature of complex human behavior itself," Turkheimer observes, where the causes tend to be local and specific, rather than generalizable. By looking to social science, Turkheimer hopes that we can have a "humbler appreciation for the possibilities" of GWAS.

In chapter 4, Northcott addresses a more general issue, not of identifying specific causal factors underlying behavioral traits, but rather how we decide whether a trait is best thought of as a 'genetic trait'. Of course, as Northcott points out, every trait is a result of a complex developmental process involving a number of genetic and environmental factors. Thus, he asks, "How then can some traits usefully be termed genetic and others not?" His answer, in short, is to develop a relational definition of genetic traits that is sensitive to context; as a result, "no trait is genetic always and everywhere." Rather, whether and to what extent a trait can be counted as genetic depends on the explanatory context.

In order to develop this definition, Northcott draws on the wider causation literature from philosophy of science. In particular, he favors a contrastive theory of

explanation where "a trait is genetic just in case it is explained by genes or it is *not* explained by environment. If genes made the difference, the trait is genetic; likewise, it is genetic if environment did *not* make the difference." Northcott uses the example of a trait, T, where T = his actual two legs (one of which is slightly bent due to a childhood accident). T is appropriately thought of as a genetic trait when the chosen contrast is having just one leg, but not when the contrast is having two straight legs. While this is an example of a specific (token) case, it can easily be applied to more general (type) case by considering a particular population of token cases. For example, Down's syndrome counts as a genetic trait on his definition given that there is no environmental input that could have led to its avoidance (unlike for PKU, for example). Of course, given the phenomenon of gene-environment interaction, genetic differences may only lead to particular trait differences in certain environments; thus, the same trait may or may not fall under the definition of a genetic trait depending on the explanatory context (or population of token cases) that is chosen.

As Northcott points out, one might understandably worry whether this puts too much weight on the choice of contrasts. However, as he goes on to explain, while it *is* the case that whether or not (and in what contexts) a trait is labeled 'genetic' crucially depends on the contrasts that are chosen, what matters is that the definition can be straightforwardly applied, as seen in the examples provided. Northcott concludes that, "Therefore it is not fatal that we have no foolproof algorithm for generating choice of contrast in every context. That merely implies that there may be no fact of the matter regarding whether a trait is genetic before contrasts are specified – which is exactly what a relational definition […] is claiming anyway." In other words, on this account, there just is no context-independent matter of fact as to whether a trait is genetic.

Northcott connects his account to the wider literature on causation in order to illustrate how it relates to previous work on genetic causation as well as to illuminate why labeling a trait 'genetic' might be useful. He asks, "First, consider why we should even care whether a trait is genetic or not. What normative punch could ever result form such a claim? This paper's account, by way of its connection to the causation literature, offers an answer – the counterfactuals that, according to [my definition of a genetic trait], comprise such claims are also exactly those that license *interventions*." For example, by labeling eye color 'genetic', part of what is being claimed is that no salient environmental intervention could change one's actual eye color (just as we saw above with Down's syndrome). On the other hand, while PKU was at once thought to be a genetic trait, Northcott's account suggests that it is probably *not* best described as genetic, given that there is an environmental intervention that makes a difference, namely drastically reducing the amount of phenylalanine in one's diet.

Towards the end of the chapter, Northcott introduces another important distinction: genetic traits versus genetic dispositions. As he explains, there are often contexts in which it doesn't make sense to explain the development of a particular trait *either* in terms of genes or in terms of environments – for instance, when talking about talent. He points to Mozart's musical ability as an example of a genetic disposition,

as it was surely influenced by Mozart's genes but would not have been realized without his unique environment. In such cases, Northcott suggests using the terminology of a genetic *disposition* rather than a genetic trait, where the disposition "is explained (in that context) by both genes and environment but we want to focus attention just on the genes side." According to Northcott, it makes more sense to talk in terms of genetic dispositions in cases where there is a potential for an ability or talent to develop, but where that potential is only realized given a particular set of environmental inputs. He concludes by pointing out that many traits that are subject to disputes as to whether they are appropriately labeled 'genetic', such as alcoholism, schizophrenia, athletic ability, and homosexuality, are actually genetic *dispositions*, and thus that gaining a better understanding of the distinction he has introduced can help in addressing controversies over a number of human behaviors.

1.2.3 Part III: Developmental Explanations of Behavior

The two papers in this section discuss developmental approaches to the study of behavior, emphasizing learning in nonhuman animals. Karola Stotz and Colin Allen address the general conceptual relationship between learning and development, ultimately drawing conclusions not only about that relationship, but also about how behavior ought to be studied as a result. Andrew Fenton focuses his analysis on a particular organism (chimpanzees) and a particular type of behavior (evidence gathering) in order to make specific claims about nonhuman animals' status as epistemic subjects. The arguments presented in both papers have implications for scientific practice – Stotz and Allen's for comparative psychology and ethology, and Fenton's for chimpanzee cognitive studies (as well as cognitive studies of other primates).

In chapter 5, Stotz and Allen aim to examine and clarify the relationships between concepts of learning, experience, and development in the study of animal behavior. In particular, they advocate for research that integrates learning and development such that they're seen not as two separate processes (learning *and* development) but rather as part of one another (learning *as* development).

They begin by summarizing the history of the two main disciplines that study animal behavior: comparative psychology, which is situated in psychology more generally, and ethology, which stems from evolutionary biology and which has split into distinct sub-disciplines including neuroethology, behavioral ecology, cognitive ethology, and evolutionary psychology. Comparative psychologists, they point out, are largely interested in animal learning in controlled conditions, and thus favor laboratory experiments that study acquired behavior, while ethologists focus on species-typical behavior in natural habitats, often conducting field studies to examine innate behavior. Thus, this disciplinary dichotomy maps onto a dichotomy between acquired and innate, a dichotomy that Stotz and Allen reject. Despite comparative psychologists' recent claims to be taking a more integrative approach, the authors argue that they fail to take development seriously, a failure that, according to Stotz and Allen, stems from the separation of psychology from biology (with rare exceptions,

such as developmental psychobiology). Thus, in order to rectify the matter, they call for a "biologically-informed psychology and a psychologically-informed biology."

Stotz and Allen go on to explicate what it means to take development seriously. Doing so, they argue, requires more than just acknowledging the importance of environmental factors; it requires the rejection of the traditional dichotomy between genetic/innate/inherited on the one hand, and environmental/learned/acquired on the other. Taking development seriously includes not just emphasizing the importance of gene-environment interaction, but refraining from drawing any distinctions between those causal factors to begin with. In support of this view, Stotz and Allen discuss various phenomena, such as developmental niche construction (where organisms actively construct their environment) and the role of environmental factors in gene expression.

The view proposed by Stotz and Allen is in line with developmental systems theory (DST), a theory they explicitly advocate and develop, in part by presenting recent scientific findings that were unavailable at the time of DST's introduction and which support and extend the theory. Furthermore, their more explicit goal is to "apply DST's framework to a new pressing question, namely how should one conceptualize the relationship between development and learning." It is just this question the authors take up in the last part of their paper by analyzing and criticizing the old distinction between learning and development.

Learning and development, they argue, are processes that should be assimilated to one another: "From a psychobiological perspective, learning appears as a category within an overall framework of development as the lifelong, adaptive construction of the phenotype out of the interaction between genes, the organism and its environment." To illustrate this, they describe various ways in which epigenetic mechanisms relate to learning and development as integrated processes.[6] Development, they say, is "the process of organismic transformation from a single cell to a differentiated, structured entity," while "learning is a specialized process of (typically neural) differentiation and structural change that supports (adaptive) modification of behavior by experience." Based on these characterizations, Stotz and Allen conclude that "learning is a kind of developmental process: i.e., learning as development" and, likewise, that development itself is a type of learning. Thus, each is a part of the other and should be studied together, rather than as distinct processes.

In chapter 6, Andrew Fenton defends the thesis that chimpanzees are "substantive epistemic subjects." In particular, he looks at two claims in support of this view: first, that chimpanzees display acts of evidence gathering, and second, that they achieve a certain amount of epistemic success in doing so. (While his analysis focuses on chimpanzees, Fenton notes that it may also be applied to other nonhuman animals.) His claim, if correct, has implications for epistemology both in

[6] Stotz and Allen note that most of the evidence for these mechanisms comes from experiments with animals that would not be practical or ethical to do on humans. However, they suggest that the evidence from animal studies is sufficient to warrant looking for epigenetic changes in humans, citing one study already underway that is examining the influence of parental care on child development.

that it offers an alternative to other accounts of animal knowledge (in particular, contemporary reliabilism and the anthropocentric stance), and in that it suggests the need for those working in naturalized epistemology to develop accounts of knowledge that include the epistemic activities of nonhuman animals.

Fenton relates the notion of a 'substantive epistemic subject' to Gould and Gould's account of an active knower, whereby "an organism plays an important role in the acquisition of knowledge (it learns by manipulating/experimenting with its environment)." He goes on to demonstrate how chimpanzees count as active knowers through their activities as evidence gatherers. Through a variety of detailed examples, Fenton shows the following: (1) that chimpanzees must be sensitive and responsive to changes in the social environment, and in doing so must gather evidence about social hierarchies and the like; (2) that they can acquire proto-linguistic, or perhaps even weak linguistic, skills in research settings (e.g., by learning and making use of sign language to communicate with other chimpanzees); and (3) that some of their evidence gathering behavior is analogous to that found in human children.

As Fenton explains, chimpanzee stone tool use includes moments of investigation, and more importantly, is a skill that is not innate (i.e., it is not just an expression of a genetic predisposition) but rather a *learned* behavior requiring particular environmental conditions to come about. In addition, chimpanzees' stone tool use, such as their nut-cracking behavior, requires the presence of "causally efficacious information states," which "enjoy a certain prominence in the [chimpanzee's] noetic structure." Furthermore, the knowledge (or something like it) that chimpanzees obtain from their evidence gathering can be passed on from one generation to the next – a point that Fenton suggests is relevant both to debates about chimpanzee culture and to analytic epistemologists who are interested in social knowledge.

What the detailed examples and subsequent analyses show, then, is that chimpanzees engage in epistemic activities and that these activities "track the accuracy of the relevant information states that inform the subsequent skilled behaviour." Fenton has illustrated not only that chimpanzees are evidence gatherers, but that their epistemic activities (of which evidence gathering is an example) can lead to a certain amount of epistemic success. Assuming his analysis is correct, this lends support to Fenton's original claim that chimpanzees (and possibly other nonhuman animals) are substantive epistemic subjects – a claim that has important implications for both philosophy and scientific practice.

In particular, Fenton notes that epistemologists ought to pay attention to possible cases of nonhuman animal knowledge, such as is illustrated here, and attend to the particular conceptions of knowledge, epistemic standards, and types of epistemic activities that they display. As he puts it, "If, as I have argued, chimpanzees are substantive epistemic subjects, epistemologists should not ignore their epistemic perspectives." Unfortunately, however, this is precisely what many epistemologists do. Furthermore, those who do include the epistemic perspectives of (at least some) nonhuman animals tend to treat them as *second-class* epistemic subjects, and they develop their epistemological accounts based on data drawn almost entirely from human epistemic activities. These are two practices that Fenton would like to see changed.

1.2.4 Part IV: Evolutionary Explanations of Behavior

The chapters collected in the fourth part of this volume focus on a variety of issues, including: the general question of how behavior can be accounted for from an evolutionary perspective (Wilson), evolutionary explanations of the production of art (De Smedt and De Cruz, Verpooten and Nelissen), the general research program of evolutionary psychology (Schulz, Ward), the consequences of evolutionary psychology for our conception of free will (Garvey), and evolutionary explanations of altruistic behavior (Clavien, Klein).

David Sloan Wilson opens the evolutionary part of this volume by describing some of the consequences of what he calls an "intellectual seismic shift" in thinking about human behavior. As Wilson points out, there is a long-standing tradition of thinking about human behavior as being determined by genes that are passed on faithfully from generation to generation. Genetic change takes place over a timescale of hundreds or thousands of generations, that is, on a timescale that is much larger than the duration of individual human lives and a few consecutive generations. Thus, on timescales that matter to us in our everyday lives, behavior can be considered as being fixed. However, Wilson argues, developments in evolutionary science that have been accumulating over the past two decades show that this view of human behavior as grounded in an unchanging genetic basis is mistaken.

Wilson's aim is to show how behavioral and cultural change fall within the scope of evolutionary science and to point to ways in which an evolutionary understanding of human behavior can help us to improve our lives and the societies in which we live. The phenomenon of phenotypic plasticity, i.e., the capability of organisms to change some of their traits in response to changes in the environment in which they live, occupies a central position in Wilson's argument. As Wilson points out, with respect to behavioral traits human beings exhibit a higher degree of phenotypic plasticity than do organisms of other species. This is for two reasons: First, many human behavioral traits are what Wilson calls "rigidly flexible", that is, they have a built-in capacity for providing different outputs in different circumstances. Second, many human behavioral traits can be conceived of as so-called "Darwin machines", that is, they themselves instantiate some form of evolutionary process that enables open-ended adaptation to environmental circumstances. (Although not a behavioral trait, the human immune system is a well-known example of an organismal trait that itself instantiates an evolutionary process based on variation and selection.) These two types of phenotypic plasticity, in combination with the human capability to transmit behavioral changes to later generations by means of cultural heredity and cultural evolution, render human beings highly adaptable to changes in their environments.

However, Wilson observes, the fact that evolutionary processes are a central factor in this adaptive capability of human beings, both in the form of Darwin machines within individual humans and in the form of cultural evolution, harbors both opportunities and dangers. Evolutionary processes can lead to outcomes that are beneficial to the organisms in question, as well as to outcomes that are very harmful.

Therefore, Wilson pleads that we should become "wise managers" of the evolutionary processes that concern human well-being. This management of relevant evolutionary processes, Wilson suggests, should take place by means of providing such environments in which human adaptation and cultural evolution will have the highest chance of producing outputs that further human well-being: "Provide the right conditions and the world can become a better place seemingly by itself. Provide the wrong conditions and even the most heroic efforts to make the world a better place can fail miserably." The idea is that under adverse circumstances human populations tend to evolve traits that are not conducive to human well-being: early reproduction in women and violent behavior in men, for example, are adaptations to highly insecure environments, Wilson argues. Removing such adverse conditions will lead the populations to evolve in directions that are more conducive to human well-being. What Wilson calls for, then, is the elaboration of social policies that are informed by considerations of how evolutionary processes can shape our behaviors and the societies in which we live, using evolutionary thinking for the benefit of humanity. Social policies should be aimed at providing environments that allow human populations to evolve desirable traits.

In chapters 8 and 9, a particular kind of human behavior is examined, namely the production of art. Johan De Smedt and Helen De Cruz, in chapter 8, explore the opposition between two kinds of evolutionary explanations of artistic behavior in humans: explanations that understand artistic behavior as an adaptation and explanations that see it as a byproduct of adaptations that evolved for different functions. De Smedt and De Cruz examine the evidence in favor of and the difficulties that arise with respect to both kinds of explanations and argue that in each case the problems are too large to accept the explanations in question.

In the case of adaptationist explanations, at least three problems occur. First, adaptationist explanations of artistic behavior seem all too easy to find and thus are faced with the question whether they are more than "just so" stories. Second, often such explanations are not focused precisely on artistic behavior, but attempt to explain a much broader range of behaviors, including rituals, imagination, humor, etc. That is, they don't explain artistic behavior as an adaptation, but as one aspect of a much more encompassing adaptive behavioral trait. Third, if artistic behavior is an adaptation with its own selective history, it would have to be rooted in a separate mental "art module". The modular organization of the mind, however, still is a highly problematic issue and it remains unclear to what extent the mind can actually be divided up into independently evolved modules. In this context De Smedt and De Cruz discuss results of recent neurobiological studies that raise doubt about the existence of a separate mental "art module".

Byproduct explanations of artistic behavior, however, fare no better than adaptationist explanations. According to one theory, for example, works of art appeal to human aesthetic and emotional preferences that evolved in relation to other functions. A problem with this theory, however, is that many artworks in fact don't seem to do this: De Smedt and De Cruz mention the sometimes haunting paintings by Francis Bacon as an example. In addition, the production of artworks costs considerable time and energy on behalf of the makers, the investment of which would make sense

only if artistic behavior would serve a clear function (and would be an adaptation) but doesn't seem to make sense for a mere byproduct.

As an alternative to evolutionary explanations of artistic behavior as an adaptation or as a byproduct of other adaptations, De Smedt and De Cruz propose an account of artistic behavior as a product of cultural group selection. Within this framework, they explore two theoretical options: artistic behavior as a marker for altruistically/cooperatively inclined members of society (so-called "green beards") or as a marker for ethnicity, that is, for adherence to a particular set of sociocultural norms. De Smedt and De Cruz examine archaeological evidence in order to make a case for the latter option, but emphasize that probably "no silver bullet theory will be able to successfully explain all forms of art production."

In chapter 9, Jan Verpooten and Mark Nelissen present an account of the evolutionary origins of art that opposes the view presented by De Smedt and De Cruz. They address how artistic behavior can be evolutionarily explained and draw attention in this context to the importance of a particular model from sexual selection theory about the selection of signals between potential mates. They review two categories of models in sexual selection theory that can be applied to the evolution of artistic behavior, namely indirect benefit models and sensory exploitation models. According to indirect benefit models, females select males with particular traits that indicate the presence of beneficial traits in the males that they can pass on to their (and the selecting females') offspring. Such selection practices by females are indirectly selected, as they hitchhike on the direct selection of these beneficial genes. According to sensory exploitation models, particular traits may evolve if they appeal to sensory preferences of organisms that are actually aimed at different phenomena. As an example, Verpooten and Nelissen mention the evolution of orange spots in guppies: female preferences for orange food items lead males exhibiting orange spots to be more attractive to these females and thus to higher reproductive success for males with orange spots.

Although indirect benefit selection and sensory exploitation selection are usually seen as intertwined, Verpooten and Nelissen argue that at least some of the sensory biases that are found in nature might be the products of sensory exploitation selection alone. They criticize evolutionary explanations of artistic behavior framed in terms of indirect benefit models (an account proposed by Miller and one proposed by Boyd and Richerson) for underestimating the role of sensory exploitation in the evolution of artistic behavior. On the account that Verpooten and Nelissen propose, sensory exploitation is a central factor in the evolution of artistic behavior. Given that on sensory exploitation models traits evolve by means of exploiting sensory biases that evolved for different purposes, on such an account art must be understood as a spandrel, that is, a byproduct of other evolved traits – a view that clearly contrasts with the view that is defended in the preceding chapter.

Whereas chapters 8 and 9 focus on the evolution of art, chapters 10, 11, and 12 examine the research program of evolutionary psychology. In Chapter 10, Armin Schulz examines a particular strategy that evolutionary psychologists use to legitimate their approach to studying the human mind. Evolutionary psychology, especially the "strong" research program as propagated by Leda Cosmides, John Tooby, David

Buss, and others (sometimes called 'Evolutionary Psychology' with a capital 'E' and 'P'), is regularly criticized for being too speculative in nature to be able to provide a useful contribution to the science of psychology. In particular, critics often point out that explanations in evolutionary psychology often lack a sufficient evidential basis, such that the research program rests too much on "just so" stories when trying to account for particular mental phenomena. Evolutionary psychology, critics argue, fails to provide good explanations of the sort that evolutionary biology does. One strategy of evolutionary psychologists to defend their approach is to claim that evolutionary psychology does not use evolutionary theory as a basis for explaining mental phenomena, but rather uses evolutionary theory as a heuristic tool. This defense, if adequate to the actual situation in evolutionary psychological research, would defuse the arguments of those critics who see evolutionary psychology as crucially resting on unscientific "just so" stories.

Schulz observes that this response of evolutionary psychologists to their critics is insufficiently supported, as no cases have been presented so far that unequivocally show that evolutionary theory serves as a heuristic tool in evolutionary psychology research. Schulz thus undertakes to examine the feasibility of this response, looking at evolutionary psychology in general (rather than just focusing on the "strong" program mentioned above). Interestingly, Schulz reaches diverging conclusions: On the one hand, taking Cosmides and Tooby's explanation of cheater detection as an example, it turns out that standard examples of evolutionary psychology research do not in fact use evolutionary theory as a heuristic tool. Thus, there is strong evidence that the evolutionary psychologists' defense fails and the critics of evolutionary psychology are right. On the other hand, however, Schulz shows that cases in which evolutionary theory plays a heuristic role can be found within evolutionary psychology, although such cases are comparatively rare. Schulz presents one such case, namely Gergely Csibra and György Gergely's work on natural pedagogy. This case shows that there are heuristic usages of evolutionary theory in evolutionary psychology, Schulz argues, although explanatory usages – i.e., those that are subject to severe criticism – are much more common.

In Chapter 11, Chuck Ward aims to deepen a widespread criticism of the "strong" program of evolutionary psychology, namely that one of the program's core assumptions – the assumption that the basic features of the human mind constitute adaptations to the Pleistocene environment in which our ancestors lived – cannot be upheld. Ward explores development-based criticisms of the "strong" program that focus on the phenomenon of neural plasticity, i.e., the phenomenon that the neural structures of organisms' brains can change in response to the environments in which they live, their experiences and their actual behavior. Ward reviews evidence for the existence of neural plastic responses of the human brain to environmental cues and argues that this evidence suggests a way of explaining human cognitive processes that constitutes an alternative to the explanation that understands these processes as adaptations to life in a Stone Age environment.

Ward considers how various authors have used the phenomenon of neural plasticity as a general argument against the "strong" program in evolutionary psychology and examines two kinds of human behavior more closely: reading and writing, and

musical training. These kinds of behavior constitute paradigmatic examples of culturally-mediated behaviors: behaviors that are inherited between generations because individual human beings are embedded in a common cultural environment in which they grow up and in which their neural structures develop. In recent empirical research, evidence has accumulated that practices of learning to read, to write, or to play musical instruments induce physical changes in the brains of humans involved in such practices. As Ward argues, "these examples demonstrate the existence of processes that can serve to introduce and reliably propagate modifications in our cognitive architecture *without genetic change*."

The existence of such processes, then, is inconsistent with the core assumption of the "strong" program in evolutionary psychology that contemporary human cognitive architecture has originated in the Pleistocene in the form of genetically-based adaptations and has been propagated to present-day humans by means of genetic inheritance. Contrary to the claim of proponents of the "strong" program in evolutionary psychology, Ward concludes, this program's way of explaining human cognitive traits is not the only game in town.

Brian Garvey also examines the "strong" program in evolutionary psychology, in chapter 12, but does so in relation to the issue of free will. Garvey argues that the modularity of mind that the "strong" program in evolutionary psychology assumes constitutes an obstacle to free will. According to Garvey, this obstacle is comparable to such restrictions on free will as addictions, compulsive behaviors, etc. – that is, factors that compel people to act in particular ways even if it is in principle possible for them to act differently.

Garvey discusses a number of accusations of sociobiology and the "strong" program in evolutionary psychology that interpret these research programs as implying that, if they are correct, our will is less free than we think it is. The standard defense against such accusations is compatibilism: the position that even if human actions are determined – in this case, by the makeup of our brains – it does not imply that they are not free. That is, research programs such as sociobiology or "strong" evolutionary psychology that pursue the reduction of mental phenomena to biological or physiological phenomena can be right while still leaving the possibility of having free will. But, Garvey argues, even if this compatibilist answer to the aforementioned accusations is accepted, it still may be the case that these programs provide other reasons for thinking that the human will is less free than we would think or hope.

In the case of the "strong" program in evolutionary psychology the culprit is the program's massive modularity thesis. According to this thesis, the human brain is made up of hundreds or thousands of modules, each of which has evolved in response to its own selection pressures, that is, each of which has evolved as a solution to a particular environmental problem. Moreover, proponents of the "strong" program in evolutionary psychology hold that the relevant evolutionary events occurred in the Stone Age, as the human brain has not undergone much further evolution since that time. Thus, the modules in the human brain constitute adaptations to the Stone Age environments in which our ancestors lived. Due to their being adaptations to particular environments, our brain modules make us act in ways that fit these environments. But, as Garvey notes, "what was adaptive in the Stone Age need

not be adaptive now, and nor need it coincide with what we want now." Thus, if evolutionary psychologists who endorse the "strong" program are right, the makeup of the human brain causes us to have less free will than we thought we did: we are tuned to act in particular ways that might have been suitable to a different environment, but which can be thought of as a kind of compulsive behavior in our present one. In the end, however, Garvey places the burden on the proponents of approaches like the "strong" program in evolutionary psychology. As Garvey points out, while proponents of such approaches often claim that humans have free will after all, as they can override those desires that have been inherited from our Stone Age ancestors, they fail to give an account of how such compulsions might be overridden. There might thus be a way out for proponents of such approaches, but only if they provide us with the required account. As long as this issue has not been resolved, Garvey concludes, we can legitimately suspect that the "strong" program in evolutionary psychology has negative implications for our notions of free will.

In chapters 13 and 14, Christine Clavien and Rebekka Klein both address the question of how altruistic behavior can be explained against the background of human motivations that are often directed toward one's own interests and self-satisfaction. Clavien examines the opposition between two deeply entrenched positions in psychology and in the philosophy of psychology with respect to the question of whether humans are capable of behaving altruistically. On the one hand there is the position of psychological egoism, that is, the claim that human beings never perform genuinely altruistic actions; although human actions may appear to be altruistic and may have positive effects for others, humans ultimately always act in ways that are directed at their own interests. On the other hand there is psychological altruism, which holds that human beings are capable of performing genuinely altruistic actions; that is, while not all apparently altruistic actions are indeed selfless, at least some can be conceived of as being genuinely altruistic.

Clavien analyzes the long-standing debate between advocates of these opposing positions and argues that so far the debate has been carried out in an unfruitful manner: the way the debate is usually framed leads to a deadlock between the two positions, she argues, in which it is not possible to decide in favor of either of the two competitors. The central notions in Clavien's analysis of the debate are the notions of 'motive' and 'motivation'. As Clavien points out, psychological altruism and egoism are claims about the motives (i.e., psychological states, such as desires, intentions and judgments) of people that underlie their actions of helping other people. But such psychological states are extremely difficult to access and both sides in the debate can always take recourse to unconscious motives, Clavien observes. In other words, defenders of psychological egoism can always argue that, even if test persons report having motives only aimed at the interests of others, and experiments do not reveal any egoistic motives underlying apparently altruistic actions, what *ultimately* underlies the actions under consideration are unconscious egoistic motives that just fail to come to the surface. Defenders of psychological altruism can take recourse to a similar line of argumentation. Thus, "the debate over altruism cancels itself out in a battle of a priori statements," that is, a priori assumptions about empirically non-accessible, unconscious psychological states.

The way out of this deadlock, Clavien suggests, is to frame the debate in terms of the relational notion of 'motivation' rather than the notion of 'motive'. A motivation is an affective state that causes someone to act; it may be based on a motive, but also on an emotion, a sensation, etc. As affective states, motivations are empirically accessible and, therefore, probably better suited to break the deadlock than are motives, which, after all, aren't necessarily empirically accessible. Framing the debate in this way enables us to examine the role of altruistic emotions in causing apparently altruistic actions. As a consequence, evolutionary arguments can enter the debate, which, Clavien argues, is then decided in favor of psychological altruism.

Chapter 14 focuses on a particular aspect of altruistic behavior, namely punishment. In this chapter, which constitutes a bridge between Parts IV and V of this volume, Klein connects evolutionary and neurobiological explanations of altruistic and cooperative behavior with the question of how punishment may be evaluated from a moral point of view. In contrast to the psychological notion of altruism (which features in Clavien's chapter) and the biological notion of altruism (which measures the altruistic content of an animal's action in terms of its effect on the Darwinian fitnesses of the animal itself and of other animals), Klein focuses on the economic notion of altruism (measured in terms of the costs and benefits for the acting individual that are entailed by the action). Klein reviews results from behavioral experiments in experimental economics and research into the evolution of social cooperation that shed light on norm-enforcing practices such as altruistic punishment (i.e., acts of punishment which entail benefits for future partners in social interactions but not for the individual who performs the act of punishment, and thus function to police social interactions).

While evolutionary explanations of altruistic punishment explain why such behavior has become widespread throughout the human population, they do not explain how an individual's motives and motivations may cause such behavior. In order to clarify this matter, Klein reviews neurobiological studies of altruistic punishment behavior which suggest that such behavior is driven by hedonic motivation and thus is connected to natural selection for the avoidance of pain and other unpleasant states. It now looks like we may have good neurobiological and evolutionary explanations of why people exhibit altruistic punishing behavior that explains such behavior as being rooted in personal motivations that are subject to selection and effects that benefit social cohesion in the group in which the individual lives (although further research is needed here). However, Klein argues, it would be too quick to value altruistic punishment as generally beneficial to society on this basis, because the personal motivations underlying punishing acts may be aimed at the welfare of society (e.g., satisfaction with keeping up the norms of society) but may also be aimed elsewhere (e.g., mere desire for revenge). Thus, Klein points out, punishment cannot generally be judged as a morally good behavior, even if the neurobiological and evolutionary explanations suggest that it furthers the welfare of society. Rather, individual acts of punishment need to be assessed by themselves, while taking into account the personal motivations underlying the act in question.

1.2.5 Part V: Neurobiological Explanations of Behavior

The two chapters collected in the final part of this volume address the question of how organismal behavior can be explained from the perspective of neurobiology. While the first of these chapters examines how neurobiologists individuate traits in need of explanation, the second chapter examines different ways of investigating brain activity and different modes of explanation associated with them.

In Chapter 15, Marcel Weber addresses two important philosophical problems that arise in relation to behavioral biology. The first is the general problem from philosophy of science regarding the theory-ladenness of observation. The second is a particular problem from the philosophy of biology, namely the problem of how organismal traits are to be individuated. In behavioral biology, Weber argues, both problems arise in connection with the identification of the explananda of behavioral biology. Behavioral biology, one might say, aims at explaining behavioral traits that organisms exhibit. But what, exactly, is a behavioral trait? What elements does a particular behavior consist of; what should be counted as part of the explanadum and what as not being a part of it; and, in particular, what determines what kind of behavioral trait a trait in question is? Organismal traits aren't simply given, but biologists have to individuate and classify them before being able to study and explain them. For behavioral traits, this is especially difficult, as behaviors often involve different parts of the organisms and often are spread out over longer periods of time.

Weber considers three ways of individuating and classifying behavioral traits: the intentional stance (according to which behavioral traits can be individuated by ascribing intentions to the animals exhibiting these traits), using proper functions to individuate traits (according to which behavioral traits can be individuated as traits that perform particular causal roles for the organisms exhibiting them, where these causal roles are the causes of the traits' presence), and using the notion of homology to individuate traits (where behavioral traits can be individuated on the basis of shared ancestry). Weber fleshes out these ways of individuating behavioral traits in more detail and ends up with five distinct theoretical notions that might constitute the basis for trait individuation and classification, but concludes that none of these does its job sufficiently well. Thus, an alternative account is needed.

Weber develops this alternative account by focusing on the notion of biological function, which in this case is conceived of by means of a version of the causal role account of functions. According to Weber, behavioral biologists (as well as biologists in other fields of work, such as experimental biology) individuate and classify the organismal traits they study on the basis of the functions that these traits perform. Here, the function of a trait is conceived of as being what a trait does or what it is capable of doing (its capacity) in the context of an encompassing system of which that trait is a part. More specifically, a trait's function is its contribution to realizing the function of the system of which it is a part. That system's function, in turn, is to be analyzed in the context of a larger encompassing system, until we reach the level of the organism. In the end, all functions are analyzed in the context

of the self-reproduction of the whole organism: a trait's function (on the basis of which the trait is individuated and classified) is what it does (or is capable of doing) to contribute indirectly or directly to realizing the self-reproduction of the organism that exhibits these traits. This, Weber holds, is what makes a function into a *biological* function.

Weber supports his case by examining how behavioral traits are individuated in the study of the nematode *Caenorhabditis elegans*. *C. elegans* worms exhibit social feeding behavior which can be explained neurobiologically by means of the effect of a particular neurotransmitter in the context of the operation of a regulatory mechanism that responds to environmental stimuli. But, Weber argues, the explanandum that is being tackled neurobiologically was identified as a kind of *behavior* in the first place (instead of, for example, an instance of simple surface adhesion) by referring to the trait's function in the light of the organisms' self-reproduction. That is, while neurobiologists were examining a particular phenomenon (the clumping of *C. elegans* worms under particular conditions), it wasn't clear from the outset that this was a behavioral phenomenon – this became clear only upon consideration of the phenomenon in terms of biological functions. This way of individuating and classifying behavioral traits, however, as Weber points out, allows explananda to be changeable: neurobiologists don't just pick out phenomena in need of explanation and go on to explain them, but modify the explananda along the way.

In the volume's final chapter, Adele Abrahamsen and William Bechtel consider two different perspectives on brain and neural system activity, the reactive perspective and the endogenous perspective, and connect two modes of explanation with them. The reactive perspective on brain and neural system activity focuses on how neuronal systems respond to external stimuli. Endogenous brain and neural system activity, in contrast, is activity in absence of stimuli from the outside. As Abrahamsen and Bechtel point out, psychological and neuroscientific research has for the most part taken the former perspective, presenting test persons with specific stimuli and investigating the activity of the brain and neural system that resulted in response to these stimuli, and disregarding the endogenous activity of the brain and the neural system.

Abrahamsen and Bechtel argue, however, that both perspectives have a long history in neuroscientific research, tracing back to the late nineteenth / early twentieth century. They provide a rich historical overview of empirical research that has been done under the two perspectives, showing how the endogenous perspective has become increasingly prominent in recent neuroscientific work. In particular, Abrahamsen and Bechtel discuss recent research on endogeneous brain activity, with the aim of showing that the lack of interest in the endogeneous perspective that many neuroscientists exhibit is unwarranted. As Abrahamsen and Bechtel point out, researchers taking the reactive perspective on neuroscientific research tend to downplay the significance of the results achieved under the endogeneous perspective, treating endogeneous activity of the brain and neural system as noise rather than useful information. However, Abrahamsen and Bechtel argue, "clearly the time for dismissing the endogenous activity as mere noise has passed."

Abrahamsen and Bechtel frame the importance of the two perspectives on neuroscientific research in terms of two modes of explanation that they provide. Both

are kinds of mechanistic explanations and thus fit well into what today is often called the New Mechanistic Philosophy. According to the New Mechanistic Philosophy, explanation in science often proceeds by specifying a mechanism that is capable of bringing the explanandum about. While proponents of the New Mechanicism endorse diverging conceptions of what exactly mechanisms are, Abrahamsen and Bechtel hold a specific view of what a mechanism is. They argue that two types of mechanistic explanation can be distinguished: basic mechanistic explanations, which explain by specifying the parts of a system, their organization, and the sequence the system goes through on its way to a final state from a particular initial state; and dynamic mechanistic explanations, which also include specifications of patterns of change in time that a system might exhibit. On Abrahamsen and Bechtel's account, which is an account of dynamic mechanistic explanations, a mechanism is "a structure performing a function in virtue of its component parts, component operations, and their organization," where "the orchestrated functioning of the mechanism, manifested in patterns of change over time in properties of its parts and operations, is responsible for one or more phenomena."

Abrahamsen and Bechtel conclude that the brain should be understood as an endogeneously active mechanism that is perturbed by stimuli, i.e., a system that changes its activity due to both its internal dynamics and its external perturbations; they end their chapter by arguing that the conception of mechanisms that they advocate best fits the specificities of this view of the brain. Neurobiological explanations of behavioral phenomena, then, are constructed by specifying the parts of a particular neural structure responsible for bringing about the explanandum, the properties and operations of these parts, as well as how changes in these relate to the explanandum under consideration. In such explanations, Abrahamsen and Bechtel argue, both reactions of the neural structure to external stimuli and the internal dynamics of the neural structure should be taken into account.

Acknowledgements We want to thank the 19 contributors to this volume, as well as the nearly 40 referees who generously spent their time and energy evaluating and commenting on the chapters that were submitted for publication. Also, we are indebted to those institutions who helped make possible the conference that led to this volume: the German Research Council (DFG), the Lower Saxony Ministry of Science and Culture, the Leibniz Universität Hannover Fund for Internationalization, the Center for Philosophy and Ethics of Science (ZEWW) at the Leibniz Universität Hannover, and the Minnesota Center for Philosophy of Science (MCPS) at the University of Minnesota. In addition, we're grateful to the conference participants for providing excellent presentations and stimulating discussions, as well as to an engaging audience.

Chapter 2
Knowledge for What? Monist, Pluralist, Pragmatist Approaches to the Sciences of Behavior[1]

Helen Longino

> *With the greatest of hubris, quantitative behavior genetics strives to traverse the molecular and psychological levels in one grand inferential leap.*
>
> (Wahlsten & Gottlieb, 1997)

> *Complex developmental processes, …, are not amenable to any microanalysis we currently know how to conduct. … [T]hus mechanistic science is unlikely to yield useful information about complex behavioral problems, ….*
>
> (Scarr, 1995)

2.1

I have been conducting a comparative epistemological and social analysis of research approaches in the sciences of human behavior. In this study, which involves analysis of research reports in journals and at seminars and conferences, meta-analyses, polemical exchanges among the researchers, and public media representations of the research and its implications, I have looked primarily at what might be dubbed, after Ernst Mayr's distinction, proximate forms of explanation. That is, I

[1]This talk was given as a keynote lecture at the Biological Explanations of Behavior Conference, Hannover, Germany, June 12-15, 2008. A revised, but similar, version was given as a keynote lecture at the Conference of the Society of Philosophy of Science in Practice, Minneapolis, MN, June, 2009. For more detailed discussion of the issues broached, readers are referred to my forthcoming monograph on understanding the sciences of behavior. I am grateful to the editors and to anonymous referees for constructive suggestion for revision of this manuscript.

H. Longino (✉)
Department of Philosophy, Stanford University, Building 90, Stanford, CA 94305-2155, USA
e-mail: hlongino@stanford.edu

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*, Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_2,

have excluded evolutionary approaches to behavior.[2] Among these proximate forms of explanation, I have investigated both single factor approaches – genetic, neurobiological, social-environment – and integrative approaches – what is known as developmental systems theory as well as a more limited approach dubbed the GxExN approach. In this essay I update arguments I have elsewhere offered for adopting a pluralist stance towards this multiplicity of approaches, but further argue that pluralism alone leaves us without a way of making use of the knowledge generated by the different approaches. Pluralism must be supplemented by a form of pragmatism that attends to what kinds of question a given approach can answer together with what kinds of question our practical experience makes salient.

2.2

Behavior genetics divides into quantitative behavior genetics (also referred to as classical behavior genetics) and molecular behavior genetics, the former drawing on methods of population genetics, the latter drawing on molecular biology. Both are interested in identifying genetic contributions to behavior. Quantitative behavior genetics attempts to correlate variation in the expression of some trait in a population with genetic variation in that population. It is interested in the question: how much of a given behavior of interest B is heritable, which translates into the question: how much of the difference in expression of B among individuals in a population is correlated with genetic difference in that population? The methods involve finding behavioral correlations and variation in correlations among biologically related individuals, and trying to separate genetic from environmental influence by studying adoptees and twins separated at birth or shortly thereafter. For example, a twin study examining a broad range of behaviors examined concordance in measures of antisocial behavior in 331 twin pairs raised together and 71 reared apart. Behaviors were identified through a self-report questionnaire (the MMPI) and included two sets of questions measuring antisocial or aggressive behavior. The concordance in answers among the twins reared apart supported a heritability estimate of .8.[3] Quantitative behavior geneticists extend their methods with a variety of techniques, including longitudinal analyses that address the question about the stability or mutability of genetic influence on a given behavior over time.

One of the values claimed for quantitative behavior genetics is that when some genetic influence is suggested by family concordances or correlations, the behavior becomes a candidate for analysis by molecular genetics whose aim is to find associations between phenotypic traits and sets of specific genes or gene regions.

[2]Proximate and ultimate (or evolutionary) explanations are answers to different kinds of question (ontogenetic and phylogenetic, respectively) and so not susceptible to the kind of comparative analysis I am conducting.

[3]Tellegen, et al. (1988). Twin study heritability results included in a meta-analysis performed by Mason and Frick (1994) range from 0 to .84.

The questions asked by molecular geneticists concern whether genetic markers, which are multi-allelic gene regions whose frequency can be observed relatively readily, can be associated with the incidence of B in a given pedigree or family lineage. The finding of markers associable with phenotypic traits suggests that a gene in the vicinity of the marker is causally influencing the incidence of the trait. In the early 1990s, 14 male volunteers of a Dutch family, all of whom experienced episodes of aggressive behavior, were found also to share allelic variation on a region of the X chromosome coding for the enzyme monoamine oxidase (or MAOA).[4] This enzyme is involved in the metabolic cycle of serotonin. The Brunner study stimulated much concern over possible genetic intervention and genetic discrimination. This has subsided and studies of the roles of irregularities of MAOA related genes and of other genes related to aspects of serotonin metabolism have proceeded apace. The investigative technologies available for studying the genome are advancing rapidly, and such techniques as Genome-Wide Association Study raise the hope that more gene regions can be identified.

Neurophysiology and neuroanatomy are interested in identifying the role neural structures and processes play in behavior.[5] One intensely studied aspect of neurophysiology has been the serotonergic system: the set of processes involved in the diffusion and reuptake of the neurotransmitter, serotonin. Variation in serotonin concentrations, in number and distribution of serotonin receptors, and in serotonin reuptake has been associated with a number of psychological/behavioral phenomena from depression to suicidality to aggression. As is often the case with physiological research, after initial findings of a relationship of some substance or process to a higher level trait, these lines of investigation initially created more puzzles than they solved. Research in the 1990s sought to elaborate the mechanisms of involvement and separate out possible physiological confounders. Was the culprit decreased serotonin production or diminished uptake of serotonin? To address this question, one study of ten subjects and five controls by Emil Coccaro and colleagues investigated the possible involvement of serotonin receptors in the causal pathway.[6] Researchers administered a serotonin antagonist that would block the serotonin receptors to the subjects but not to the controls. They then administered an agent, buspirone, that physiologically mimics serotonin. Receptor sensitivity was assessed by measuring prolactin levels before and after administration of buspirone. Prolactin is released when serotonin or one of its agonists bind to serotonin receptors. Lower levels of prolactin have been associated with higher levels of aggression/irritability. Prolactin levels in subjects whose receptors were blocked were lower in relation to individual baselines than in controls. This experiment implicates receptor function rather than serotonin production in serotonin's behavioral effects.

[4]Brunner, et al. (1993). Five members of the family exhibited extreme levels of violence, while nine others exhibited more moderate, but still higher levels of violence.

[5]I deliberately use the broad locution, "play a role in", and avoid causal locutions such as "produce" as there are very different kinds of causal relation that can be investigated. And in the case of neurophysiology, there is a very live question as to whether what is investigated is causation or constitution.

[6]Coccaro, Gabriel, and Siever (1990).

Other kinds of question addressed in this research approach include whether the neural processes associated with a behavior are distributed or local,[7] with what other neural and organic processes the processes associated with the behavior interact, and so on. Neurobiological research also includes the use of the various neuroimaging techniques, as well as autopsy, to identify neural and brain structures involves in various behaviors.

Social/environmental approaches seek to understand the role that environmental and other exogenous factors play in a given behavior. They may investigate the role gross or macro-level social variables (social class, ethnic, racial, and cultural identity, urban/suburban, immigrant/native, etc.) play in the expression/frequency of a behavior of interest. They may investigate the role micro-level variables such as family, school, peers, media exposure, play in the expression of the behavior. Does one of these predominate in its expression? Other research questions include: Do micro- and macro- level variables interact in the expression of the behavior? If so, how? How do differences within a family influence the expression of B by its members? They may employ large databases such as are made available from courts and other governmental institutions, or may conduct more fine-grained laboratory observation of behavior.

In one study, Cathy Widom and colleagues employ the first strategy in efforts to link adolescent and adult violent and antisocial behavior to abuse in childhood. In one of their studies, they compared the records of 416 adults with histories of physical and sexual abuse in childhood with those of a control group of 283 adults with no documented history of abuse.[8] The rate of antisocial personality diagnosis in the group with histories of abuse was 13.5% as compared with 7.1% in the control group. The researchers conclude from this (and other studies in similar vein) that experience of abuse as a child is a significant causal factor in adult violence, and that special prevention efforts directed towards victims of abuse could reduce later criminal behavior.

In a study conducted at a finer level of granularity, researchers sought to correlate familial interaction patterns with long-term disruptive behavior in eight and nine year old boys.[9] The boys chosen for the study were identified by teachers who completed Social Behavior Questionnaires on their students. Interactions in 44 families were studied by observing the parents and child in question engaged in joint tasks in the researchers' laboratory. Observers used checklists in rating dyadic interactions between father and child, mother and child, and between the parents. Researchers found that negative behaviors (such as verbal abuse or attacks) and positive behaviors (such as endearments) in the parent-child dyads were not reciprocal, but that negative behavior of one parent toward the boy was correlated with negative behavior on his part toward the other parent. In addition, negative behavior of boys toward their mother was correlated with fathers' negative

[7] A distributed process being one that involves neuronal structures throughout the brain, while local ones are specific to a single region or even a single neuron.

[8] Luntz and Widom (1994).

[9] Lavigueur, Tremblay, and Saucier (1995).

attitudes toward their female spouses. The researchers speculate that coaching the parents in alternative styles of interaction could reduce the chances that their child's disruptive behavior will later develop into more serious anti-social behavior.

Most researchers accept that observable behaviors are outcomes of interactions among all these factors. The points of contention concern not whether any of these factors are real or contribute, in some way, to a given behavior, but 1) which predominate, 2) how to quantify their relative contributions to behavioral outcomes and 3) how to represent the interactions among them. Hence, researchers don't need an argument that one or another factor plays a role, but rather a way of measuring and calculating their respective roles. Competition among / uncertainty about the approaches concerns whether any one has the tools required to calculate values for the factors stressed by the others. The debates, then, are less about ontology than about methodology: given that all the factors identified in the various approaches play some role, which approach is likely to be most informative about the etiology of behavior?

2.3

All approaches must assume that the traits under investigation are well-defined. By this I mean that the traits have clear criteria of identification, of operationalization, and of measurement. This may seem a trivial requirement, but I have elsewhere shown that this assumption is not satisfied in the case of aggression or of sexual orientation, two families of behavior that have received extensive study.[10] Because the research interest consists in understanding relatively enduring traits, the object of investigation is dispositions to behave in certain ways in certain conditions, rather than episodes of behavior. Episodes are taken to be indicative of dispositions.

More to the point for the present analysis, all select from a range of possible types of cause. This range is what I call the potential causal space, or space of potential causes, and it can be displayed in a grid, as in Fig. 2.1.

Genotype 1 [allele pairs]	Genotype 2 [whole genome]	Intrauterine environment	Physiology [hormone secretory patterns; neurotransmitter metabolism] Anatomy [brain structure]	Non-shared environment [birth order; differential parental attention; peers]	Shared (intra-family) environment [parental attitudes re discipline; communication styles; abusive/nonabusive]	Socio-Economic Status [parental income; level of education; race/ethnicity]

Fig. 2.1 Undifferentiated causal space

The specificity of assumptions informing and shaping the individual research approaches and the methods of observation and measurement they employ means that this range or space of potential causes, all members of which are implicitly agreed to play some role, is only partially activated in any given research approach.

[10] Longino (2001) and forthcoming.

These assumptions, it should be stressed, are not explicit, but rather assumptions required to confer evidential import on the data.

Assumptions of the behavior genetic approach include the following:

1. The causal contributions of genes to the inculcation of a behavioral disposition are separable from other causal influences on the inculcation of that disposition, that is, given that there is interaction between genetic and non-genetic factors, it is possible to distinguish their respective contributions to the variation in the disposition to exhibit some particular behavior.
2. Conversely, the effects of genes are separable from effects of other factors, that is, it is possible to distinguish at the phenotypic level what, or how much, of a trait is produced by genetic factors and how much by non-genetic.
3. Heritability is an appropriate measure of that genetic contribution, that is, appropriately designed studies of variation in the expression of behaviors in stipulated populations, will reveal the genetic contribution to variation in those behaviors.[11]

Other assumptions, built into the methods of heritability studies (twin and adoption research that attempts to separate similarity of genetic structure from similarity of rearing environment), include:

4. The available causal space can be represented as including genetic and environmental causes (with a noise factor built in to the equation).
5. The environment is distinguishable into shared and non-shared environment, thus accounting for variation accounted for neither by genetic factors or by shared environmental factors.

These assumptions mean that the causal space open to investigation by the methods of classical behavior genetics takes the form of Fig. 2.2:

Diagonal lines = active space (although in principle could include features of shared (intra-family) environment, in practice these are not taken into account or are subsumed under the SES categories)

Solid black = inactive space (empty: either randomly distributed or effect of genotype)

Fig. 2.2 The causal space for behavior genetics

[11] There is a certain amount of equivocation in the representation of conclusions from heritability studies, a slide from thinking about the genetic contribution to *difference in a population* in the expression of a trait to expression of a trait *simpliciter*.

Molecular geneticists first identify a population both sharing a trait and likely (by pedigree, or familial, analysis) to share genetic configurations. They then, using additional hints provided by the pedigree analysis, seek evidence of shared allelic variation. Assumptions of this approach include:

1. The base rate of the trait in the general population is both determinable and high or low enough to establish significance of the allelic variation correlated with trait variation in the sample.
2. The sample size in any particular study is sufficient for the detection of relevant allelic variation.
3. The causal space of interest is the variety of possible alleles and/or the whole genome.

This yields the following selection (Fig. 2.3) from the grid:

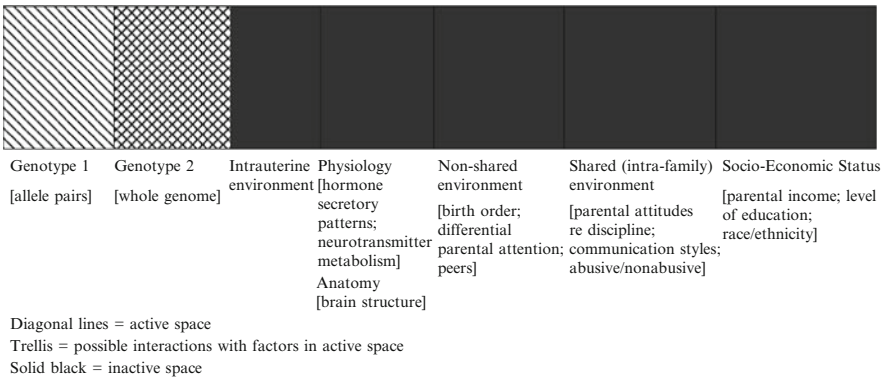

Fig. 2.3 The causal space for molecular behavior genetics

Social-environment researchers are interested in identifying the factors in individuals' environments that incline them towards one behavioral pattern rather than another. Their assumptions include:

1. Social and familial factors are causally independent of the subjects whose behavior is the object of study and for whom they constitute an environment.
2. Subjects are sufficiently endogenously uniform or genetic variation among subjects is randomly distributed and averages out in the population enabling variation in behavior to be correlated with variation in environment.
3. The causal space of interest is the variety of environmental factors that can impinge on behavior and the development of dispositions.

This assumption yields this quite different selection (Fig. 2.4) from the grid:

Fig. 2.4 The causal space for social/environmental approaches

Neurobiological approaches also require assumptions related to the investigative methodologies they have at their disposal. These yield Fig. 2.5 and include:

1. Brain areas showing greater glucose metabolism during a particular thought process are causally (or constitutively) involved in that thought process.
2. Anatomical correlates of behaviors are functionally related to the behaviors with which they are correlated.
3. The development of these anatomical correlates preceded rather than followed the relevant behaviors.
4. The causal space of interest is structures and processes in the brain and nervous system.

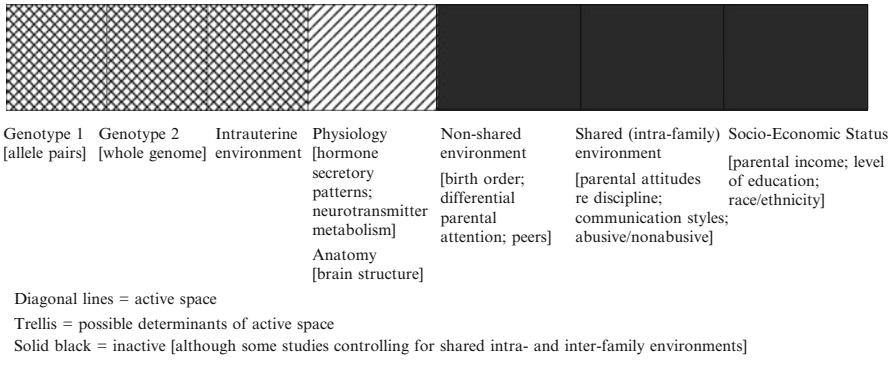

Fig. 2.5 The causal space for physiological and anatomical approaches

As the grids illustrate, each approach effectively situates itself in a different causal universe, making comparative assessment impossible. Two kinds of measurement are in play. One is measurement of the phenomenon to be accounted for (the "explanandum" or the "dependent variable"): a particular behavior pattern/disposition or variation in a behavior pattern. Here the approaches use similar measurement techniques. The other kind of measurement is of the factors an approach investigates as accounting for the phenomenon to be explained, the "explanans" or "independent variable". Measurements of the same factor being treated as explanans or independent variable conducted under one assumption concerning the structure of the causal space need not be consistent across approaches. From a god's eye point of view we may see the whole space, but if immersed in research, factors that are unmeasurable within a given approach may exert an influence that the measurement strategies either fail to pick up or attribute to different categories. Uterine factors, for example, will get classified as environmental factors under a genetic approach that focuses on biological relatedness or genetic similarity. Under an environmental approach that measures relevant differences in the social environment they do not appear at all, fading into the undifferentiated biological background. And both genetic and environmental approaches can say of the other that it fails to pick up causal relations identifiable by the one. Given that all acknowledge the interactivity of multiple causal factors in the inculcation of behavioral dispositions, a more comprehensive approach looks more promising. But here we encounter different difficulties.

Developmental systems theory (DST) is the name given to a bold set of claims about organismic development, including the development of behavioral dispositions. It has set itself up as a challenger to orthodox evolutionary theory as well as to developmental genetics.[12] The unit of evolution and of development is the developmental system, a set of complexly interacting factors whose effects coincide in the individual organism, but are not wholly contained within its skin. These include, for example, the environment of rearing and aspects of the system of nurturance of newborns and infants typical of any given species. The developmental system is not just the individual organism but the organism in its environment. The questions typical of this approach include: how does a given behavior B come to be expressed in individuals? what developmental trajectories (that is, sequence of changes in the developmental system) can be identified that culminate in B? is the disposition to B canalized? if so, how? at what levels of organismic integration and organization do the causal/developmental processes relevant to B occur (at the genetic level? at the cellular level? organic? environmental? some combination of these?)? how do complexity of organization and specialization of function develop in the individual organism? given that different types of causal factor are not separable, how can intra-level and inter-level interactions be studied?

[12]Primary expositors of Developmental Systems Theory have been Susan Oyama and the late Gilbert Gottlieb. See Oyama (1985); Wahlsten and Gottlieb (1997); Gottlieb (2001).

The assumptions of this approach include:

1. The interactivity of causes means that separation of causes is never possible.
2. The only interesting biological question is a developmental question.
3. Methods to support the central claim about the parity and interaction of causal factors will be found.
4. The unit of analysis must be the developmental system.

Given assumption 1), the causal space of DST includes all the types of factor, i.e. the entire grid, and assumption 4), what changes is not really a single property or propensity of an organism but the entire system configuration. In contrast with the preceding approaches, the entire set of interacting factors, both cause and effect, belong to the same universe and are distinguished as one stage of the system from another stage of the system represented in Fig. 2.6. A more complete representation would show how each type of factor can affect each other type of factor and affect how each other type affects higher level states of the organism.

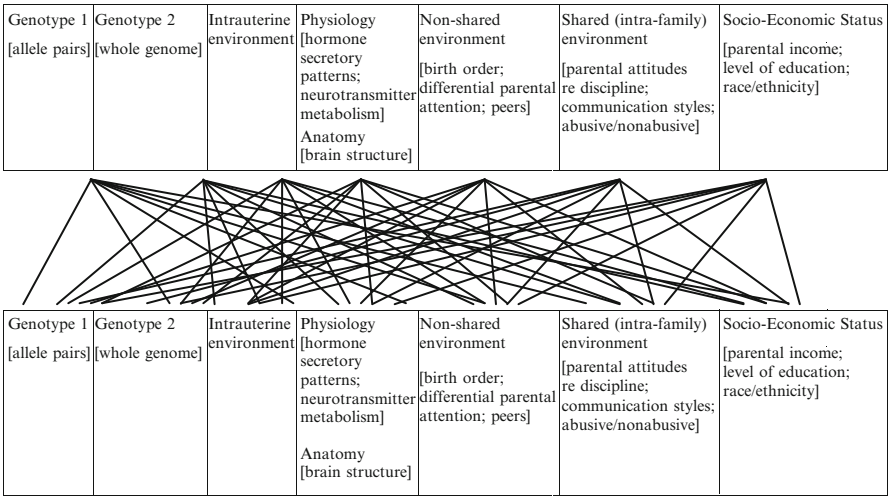

Fig. 2.6 A partial representation of the causal relations posited by Developmental Systems Theory

There is probably some sense in which this, or something like it, is the correct picture. Organisms are complex objects, and organisms in environments, even more complex. But this is not a parsing of the causal space that lends itself to empirical investigation. Furthermore, in order to be evaluated empirically in relation to any of the single factor approaches, the values of all factors and the strength of their interactions and mutual modifications would have to simultaneously measured. Even if one could construct computer simulations showing how a hypothetical system might work, an empirical determination exceeds the capabilities of present measuring systems. Thus, even if this is a correct picture, we are not entitled to claim so on the basis of empirical evidence. Empirical research does demonstrate the inadequacy of

any single approach, but this is not the same as demonstrating the adequacy of this particular representation of the causal relations.

Recently a more restricted integrationist approach has garnered a great deal of attention. The team of Avshalom Caspi and Terri Moffitt and their collaborators have worked out a specific model representing the interaction of genes and environmental factors in the etiology of specific behavioral and psychological disorders.[13] Their model posits a neural substrate for any given disorder that is acted on by both genes and environmental stimuli. The research questions of this approach include: For some specific psychiatric disorder D_P, what is the neurological substrate N of psychiatric disorder D_P? What is the specific disorder D_N of N underlying D_P? How do G and E interact in affecting N to induce D_N? The empirical information on which the model is based consists of both behavior/psychiatric genetic research showing some correlation between D_P (e.g. depression) and some allelic configuration and environmental research showing some correlation between D_P and exposure to some environmental stressor (the death of a spouse). What Caspi, Moffitt, and collaborators have done is to find that individuals characterized by overlaps (both the allele and the environmental stressor) show a much higher incidence of the particular disorder or problematic behavior, than individuals characterized by one factor alone. This, and the assumption of neural involvement, leads them to posit the following model (Fig. 2.7):

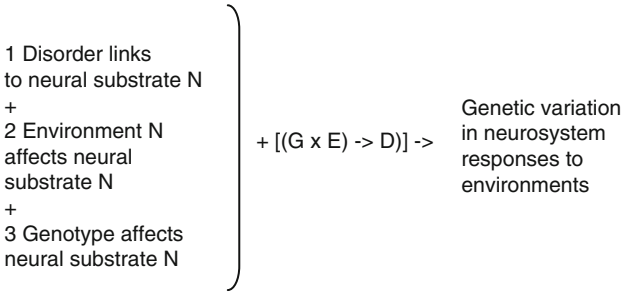

Fig. 2.7 GxExN model, modified from Caspi and Moffitt 2006, p. 585

The hope is that a specific psychiatric disorder can be linked to some specific neurobiological deficit or disorder, and that the neurobiological disorder can be linked to a genetic configuration. The neurobiological contribution will be identified by some kind of triangulation involving genes (identified through heritability and linkage studies) and environments studiable through socio-environment methods.

[13] Caspi, Sugden, and Moffitt (2003); Caspi and Moffitt (2006). About nine months after this talk was given in Hannover, Neil Risch, Kathleen Merikangas and colleagues published a meta-analysis casting doubt on the gene-depression connection that was one of the main empirical supports for the Caspi and Moffitt integrationist approach (Risch, Herrell, Lehner, et al. 2009).

Application of the model assumes that genes moderate the effect of environmental pathogens (their terminology) on disorder (i.e. that the higher frequencies in the overlap of genetic configuration and environmental stressor is accounted for by a genetically influenced sensitivity to environmental stressors), that experimental neuroscience can specify the proximal role of nervous system reactivity in the gene–environment interaction (i.e. will be able to identify the nature of increased sensitivity) and that it is possible to overcome the challenge of small sample sizes (through, for example, idealizations and analogues).

With these assumptions the potential causal space is somewhat reconfigured as in Fig. 2.8:

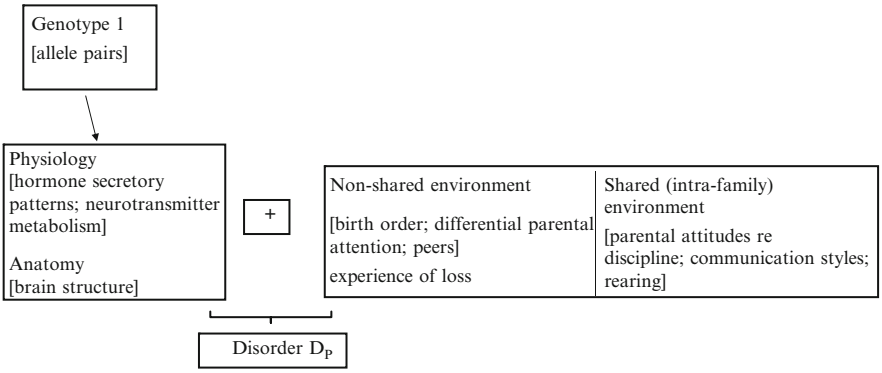

Fig. 2.8 Causal space for GxExN. Only a selection of the potential factors is studied, and they interact in producing the particular disorder

2.4

Analyzing the causal presuppositions and methods of these approaches, then, reveals that each operates in a distinct causal universe. By "distinct causal universe," I don't mean separable, ontological distinct spheres of causality, but conceptually constructed spheres of investigation. Philosophy offers several ways to respond to such a situation. Let us, for ease of consideration, limit ourselves to an epistemological response. Monism, as an epistemological view, holds that there is one, correct, comprehensive account and that it is possible to engage in comparative evaluation of alternatives in order to identify which it is. Inquiry ought to be directed to finding that one correct account.[14] Pluralism holds that given any given set of alternative accounts of a phenomenon, while some may well be false or deficient, it is nevertheless possible that there are multiple correct accounts, that none should be expected to be

[14]For more discussion of monism and pluralism see Longino (2002.pp. 93-95, 175-202) and Kellert, Longino, and Waters (2006).

comprehensive, and that it is possible to engage in intra-approach comparative evaluation, but not in inter-approach comparative evaluation (among approaches that meet some minimal empirical requirement).[15] Pragmatism suggests that alternative approaches should be judged in relation to practical goals of action with respect to the objects of the research in question. Each of these has advantages and disadvantages. In the end, I think some combination of pluralism and pragmatism offers a way of treating the variety of explanatory approaches that acknowledges the contribution each makes to the overall goal of understanding behavior.

Monism honors the impulse to unity and comprehensiveness that seems to drive many researchers, especially theoretical researchers. It makes for a relatively straightforward epistemology (true or false, correct or incorrect), and it makes sense of the debates among proponents of different and incompatible approaches to the same phenomenon or class of phenomena.[16] However, it presupposes that the data that would be used to adjudicate among approaches can, at least in principle, be completely and univocally described. In the case at hand, one has to ask: is research focused on one parsing of the causal space adequate to assign values to elements in the others? I hope the above illustrations of the causal spaces presupposed by the single factor approaches suffice to give a negative answer to this question. But, one might then suppose that a different parsing, indeed, one that includes all relevant factors should do better. Of the two integrationist approaches, however, one, the DST approach, is empirically intractable, while the other is limited in its scope to disorders, not to behavior generally. Monism, pace the debates swirling in research and philosophical circles about nature vs. nurture, requires conditions not satisfiable by the approaches currently practiced. This is not to say that some approach in the future might satisfy the conditions. But the problem with monism is that it legitimates forms of argument directed to elimination of all but one of a set of contesting approaches any time such a set exists.

The pluralist is more impressed by the (apparent) fact that each of the approaches has generated productive and useful research. Single factor and integrationist approaches can muster evidential support for their claims. The pluralist will propose that our task as philosophers is not to participate in debates about which of these approaches is the correct one, but to understand and help to articulate their scope, their evidential requirements, and their limitations. But pluralism has different problems: What's the sense in which each is correct? I have proposed conformation as an umbrella term for varieties of semantic/epistemic success (including truth, similarity, approximation, isomorphism, homomorphism) that enable us, as epistemologists, to countenance multiple non-congruent accounts of the same phenomenon.[17] Is this too coarse-grained a form of evaluation? How, if multiple approaches are correct, would

[15] See Longino (2006).

[16] See the debates from which the opening quotes to this paper are drawn. Also Turkheimer and Gottesman (1991) versus Gottlieb (1991) and also McGue (1994); Maccoby (2000).

[17] Longino (2002).

we determine which to use in practical situations? Doesn't application of scientific models/theories presuppose their epistemic superiority to alternatives?

Here, it seems to me, is the appropriate place for pragmatism, as a higher order sorting procedure for approaches that meet the standard of conformation mandated by the pluralist. Pragmatism is often accused of recommending acceptance of hypotheses and theories solely on the basis of their utility, which conjures images and memories of racist and otherwise faulty science. But paired with the kinds of empirical requirement that are central to pluralism, pragmatism can help address the problem about applicability of incompatible but equally empirically adequate approaches. Pragmatism, as a second order sorting procedure, recommends that we evaluate theories and models with respect to the specific questions they set out to answer and the kinds of intervention in the world the answers make possible.[18]

Each of these approaches does specific kinds of work, reveals particular families of causal dependencies, knowledge of each of which serves useful purposes. Behavior genetics provides clues to the function of particular genes or gene complexes and narrows the search for intermediate physiological processes. (This capacity is on display in the Caspi and Moffitt work, among others.) Behavioral neuroanatomy and neurophysiology provide clues to the interrelation of neural structures and processes. That is, regardless of the extent to which they account for the expression of any given behavior, research conducted with those frameworks is likely to have cognitively and practically useful outcomes. Research conducted within the social-environment framework enables comparisons of the effectiveness of different environmental interventions in modifying behaviors. The Developmental Systems approach at least helps apply brakes to overhasty application of single factor frameworks as well as encouraging, if not research that could directly test the full set of interactions in any given instance, research that tries to identify specific (mostly pairwise) interactions. Finally, the obvious value of the G+E+N approach is that, when it achieves results, it helps to identify proximate causes of identifiable psychiatric disorder (in those cases that fit the model) and, thereby, a strategy for therapy. The answer to the question: on what approach should we rely for application in intervention and policy? must be: it depends on the kind of intervention needed and the kind of policy required.

2.5

A pluralist stance has informed the approach to analysis of this research, but it was suggested by a preliminary investigation that revealed that all approaches were home to research efforts that could claim empirical success. Pluralism is a way of trying to make philosophical sense of this situation. I have tried to show how it is that these different approaches could all be successful by showing that there is no

[18]For further discussion, see Longino (forthcoming).

common basis of evaluation, even though methods internal to the approaches are adequate to separate empirically adequate from inadequate. But, pluralism without the kind of second order pragmatism outlined above is incomplete.

References

Brunner, H.G., et al. (1993a): "Abnormal Behavior Associated with a Point Mutation in the Structural Gene for Monoamine Oxidase A." *Science* 262: 578–580.

Caspi, A. & Moffitt, T. (2006): "Gene-environment Interactions in Psychiatry: Joining Forces with Neuroscience." *Nature Reviews: Neuroscience* 7: 583–590.

Caspi, A., Sugden, K., Moffitt, T. et al. (2003): "Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene." *Science* 301: 386–389.

Coccaro, E., Gabriel, S. & Siever, L. (1990): "Buspirone Challenge: Preliminary Evidence for a Role for Central 5-HT-1a Receptor Function in Impulsive Aggressive Behavior in Humans." *Psychopharmacology Bulletin* 26: 393–405.

Gottlieb, G. (1991): "Experimental Canalization of Behavioral Development: Theory." *Developmental Psychology* 27: 4–13.

Gottlieb, G. (2001): "A Developmental Psychobiological Systems View: Early Formulation and Current Status", in: Oyama, S., Griffiths, P.E. & Gray, R. D. (eds.): *Cycles of Contingency*. Cambridge (MA): MIT Press, pp. 41–54.

Kellert, S., Longino, H. E. & Waters, C. K. (2006): "The Pluralism Stance", in: Kellert, S., Longino, H. E. & Waters, C. K. (ed.): *Scientific Pluralism (Minnesota Studies in Philosophy of Science vol. XIX)*, Minneapolis (MN): University of Minnesota Press.

Lavigueur, S., Tremblay, R. & Saucier, J.-F. (1995): "Interactional Processes in Families with Disruptive Boys: Patterns of Direct and Indirect Influence." *Journal of Abnormal Child Psychology* 23: 359–378.

Longino, H. E. (2001): "What Do We Measure When We Measure Aggression" *Studies in History and Philosophy of Science* 32, 4: 685–701.

Longino, H. E. (2002): *The Fate of Knowledge*. Princeton (NJ): Princeton University Press.

Longino, H. E. (2006): "Theoretical Pluralism and the Sciences of Behavior", in: Kellert, S., Longino, H. E. & Waters, C. K. (ed.): *Scientific Pluralism (Minnesota Studies in Philosophy of Science vol. XIX)*, Minneapolis (MN): University of Minnesota Press.

Longino, H. E. (forthcoming): *Understanding the Sciences of Human Behavior* (working title). Chicago, IL: University of Chicago Press.

Luntz, B. & Widom, C. (1994): "Antisocial Personality Disorder in Abused and Neglected Children Grown Up." *American Journal of Psychiatry* 151: 670–674.

Maccoby, E. (2000): "Parenting and its Effects on Children: On Reading and Misreading Behavior Genetics." *Annual Review of Psychology* 51: 1–27.

Mason, D. & Frick, P. (1994): "The Heritability of Antisocial Behavior: A Meta-Analysis of Twin and Adoption Studies." *Journal of Psychopathology and Behavioral Assessment* 16: 301–323.

McGue, M. (1994): "Why Developmental Psychology Should Find Room for Behavioral Genetics", in: Alexander, C. N. et. al. (eds.): *Threats To Optimal Development: Integrating Biological, Psychological, and Social Risk Factors*, Hove: Lawrence Erlbaum Associates, pp. 105–119.

Oyama, S. (1985): *The Ontogeny of Information*. New York: Cambridge University Press.

Risch, N., Herrell, R., Lehner, T. et al. (2009): "Interaction Between the Serotonin Transporter Gene (5-HTTLPR), Stressful Life Events, and Risk of Depression: A Meta-analysis." *Journal of the American Medical Association* 301: 2462–2471.

Scarr, S. (1995): "Commentary" *Human Development* 38: 154–157.

Tellegen, A., Lykken, D. T., Bouchard, Jr., T. J., Wilcox, K. J., Segal, N. L. & Rich, S. (1998): "Personality Similarity in Twins Reared Apart." *Journal of Personality and Social Psychology* 54: 1031–1039.

Turkheimer, E. & Gottesman, I. (1991): "Individual Differences and the Canalization of Human Behavior." *Developmental Psychology* 27: 18–22.
Wahlsten, D. & Gottlieb, G. (1997): "The Invalid Separation of Effects of Nature and Nuture: Lessons From Animal Experimentation", in: Sternberg, R., Grigorenko, E. et. al. (eds.): *Intelligence, Heredity, and Environment*, Cambridge: Cambridge University Press, pp. 163–192.

Part II
Genetic Explanations of Behavior

Chapter 3
Genome Wide Association Studies of Behavior are Social Science

Eric Turkheimer

3.1 GWAS and Its Discontents

More than a decade ago, as a half-century of population-based modeling of twin and adoption studies was giving way to the Human Genome Project and the era of measured DNA, I wrote:

> Population-based behavioral genetics has demonstrated that genotype and behavior can be expected to covary. Although the epigenetic developmental pathways linking gene products to complex behavior will in general be almost unimaginably complex, modern molecular genetics has made it possible to detect small covariations between alleles and behavior that span the complexity of the causal network….. Such associations are real and potentially interesting, but they remain correlations— and small ones— not evidence of substantial causal pathways between individual alleles and complex behavior or evidence of genes for extroversion or intelligence or evidence that future scientific efforts will be most productively applied at a genetic level of analysis. If the history of empirical psychology has taught researchers anything, it is that correlations between causally distant variables cannot be counted on to lead to coherent etiological models. (Turkheimer, 1998, p. 789)

At the time, my prediction had a distinctly Luddite ring to it. Why would anyone bet against the inexorable progress of science? My gloominess on the topic was in sharp contrast to the optimistic, not to say hegemonic, claims of most researchers at the time. Here, for example, is Plomin and Crabbe (2000) in an article entitled, "DNA": "The authors predict that in a few years, many areas of psychology will be awash in specific genes responsible for the widespread influence of genetics on behavior." (p. 806)

These predictions were made at the turn of the present century, as the Human Genome Project was realized, as human genetics made the transition from statistical

E. Turkheimer (✉)
Department of Psychology, University of Virginia, PO Box 400400, Charlottesville, VA 22904-4400, USA
e-mail: ent3c@virginia.edu

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*, Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_3,

accounting of biologically related family members to the analysis of actual DNA. We are now at the end of that era, or at least it's first chapter. The technology available to genomic scientists has increased exponentially, and lately reached an apotheosis in the form of Genome Wide Association Studies, or GWAS, which allow us search through the entire genome for the bits of DNA that are more closely associated with disease or variation in normal behavior. GWAS, like so much human genomics before it, has produced somewhat paradoxical results: we are indeed, as Plomin predicted, awash in associations between human characteristics and genetic variation. At the same time, as I predicted, it is widely agreed that real scientific progress has been far more difficult than anyone expected; most, I think, would agree that new era of human genomics has been a disappointment so far. This essay will attempt to resolve this paradox, to understand how human genomics can fill libraries with "results" that nevertheless seem to fail to make progress toward the goals they were designed to reach.

3.2 Background

Genome wide association studies cannot be understood without seeing them in the historical context of behavioral genetics, which has its origins in the practical science of animal breeding. People have been breeding animals for complex characteristics, including behavioral ones, for thousands of years. The first comprehensive text about behavioral genetics, Fuller and Thomson (1960) was primarily about temperament in dogs.

Animal breeding predates both Darwin and Mendel, so much of it, whether on the farm or in the lab, was conducted without reference to anything like modern genetics. That started to change in the 20th century, although most of the traits bred in lower animals do not fit a Mendelian model of inheritance. The characteristics in Mendel's peas segregated from generation to generation: crossing wrinkled peas with smooth peas did not produce moderately wrinkled peas, but rather a mix of wrinkled and non-wrinkled, in proportions determined by the laws of classical genetics. Crossing cows high in milk production with cows low in milk production *does* produce cows with moderate levels of milk production, and selecting the highest milk producers for reproduction produces a steady increase across generations.

The classical genetics of Mendel and the genetics of complex characters like milk production was integrated (still long before anything was known about DNA) by R. A. Fisher (1918), who showed that a large number of independently segregating genes of small effect could be summed to produce a normally distributed trait that was inherited but which did not segregate. The statistical underpinning of the synthesis was based on the concept of variation. Differences among animals in milk production are associated with the degree of genetic similarity among them, as opposed to where they are raised or how they are fed, which would normally be held constant by the experimenter. The proportion of observed variation in a trait that is associated with variation in genetic relatedness is known as heritability. Heritability

is a useful concept to animal breeders, because it is related to the rate of change produced by selective breeding.

The concept of heritability can be extended to the study of humans, with some important caveats. The basis for the extension is the study of groups of people with known differences in degree of genetic relatedness, most famously identical and fraternal twins, but also siblings, parents and children, adopted (and therefore genetically unrelated) siblings, cousins, and so forth. Just as in farm animals, one can estimate a proportion of variation associated with genetic differences to the total proportion of a trait, and compute heritability between zero and one.

The crucial difference between notions of heritability in controlled studies of lower animals and studies of natural variation in humans is that for animals, the genetic and environmental variances are under the experimenter's control, and therefore fixed and uncorrelated with each other; in humans variation cannot be controlled, for obvious ethical reasons. Once variances become uncontrolled and correlated with each other, heritability coefficients no longer depend exclusively (not even primarily) on the biological characteristics of the trait in question. Instead they depend on the variability of the trait and the variation and covariation of the genes and environments that underlie it, in the particular population being studied. Having two arms notoriously has a heritability of zero, for example, because the genetic mechanisms that cause us to have two arms don't vary among individuals. Although developing two arms is intuitively and sensibly a biological process, variation in arm-number is primarily due to environmental events like accidents. One could not selectively breed cows for three-leggedness, and the reason is not that leg-number in cows is somehow essentially environmental. Rather, the genetic mechanisms involved in leg-number do not vary among cows, so it is not possible to select for them.

It is therefore not a good idea to cite heritability coefficients as a measure of how "genetic" or "environmental" something is, height included, and the high heritability of height in modern populations does not mean that it is genetically determined. One can imagine circumstances under which the heritability of height would be substantially lower (for example, under circumstances in which there were radical differences in access to adequate nutrition), and height has undergone obvious changes in recent historical time that cannot be the result of genetics. I cite the heritability of height here simply to say that height has the characteristics that lead people to think that it *ought* to be amenable to GWAS.

In any case, once the statistical means for computing the heritability of human characteristics was established, it was open season. Thousands upon thousands of family studies (mostly twin or adoption studies) were conducted, and heritabilities were computed for the usual behavioral suspects: intelligence, personality and mental illness. And to the surprise of all concerned, the studies all came out the same way: everything was heritable. Not perfectly heritable, of course, but substantially and significantly heritable. Ignoring the caveats about the interpretability of heritability coefficients in free-ranging humans, this outcome was generally taken as a great victory for genetic explanations of behavior, either to be celebrated or lamented, depending on the predisposition of the writer. I have written elsewhere (Turkheimer, 2000) about

why such conclusions turned out to be premature. The reasons can be summarized as follows, and they have resonance for the contemporary problem at hand:

1. Not only the major and established dimensions of behavior turned out to be heritable, but so did everything else. Depression is heritable, but so is marital status; intelligence is heritable, but so is how much TV people watch.
2. Heritabilities, as one might have predicted from the forgoing discussion, didn't replicate very well from study to study. They were almost never zero, but whether they were relatively high or low seemed to vary from study to study and situation to situation.
3. Largely as a consequence of (2), it is difficult to identify any major scientific advances that were produced by the twin studies, beyond the establishment that heritability is greater than zero. For example, what do we know about personality on the basis of twin studies that we did not know without them? We know that personality is moderately heritable, a fact that is not without consequences (Turkheimer, 2000), but hopes that twin studies would elucidate the causal processes underling the development of personality went mostly unfulfilled.

Such was the state of behavioral genetics at the dawn of the human genome project, which was widely viewed as a panacea for the epistemological shortcomings of twin studies. We may not have learned all that much from partitioning variance in family data, we were told, but wait until we get our hands on the actual DNA! With heritability computed in family studies as a guide (a mistaken strategy, by the way, given the inherent variability of heritability coefficients) we can now proceed to piece together the genetic processes leading to complex human traits from the ground up.

There were two main research strategies available at the outset. Linkage studies search through the genome in family pedigrees for genetic markers (locations on the genome smaller than a gene) that segregate within families in the same way as a trait of interest. Linkage studies have the advantage of being able to search the entire genome, and the disadvantage of only identifying regions, as opposed to specific locations, of interest. Association studies target specific and pre-identified genetic markers, called candidate genes, and ask whether they are correlated with the expression of a trait in the population. Association studies have the advantage of identifying relations with specific genes as opposed to regions, but are limited by our ability to decide on the candidate genes to investigate.

The newest technology, genome wide association studies, are what everyone had in mind when the genome project got underway. GWAS is the apotheosis of contemporary gene-hunting, combining many of the features of linkage and association studies. Inexpensive chips now make it easy and cheap to test for a million genetic markers in the form of single nucleotide polymorphisms, or SNPs, individual units of DNA that only take two of the four possible values of ACGT. It is thus possible to scan practically the entire genetic sequence for associations between alleles and complex traits, with a simplicity and low cost that makes it possible to include tens of thousands of research participants. Because there are so many markers across the genome, the poor focus of linkage studies has been greatly (but not completely) ameliorated, and for better or for worse one does not have to make prior identification of

the candidate genes. All that needs to be done is to find a sample of people with schizophrenia, a control group without schizophrenia, print out their genomes and look for the differences. Why wouldn't that work? But progress has been, it is safe to say, disappointing. It is not that no associations between individual alleles and specific behaviors have been found. To the contrary, we are indeed awash in them: thousands have been identified. However, the "specific" and "responsible for" clauses in Plomin and Crabbe's daring prediction have proven more difficult: despite the myriad linkages and associations between alleles and complex human traits that have been reported, three persistent limitations have proved very difficult to overcome, and they should sound familiar:

1. The reported associations are very small, in the sense that they each explain a tiny proportion of the overall variability, and collectively not much more than that;
2. The associations don't replicate very well; and
3. In part as a consequence of the first two, the various small associations between genes and behavioral outcomes haven't added up to etiological *explanations* of behaviors and especially behavioral disorders.

In other words, we are back where we started.

3.3 The Missing Heritability Problem

Others may take a rosier view than I do of the general progress that has been made toward genetic theories of behavioral syndromes, but I will save that argument for another paper. Here, I would like to discuss a remarkable series of papers published recently in *Nature Genetics*, concerning not depression or schizophrenia, not IQ or extraversion, but height. Height, that is, with near-perfect reliability of measurement and a heritability of .9 (Silventoinen et al., 2003). Height, for which there should be little problem with complex causal feedback loops. (Tall parents don't expose their children to special height-inducing environments.) Height, which has obvious biological analogs in the simplest of organisms. The genomic research paradigm, in which heritability is the gateway to identifying the specific genes composing the genetic etiology of a trait, may have turned out to be more complex than expected for juvenile delinquency, but surely it ought to work for height?

A single issue of *Nature Genetics* contained three empirical reports of genome wide association studies of height (Gudbjartsson et al., 2008; Lettre et al., 2008, Weedon et al., 2008) and a summary article describing their conclusions (Visscher, 2008). At bottom, GWAS is a search algorithm for correlations. The height studies each produced something under a half a million of them. From the outset, consideration of such results poses a problem that has been faced many times by any non-experimental social scientist: given a vast array of results that are presumably a joint reflection of some underlying process of interest, other processes of less interest that have not been controlled experimentally, and some amount of sampling error, how do you tell them apart?

The answer, of course, is null hypothesis significance testing (NHST). For any given individual association, one can compute the probability that an effect of that magnitude would occur in the sample, given a null hypothesis of no association in the population. If that probability is lower than some agreed upon "alpha" probability, one declares the null hypothesis of no association false. The alpha probability is therefore an error rate, the proportion of errors one is willing to tolerate when declaring null hypotheses false. There is, of course, another error rate involved, the "beta" or "Type-II" error rate, which describes the probability of being in error when failing to declare a null hypothesis false.

NHST is greatly complicated when there is more than one result (in this case, 400,000 results) to test. If the probability of being incorrect about any single hypothesis test is equal to α, then the probability of being incorrect on at least one of k hypothesis tests equals $1-(1-\alpha^k)$, which approaches 1.0 very quickly. Social science has developed modest technologies for dealing with the problem, like the familiar Bonferroni correction[1], but such methods do not begin to apply to the enormous number of tests conducted in GWAS, for which somewhat more sophisticated methods have been developed.

Significance testing in GWAS incorporates several steps. First, the full distribution of test probabilities is plotted against the expected distribution under the null hypothesis, to establish that *something* is disturbing the null distribution. In Weedon et al. (2008; see their Figure 1) there was an unmistakable overrepresentation of low probabilities. In the largest sample (combined meta-analytically across several studies), for example, there were 27 tests with significance levels less than 10^{-5}, compared to the four that would be expected on the basis of sampling error under the null hypothesis of no association. Weedon et al. conclude, "Approximately 23 of these loci are therefore likely to represent true positives." (p. 576)

The associations are then subjected to an even more stringent test. Thirty-nine of the original 400,000 SNPs (the 27 that exceeded the 10^{-5} criterion plus 11 that exceeded a 10^{-4} criterion, plus one more identified as a candidate in another study) were retested in an independent sample of 16,482 participants. Twenty of these 39 achieved $p<.005$ in the independent test. Combining the screening and the cross-validation, twenty SNPs had p values lower than 5×10^{-7}, 17 were lower than 10^{-8}, and 10 were lower than 10^{-10}. That's pretty significant!

But as we proceed through Weedon et al. or the other empirical reports, we find there is a second problem lurking behind the familiar one of significance testing. The statistically significant associations are further tested for something called "population stratification," and once it is found to be absent, Weedon et al. can declare, "This means that the associations are likely to reflect true biological effects on height." (p. 580) Now we would appear to be getting somewhere, although it will turn out to be problematic that no one pauses to explain what "true biological effects" are, and how they can be distinguished from biological effects that are not true or true effects that are not biological.

[1] In which the required significance level, usually $p<.05$, is divided by the total number of tests to be conducted in the experiment.

What is population stratification? The classic example of population stratification involves the discovery of a "chopsticks gene" (Hamer & Sirota, 2000). Suppose you are seeking a gene contributing to the use of chopsticks in a sample that happens to include both Asian and American participants. Any gene that differs in frequency between the Chinese and American populations will be associated with use of chopsticks, but the associations will be causally spurious. Chopstick use is *caused* by exposure to the rearing practices of Asian families; exposure to the rearing practices is *correlated with* gene frequencies, and this correlation induces a spurious one between the genes and chopstick use.

As is often the case when difficulties of this kind arise in situations where genetic methods are employed in the service of social scientific ends, the technical-sounding name that is given to the problem and to the various statistical methods that are developed to cope with it foster the impression that population stratification is essentially a technical problem in molecular genetics, to be overcome in the same way that so many other problems in genetics have been overcome, by burying them under the relentless forward momentum of contemporary genomic technology. If we can put half a million SNPs on a single chip, how big a problem can population stratification be?

But in fact, population stratification is a very old problem, and has little to do with genetics per se. Notice that population stratification doesn't arise in studies of non-human animals. That is because we have experimental control over the environments to which laboratory organisms are exposed, so we can determine that environments are either invariant or random, and there are no potential correlations between the occurrence of alleles and exposure to environments. In a horrific world in which it were possible to control the environments of humans so they could be raised identically, or randomly assigned to environments of the experimenter's choosing, population stratification would not be as severe a difficulty.

Population stratification is a problem in non-experimental causal inference, and as always, definitive attribution of causation is a matter of experimental design, not statistical analysis. A wide variety of tests, corrections and workarounds have been developed to ameliorate the effects of population stratification on GWAS. Like the original problem itself, these fixes are overlaid with a veneer of genetic technology that may lead the unwary interpreter to believe that the problem has been licked, that the science of genomic association has moved on from population stratification just as the newest SNP chip is bigger and cheaper than the last. But methodological problems in scientific inference are not so easily overcome by the next wave of technology. The fixes, moreover, are reworkings of statistical methods that have been available to social scientists for many years. And as any working social scientist is all too well aware, although the methods are sophisticated and interesting as statistical devices and useful enough as halfway measures, they don't work to discriminate true causal effects from extraneous processes that have not been controlled by the experimental method. In the long run, statistics cannot replace the causal rigor of the experimental method, no more so in genomics than in sociology.

3.4 Why not EWAS?

To a remarkable degree, GWAS was foreshadowed in a domain that might at first seem quite remote: the human social environment, and the quasi-scientific methods that have been developed to study it. The twin inferential issues in GWAS– distinguishing "true" associations from those expected on the basis of sampling error, and then distinguishing "true" causal processes from mere associations– are the bread and butter of social scientists working as far from genomics as it is possible to work. If you are a developmental psychologist trying to identify the environments that predispose some adolescents to become delinquents, what do you do? Most of the time, random assignment to environmental conditions is out of the question. So you gather as much data as possible about neighborhoods, schools, families and peers, measure delinquent outcomes in the children, and endeavor to show that some aspects of the environment *predict* (read: *are correlated with*) delinquent behavior. If you are comprehensive in your measurements of relevant environments, you might be tempted to say that you conducted an Environment Wide Association Study, or EWAS.

Of course, no self-respecting social scientist would announce such a thing because the methodological connotations are so dreadful, conjuring images of vast correlation matrices with circles around the few of them that have exceeded some magical level of statistical significance. But there is no need to be unduly derogatory about the fundamental state of affairs: in most of human behavioral science experimentation is not possible, and because it is not, scientists resort to other means, most prominent among them the analysis of systems of statistical associations. Presented with an interesting association between an environmental risk factor and a behavioral outcome, but lacking possibilities for randomized experimentation that might establish the association as causal, what would the traditional social scientist do?

The first the thing the scientist would do, of course, is to test the association for significance. For the better part of a century, far from the high-tech world of the Human Genome Project, psychologists of all persuasions have been testing their associations with NHST. From social psychologists running college students through elaborate randomized experimental conditions, to developmentalists analyzing enormous uncontrolled correlation matrices arising from observations of families, to cognitive psychologists giving repeated trials of memory tasks, to psychobiologists taking single-neuron recordings from hamster brains, to clinicians trying to establish the efficacy of psychotherapy, only two things have tied together the impossibly diverse collection of researchers that make up a psychology department: a commitment to collecting data one way or another, and an intention to test the resulting associations with NHST.

The reasons NHST has failed as a basis for scientific psychology are deep, wide, no longer a matter of serious controversy, and not the main point of this paper (see, among many others, Cohen, 1994; Schmidt, 1996). The probability levels that are computed compulsively to five decimals depend on assumptions that cannot be tested, let alone confirmed; their binary, reject or fail-to-reject formalism does violence to the subtleties of actual evaluation of scientific hypotheses in the laboratory; the tests

depend ineluctably on sample size; they encourage attention to Type I errors at the expense of attention to statistical power; the probabilities themselves represent the converse of what we really want to know, telling us the likelihood of our data given our hypothesis, when we really want the likelihood of our hypothesis being correct, given our data. These failures have been well-catalogued elsewhere and I won't do so again here (see Cohen, 1994; Harlow, Mulaik & Steiger, 1997).

In the end, the failure of NHST can be seen as a failure to solve the central dilemma of scientific psychology: for researchers working in one of the many psychological domains where randomized experimentation is impossible for practical or ethical reasons, NHST has not succeeded in discriminating actual causal processes from spurious correlations and non-causal associations. And even when experimentation *is* possible, the causal pathways leading to complex human behavior are often so diverse that empirical science seems all but helpless to unpack them, and here too NHST has provided no help.[2]

3.5 Searching for Causes in Social Science

This brings us to the next and more important, because less examined, step in the inferential chain. Given an association that passes a test of significance, how do we know if it is really causal, as opposed to the result of spurious confounds, of "population stratification"? The two broad classes of methods that are brought to bear are multivariate statistics and quasi-experimental research methods. The most basic statistical approach is multiple regression, in which possible confounds are measured and included as predictors along with the alleged causal factor. Under some restrictive conditions, the estimated regression coefficient for the factor of interest then represents its association with the outcome with values of the measured covariates "held constant" statistically. In some contexts (traditionally including situations where the effects of interest are categorical, and the potential confounds are continuous) this method is referred to as Analysis of Covariance or ANCOVA. The biggest shortcoming of multiple regression is that it requires measuring (and measuring well) all of the potential confounds of the alleged causal relationship. It is not generally possible to know if this has been accomplished successfully. Most of the multivariate alternatives to multiple regression can be characterized as attempts to circumvent the need to measure every single individual variable that might confound a causal relationship.

Principle Component Analysis, or PCA, uses the multivariate structure of the covariances among uncontrolled variables to define one or several dimensions that jointly determine the multivariate domain. So if one has measures of parental

[2]The greatest proponent of such ideas was the great theoretical psychologist Paul Meehl. The interested reader is directed to his many papers on the subject, most especially, Meehl, 1978, which should be required reading for GWAS researchers.

income, housing quality, neighborhood quality, and academic levels of local schools, one could use the positive associations among them to define a "latent variable" called *poverty*.[3] Once again under fairly restrictive assumptions, controlling for the multivariate construct succeeds in including not only the measured variables that were used to estimate it, but also the unmeasured indicators that could have been measured but weren't.

A more advanced classical method is called instrumental variable regression (Angrist, Imbens & Rubin, 1996). Given an observed association between a purported cause and an outcome, an instrument is a third variable which is correlated with the purported cause and the potential confounds, but not with the outcome, conditional on the cause and the confounds. Suppose a scientist observes an association between father-absence in families and delinquency in children: Is the relationship causal? One way to answer the question is by finding an *instrument*. In the classic example, the government might establish a new tax policy that has the effect of keeping families intact, but which would not plausibly affect rates of delinquency on its own, except by way of its correlation with intact families. Under these conditions and several other assumptions, it is possible to estimate the causal effect of intact families independent of the confounds.

A third statistical method is called propensity score analysis (Rosenbaum & Rubin, 1983). Propensity scores are a method for summarizing all of the available information about confounds of a potential cause. Returning once again to the absent father example, one way to state the problem is that because we cannot randomly assign children to absent father conditions, children with an absent father differ in many uncontrolled ways other than the father absence itself. If we collect as many possible predictors of father absence that we can think of and load them all into an equation predicting father absence, the modeled probability summarizes the overall tendency for father-present and father-absent families to be non-randomly assigned. We can match families for the overall *propensity* to have an absent father, allowing us to estimate the causal effect of absence without bias.

3.6 Within Family Designs and the Nonshared Environment

An alternative to statistical methods for establishing causation in non-experimental data is to use *quasi*-experimental designs. The range of possibilities is vast and beyond the scope of this paper (Campbell, Stanley & Gage, 1963; Rutter et al., 2001). One particular form of quasi-experimentation is particularly relevant to GWAS and

[3] A latent variable is a hypothetical process that cannot be observed directly, but which serves to explain relationships that can be observed among actual measurements. If one observes that many aspects of deprived environments—crime, poor schools, inadequate nutrition, unstimulating surroundings—tend to co-occur, the latent variable *poverty* can be invoked to explain why. The relevant statistical procedure is known as factor analysis. See MacCorquodale and Meehl (1948), or for an accessible statistical treatment, Loehlin (1992).

EWAS: within-family comparisons. Suppose you have a large sample of pairs of monozygotic (identical) twin children. Among these twins you will be able to find the occasional pair for which one member is exposed to a risk factor for delinquent behavior and the other is not. Suppose the twin who is exposed to the risk factor is indeed engaging in delinquent behavior. Is delinquency a causal consequence of the risk factor? Now at least you have an interesting control group: What is the non-exposed co-twin doing? If he is engaging in delinquent behavior to the same extent as the exposed twin, it doesn't seem likely that the risk factor *per se* is the decisive causal factor; on the other hand, if the non-exposed cotwin is not delinquent, then there may reason to expect that the risk factor *is* causing the delinquency, although as we will see below, twin designs are not capable of producing true causal inference from non-experimental data.

Within-family designs are important in many areas of psychology (Rodgers et al., 2000), and play an especially important role in behavioral genetics (Dick, Johnson, Viken & Rose, 2000), although it might be more accurate to say that within-family designs are the link between traditional behavioral genetics and the mainstream of developmental psychology. When twin studies first convinced the world of the importance of genetics in the development of human behavior (e.g., Bouchard et al., 1990), genetic variation shared supremacy with another biometric component. Although identical twins are universally more similar in behavior than fraternal twins, it is also the case that identical twins are substantially less than perfectly similar. This residual variability cannot be genetic, as identical twins are just that genetically, and it cannot be the result of differences in rearing environment, since twin pairs in these studies are raised together. The term came to be called the "nonshared environment," denoting differences among siblings or twins that arise because of environmental *differences* among children raised in the same family, as distinguished from the more intuitive "shared environment" which represents traditional socioeconomic and familial forces making family members more similar to each other. (For a philosophical treatment of the nonshared-shared environment distinction, see Plaisance, unpublished dissertation.)

In 1987, Robert Plomin and Denise Daniels published a paper with the title, "Why are Children Raised in the Same Family So Different from One Another?", in which they tried to formulate the causal processes that might underlie this variance component. Plomin and Daniels hypothesized, straightforwardly, that the characterization of the residual variance component as the nonshared environment was apt, that children raised in the same family were different from each other because their environmental experiences were different, and moreover that the specification of those differences should form the basis of environmentalist developmental psychology. They formulated a three-step program that succeeded in becoming the basis of a research program that extended over more than a decade and continues to this day:

1) Quantify the magnitude of the nonshared environmental variance component at the population level.
2) Identify environmental events that are experienced differently by children in the same family.
3) Specify the causal relations between nonshared environmental events and developmental outcomes.

In research of this kind, environmental differences between pairs of siblings or twins are used to predict differences in outcome. Perhaps most clearly in identical twin pairs, any relations that are identified cannot be attributed to genetic differences either between or within families, since the twins are genetically identical, or to environmental differences between families, like culture (chopstick use!) because the twins were raised in the same family, in the same cultural milieu. Another way of saying this is that quasi-experimental within-family designs control (imperfectly, of course) for population stratification. So the research mandated by Plomin and Daniels had two aspects that parallel the goals of contemporary GWAS. On the one hand, it was an attempt to decompose a population level variance component– the nonshared environment– into the actions of the individual environmental events it comprised; on the other, it was a quasi-experimental attempt to sift the myriad and easily-observed *associations* between environment and outcome for some smaller set that are potentially causal.

3.7 The Missing Environment Problem

In a way that once again foreshadowed the recent difficulties of the genome project, the outcome of the research mandated by Plomin and Daniels' program was disappointing. Mary Waldron and I (Turkheimer & Waldron, 2000) conducted a comprehensive meta-analysis of the research that had been conducted under the banner of the nonshared environment. In the studies we reviewed, the environment was actually measured for each member of a twin pair, rather than inferred from the twin design; just as in GWAS, DNA is now measured, as opposed to inferred from population genetics. So, for example, one might measure differences in the harshness of communications directed at siblings by their parents, and use these differences to predict differences in delinquency in the siblings. Plomin and Daniels' hypothesis can once again be stated in terms of the two aspects of the research. They hypothesized that the population-level nonshared environmental variance component could be decomposed into individual effects such as these, or equivalently, that the many non-experimental associations that are observed between risk factors and outcomes can be shown to be plausibly causal by exposing them the to within-family design.

Either way, our review demonstrated that the hypothesis could not be supported. Although the nonshared environment accounted for upwards of 50% of the variability in the studies we reviewed, the median percentage explained by any individual measured environment was under 2%. The review showed that the nonshared environmental variance component could not be decomposed into many small causal environmental events. There are substantial differences in delinquent behavior between pairs of siblings, even pairs of identical twins reared together in the same family, and the twin design can be used to establish that these differences are broadly environmental in origin. But when the investigator selects "candidate environments" that differ between siblings, for example the emotional quality of their interactions with mother, the individual effects of the candidate environments don't come close

to adding up to the total effect of "the environment" as estimated by the twin studies. Another way of saying the same thing is that observed associations between environments and outcomes—in the population, without controlling for the between-family effects of genes and shared environment, children who have more negative interactions with their mothers are more likely to be delinquent— do not stand up to the more rigorous quasi-experimental test of comparisons of siblings or twins raised together. Within families, the sibling with more negative maternal interactions is not more likely to be delinquent than the brother or sister with more positive interactions, at least not sufficiently so to account for a substantial portion of the variance component called nonshared environment. The problem of the missing variance in the nonshared environment, which was never christened as "the missing environment problem", although that is exactly what it is, remains unsolved; I remain gloomy.

The answer to the question, "Why not conduct EWAS?" is that social scientists have been conducting EWAS for 100 years. I would go so far as to assert that the history of social science before the genomic era was essentially an extended attempt at EWAS. How has it come out? The answer depends on one's opinion of the incomprehensibly large body of studies, results and evidence that environmentally-oriented social science has produced, a full evaluation of which would take us far afield. This much can be said: although environmental social science has made many interesting discoveries, and described innumerable developmental processes, some of them plausibly causal, it has not formulated comprehensive explanations of the kinds of complex human characteristics it set out to understand. There is much to learn from the thousands of environmentally-oriented studies of juvenile delinquency, divorce, depression– the list is endless– but the reader who seeks a *theory* of juvenile delinquency, or put another way, who wishes to explain, to specify, a substantial chunk of the variability in juvenile delinquency that is broadly attributed to "the environment" will not be satisfied.

There is a subtle distinction to be made here about the kinds of explanations that are possible in social science. On the one hand, to the extent the goal is to explain the environmental etiology of something like juvenile delinquency in a general sense, to identify the specific factors that cause delinquency across a broad range of contexts, only the most general, if not platitudinous, explanations can be found: poverty is bad, stable families are good. But if the question then becomes, what is it about poverty that causes delinquency, is it schooling or peer groups or diet or environmental toxins, the missing environment problem asserts itself: it is at once all of these things and none of them. Together, they all add up to the construct we call poverty, which has a demonstrably negative effect; but one at a time, their effects are too small, and too dependent on context, to be quantified reliably or added together meaningfully.

Still, the content of social science would appear to comprise more than mere repetitions of associations among generalities, although there is certainly plenty of that. Any given study of delinquency, located in a particular time and place, produces its own set of findings, in the form of particular associations among individual variables, the ones that happen to have made it over the hurdle of statistical significance in this one particular study. They may have done so simply as a result of chance, or because they really were potent causes of delinquency in the particular

socio-temporal context embodied by the sample. We usually have no way of knowing which, but either way, social science has seen so many of these significant but ephemeral associations come and go that we no longer expect very much of them.

So in social science, we have a choice. We can characterize associations among very general constructs like poverty and delinquency, which may be expected to "replicate" from one situation to the next but don't actually tell us very much about the specific causal processes that are involved. Alternatively, we can immerse ourselves in the minutiae of the particular variables that seem to be associated with delinquency in a particular time and place, which offers a satisfying sense that we are actually explaining why something happened, but frustrates us with a maddening tendency not to replicate in the next study, conducted in a subtly different context. The result is either complacent satisfaction with predictable generalities, or endless Ptolemaic theorizing about finer and finer distinctions about the outcomes of different studies, until the field gets tired of the exercise and moves on to a new phenomenon. (See Meehl's 1978 account of theorizing about the "risky shift" in the 1950s).

3.8 GWAS and EWAS

I hope that the parallels between this situation and modern genomics are now obvious. For many years in genomics, twin studies were used over and over again to re-establish the vague generality that variation in genes is correlated one way or another with variation in phenotype, with variation in *every* phenotype. After a few decades, it became clear that reasserting the heritability of something had no more actual causal content than asserting that children who live in deprived neighborhoods do worse in school, or that older children do better on developmental tests than younger children. Then modern genomics arrived, finally permitting the attempt to break down the vague concept embodied by "heritability" into the tiny molecular processes that compose it, and in the human domain we are forced to do so without the methodological advantage of randomized experimentation. The unhappy returns of GWAS are the result.

The parallel failures of EWAS and GWAS suggest that these apparent shortcomings of old-fashioned social science never did reside in the genetic naiveté of traditional environmentalists, as so many prideful behavioral geneticists have led us to believe. Instead, the problem lies in the nature of complex human behavior itself, and as such it is not really a shortcoming. We do not have a general theory of juvenile delinquency because in an important sense juvenile delinquency will not bear general theorizing. Obviously, every delinquent teenager is delinquent for some set of reasons, but the causes of one teenager's delinquency do not generalize well to the delinquency of another. (For further discussion of these ideas, see the discussion of Meehl's concept of "specific genetic etiology" in Turkheimer, 1998, and the relevant Meehl papers referenced there.)

Considering the methodological parallels between the nonshared environmental and the genomic projects promotes a humbler appreciation of the possibilities for

the latter. There is, for starters, a deep irony underlying the genome project's obsession with tiny p levels. After a century of feckless application of NHST in the face of ever-increasing philosophical and statistical condemnation of the practice, traditional social science appears finally to be giving up the ghost on significance testing. At the same time, at the outer limits of our extraordinary ability to quantify the genetic sequence, NHST is rising again. Why? Is there something about genomics that we expect to vindicate a practice discredited by half a century of unsuccessful social science?

The meager contribution of NHST to classical social science focuses our attention on exactly what is proved by the atomically small p levels achieved by the height researchers. They demonstrate, and this much we can take as conclusive notwithstanding the attendant statistical assumptions, that the observed associations between SNPs and height are very unlikely to have occurred because of sampling error. The null hypothesis that human height is unrelated to SNPs, and by extension to allelic variation, has been busted. Unfortunately, nobody ever thought such a thing in the first place, so it's a pyrrhic victory. We stand reminded: associations between SNPs and distant outcomes are associations, that is to say correlations, and absent further evidence they are nothing more than that. NHST does not provide further evidence.

So after all of the extraordinary technology of modern genomics has done its work, the study of the genetics of complex human characteristics finds itself in the same unsatisfactory scientific stance as a sociologist in 1955, trying to make sense out of a vast catalog of non-experimental survey data that purports to explain why some juveniles become delinquents while others do not. Except that the geneticist's database is even larger, and the individual associations are, if anything, smaller. The tool that is supposed to help fix things doesn't work, having been designed for the task of discriminating sampling error from population variation, rather than the identification of causal needles lost in a haystack of correlations. The tool that might actually help—randomized experimention—isn't available for ethical reasons.

In the same way, the methods of controlling for population stratification in genomics correspond point by point to the statistical and quasi-experimental methods that social scientists have been using for a century: PCA (Price et al., 2006), instrumental variables (Lawlor et al., 2008) and propensity scores (Epstein, Allen & Satten, 2007). Like their social scientific counterparts they work, more or less, but are ultimately unable to solve the broad and deep problems of causal inference that necessitated them in the first place. If a confound to an association between an allele and height is as well-behaved as the model confound of chopstick use by Asian culture, then the extant methods will identify and control for it. But what if the allele is part of a developmental process that produces a child who is more successful in demanding nutritional resources from his or her parents? Is that a height gene, a marker of a "true biological effect" on height? The variety of causal pathways that could potentially be involved in a tiny uncontrolled association is so enormous that focusing on one class of them that can be identified with some reliability borders on the futile. The point is not that the relatively small magnitude of population stratification effects should promote a sanguine view of the possibilities for raw, uncorrected GWAS, as some papers have recently suggested (Hutchison et al., 2004), but

rather that fixed statistical procedures for controlling population stratification are no more likely to correct the real problem than highly stringent significance levels.

It would be unfair not to point out that these statistical methods have some advantages when they are used in genomics, compared to their traditional use in the social sciences. The one parameter that is generally constrained by theory in twin studies- the correlation of either 1.0 or.5 between the latent genotypes of monozygotic or dizygotic twins - is exactly one parameter more than is constrained in non-genetic analyses of the same kind of behavior. The predictors, predictions, and outcomes of non-experimental social science can multiply virtually without constraint, and the modest correlational structure imposed on them by population genetic theory explains the appeal genetic modeling has for its practitioners. In addition, GWAS allows geneticists to approach an empirical standard that environmental researchers cannot match, i.e., to catalog a nearly complete record of the genetic material of individual research participants. (Contemporary methodology based on SNPs is still a step removed from the actual genetic sequence, but those remaining barriers will probably come down soon.) One reason EWAS is not possible is that the complete environmental inputs of real humans are unrecordable in principle, and also because there is no discrete environmental theory that corresponds to the intricate modern synthesis of molecular genetics, population genetics and evolutionary biology. It is hard to imagine there ever will be.

Finally, just as with the nonshared environment, within-family designs have a special place in the molecular genetics of complex phenotypes. Comparisons of parents and children or pairs of siblings offer the single most reliable way to control for population stratification. If a pair of siblings reared in the same family differs at a genetic marker and also differs in chopstick use or delinquent behavior, the association between the allelic and the behavioral differences cannot be the result of a confound resulting from exposure to different cultural environments.[4] The analogy between social scientific and genomic applications of sibling difference designs helps to show population stratification for what it is: a shared-environmental confound of an observed association. Unfortunately, the same papers that have declared population stratification a “red herring” that can safely be ignored in GWAS have specifically concluded that sib-pair analyses are too demanding (Cardon & Palmer, 2003). Collecting 65,000 individuals for a GWAS study is one thing; collecting 30,000 sibling pairs is another.

Abandoning sib-pair comparisons would be a serious error. Environmentally-oriented social science has demonstrated quite conclusively that the sibling design is a far more effective way to weed out non-experimental confounders than its statistical competitors. That so many observed associations are discounted by the sibling

[4]As was the case for within-family studies of the environment, however, the existence of within sib-pair genetic associations still do not *prove* a causal relationship between the gene and the outcome. There still might be uncontrolled confounds within pairs (one member might be sent to a Japanese school where chopstick use is encouraged, while the other goes to an American school). The within-pair association controls for a class of confounds that vary between sibling pairs, which is a big help but not a panacea for the shortcomings of non-experimental science.

comparison is not a reason to discontinue its use, but is a measure of its success. It's too bad that so many associations turn out to be non-causal when exposed to risk of disconfirmation by the within-family design, but that's the way it goes. Even the limitation on statistical power imposed by the less than astronomical size of sibling samples is probably a good thing. As the magnitude of associations gets smaller and smaller, so does the probability that we will be able to make any developmental sense out of them (Turkheimer, 2006).

3.9 Genomic Social Science and Social Scientific Genomics

At several places in this essay I have compared GWAS to something called social science. What do I mean by that? Here is a working definition: social science is a domain of inquiry into human behavior is characterized by the following:

1) There are a large number of potential causes, individually small in their effects.
2) The causes are non-independent and non-additive.
3) Randomized experimentation is not possible.

It has been widely and sometimes triumphantly noted that to remain relevant, contemporary social science must be informed by genomics and affiliated biomedical sciences like neuroanatomy and pharmacology. It is less widely recognized that the road between social science and genomics runs both ways. Old modes of explanation in the social sciences have certainly been challenged by the introduction of genetic pathways into traditional causal models, but at the same time, the glittering technologies of modern genomics are finding their limits in the centuries-old methodological complexities of human science.

The three defining characteristics of social science magnify each other in complex ways. It is not necessarily a problem, for example, that a scientific domain consists of many small causal elements. Certainly many parts of human and non-human biology are built up out of very intricate networks of small causal effects. But how are such causal processes established? They are established via randomized scientific experimentation, much of it unspeakably gruesome if breathed in the same sentence as the word "human." (William Wimsatt, 1997, tells a story of a biophysicist challenged to define his field. He said, "take an organism, homogenize it in a Waring blender, and the biophysicist is interested in those properties that are invariant under that transformation.") Much (it would be interesting to speculate about how much) of the mystery that is human behavior might be elucidated if the full experimental armamentarium of the biologist were available to the psychologist, but even considering the possibility borders on the horrific.

GWAS of complex human characteristics is social science. It is possible to conduct meaningful science under such conditions, but there are strict, and sometimes crippling, limitations on the scope of the conclusions that can be drawn. In traditional social science, successful outcomes have been produced not by mechanical application of statistical procedures to vast correlation matrices in the hope of

finding "true" effects, but rather by careful administration of quasi-experimental methods across multiple domains to detect limited instances of local regularity. This is the strategy that will be successful in human genomics as well, but it is difficult to be optimistic based on current evidence. Most GWAS research remains intent on finding "genes for" one thing or another, based on the belief that there are "true biological effects" out there to be found.

On a more optimistic note, the recent popularity of GE interaction studies represents a step in the right direction. These studies begin with one of the small associations that are detected by GWAS, and proceed to refine it by identifying environments that modify it. In the paradigmatic study of the association between a gene encoding metabolism of MAOA and antisocial behavior (Caspi et al., 2002), for example, a variant known to be associated with antisocial behavior was shown to display the effect only in the presence of a stressful rearing environment. What is interesting in terms of the argument that has been made in this paper is that such a finding represents a *restriction* on the behavioral consequences of the allele, a step back from an attempt to promulgate a general theory of the causes of violent behavior or the consequences of stressful environments or MAOA. Of such small steps successful social science is made. The extraordinary impact of this study and others like it is testimony to the need to get beyond "gene finding" and the false hope, discouraging in the long run, that genomics will bring change to the long record of slow and imperfect partial explanation in the social sciences. (For a philosophical discussion of G×E interaction, see Tabery (2009).)

3.10 Conclusion

We have yet to conclude our account of the GWAS of height. When all was said and done, across the three papers, each comprising multiple studies totaling 65,000 participants and 400,000 SNPs, assessing a trait with a heritability of.9 and a reliability of measurement greater than that, the three studies identified 20, 10 and 21 "significant" SNPs, jointly accounting for 2.9%, 2.0% and 3.7% of the total variation in height. Of the 51 SNPs identified in at least one of the three studies, eight were found in two of them, and two were found in all three. Some of the SNPs replicated those found by earlier studies, some did not; some earlier linkages were replicated, some were not.

Yet despite what one might take to be fairly discouraging results, the study authors, and especially the accompanying editorial summarizing them, adopt an upbeat and even triumphant tone. In the editorial, Visscher concluded,

> The main conclusion emerging from the current studies is that GWAS are able to robustly identify common variants that are associated with height but that the effect sizes of individual variants are small, so that very large sample sizes are needed to detect associations reliably. Single laboratories are unlikely to have sufficient sample sizes to do powerful studies on their own, and the trend in human complex trait mapping has been to create consortia of research groups and even consortia of consortia. It remains unclear at this stage how

> much genetic variation can be explained through the GWAS approach. However, if the samples in these three studies were combined together with other datasets that have been collected on height and genome-wide SNP data, then this question could be answered empirically. Genome-wide studies on, say, 100,000 individuals, unthinkable only a few years ago, will be soon be a reality. (2008, p. 490)

And what then, in the coming era of consortia of consortia? Will we be more successful in combining causally ambiguous associations each explaining a tenth of a percent of the variance than we are now when they each account for one percent?

This implacable scientific optimism has been typical of behavioral genomics since its inception. The prescribed cure for the vanishingly small effect sizes typical of genomics has always been more statistical power, in the form of ever-larger sample sizes. But at some point, the field is going to have to grapple with the possibility that the difficulty is not statistical power at all, and therefore cannot be remedied by enormous sample sizes and stringent *p* levels. No one is prone to think anymore that the answer to the environmental etiology of juvenile delinquency is to be found in larger and larger samples, allowing detection of tinier and tinier associations with environmental risks. Environmental social science has learned a bitter lesson: the explanation of behavior is difficult not because the relevant causes, though countable and essentially additive, are small and difficult to detect; rather, social science is difficult because causes are innumerable and essentially *non*-additive (Turkheimer, 2004). What causes juvenile delinquency in one place or even one person doesn't necessarily cause it in another, and whether or not a particular environmental risk causes delinquency in a particular instance depends on so many other factors, environmental and genetic, that wide-ranging scientific explanations of important phenomena are not possible.

For most complex human characteristics, the optimistically expressed but largely unexamined claims of the discovery of "true biological effects" are quixotic. Effects can be true in the sense that they have a low probability of having resulted from sampling error, as demonstrated by significance testing, but the null hypothesis that allelic variation is unrelated to complex variation is not the real issue in GWAS any more than it is in EWAS. Of course allelic variation is associated with complex outcomes: the null hypothesis is always wrong.

The claim that an effect is truly "biological" is more difficult to understand. In the limited context of population stratification, the claim presumably means that a restricted set of competing causal claims related to the actions of other alleles or environmental exposures related to them has been ruled out or corrected for, but the range of competing causal claims that might actually be made is so wide that the remediations are unconvincing and (based on evidence to date) ineffective. But in practice, the claim of a "true biological effect" is intended to connote more than a careful exclusion of a few competing causal hypotheses. The unspoken claim is that assiduous attention to statistical significance and population stratification will lead to discovery of an allele with an *identifiable biological pathway* extending through the many levels of analysis separating the allele from the complex phenomenon it is purported to explain. If I am correct that this is what the GWAS researchers intend, it is no wonder that they don't unpack the content of the claim, because on minimal

examination it is so obviously false, false even for something not-really-so-complex as height, never mind delinquency.

In the same paper that produced the quotation at the beginning of this paper (Turkheimer, 1996), I introduced a distinction between two forms of biological explanation that I called weak and strong biologism. Weak biologism is the claim, which needs nothing more than a belief in philosophical materialism to establish it, that "biology" in one form or another (usually genes or brains) underlies all complex characteristics of organisms. In the modern era, almost everyone recognizes that weak biologism is universally true: there are few vitalists or spiritualists left anymore. Weak biologism, I suggested, is why everything is heritable; it is also why everything shows a complex pattern of small associations with individual genetic markers.

Strong biologism is the claim that a complex characteristic is a consequence of a "true biological effect," the specific result of a specific event at the genomic or neurological level of analysis. The relationship between Trisomy 21 and Down Syndrome, or between a stroke lesion in the left hemisphere and a resulting aphasia, are instances of strong biologism. Strong biologism is rare and scientifically compelling. Genetically oriented behavioral scientists (in those days mostly twin researchers) I argued, had identified a fool-proof move: claim strong biological explanation on the basis of weak biological relations that depend only on the inevitable instantiation of behavior in the brain and genome.

GWAS is a reassertion of this old strategy at the molecular genetic level. The endless repetitions of genome scans that identify a few weak-to moderate signals which then don't replicate very well in the next study is simply a rediscovery on the molecular level of what I (Turkheimer, 2000) have called the First Law of Behavior Genetics: everything is heritable. Everything is heritable because of weak biologism, GWAS is always bound to produce a few "results" because everything is heritable, and heritability is instantiated in the genome, in the same not very useful sense that cognition is instantiated in the brain. The solution to the missing heritability problem is to be found in the gaps between these universal but vague concepts of physical instantiation and actual mechanistic explanation of the complex characteristics of organisms.

References

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996): 'Identification of causal effects using instrumental variables'. *Journal of the American Statistical Association* 91: 444–455.

Bouchard, T. J., Lykken, D. T., McGue, M. Segal, N. L. & Tellegen, A. (1990): 'Sources of human psychological differences: the Minnesota study of twins reared apart'. *Science* 250: 223–228.

Campbell, D. T., Stanley, J. C. & Gage, N. L. (1963): *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Cardon, L. R. & Palmer, L. J. (2003): 'Populations stratification and spurious allelic association'. *The Lancet* 361: 598–604.

Caspi, A., McClay, J., Moffitt, T. E., Mill, J., Martin, J., Craig, I. W., Taylor, A., Poulton, R. (2003): 'Role of genotype in the cycle of violence in maltreated children'. *Science* 297: 851–854.

Cohen, J. (1994): 'The world is round (p<.05)'. *American Psychologist* 49: 997–1003.
Dick, D.M., Johnson, J.K, Viken, R.J. & Rose, R.J. (2000): 'Testing between-family Associations in within-family comparisons'. *Psychological Science* 11: 409–413.
Epstein, M. P., Allen, A. S., & Satten, G. A. (2007): 'A simple and improved correction for population stratification in case-control studies'. *The American Journal of Human Genetics* 80: 921–930.
Fisher, R. A. (1918): 'The correlation between relatives on the supposition of Mendelian inheritance'. *Transactions of the Royal Society of Edinburgh* 52: 399–433.
Gudbjartsson, D. F., Walters, D. F., Thorleifsson, H. S., Halldorsson, B. V., Zusmanovich, P. et al. (2008): 'Many sequence variants affecting diversity of adult human height'. *Nature Genetics* 40: 609–615.
Hamer, D. & Sirota, L. (2000): 'Beware the chopsticks gene'. *Molecular Psychiatry* 5: 11–13.
Harlow, L., Mulaik, S. A. & Steiger, J. H. (eds.) (1997): *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum.
Hutchison, K. E., Stallings, M., McGeary, J. & Bryan, A. (2004): 'Population stratification in the candidate gene study: Fatal threat or red herring?' *Psychological Bulletin* 130: 66–79.
Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. & Davey Smith, G. (2008): 'Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology'. *Statistics in Medicine* 27: 1133–1163.
Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I. et al. (2008): 'Identification of ten loci associated with height highlights new biological pathways in human growth'. *Nature Genetics* 5: 584–591.
Loehlin, J. C. (1992): *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*. Lawrence Erlbaum Associates.
MacCorquodale, K. & Meehl, P. E. (1948): 'On a distinction between hypothetical constructs and intervening variables'. *Psychological Review* 55: 95–107.
Meehl, P. E. (1978): 'Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology'. *Journal of Consulting and Clinical Psychology* 46: 806–834.
Plaisance, K. S. (2006): *Behavioral Genetics and the Environment: The Generation and Exportation of Scientific Claims*, unpublished dissertation, University of Minnesota.
Plomin, R. & Crabbe, J. (2000): 'DNA'. *Psychological Bulletin* 126: 806–828.
Plomin, R. & Daniels, D. (1987): 'Why are children in the same family so different from one another?' *Behavioral and Brain Sciences* 10: 1–16.
Price, A.L., Patterson, N.J, Plenge, R.M., Weinblatt, M.E, Shadick, N.A. & Reich, D. (2006): 'Principle components analysis corrects for stratification in genome-wide association studies'. *Nature Genetics* 38: 904–909.
Rodgers, J. L., Cleveland, H. H., van den Oord, E. & Rowe, D. C. (2000): 'Resolving the debate over birth order, family size, and intelligence'. *American Psychologist* 55: 599–612.
Rosenbaum, P. R, & Rubin, D. B. (1983): 'The central role of the propensity score in observational studies for causal effects'. *Biometrika* 70: 41–55.
Rutter, M., Pickles, A., Murray, R,. & Eaves, L. (2001): 'Testing hypotheses on specific environmental causal effects on behavior'. *Psychological Bulletin* 127: 291–324.
Schmidt, F. L. (1996): 'Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers'. *Psychological Methods* 1: 115–129.
Silventoinen et al. (2003): 'Heritability of adult body height: A comparative study of twin cohorts in eight countries'. *Twin Research and Human Genetics* 6: 399–408.
Tabery, J. (2009): 'Difference mechanisms: Explaining variation with mechanisms'. *Biology & Philosophy* 24: 645–664.
Turkheimer, E. (1998): 'Heritability and biological explanation'. *Psychological Review* 105: 782–791.
Turkheimer, E. (2000): 'Three laws of behavior genetics and what they mean'. *Current Directions in Psychological Science* 9: 160–164.
Turkheimer, E. (2004): 'Spinach and ice cream: Why social science is so difficult'. In. L. DiLalla (ed.): *Behavior Genetics Principles: Perspectives in Development, Personality, and Psychopathology*. Washington, DC, US: American Psychological Association, pp. 161–189.

Turkheimer, E. (2006): 'Interaction and play'. *PsycCRITIQUES* 51: 43.
Turkheimer, E. & Waldron, M. C. (2000): 'Nonshared environment: A theoretical, methodological and quantitative review'. *Psychological Bulletin* 126: 78–108.
Visscher, P. M. (2008): 'Sizing up human height variation'. *Nature Genetics* 40: 489–490.
Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M. et al. (2008): 'Genome-wide association analysis identifies 20 loci that influence adult height'. *Nature Genetics* 40: 575–583.
Wimsatt, W. (1997): Transcripts from "modularity of animal form". Proceedings of the evolvability of developmental mechanisms short course, http://celldynamics.org/evolvacourse/transcripts/BillWimsatt.html

Chapter 4
Genetic Traits and Causal Explanation

Robert Northcott

4.1 Introduction

Many traits and dispositions are labeled *genetic*, and such labeling is both widespread in science and apparently far from arbitrary. Yet it is a truism that every trait is the end-product of a complex developmental process involving both many different genes and a multitude of environmental and other epigenetic factors. How then can some traits usefully be termed genetic and others not?

In this paper, I import influential recent theories of causal explanation to develop an account of genetic causation, and hence of genetic traits and dispositions. My thesis is that the latter two are best seen as *explanatory claims*; in particular, a trait is genetic, roughly speaking, just in case it is explained by genes or it is *not* explained by environment. The greater novelty lies in how this idea is made precise in the form of a relational definition, an implication of which is that no trait is genetic always and everywhere. Rather, every trait may be either genetic or non-genetic, depending on explanatory context.

Precisely because genes are part of the causal history of any trait, merely tracking which traits have genetic causes is insufficient for picking out some rather than others to be 'genetic'. Some authors have therefore sought to supplement that by defining genetic traits to be those where there is some special kind of connection between gene and trait over and above mere causation. Examples include: that genetic traits are those that are caused by genes especially 'directly' (Hull 1981), or that there is a special kind of necessity linking them to genes (Gifford 1990). A stronger tradition has been to analyze the matter in terms of *differences* – this gene rather than that

R. Northcott (✉)
Department of Philosophy, Birkbeck College, Malet Street, London WC1E 7HX, UK
e-mail: r.northcott@bbk.ac.uk

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*,
Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_4,

one is the cause of this trait rather than that one (Sterelny and Kitcher 1988). Indeed, a focus on differences lies at the heart of the statistical techniques standard in behavioral genetics, which look for correlation between *variation* in phenotype and *variation* in genes. That still leaves it to be decided though exactly what pattern of correlation or variation is taken to be characteristic of genetic traits in particular. One crude approach, seldom advanced explicitly, would be to look simply for high statistical heritability. More sophisticated definitions can also be framed in terms of such statistical relations though, such as Smith (2007); these likewise focus on differences. (All these previous approaches are discussed in section 5.)

This paper's approach too focuses on the causal connection between gene and trait, and is also framed in terms of differences. But it is connected explicitly to the wider literature from philosophy of science on causal explanation. Among the benefits resulting from this, a major one is that it yields an analysis of the relation between genetic *dispositions* and traits (section 6). These two concepts are clearly distinct and as a result used differently but, to my knowledge, their relation has not been analyzed previously. It turns out that a well founded account of genetic dispositions is crucial to understanding whether to call genetic or non-genetic the many traits that have statistical heritabilities around 0.5 in the populations in which they have been studied.

This paper is theoretical in the sense that it is initially couched in terms of theories of causation rather than in terms of biological or medical practice. Nevertheless, there is value in first investigating just what sense can be made of genetic causation or explanation in the abstract, so to speak, before then examining the applicability of whatever version of those notions turns out to be theoretically defensible.

4.2 Contrastive Explanation

The core of this paper will be to analyze genetic explanation. To that end, I shall adopt the leading contemporary theory of causal explanation, which attributes to such explanations a *contrastive* structure – a cause-rather-than-contrast explains an effect-rather-than-contrast. To illustrate, consider the claim 'Socrates sipping hemlock explains why he died'. That sounds plausible enough, but consider two possible clarifications:

1) 'Socrates sipping *rather than guzzling* hemlock explains why he died.' (Seems wrong.)
2) 'Socrates sipping hemlock *rather than wine* explains why he died.' (Seems right again.)

The lesson is that explanatory properties are sensitive to choice of contrast. A similar lesson applies to effects as well as causes. Imagine that a short circuit ignites a mixture of wood and potassium salts, yielding a purple fire. Then:

1) The short circuit explains the purple fire rather than *no* fire.
2) But it's the potassium salts that explain the purple fire rather than *yellow* fire.

The contrastive view dates from Dretske (1972). Notable developments of it include van Fraassen (1980), Garfinkel (1981), Achinstein (1983), Hitchcock (1996), and – most influential recently – Woodward (2003). I leave further details and justifications to those works, and here focus instead on how to apply the contrastive apparatus to genes.

Begin with the explanandum, i.e. with the object of a genetic explanation. Historically, this has variously been taken to be a behavior, trait, disposition, or piece of knowledge. I shall take behaviors and knowledge to be examples of phenotypic traits or dispositions. The analysis to be developed here will therefore concern just the latter two. To begin with I shall focus on traits, before turning in section 6 to dispositions.

Label the particular actual trait of interest T_a. In accordance with the contrastive view, it is crucial also to specify a contrast to T_a, so label that T*. To see intuitively how choice of T* matters, consider T_a = my actual two legs, and T* = I have only one leg. This represents a paradigm case of a genetic trait – what about me could be more genetically influenced than that I have two legs rather than one? Now, as it happens, one of my legs is actually slightly bent due to a childhood accident. This suggests an alternative contrast of T^* = my two legs are both straight.[1] In order to explain why my leg is bent rather than straight, i.e. T_a rather than this new T*, we would now appeal to my accident and *not* to genes.[2] That is, for the first choice of T* we deem genes to be explanatory, while for the second choice we do not.[3] Yet T_a is identical in both cases, namely my actual legs. So just specifying T_a alone is insufficient.

Matters become a little more complicated when we turn to the explanans, i.e. to a trait's causal history. The contrast for the explanans is some alternative version of that history. More precisely, it is some alternative event within that history plus all the causal consequences of that alternative event. I shall be interested only in a

[1] The different choices of contrast here can be seen as a formal device to pick out different *aspects* of my legs – respectively, their number and their straightness. Choice of T* can also serve to represent the distinction (Sober 1998, 795) between a trait's initial development and its subsequent modifiability.

[2] Moreover, in some contexts even being born with a defect may also be explained environmentally – see the discussion of a thalidomide example in section 5 below.

[3] The objects of ascriptions of genetic explanation may form a broader class than 'traits' as that term is customarily used in biology. On some views, for instance, the single event of my currently having my actual legs can, strictly speaking, correspond to more than one trait, depending on whether we are focusing on my legs' number or straightness. In effect, this would be to individuate traits in part via explanatory context. I shall ignore controversies over exactly how traits should be defined, and instead shall use the term rather liberally. Generally, I shall understand the relata of a causal explanation to be two pairs of events, or more precisely, two pairs of an actual and contrast event, and my eventual definitions will be framed accordingly. Therefore I retain the text's specifications of T_a and T*, even at the cost of occasional conflict with some views of trait individuation. (The substance of this paper's case could be made even given other choices of explanatory relata than pairs of events, and so for ease of exposition I shall sometimes leave it underspecified exactly what kinds these relata are.)

trait's causal history since the relevant organism's conception. Of particular importance to us is that we may partition this history into two types of factor, colloquially labeled 'genes' and 'environment'.[4] More precisely, the partition of the causal history is into two highly asymmetric portions – first, the particular genome at the moment of the particular organism's conception; and second, the entire rest of that organism's developmental history. Corresponding to this division, contrasts associated with the explanans are also of two types – those that cite alternative particular genomes at conception, and those that instead cite alternative versions of the rest of the developmental history.[5]

4.3 A definition

We may now formulate an explicit definition. It is intended to apply to token cases, i.e. to particular traits of particular organisms. Let T be a function that takes an organism's causal history as input, and yields trait values as outputs. Formally, let T_a = the actual trait value; G_a = the actual genome at conception; and E_a = the rest of the actual developmental history. The organism's complete history is therefore $G_a\&E_a$. And let T* = the salient contrast trait value; G* = the salient contrast history resulting from the substitution[6] in of an alternative genome at conception; and E* = the salient contrast history resulting from the substitution in of some other alternative event in the developmental history.[7]

Some clarifications from the start will be useful. In all cases, the contrast-effect will be a T*, and the contrast-cause either a G* or an E*. (As noted, in any one case the contrast-cause will be of either G* type or E* type, never both simultaneously.) T*, G*, and E* are all counterfactuals. For both G* and E*, the only differences from the actual history are, to repeat, the initial substitution itself plus that substitution's causal consequences. Intuitively, this amounts to a *ceteris paribus* provision – we are

[4] I ignore the occasionally fuzzy borderline between these two categories because, with regard to genetic traits, that has not been the salient locus of philosophical dispute. Rather, the strengths and weaknesses of various analyses have concerned other matters (see section).

[5] In principle, of course, one could formulate contrasts that vary *both* these aspects of the explanans. As a matter of fact, such 'combined' contrasts do not ever seem to be salient in actual disputes (see also van Fraassen 1980, 126). But if they were, in my view consideration of them would give us no information regarding genetic explanation.

[6] 'Substitution' is intended here as a neutral term that may be taken to correspond either to a 'miracle' in Lewis-style possible-worlds semantics, or to an 'intervention' in the semantics of the causal modeling literature. Again, for the purpose of elucidating genetic traits I do not endorse any particular semantics for counterfactuals in general, since that is not the salient locus of philosophical interest.

[7] In general, T*, G* and E* may refer to *sets* of contrasts. For ease of exposition, I shall often assume them to be singletons. A definition analogous to [GT], given below, can easily be formulated for non-singleton cases too. (See Northcott 2008b for more on non-singleton contrasts.)

interested in what would have occurred had G_a or E_a been different but with nothing else changed. For example, what would have happened had a mutated version of a gene been present, keeping the rest of the genome and all epigenetic factors the same? In this way, we would isolate the causal contribution in this particular situation of that gene alone.

For ease of exposition, I shall often denote a G* or E* just by the alternative substituted in, for instance 'G* = alternative genome X'. Then, our definition is:

1) When a G* is salient, T_a is *genetic iff*: $T(G^*) = T^*$
2) When an E* is salient, T_a is *genetic iff*: $T(E^*) \neq T^*$ [GT]

T(G*) denotes the trait that would have resulted from the alternative causal history represented by G*. On a contrastive view, the conditions for full explanation are: that the actual cause yields the actual effect; *and* that the contrast-cause would have yielded the contrast-effect. The first condition is satisfied automatically here, since by assumption $T(G_a\&E_a) = T_a$.[8] When a G* is salient, the second condition is satisfied *iff* T(G*) = T*; when an E* is salient, it is satisfied *iff* T(E*) = T*.

In words, definition [GT] amounts to saying that a trait is genetic just in case it is explained by genes or it is *not* explained by environment. If genes made the difference, the trait is genetic; likewise, it is genetic if environment did *not* make the difference. Formally, these correspond respectively to the conditions T(G*) = T*, and $T(E^*) \neq T^*$.[9]

It is important to be clear from the start on the type-token distinction. Definition [GT] applies only to the token case of a particular trait of a particular organism. Which contrasts are salient will obviously vary with context. Moreover, the function T is essentially a device for representing causal relations, and such relations are also context-dependent. Whether striking a match causes fire, for instance, depends on whether the match is wet, whether there is sufficient oxygen, whether it is windy, etc.

Claims that a trait is genetic are often made at the *type* level – 'Down's syndrome is genetic' or 'scars are not genetic'. How then can [GT] be applied to them? My answer is that such type claims are implicitly about particular populations of token cases. In particular, usually they tacitly assume 'typical' or 'normal' populations and explanatory concerns, corresponding to particular (collections of) choices of T* and G* or E*. For each of the token cases within the population, the conditions in [GT] are deemed to be satisfied. Any given case of Down's syndrome, for

[8] Strictly speaking, this assumption is reliable only given determinism. For simplicity, I shall consider here only the deterministic case. Evaluation of indeterministic counterfactuals, and of probabilistic causal-explanatory claims more generally, goes beyond the scope of this paper. In practice though, uncertainty due to indeterminism seems only rarely to be the focus of actual disputes about whether a trait is genetic.

[9] Given that $T(G_a\&E_a) = T_a$, claims that a trait is genetic thus boil down to evaluations of particular counterfactuals. Of course, as with all counterfactuals, these ones may be vague or indeterminate. In so far as they are, then so likewise I claim are the associated genetic claims.

example, could not have been avoided by any salient alternative environmental input – or at least that is what is being asserted when that trait is described as 'genetic' at a type level.

A common worry is whether it makes sense to talk of contrasting either genes or environment in isolation, given the two factors' obvious continual interaction. Both gene-environment interaction and correlation are standard difficulties in population-level analyses, but at the token individual level the worry is misplaced. For example, suppose I had been born with greater athletic talent.[10] Possession of, so to speak, a more athletic genome would presumably have led to a different environment too relative to the one I actually did experience – perhaps it would have led me to seek, and to have been given, more intensive athletic training.[11] But that does not render this counterfactual any harder to evaluate than counterfactuals in general. For any genetic contrast G* (or environmental one E*), the history need be held fixed only up to the time of the antecedent, i.e. up to the time of the relevant substitution. The subsequent history can then unfold as it may. In causal modeling terms, there is no problem if the subsequent history changes too, so long as it does so only as a causal consequence of the initial intervention (Woodward 2003).

As is well known, the very same genome may yield dramatically different phenotypic outcomes depending on environment. This is just gene-environment interaction at work. One consequence is that the phenotypic significance of a particular genetic difference may also vary with environment. For instance, anticipating an example below, even though genes may explain why I am taller than my father, it is also possible that in a very nutrient-poor environment both of us alike would have grown only to five feet. In other words, our genetic difference might yield a height difference only in some environments not others.

Formally, such context-dependence is already incorporated into [GT], via the sensitivity of the T function to environmental conditions. Nevertheless, the issue can still lead, intuitively at least, to problems for any causation-based analysis of genetic traits. A familiar example, discussed for instance in Sesardic (2005), illustrates how: suppose that all red-haired children are discriminated against because of their hair color, receiving no formal education and thus scoring poorly on scholastic tests. Suppose also (plausibly) that red hair is caused by the presence of certain genes (or, strictly, by the presence of certain alleles of those genes rather than others). It follows that, given the discriminatory social environment, those genes will also

[10] On some views of personal identity this more athletic creature would no longer be 'me' at all. For our purposes, the difficulty is not a deep one. If desired, the discussion can be re-phrased without loss as whether a counterpart with the particular alternative genome at conception would have ended up with the same traits as me.

[11] This would be an instance of gene-environment *correlation* – the more athletic genotype is correlated with a more intensive training environment. In addition, more intensive training presumably itself causes better athletic performance. If this extra effect on performance is greater in the case of the more athletic genotype than the less athletic one, it would be an instance too of gene-environment *interaction* – i.e. when the impact of a particular environment varies depending on the genotype.

cause reduced scholastic achievement. Yet presumably, given normal education red-haired children would have done just as well as other children. Thus we naturally explain the low scholastic scores by appeal to the social discrimination, not to the 'red-hair genes'. Yet, the objection concludes, formula [GT] will count the trait as explained by genes, and thus (sometimes) as genetic, nevertheless.

In reply, first note that this is an issue for everyone (as Sesardic's discussion makes clear). Any definition based on statistical correlation, for instance, will likewise endorse a large genetic role here. Next, there is no dispute that in such a scenario the 'red-hair genes' *are* causes of low scholastic achievement, via the intermediate cause of the discrimination that they induce. The problem is thus a version of Mill's classic one of causal selection: there is no ontological preference for the 'good' cause (social discrimination) over the 'awkward' one (i.e. the genes). The conventional wisdom is that distinguishing between the two causes can therefore only be done by appeal to pragmatic factors. The issue here is that our common emphasis on the discrimination rather than the genes is presumably the result of a moral judgment, perhaps intertwined with awareness of possibilities for intervention.

It may well be that, as a result, our notion of 'genetic trait' is infused with these extra pragmatic considerations, in which case no purely logical analysis of the term will ever be fully satisfactory. On the other hand, a recent strand in the causation literature addresses this very issue. Backed by empirical evidence, it suggests that a 'cause' is, by definition, roughly speaking that particular counterfactual dependency that is rendered salient by normative considerations (Hitchcock and Knobe 2009). So understood, we might, as desired, pick out as a cause in this example the dependency on social discrimination but not the dependency on red-hair genes. A causal definition of a genetic trait would therefore be endorsed again. Once full resolution of these theoretical issues is reached, the formulation in [GT] might be adjusted appropriately.

4.4 Objectivity and Context

It follows from definition [GT] that being a genetic trait is a relational property, of the form T_a *is genetic relative to T* and G** (or to T* and E*, as the case may be). There is no absolute fact of the matter, independent of explanatory context. More formally, whenever we ask whether some trait is genetic, on my view a presupposition of the question is a particular specification of contrasts. That these relativizations are often not explicit does not show that they are not present, only that they are tacit. The intuition against the thought that the same trait could be both genetic and not genetic, is explained as being the result of a violation of pragmatic maxims dictating relevance to our conversational presuppositions (in particular, to the presupposition fixing only a particular specification of contrasts as salient). Similarly, explanatory claims in general often have a non-relational surface form even though really they are relational.

Such a relational property is not arbitrary. In particular, once (but only once) given a specification of contrasts, the truth of whether a trait is genetic is clearly objective – or anyway as objective as the evaluations of the relevant counterfactuals. It is thus certainly possible to rule out some claims as erroneous. For example, assertions that some individual's criminality is genetic may, in context, be assertions that given the same upbringing or schooling an individual with a different genome would not have committed crimes. That claim is an objective causal hypothesis.

Often, contrasts are not specified explicitly. Therefore we still require an account of how, in those cases, they are determined implicitly. The answer, as already mentioned and, as for any conversational presupposition, is conversational context. Can we flesh that out in more detail? One general constraint we have already seen – contrasts are counterfactual, i.e. $T^* \neq T_a$, and neither G* nor E* replicates the organism's actual developmental history $G_a\&E_a$. A further general constraint is that contrasts must, so to speak, genuinely contrast. I propose that T_a and $G_a\&E_a$ must each be nomologically incompatible with their associated contrasts. (This ensures that T_a and $G_a\&E_a$ constrain the definition at all.) However, it also remains true that, beyond this, there exists no generally agreed-upon formal procedure for nailing down in every case exactly how circumstances do specify choices of contrast.[12] How then can we ever be confident what the intended contrasts are, and so how can [GT] ever be tested? The best way, it seems to me, and therefore the best evidence for [GT], is deliberately to manipulate contrasts and then to track whether our judgment that a trait is genetic indeed consequently varies in the manner predicted. Obviously, no exhaustive catalogue of cases is possible so I can only appeal to particular illustrative examples.

To this end, consider T_a = my height, and $G_a\&E_a$ = my own life history. Suppose I ask the explanatory question, why am I taller than a neighbor of mine who grew up in a poor family? This immediately suggests the contrast T* = my neighbor's height, and, roughly speaking, either G* = my neighbor raised in my environment, or E* = me raised in my neighbor's environment. If we judge that the neighbor would not have grown any taller even given my wealthier upbringing, i.e. $T(G^*) = T^*$, then sure enough we correspondingly judge that my greater height is indeed genetic – because explained by our different genes. And likewise if, in the E* case, we judge that if raised in his environment I too would have attained only my neighbor's height, i.e. $T(E^*) = T^*$, then, as per [GT], we would *not* judge my greater height to be genetic – because explained instead by my upbringing.

Next, suppose I set, for the same actual trait T_a = my height, a different conversational context by asking, why am I taller than my father? Now the contrasts are naturally T* = my father's height, and, roughly speaking, either G* = my father raised in my environment, or E* = me raised in my father's environment. If I judge that my father would likely have reached my height if he'd had my nutritionally superior upbringing, i.e. that $T(G^*) \neq T^*$, then our height difference is not explained by

[12] See Schaffer (2005), Maslen (2004) and van Fraassen (1980) for further discussion.

genes – and sure enough is not judged genetic. Similarly, if I believe that I would have attained only my father's height given his upbringing, i.e. that $T(E^*) = T^*$, then our height difference is explained by environment and once more we accordingly judge it not genetic. Again, judgment tracks the predictions of [GT].

Notice how the very same trait, namely my height, is thus deemed genetic in some explanatory contexts but not in others. Indeed, according to [GT], a similar fate awaits *any* trait. We already saw this also for T_a = my legs (section 2).

Does this not all leave too much weight resting on which contrasts are deemed 'salient'? No, and it is important to appreciate in exactly what way that issue matters – and in what way it doesn't. For sure, according to this paper's account whether a given trait is deemed genetic depends critically on choice of contrasts. But in order to *test* that account, what is relevant is that judgments of whether a trait is genetic track changes in contrasts in the manner claimed. As noted, we may test that in turn by manipulating choice of contrast independently and then seeing whether our judgment follows as predicted. Therefore it is not fatal that we have no foolproof algorithm for generating choice of contrast in every context. That merely implies that there may be no fact of the matter regarding whether a trait is genetic before contrasts are specified – which is exactly what a relational definition such as [GT] is claiming anyway. The demand for an objective determination of contrasts perhaps betrays an unspoken but incorrect assumption that there should always be some fact of the matter regarding whether a trait is genetic. But if context leaves choice of contrast indeterminate, the existence of such a contrast-independent fact of the matter is precisely what [GT] denies.

4.5 Relation to Previous Literature

Many of the conclusions from this paper are familiar. Sterelny and Kitcher (1988) long ago argued that genetic causation should be understood context-specifically. We can speak of a gene causing a phenotypic trait in a particular case even though that gene may not do so in every circumstance. In this paper, that idea is generalized and formalized by defining the explanans of a genetic explanation to be $G_a\&E_a$-rather-than-G*, i.e. one genome rather than another. The sensitivity to extra-genomic environment is represented via the sensitivity of the trait function T to E_a. Sterelny and Kitcher's explication of the 'gene for' locution can be represented similarly. Generally, the common emphasis on genetic *differences* explaining trait *differences* is captured naturally by a contrastive apparatus. This paper's account of genetic explanation shares features with other accounts too. It shares the emphasis on pragmatic relativization in Gannett (1999), for instance, and the emphasis on the token case of Waters' (1994) 'Difference Principle'. The point here is to demonstrate how these conclusions can be expressed in terms of, and endorsed by, the wider theory of causal explanation.

As noted earlier, one thing that has proved elusive in philosophy generally is an objective algorithm for determining what contrasts are implied by any given context.

Nevertheless, that does not mean that we cannot examine which pragmatic factors tend to influence choice of T*, G* and E*, or, more particularly, that influence whether it's a genetic cause that is picked out and thus a trait declared 'genetic'. Indeed, in effect much of the literature has focused on exactly this issue. Often, what makes a certain cause salient is the research program or goals of the particular scientist. Some may focus only on those causes that actually vary in the population for instance (Waters 2007), others on those that do not actually vary but that are nevertheless potentially the target of efficacious interventions. But, according to [GT], these disputes regarding goals are not disputes regarding the *definition* of whether a trait is genetic.

Smith (2007), focusing particularly on the example of genetic diseases, surveys many causal selection criteria that have been used to justify picking out traits as genetic, such as those mentioned in section 1: unusually 'direct' causation by genes; and genes being the only 'abnormal' factor in a trait's causal history. In this paper's terms, these criteria can be seen as criteria for selecting some rather than other contrasts. Smith argues convincingly that none of them is adequate in the face of actual biological complexity. One common weakness is their lack of relationality – that is to say, a given trait will be declared either genetic or non-genetic *simpliciter*, with no allowance for variation with context.

Standard methods in behavioral genetics and other sciences can be used to define statistical heritability – roughly speaking, the proportion of phenotypic variance in a given population that is 'due to' genetic variance. Glossing over many details, this essentially tracks statistical correlation between genetic and phenotypic variance. One attractive feature of statistical heritability in this context is that it incorporates a focus on differences, and thus is able to allot different scores to different traits despite the fact that all traits alike have genetic causes. Moreover, because it is population-specific, statistical heritability also incorporates a certain relationality. Nevertheless, notwithstanding its possible usefulness for other purposes, high statistical heritability is a poor candidate for a definition of whether a trait is 'genetic'. Notoriously, a trait commonly thought genetic, such as number of legs, will often score very low for heritability simply because it is almost universal and what little variation there is in the population is due to environmental factors such as accidents. As it were, the relationality here is of the wrong sort – it is not sensitive to the appropriate counterfactuals, but rather to other members of an actual population. As a result, it is much disputed whether any causal inferences follow from heritability scores at all.[13] Moreover, population-level statistics such as heritability are inapplicable to individual-level cases.[14] Thus it is necessarily beyond its purview whether my own eye color, for instance, is genetic.

[13] For a sample of such attacks, see Lewontin (1974), Shipley (2000), Spirtes et al. (2000), and Northcott (2008a). In the philosophical literature, Sesardic (2005) mounts the most vigorous defense of heritability against this consensus.

[14] Sober (1988) argues otherwise, although see Northcott (2006) in response.

Smith (2007) offers his own proposed analysis of genetic traits, in particular of genetic diseases, also based on population frequencies but much more sophisticated than simple statistical heritability. Nevertheless, the same basic worries apply, namely the lack of connection to the wider causation literature and the inapplicability to individual-level cases.

Another proposal might be to define genetic traits in terms of evolution, perhaps to be those that make a difference to evolutionary outcomes.[15] After all, evolution is often defined as change in gene frequencies within a population, and traditionally one of the necessary conditions for a trait to evolve is that that trait be heritable. But a closer look gives pause. First, every trait is influenced by genes and thus potentially makes a difference to evolutionary outcomes, so that alone is insufficient to distinguish between traits in a principled way. Further, some genetic diseases, such as Klinefelter's syndrome, are not heritable and thus presumably do *not* make a difference to evolutionary outcomes in the way envisioned.[16] Moreover, finally, much contemporary theory emphasizes the importance of non-genetic mechanisms in evolution anyway (Sterelny 2003, Oyama et al. 2001).[17]

I believe it is an open question whether a purely logical analysis exists that does successfully track our every usage of 'genetic' with respect to traits. There is certainly no guarantee that any should. (Indeed the red-haired example in section 3 suggests, consistent with [GT], that some pragmatic relativization is unavoidable.) But the point of this paper is something different, namely to formulate a connection to wider philosophical literature. How exactly the contrasts in [GT] should be filled in is a separate matter, about which [GT] is not directly concerned since it remains neutral on where contrasts come from. As noted earlier, the 'test' of [GT] is, rather, that once contrasts have been determined, matters then track [GT] as predicted. Embedding our analysis into the theory of explanation, meanwhile, offers several benefits of its own, to which I turn now.

First, consider why we should even care whether a trait is genetic or not. What normative punch could ever result from such a claim? This paper's account, by way of its connection to the causation literature, offers an answer – the counterfactuals that, according to [GT], comprise such claims are also exactly those that license *interventions*. Declaring my eye color to be genetic, for instance, asserts that – in context – no salient environmental intervention could have altered it in the past, or could alter it now. Calling a trait genetic thus serves to pick out which interventions are (or would have been) efficacious, and which not. Implications follow immediately

[15] I thank an anonymous referee for this suggestion.

[16] Klinefelter's syndrome describes XXY individuals, i.e. males with a third sex chromosome. Its cause is an error during meiosis rather than a father with the condition. (Indeed, before modern reproductive technology one of the condition's symptoms was infertility.)

[17] An alternative proposal is that genetic traits are those that are evolutionary adaptations. But this again seems to track actual usage badly. For example, many genetic diseases, such as cystic fibrosis or Down's syndrome, are hardly adaptive.

for explanation and blame. For example, deeming a disease to be genetic implies that it cannot be blamed on any (salient) aspect of parental care.

As ever, the emphasis on salience is critical here. Take the disease phenylketonuria, or PKU. As is well known, this was traditionally thought a 'genetic disease' in that it resulted from a particular genome, possession of which made the onset of the disease almost inevitable. That is, for any normal genome G*, T(G*) = T*, where T* = a normal child, in contrast to a child suffering from PKU. Moreover, it was not thought that any environmental intervention could avert this outcome, hence T(E*) ≠ T*. Famously though, it is now known that a special diet can alleviate (most of) the impact of the disease. In this sense, it is no longer a 'genetic disease', and the status of PKU has become ambiguous, as given the contrasts that are likely salient now, cases of it may be explained by either genes or environment. Depending on which particular contrast is involved, the disease thus may or may not be appropriately described as genetic. (Or, in particular circumstances, it may now be best described as a genetic *disposition* – see section 6.)

An explicitly causal scheme also highlights an ambiguity that has tended to be overlooked – namely, whether we are talking about genetic *causation* or merely genetic *explanation*. In the metaphysics literature, a contextual-contrastive view is relatively standard now with regard to explanation, but with regard to causation itself it is much more controversial (e.g. Davidson [1967] 1980). In particular, relativization to contrasts introduces a pragmatic element that has traditionally been thought characteristic only of explanation. On the other hand, there is a growing segment of the literature that does endorse an explicitly contrastive view of causation itself (Northcott 2008b, Schaffer 2005, Maslen 2004).[18] This view is also endorsed by the contemporary Bayes net and causal modeling literatures (Pearl 2000, Spirtes et al. 2000), as well as, as already noted, by Woodward (2003). On such a view, choices of contrast influence not just whether genes *explain* some trait but also whether they *cause* it. The point is that much talk of genetic causation so far has carried an (unacknowledged) metaphysical commitment to such causal contrastivism. It is better to be open about this.

Next, definition [GT] easily handles borderline cases. For example, male bellies tend to expand with age. This phenomenon is genetic in the sense of occurring in a wide range of human environments, and because it is often explained by having a male rather than female genome. But it is also not genetic in the sense that it could be avoided by eating less. Formally, for T_a = my middle-aged male cousin's large belly, and T* = a smaller belly, it is true that for G* = a typical female genome, T(G*) = T*, i.e. a woman with the same lifestyle as my cousin would have had a smaller belly. Moreover, for E* = many alternative environments than the one my cousin has experienced, such as that in many other cultures, $T(E^*) = T_a$, i.e. the large belly is not explained by my cousin's particular environment. So the trait seems

[18] And also of probabilistic causation (Hitchcock 1996).

genetic. But for *other* choices of E*, such as E* = he ate less, now T(E*) = T*, i.e. now environment *is* explanatory, and so the trait no longer seems genetic. The point is that either choice of E* will likely be salient quite often. As a result, frequently the size of the belly will seem genetic but frequently also it will not. The larger point is that, as with many traits, I have no clear intuition as to whether expanding male bellies *simpliciter* 'are genetic'. Rather, my judgment only *becomes* clear once given a particular explanatory context, just in the manner that [GT] predicts.

Perhaps we can now also see a connection between a trait being genetic and it being *innate*. In particular, might the concept of innateness be equivalent to genetic explanation and lack of environmental explanation? To investigate that would take us beyond this paper and into the huge literature on innateness and the many complications to measuring 'nature versus nurture'. The point here is merely to flag the possibility that innateness might be elucidated by the literature on causation, and perhaps also the relation between 'innate' and 'genetic'.

Lastly, a final advantage of analyzing genetic traits via an explicit theory of causation is that this also yields us an analysis of the relation between such traits and genetic *dispositions*. I turn to that final advantage now.

4.6 Traits Versus Dispositions

> *The conditions of adult membership to the Philharmonic Academy in Bologna required a candidate to write an elaborate motet in six parts, founded upon a melody assigned from the Roman Antiphonarium, the work to conform to the strictest rules, with double counterpoint and fugue. In the summer of 1770, the Academy was visited by a 14-year-old boy who tried the test. In less than three-quarters of an hour he rapped at his door and asked to be let out. The authorities sent him word not to be discouraged, but to keep on trying, as he had yet three hours, and might accomplish it. They were greatly astonished on finding that he had already finished, having produced a complete master work, abundantly up to all requirements, the whole written in a peculiarly neat and accurate manner. The 14-year-old boy was the young Mozart.*[19]

Was Mozart's dazzling skill a genetic trait? It is commonly supposed so. The thought behind this judgment is easily captured formally: set T_a = Mozart's feat of writing the motet in 45 minutes, compared to T* = needing longer as is typical for the rest of us, and let G* = some typical non-Mozartian human genome. Then it seems clear that T(G*) = T*, in other words that Mozart's feat is explained genetically. But there's a catch. Like any trait, T_a was the product of both genetic and environmental inputs. And this particular T_a required not just normal environmental inputs of nutrients, physical nurture and so on but also something much less commonplace, namely that almost from infancy Mozart was hot-housed as a

[19] Adapted from Mathews (1891, 295–296).

prodigy by his musician father.[20] So for many salient E*, this T_a is explained environmentally – without the hot-housing, Mozart could not have pulled off his dazzling feat in Bologna. Intuitively, and according to [GT], that renders T_a not genetic after all. This variability of verdict with explanatory context is not unique, indeed we have seen that it is true of all traits. What is unusual about the Mozart case is the salience of environmental explanation even though Mozart is often perceived in nativist terms as a unique 'genius'.

A telling detail is that Mozart is more usually described as having a God-given *talent* than God-given *traits*. Reading 'God-given' as 'genetic', the solution, I shall argue, lies in the notion of genetic *dispositions*. In particular, talents must be explicated as dispositions rather than traits, for like dispositions they are only potentials and so might never be realized. (To be clear on terminology here: I shall describe the trait of the 14-year-old Mozart's musical skill in Rome by his 'feat', and his initial disposition to be able eventually to reach that level of skill by 'talent'.) Whereas my having two legs is normally described as a genetic trait rather than a disposition, in the case of Mozart's musical ability it is the other way round. Why? And what, if anything, of interest does the distinction mark?

Begin with the analysis of dispositions generally. Paradigmatically, a given vase is fragile *iff* it shatters when struck. What is the biological analogue? Label the relevant disposition here, corresponding to fragility, D = Mozart's musical talent. Corresponding to the shattering, i.e. to the realization of the disposition, we have the trait T_a = Mozart's feat of writing the motet in 45 minutes. Corresponding to the striking of the vase, i.e. to the relevant environment, we have E_a = Mozart's actual childhood hot-housing. Then, filling in the definition of a disposition, Mozart has the disposition D *iff* T_a when E_a.[21]

So far, that just expresses that D is a disposition. What does it mean to say that it is a *genetic* one? I propose to define the latter to be a dispositional property of a genome at conception.[22] Thus genetic dispositions, unlike most genetic traits, are things we do possess literally from conception. Of course, because genes feature in the causal history of any trait, it now follows that *all* traits – genetic and non-genetic alike – are realizations of some genetic disposition or other.[23] So on the face of it, appeal to such dispositions seems rather vacuous. Why then should such appeals

[20]Later in life, Mozart himself commented on the necessity of an intensive environmental input: "People make a mistake who think that my art has come easily to me. Nobody has devoted so much time and thought to composition as I. There is not a famous master whose music I have not studied over and over."

[21]I stay neutral here on the vexed metaphysical issue of dispositional realism. Thus I take no position on whether a trait is also *explained* by an underlying disposition in addition merely to realizing it.

[22]Environments too have dispositions, of course. For instance, a particular kind of schooling may tend to produce particular kinds of graduates. Such dispositions do not seem relevant here though.

[23]Therefore disposition is the broader category here: all genetic traits are realizations of genetic dispositions, but genetic dispositions – when realized at all – may be realized by traits that are either genetic or non-genetic.

ever be made? Because, it turns out, in particular circumstances they *can* be informative after all. Tracing exactly how will prove to be intricate work.

Begin on the genetic side. Continuing with the Mozart case for illustration, recall that D is a property of Mozart's actual genome G_a, that T_a = his musical skill, and E_a = his hot-housing upbringing. What of the salient contrast genome G*? There are two possibilities: either G* has D too, or it does not. Only in the latter case, it turns out, will invocation of a genetic disposition be useful and hence conversationally apt. Assume to start with that it is indeed the latter case. Because therefore G* does not have D whereas G does, it follows from the definition of a disposition that:

1) G_a implies T_a when E_a; and
2) G* implies not-T_a when E_a.

Labeling as usual by T* the salient contrast to T_a, and assuming the actual environment E_a[24] in both cases, these just amount to:

1) $T(G_a) = T_a$; and
2) $T(G^*) = T^*$.

But we already know that $T(G_a) = T_a$, since that is just the actual case. So the pragmatic import of the claim that D is genetic boils down to claim 2 regarding a counterfactual. And claim 2 is of course just one half of our definition of a genetic *trait*. In other words, *given that the salient contrast for the explanans is genetic*, asserting that a trait T_a is genetic and asserting that the underlying disposition D that it realizes is genetic, *amount to exactly the same claim* – namely, that $T(G^*) = T^*$.

All this, recall, was on the assumption that G* does not possess D. Now suppose by contrast that G* *does* possess D. It would follow that G* implies T_a when E_a, and hence that $T(G^*) \neq T^*$. In other words, with respect to the explanandum T_a-rather-than-T*, having G_a rather than G* no longer makes a difference and thus appealing to genes is no longer explanatory. In such circumstances, the assertion that D is genetic no longer serves any pragmatic purpose, as D is a property of G and G* alike. For that reason, such assertions are only made in the earlier case, i.e. when $T(G^*) = T^*$.

Turn next to the second half of the story, so to speak, to when the salient contrast class for the explanans is environmental, i.e. some E*. The key point is that a genome either will or will not have a given disposition – *regardless* of environment. Intuitively, for instance, Mozart would still have been conceived with his musical talent regardless of whether his upbringing subsequently enabled him actually to fulfill it. Thus G_a will still have disposition D given either E_a or E*. Recall, the definition of D is: T_a when E_a. To say that D holds given E* is therefore merely to assert the conditional 'T_a when E_a' when that conditional's antecedent is false, i.e. when E*. Accordingly, this assertion will be (vacuously) true for *any* E*. It follows that possession of D in itself implies nothing about what would have happened given E*. In particular, it is therefore left open whether or not $T(E^*) = T^*$.

[24] Modulo any changes in environment that are causally downstream of the substitution of G* for G.

To appreciate the significance of that, recall in turn that when an E* is salient, [GT] tells us that asserting a trait to be genetic amounts to the claim $T(E^*) \neq T^*$. But that latter claim is not true for Mozart, for instance. As we saw, given that Mozart's musical feats depended on his childhood hot-housing, for a typical E* we find on the contrary that $T(E^*) = T^*$. To capture what is genetic about Mozart's talent we must contrast only alternative genomes, *not* alternative environments. Formally, his case satisfies only one half of our definition of a genetic trait, namely the first half. But we see now that it satisfies *all* of the definition of a genetic *disposition*, because the latter is conveniently silent about T(E*). And that is why, when expressing the genetic aspect of Mozart's talent, we appeal to disposition rather than trait.

The pay-off from this, finally, is that it now shows us exactly when and why appealing to a genetic disposition is useful. In particular, we assert a disposition rather than a trait to be genetic *iff* the following pragmatic (1 and 4) and metaphysical (2 and 3) conditions are satisfied:

1) The explanatory context does not make it clear that G* rather than E* is the salient contrast
2) G* does not possess D, i.e. $T(G^*) = T^*$
3) $T(E^*) = T^*$
4) We wish only to assert $T(G^*) = T^*$ [GD]

Here is the reasoning. First, often we shall want to assert $T(G^*) = T^*$ (assuming it is true) as doing so may have scientific value, i.e. often conditions 2 and 4 hold. And often we achieve this simply by asserting the *trait* T_a to be genetic since, when a G* is salient, that is just to assert that $T(G^*) = T^*$. But given conditions 3 and 1, this usual strategy fails, since asserting a trait to be genetic is also to assert that $T(E^*) \neq T^*$ in E* cases, and this latter implication now becomes both false and potentially salient. In such circumstances we may still assert $T(G^*) = T^*$, but now only by asserting instead that the underlying *disposition* is genetic. Given condition 2, asserting the disposition to be genetic implies that $T(G^*) = T^*$. In G* cases, this merely replicates the implication of asserting the trait to be genetic. But given also conditions 3 and 1, the situation changes, as now the silence of genetic dispositions regarding T(E*) – which distinguishes them from genetic traits – becomes decisively useful.

Thus the metaphysical implications of an appeal to genetic dispositions are derivable only indirectly, via pragmatic considerations. Intuitively, *we appeal to dispositions when a trait is explained (in that context) by both genes and environment but we want to focus attention just on the genes side*.

The best evidence for this account is again examination of how actual usage tracks explanatory context in just the way predicted. With Mozart, often $T(G^*) = T^*$ and $T(E^*) = T^*$, i.e. conditions 2 and 3 hold. In many contexts, such as when discussing his hot-housing, it will be unclear that G* rather than E* is salient, i.e. condition 1 holds too. And Mozart's uniqueness, finally, is captured by $T(G^*) = T^*$, where G* = other human genomes. If we wish to assert that uniqueness then we are satisfying condition 4, and hence now all four conditions. Thus it is that we naturally appeal to Mozart's 'innate genius' or 'God-given talent', i.e. appeal now to a genetic disposition rather than a trait.

Here is another, more biological example. We would typically say that my having two feet is a genetic trait. But if it is pointed out that thalidomide in the womb during early pregnancy may lead to no feet developing, we revise our claim to saying only that I have a natural *tendency* to develop two feet, i.e. now appealing to a genetic disposition. Why this switch? Because, on this paper's view, in the typical conversational context we do not have in mind those unusual environments in which two feet will not develop. Rather, we have in mind some E* such that T(E*) ≠ T* (for T* = I do not have two feet). But in a thalidomide context, now E* = thalidomide in the womb, and so T(E*) = T*, i.e. the environment becomes explanatory. In both cases, genes are explanatory, i.e. T(G*) = T* (say, for G* = a non-human genome). Formally, the thalidomide context, but not the typical context, fulfills [GD]'s conditions 2 and 3. Mere mention of thalidomide presumably may suggest E* rather than G* to be salient, thus fulfilling condition 1. Therefore if condition 4 also holds, i.e. if we wish to assert that my having a human genome explains my two feet, we appeal to the tendency/disposition. Thus our account successfully explains the original datum, namely that when switching from the typical to the thalidomide conversational context we switch from trait to disposition.[25]

One more example, briefly: Tiger Woods hits the ball further than most other professional golfers. Announcers often mention his 'natural length', i.e. a genetic trait. But when comparing the constantly training adult Woods to his ten-year-old childhood self, thus switching to a context in which an environmental input (i.e. his training) is now explanatory too, the comment becomes instead how he has fulfilled his 'natural talent', i.e. a genetic disposition.

In distinguishing between genetic traits and dispositions, this final section has staked out virgin territory. Yet how else to explain why when regarding Mozart we invoke genetic dispositions, but when regarding my legs we usually invoke genetic traits? Our account explains why appeal to dispositions is likely when T(E*) = T* and when G* does not possess D. Prime generators of such contexts are human traits that are non-universal and environmentally sensitive. And such traits often turn out to be precisely the subjects of famous disputes. Besides musical ability, examples include homosexuality, alcoholism, schizophrenia, high scores on IQ tests, athletic ability, and many cancers. A good analysis of genetic dispositions is therefore relevant to precisely those hot-button controversies that are one of the main things we want an account of genetic traits *for*.[26]

---

[25] We are also licensed to say that, in contrast to some hypothetical thalidomide-resistant creature, humans have the unfortunate genetic disposition to develop *no* feet in a thalidomide-bathed prenatal environment.

[26] For better or worse, moral charge is attached to some environmental explanations but rarely to genetic ones. This paper's analysis explains why there is therefore motivation (for some) to describe homosexuality or alcoholism as genetic dispositions rather than genetic traits even in advance of the full scientific story – because we are thereby leaving open the possibility of an environmental explanation.

Acknowledgments I would like to thank an anonymous referee and especially Katie Plaisance for useful comments. I would also like to thank Katie, and Thomas Reydon, for organizing the stimulating conference at which this paper was presented.

References

Achinstein, P. (1983): *The Nature of Explanation*. Oxford: Oxford University Press.
Davidson, D. ([1967] 1980): 'Causal relations'. In Davidson, D.: *Essays on Actions and Events*, pp. 149–162.
Dretske, F. (1972): 'Contrastive Statements'. *Philosophical Review* 81: 411–437.
Gannett, L. (1999): 'What's in a Cause?': the pragmatic dimensions of genetic explanations', *Biology and Philosophy* 14: 349–374.
Garfinkel, A. (1981): *Forms of Explanation*. New Haven (CT): Yale University Press.
Gifford, F. (1990): 'Genetic Traits', *Biology and Philosophy* 5: 327–347.
Hitchcock, C. (1996): 'The Role of Contrast in Causal and Explanatory Claims', *Synthese* 107: 395–419.
Hitchcock, C., and J. Knobe (2009): 'Cause and norm', *Journal of Philosophy* 106.
Hull, D. (1981): 'Units of Evolution: A Metaphysical Essay'. In U.L. Jensen and R. Harré (eds.): *The Philosophy of Evolution*. Brighton: Harvester Press.
Lewontin, R. (1974): 'Analysis of variance and analysis of causes', *American Journal of Human Genetics*: 400–411.
Maslen, C. (2004): 'Causes, Contrasts and the Nontransitivity of Causation'. In J. Collins, N. Hall, and L.A. Paul (eds.): *Causation and Counterfactuals*. Cambridge (MA): MIT Press, pp. 341-357.
Mathews, W.S.B. (1891): *A popular history of the art of music, from the earliest times until the present*. (Text available on Google books.)
Northcott, R. (2006): 'Causal efficacy and the analysis of variance', *Biology and Philosophy* 21: 253–276.
Northcott, R. (2008a): 'Can ANOVA measure causal strength?', *Quarterly Review of Biology* 83: 47–55.
Northcott, R. (2008b): 'Causation and contrast classes', *Philosophical Studies* 39: 111–123.
Oyama, S., P. Griffiths, and R.D. Gray (2001): *Cycles of Contingency. Developmental Systems and Evolution*. Cambridge (MA): MIT Press.
Pearl, J. (2000): *Causality*. New York: Cambridge University Press.
Schaffer, J., (2005): 'Contrastive Causation', *Philosophical Review* 114: 297–328.
Sesardic, N. (2005): *Making Sense of Heritability*. Cambridge: Cambridge University Press.
Shipley, B. (2000): *Cause and Correlation in Biology*. Cambridge: Cambridge University Press.
Smith, K. (2007): 'Towards an Adequate Account of Genetic Disease'. In: Kincaid & McKitrick (eds.): *Establishing Medical Reality: Essays in the Metaphysics and Epistemology of Medicine*. Dordrecht: Springer, pp. 83–110.
Sober, E. (1988): 'Apportioning causal responsibility', *Journal of Philosophy* 85: 303–18.
Sober, E. (1998): 'Innate Knowledge'. In: E. Craig and L. Floridi (eds.): *Routledge Encyclopedia of Philosophy*. London & New York: Routledge, (pp. 794–797.
Spirtes, P., C. Glymour, and R. Scheines (2000): *Causation, Prediction, and Search* (2nd ed.), Cambridge (MA): MIT Press.
Sterelny, K., and P. Kitcher (1988): 'The Return of the Gene', *Journal of Philosophy* 85: 339–361.
Sterelny, K. (2003): *Thought in a Hostile World*. Oxford: Blackwell.
van Fraassen, B. (1980): *The Scientific Image*. Oxford: Oxford University Press.
Waters, C. K. (1994): 'Genes Made Molecular', *Philosophy of Science* 61: 163–85.
Waters, C. K. (2007): 'Causes that make a difference', *Journal of Philosophy* 104: 551–579.
Woodward, J. (2003): *Making Things Happen*. Oxford: Oxford University Press.

Part III
Developmental Explanations of Behavior

Chapter 5
From Cell-Surface Receptors to Higher Learning: A Whole World of Experience

Karola Stotz and Colin Allen

5.1 Introduction

Animal behavior has been the contested subject of study between two quite distinct disciplines, ethology, with its roots in evolutionary biology, and comparative psychology, rooted in the methods and goals of psychology. The main differences between the two fields were their methodologies, field studies versus laboratory experiments, and their subject of study, innate versus acquired behavior. For several decades in the middle of the last century the ethological tradition in Europe and the psychological tradition in America fought an intellectual war in whose center stood the ethologists' 'instinct' concept (Griffiths 2004). The ethologists, especially Konrad Lorenz, believed that by treating instinctive and learned components of behaviors independently they could uncover evolutionary relationships. But Daniel Lehrman (1953) strongly criticized the ethologists for being too quick to label behaviors as 'instinctive'. He accused them of using the term ambiguously and argued that they ignored the role of learning and development in many of the behaviors they considered instinctive. Niko Tinbergen's acceptance of some of the major criticisms of this concept and then Robert Hinde's remaking of ethology in the sixties, particularly through Hinde's effort in attempting to synthesize psychology

K. Stotz (✉)
Department of Philosophy, University of Sydney, Sydney, NSW 2006, Australia
e-mail: karola.stotz@sydney.edu.au

C. Allen
History and Philosophy of Science, and Cognitive Science Program, Indiana University, Bloomington, IN 47405, USA
e-mail: colallen@indiana.edu

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*, Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_5,

and ethology (Hinde 1966), officially put an end to the war.[1] Nevertheless, both traditions remained largely separated, with comparative psychology pursuing its own agenda on the one hand, while ethology splintered into the fields of neuroethology, behavioral ecology, sociobiology, evolutionary psychology, and cognitive ethology on the other. The disciplines that arose from ethology retained their interest in the species-typical behavior of animals in their natural habitat and the evolutionary contexts in which different behavioral phenotypes are selected, while the comparative psychologists continued to focus on general mechanisms of learning in tightly controlled conditions.

In the last decade it has become en vogue for cognitive comparative psychologists to study animal behavior in what they claim is a more 'integrated' fashion: While the acquisition of knowledge under controlled laboratory conditions remains the primary target of scientifically rigorous investigation in these new studies, the natural habitat and evolutionary endowment of the organism are taken into account, and explanations are sought at both the behavioral and the cognitive level. Nevertheless, we will argue that these studies, instead of really integrating the concepts of 'nature' and 'nurture', rather cement this old dichotomy. They combine empty nativist interpretations of behavior systems with blatantly environmentalist explanations of learning, based on the assumption that the innate must be there before learning begins. (We will describe below how the field of developmental psychobiology with proponents like Daniel Lehrman and Gilbert Gottlieb may provide a notable exception to the general lack of integration among different approaches.)

We argue that the main culprit blocking full integration remains the failure to really take development seriously if it is taken into account at all, thus echoing Lehrman's original critique of ethology over half a century ago (Lehrman 1953). In some areas of biology, particularly evolutionary developmental biology, interest in the relationship between behavior and development has surged through topics such as parental effects, extragenetic inheritance, and phenotypic plasticity. However, this has gone almost completely unnoticed in the study of animal behavior in comparative psychology, and is frequently ignored in ethology too. Reasons for this may include the traditional focus on the functions of behavior in its *species-specific* form in *adult* animals, a preformationist or deterministic conception of development, and generally the separation of psychology from biology. In psychology, development is often understood as a process of the maturational unfolding of the young to the adult that is distinct from learning, rather than treating both learning and development in a more integrated fashion as part of an overall,

[1]The Dutch zoologist Nikolaas Tinbergen was one of the founders of the study of comparative ethology, for which he shared the Nobel Prize for Physiology or Medicine in 1973 with the Austrian Konrad Lorenz, with whom he shared a strong working relationship, and the German Karl von Frisch, who worked independently on the dance language of honey bees. Robert Aubrey Hinde is a British ornithologist, ethologist and psychologist whose doctoral studies at Oxford University coincided with Tinbergen's arrival there after the Second World War.

life-long process by which an organism integrates environmental information. Quite interestingly, the shortcomings of treating learning and development separately have been seen quite clearly by two behavior analysts working in the tradition of comparative psychology and behaviorism:

> That behavior analysis does not always take development seriously is exemplified in its distinction between phylogenic and ontogenic contingencies. (…) Behavior analysis thereby includes both (a) species-typical behavioral development via natural selection and (b) individual behavioral development via contingencies of reinforcement (…) Where behavior analysis has included biological ontogenesis, however, it is treated as a relatively "automatic" process governed by the genes. (…) we have put ourselves in the position of maintaining - not rejecting - the nature-nurture dichotomy. (…) As a result, behavior analysis overlooked an opportunity to extend what we take to be its inherently developmental perspective. (…) What remains is for behavioral analysis to recognize that all biological, behavioral, and bio-behavioral phenomena are developmental processes. (Midgley and Morris 1992:235-7)

Studies by developmental psychobiologists (Lehrman 1970; West, King, and Arberg 1988; Michel and Moore 1995; Gottlieb 1997, 2001) and some social neuroscientists (Cacioppo and Berntson 2004) to be described further below, show that the generalizations of the previous paragraph are not universally true. While some focus on naturally occurring individual differences that may or may not be transgenerationally transmitted, others study the necessity of individual experience in explaining a species-typical outcome. Developmental psychobiology goes back a long way in its attempt to employ both biological and psychological concepts in explaining development and its relation to evolution. Its proponents have always fought against the reification of scientifically unhelpful dichotomies such as 'nature versus nurture' and 'innate versus acquired', and presented strong criticisms of behavioral genetics and evolutionary psychology. Social neuroscience is a relatively new research field that examines the role of the central nervous system in the development and maintenance of social behaviors. For example, Meaney and colleagues have studied how individual differences in maternal care in rats can alter an offspring's neural development, as well as its ability to cope with stress later in life. The team elucidates the molecular mechanisms that modify the expression of genes regulating hippocampal synaptic development as well as behavioral and neuroendocrine responses to stress (Meaney 2001a). Both fields have gone to some length in showing how traits can be inherited and still be acquired through nongenetic inheritance mechanisms (Szyf, McGowan, and Meaney 2008; West, King, and Arberg 1988).

Aside from these promising research areas, the failure to pay sufficient attention to developmental questions in the origin of behavior is widespread. Here we provide a few prominent examples of this failure from comparative psychology, particularly accounts of the relationship between humans and their closest relative.

In a recent paper Penn, Holyoak, and Povinelli argue against what they see as a "trend among comparative researchers … to construe the uniquely human aspect of these faculties in increasingly narrow terms" (Penn, Holyoak, and Povinelli 2008, 110). Daniel Povinelli is quite well known for his controversial denial of advanced cognition in chimpanzees, based upon his experiments with a group of seven chimpanzees that he raised from 4 years of age in circumstances that are developmentally

unusual, namely a peer-group laboratory setting without any interaction with adult chimps (Povinelli 2000). In their 2008 article, Penn, Holyoak, and Povinelli urge that discontinuities between humans and apes are more significant than is typically admitted by primatologists whom they accuse of especially ignoring the higher-order, systematic, relational capabilities of human users of physical symbol systems. Their "relational reinterpretation hypothesis (RR)" states that "only human animals possess the representational processes necessary for systematically reinterpreting first-order perceptual relations in terms of higher-order, role-governed relational structures akin to those found in a physical symbol system (PSS)" (Penn, Holyoak and Povinelli 2008, 111).

> Indeed, we will argue that the functional discontinuity between human and nonhuman minds… runs much deeper than even the spectacular scaffolding provided by language or culture alone can explain. (…) only humans appear capable of reinterpreting the higher-order relation between … perceptual relations in a structurally systematic and inferentially productive fashion. (…) our ability to do so relies on a unique representational system that has been grafted onto the cognitive architecture we inherited from our nonhuman ancestors. (Penn, Holyoak and Povinelli 2008, 110-111)

However, just as with most of traditional cognitive science, they confound cultural symbolic achievements with individual cognitive competencies. Their argument for a large discontinuity between human and non-human primates rests on a hybrid symbolic-connectionist model of cognition which does not provide any explicit role for learning, and only a diminished role for development, cultural context and direct tuition, as is commonly the case with models couched in "the currency of symbol manipulation" (McGonigle and Chalmers 2002). There exist two quite different stances towards the evolution of human cognitive capacities. The nativist stance attributes the origin of human social capacities such as folk psychology or theory of mind to the sudden appearance of genetically determined mental modules or representational systems and is favored for instance by evolutionary psychologists. The embodied and extended cognition approach points out the importance of cultural scaffolding through social, cognitive and developmental niche construction and presupposes only very simple and modest biological preadaptations, e.g. in the perceptual realm (Tomasello 1999; Donald 2000; Sterelny 2003; Wheeler and Clark 2008; Stotz 2010). The former approach, which includes Penn et al's hypothesis, is polemically dubbed the 'Rational Bubble' stance and belongs to a class of models that have in recent years come under increasing criticism from those taking an embodied stance as a quite unrealistic model of cognitive growth (McGonigle and Chalmers 2008, 143).

The difficulty of being fully sensitized to the developmental dimension of cognition is highlighted by another example. Tomasello and collaborators have proposed the 'Cultural Intelligence Hypothesis' (CIH) about the particular role that 'ultra-social' learning through cultural participation, instruction, and formal schooling played in the development and evolution of human cognition (Herrmann et al. 2007; see also Tomasello 1999). In essence Tomasello argues that the particular cognitive skills that set us humans apart from other primates "result from a variety of historical and ontogenetic processes that are set into motion by the one uniquely

human, biologically inherited, cognitive capacity", namely "understanding others as intentional (or mental) agents (like the self)" (Tomasello 2000: 15). The CIH gains support from an empirical study that compared the capacities of children with apes. But although the CIH is in large part a developmental hypothesis, the investigators neglected to address this developmental dimension in their experimental design. The apes used in this study, though compared with human children of all the same age of 2 ½ years old, were of a wide range of *adult* ages. Further, no information is provided on the rearing conditions and former experiences with similar tasks of the apes. This is quite surprising from a laboratory that has also put forward the 'Enculturation Hypothesis', an epigenetic model of the effect of human rearing on the cognitive development of apes. Research reports supporting this hypothesis, and later reports that led the investigators to believe that even 'normal' apes are socially more competent than originally thought, stand in stark contrast to the findings that support the CIH (Call and Tomasello 1998; see Tomasello and Call 2004 for further references).

These examples raise four worries: 1. Many skills that are tested in stand-alone experiments have developmental dimensions that most test designs miss or deliberately ignore. 2. There may be a range of tasks that younger ages in both apes and humans generally perform better than adults. 3. One needs the comparative context of the test results in order to interpret them properly, so for instance testing untrained and unenculturated apes against enculturated apes on the one hand and humans on the other, and sampling all three groups at different but developmentally comparable ages, would be necessary. It is important, however, to recognize that different aspects of development may proceed at different rates in different species, thus it is not ever possible to perfectly match developmental ages between two species (e.g. Gácsi et al. 2009). An experiment testing three pairs of mother and offspring chimpanzees against university students in a memory task provides a case in point (Inouea and Matsuzawa 2007). The young chimps far outcompeted both ape and human adults. This result suggests a developmental component within the tested memory faculty rather than a species difference. In order to test this, a superior experimental design would have included human children in the study. 4. It is not clear to what extent the ape and human subjects have been treated similarly. Presumably, both humans and adults have been tested by human experimenters, but apes are likely much better at reasoning about the mental states of conspecifics.

Much of comparative psychology takes place against an assumed background of animal learning theory, which treats associative learning mechanisms as strongly conserved across vertebrate species. On this conception, the capacity for learning is merely a product of development, with learning processes regarded as the successor of, rather than part and parcel of the continuous developmental process. Our main aim in this paper is to forge a closer relationship between the concepts of learning and development, and to investigate whether and how the two concepts together can be usefully deployed in the study of (human and non-human) animal behavior. This will first require a *biologically informed* comparative psychology, and second, we suggest the formulation of a broadened concept of 'experience' which may help to bridge between learning and development by including all aspects of environmental stimuli that lead to long-term, adaptive changes of behavior, including 'learning' in

its usual narrower sense. The introduction of this term implicitly questions why 'instinct' and 'learning' should be the only two choices available to us for understanding behavioral development.

In other words, our use of the concept of experience is not limited to sensory processing but includes a quite heterogeneous mix of environmental resources influencing the system's behavior. While this concept is not new, it unfortunately is very rarely used in scientific investigations, other than perhaps in its fields of origin (early ethology and developmental psychobiology) (compare also behavioral geneticists Turkheimer and Gottesman 1991, 21, who propose its use instead of 'environment'). We will provide some further clarification of the concept below.

Our understanding follows Schneirla's[2] original definition of experience, emphasized by his student Daniel Lehrman: Experience is "the contribution to development of the effects of stimulation from all available sources (external and internal), including their functional trace effects surviving from earlier development" (Schneirla 1957, 1966, cited in Lehrman 1970, 30). Within this wide range of processes "learning is only a relatively small part" (Lehrman 1970, 30). To take this really on board one needs to acknowledge that physiological regulation and the regulation of behavior cannot be sharply separated, since their underlying mechanisms do not necessarily belong to distinctly different classes. This is especially so in early development. (Re-)Introducing the concept of experience is not another way of saying that all behavior is learned, but a vehicle to bring home the inadequacy of the distinction between innate and acquired.

At this point it may be important to address the concept of 'behavior', which is notoriously hard to define. A recent study by Levitis et al. (2009) addressed this shortcoming in the field of behavior studies. From survey responses the authors tried to deduce what characteristics a scientific conception of behavior should have:

> Behavior is: The internally coordinated responses (actions or inactions) of whole living organisms (individuals or groups) to internal and/or external stimuli, excluding responses more easily understood as developmental changes.

They go on to say that:

> Developmental processes are largely excluded from the definition, as they are generally much slower than phenomena considered as behaviour, and *are primarily based on ontogenetic programmes specified by the individual's genetic makeup*. (Livitis et al. 2009, 108; emphasis added)

This sentiment is of course exactly what we are criticizing here. Taking development seriously may not change the underlying definition of behavior, but it should change the way one may distinguish between physiological changes, learning and development. All three are not behavior per se but are mechanisms that predispose toward behavior.

[2]Theodore Christian Schneirla was an American animal psychologist from the '30s and '40s who greatly influenced Daniel Lehrman and other developmental psychobiologists.

In section 2 we will identify and criticize two received views of development, predeterminism and the so-called Modern Consensus. In their place we propose an 'epigenetic systems view of development' encompassing the organism in its developmental niche, which takes seriously the idea that all traits, even those conceived of as 'innate', have to develop out of a single-cell state through the interaction between genetic and non-genetic (experiential) resources of development. The message of this section will not only be that one should dispense with old dichotomies when attempting to explain the development of a phenotypic trait, physiological and behavioral. We go further to claim that the different dichotomies, such as innate-acquired or nature-nurture, are not only inappropriate labels in themselves, mere placeholders for a real causal analysis of development; they also do not, as is commonly held, map neatly onto each other: genes do not equal nature, nor does environment stand for nurture. As a matter of fact, no developmental factor corresponds to either nature or nurture. Instead we want to promote an understanding of nature that shifts attention from allegedly fixed genetic causes to the range of natural phenotypic outcomes, and a conception of nurture as the developmental processes leading to those outcomes.

Section 3 will look at several conceptions of learning and cognition in psychology and how they are employed in the study of a wide range of organisms. We place an emphasis on simple systems approaches, such as invertebrates, the spinal cord, single cell organisms, and even eukaryotic cells in a multicellular organism, in which the boundary between learning and other kinds of experience becomes fluid. This is an important step toward reconciling accounts of learning with our conception of epigenetic development that necessarily includes some form of experience in the construction of any physiological or behavioral trait.

Section 4 will attempt a synthesis of the concepts of development, experience and learning. We lay out how, in a systems view of development, learning may appear as just one among many processes in which the experience of an environmental input generates an appropriate response and hence influences the behavioral phenotype. We consider whether development and learning are two fundamentally different kinds of processes that happen to have a similar temporal relationship to experience, or whether these terms should be more tightly assimilated to one another, and we argue for thinking not in terms of learning *and* development, but in terms of learning *as* development. This is followed by a discussion of the concept of ontogenetic niche and the kinds of experience it affords. We also discuss how such a new synthesis should help to overcome the age-old dualism between innate and acquired and thereby open up the possibility of developing scientifically more fruitful distinctions. Finally, section 5 summarizes the argument of the paper and draws some conclusions for philosophy.

5.2 Taking Development Seriously

Scientific understanding of the nature and history of living things, including their cognitive capacities and behavioral phenotype, depends crucially on having a proper understanding of the most basic of biological processes that brought them

about: development. Since ancient times this process has captured the imagination of scholars but has eluded a satisfactory explanation or consistent framework until today. The main problem in the interpretation of development has from the beginning been the question of whether organisms merely unfold or mature out of something already formed from the beginning, or whether they emerge as something qualitatively novel from an undifferentiated and unformed state. Despite being declared dead many times, this debate is alive and well today in the dichotomy of nature, or genetic determinism, and nurture, or environmental plasticity. In the context of this paper, our criticism of the failure to take development seriously is mainly directed to (cognitive) ethologists, comparative and developmental, particularly nativist, psychologists, behavior analysts and other researchers in the field of behavioral sciences. In our view, taking development seriously doesn't amount to the mere recognition of the existence of environmental influences, but to the questioning of the traditional dichotomization of developmental causes into inherited, innate or genetic as so-called "nature", on the one hand, and acquired, experienced or learned as so-called "nurture", on the other hand. It is not only the discovery of inherited epigenetic variation that is now blurring this traditional line, but also new studies of cognitive and psychological development from other fields. Nevertheless, Mark Blumberg cautions that,

> while a developmental perspective is vital to any satisfying explanation of the complexities of behavior, it would be misleading to suggest that individuals who study development are more likely to eschew facile appeals to instinct and nativism. (…) wearing the developmental badge provides little protection against nativism. (Blumberg 2005, 13)

5.2.1 Preformationism, Epigenesis, and the Modern Consensus

Preformationism, one of the ancient conceptions of development that goes back to Hippocrates, held that the organism is formed from the beginning, with the developmental process bringing about no qualitative change but merely unfolding. Some preformationists considered gametes as minuscule organisms, tiny homunculi they actually claimed to make out under the microscope, needing just to grow or unfold themselves. In the 19th century, preformationism was recast as predeterminism, the idea that development consists of an orderly progression of qualitative change to a predetermined endpoint. According to both preformationists and predeterministes, environmental factors are understood as a mere background of supportive and permissive factors. In the modern incarnation of this view, behavior appears as an 'epiphenomenon of neural maturation' (Gottlieb 2001).

The main rival to the preformationist and predeterminist conceptions was the idea of epigenesis that dates back to Aristotle and maintained that development is a contingent process of differentiation out of a homogeneous and undifferentiated state with no predetermined endpoint. Without an easy preformationist interpretation for the seemingly orderly progression of developmental events, however, epigenesists needed to appeal to either internal or external teleological or vital forces or a formative drive, like Aristotle's male formal and female material cause. Due to the

development of mechanistic science in the 17th and 18th centuries, which rejected all but the efficient cause, and subsequently the mechanistic spirit of embryology in the late 19th century, vitalism, and with it epigenesis, has fallen into deep disregard (Robert 2004; Maienschein 2005). However, one should understand both positions as necessarily ill-informed attempts at a materialist and scientific explanation of development. The preformationists desired to understand organisms within the then accepted worldview, conceiving of organisms fully as the result of known physical forces given the 'auxiliary' assumption that cells within the organism existed in a preformed state and had existed in this state since the creation. Epigenesists conceived of development without this deistic assumption, and hence needed to propose an unknown (but not necessarily mystical) vitalist force acting on biological objects, which they saw as comparable to Newton's forces that applied to physical objects (Roe 1981). Arguably, the early 20th century with its transmission genetics vindicated the former position, while both the dawn of systems biology and the science of self-organization and complexity brought the new preformationism into disrepute and vindicated a reformulated epigenetic position that will be described in this section.

Evelyn Fox Keller argues that *The Century of the Gene* referred to in her book title (Keller 2000) brought about a new and more sophisticated preformationism that replaced 'preformation' with the notion of 'information' encoded in the genome – a substitution that she rejects. Nevertheless, informational notions continue to be predominate; true to the spirit of today's interactionism the mainstream modern consensus can be "standardly construed as the epigenesis of something preformed in the DNA" (Robert 2004, 34). Instead of avoiding the unscientific dangers of both preformation and vitalist epigenesis, however, this view rests ultimately on an unscientific conception of gene and gene action. In our view, the 'genetic program' with its evasion of the responsibility to give a causal-mechanistic explanation of the problem of development is tantamount to a materialized vital force. Hence, according to Jason Scott Robert, the new conception of genes that 'program' outcomes is in this sense equivalent to an 'animistic' predeterminism.

In its place we want to promote what others have called 'probabilistic, contingent, or constitutive epigenesis', a systems view that understands development as an epigenetic process of qualitative change based on the orderly emergence of novel behavioral traits during development without recourse to a preexisting plan. The contingent nature of development, due to the immense importance of experiential factors at all stages of development, from the regulation of gene expression to the learning of tool use or language, demands that we take it seriously (Gottlieb 2001; Michel and Moore 1995; Oyama, Griffiths, and Gray 2001a; Robert 2004).

5.2.2 Beyond Nature and Nurture

One of the foremost aims of a new conception of development is to challenge the widely held view that the physiological or behavioral phenotype derives *neither* exclusively from nature *nor* from nurture, but rather from *both* nature *and* nature,

understood in the traditional sense as genes or heredity and environment or conditions of existence. This is the mainstream view of 'interactionism' that has previously been described and criticized by other authors (e.g., Sterelny and Griffiths 1999; Oyama 2001). Neither the exclusive nor the additive models make any biological sense whatsoever, since no genetic factor can properly be studied independent of, or just in addition to, the environment (Meaney 2001b). The same is true for the environment, which in itself is a concept that includes a wide variety of very different causes and factors, from the genomic environment of a gene, over its chromatin packaging and cellular context, up to ecological, social and cultural influences on the whole organism. So-called innate traits may also be effects of epigenetic factors which are reliably reproduced with the help of 'ontogenetic niche construction' (see below). There is a longstanding debate between practitioners of developmental psychobiology and behavioral geneticists, but while we are referring to the former tradition as a paradigm example of how to take development seriously, our criticism here is not specifically addressed to behavioral geneticists but rather to comparative psychologists working in the behaviorist tradition, and to the more or less committed nativists found in various branches of cognitive science.

To resolve the nature-nurture debate a new view of development is needed to address several distinct but related sub-problems: 1) It needs to systematically question preconceptions of 'explanatory' categories of behavior, such as innate, acquired, genetically determined or programmed, which obscure the necessity of investigating developmental processes in order to gain insight into the actual mechanisms of behavior. In addition such preconceptions are prone to committing the 'phylogenetic fallacy', which conflates evolutionary and developmental explanations. 2) Such a new account needs to promote a new understanding of the nature of inheritance, which includes maternal effects on gene expression, epigenetic factors such as genetic imprinting, behavioral, cultural and symbolic inheritance systems, and ontogenetic niche construction. 3) A realistic view of gene action and activation is of pivotal importance to a theory of development since it helps to distinguish between explanations of the role of genes in development on the one hand and of the complete process of development on the other. 4) A new epigenetic understanding of development should ultimately resolve the dichotomy between preformationism and epigenesis, and between 'maturation' and 'learning'. These four aims are preconditions for the integration of the concepts of 'development' and 'learning' in biological and psychological research into behavior and cognition.

5.2.3 Explanatory Categories of Behavior

The main problem with allegedly explanatory categories of behavior such as *instinctive* or *learned* is that they pretend to offer a – albeit vacuous – explanation of the cause of the behavior in question and thereby effectively suspend further investigations into the real ontogenetic causes of the behavior. They do this by their very nature of purporting to explain while actually merely labeling the phenomenon. So in this sense one could

argue that it is a terminological problem that misguides research. After careful and often arduous empirical investigation, all apparently 'innate' or 'instinctive' behavior patterns, which *by definition* exclude experience or learning as their cause, have turned out to involve epigenetic or experiential factors (Blumberg 2005). Of course, everybody is an interactionist and admits the importance of environmental factors. But a developmental systems perspective attempts to move beyond "the continuing dispute of contributions of evolutionary processes to our nature and individual experiences to our nurture" (Blumberg 2005, 13). It maintains that all bio-behavioral phenomena – "innate" and "acquired" – are the products of a continuous developmental process. Instead of treating nature and nurture as distinctive causes of development, Oyama has suggested that nurture is the *process* (of development) and nature is the *product* (Oyama 1985, 125). However, the more conventional interactionist views continue to rely too uncritically on the notion of innateness. Griffiths has argued first on conceptual grounds and then by an empirical analysis of scientific practice that the vernacular concept of innateness can imply three different and unrelated things, namely (a) the developmental fixity (non-involvement of experience), (b) species-typicality, or (c) adaptedness of a trait (Griffiths 2002). All of them are standardly equated with the label 'genetic' for this innate behavior (Griffiths 2002; Linquist et al. 2011).

So, on the one hand we want to argue against the existence of any genetically *determined* traits, Even a so-called classic genetic disease such as PKU (phenylketonuria) only produces a disease phenotype in the *context* of a diet high in phenylalanine, but is without effect in diets without it. If context is ignored, all humans could be said to be genetically determined to have scurvy because we cannot produce vitamin C. However, during most of our evolutionary history humans lived in environments that provided enough food high in vitamin C, so we were free of disease before the advent of seafaring. Conrad Waddington introduced the term "canalization" to distinguish between traits that predictably develop against the backdrop of a wide range of *both genetic and environmental* variation (Waddington 1942). And even earlier, William Johannson when introducing the term 'genotype' acknowledged that all traits are the result of a certain "norm of reaction", which describes the response of a genotype to varying environments (Johannsen 1911).

On the other hand we maintain that a deeper investigation can show the relative independence of all three characteristics (a)-(c) identified by Griffiths as commonly implied by the term 'innate'. Evolutionary adaptations need not be developmentally fixed, independent of life experience. Neither must they be hard to change, but can instead be phenotypically plastic. This is the case with many highly environmentally sensitive polyphenisms, distinct phenotypes that are elicited by different environmental conditions (see below, section 4). Nor do adaptations need to be species-typical or universal. They can result from frequency-dependent selection, where the trait is only adaptive if a certain percentage of the population carries it. Species-typical or universal traits are not necessarily the result of natural selection but can be dictated by strong physical or developmental constraints that render them hard or even impossible to change. This has been shown by many examples uncovered by the new 'physicoevolutionary' approach, or by research into the generation and fixation of phenotypic organization (homologies) of organisms (Newman 2003; Gilbert 2003). The former

advances the hypothesis that "tissue forms emerged early and abruptly because they were physically inevitable ['self-organization'] – they were not acquired incrementally through cycles of random genetic change followed by selection (Newman 2003, 221). Last but not least, universality need not be and often is not due to the developmental fixity or experience-independence of a trait. It may be and often is due to the reliable availability of certain experiences to which the organism must be exposed to develop a trait. Here it is not just important to recognize the importance of experience. To do so is often hard enough because it may involve unexpected and therefore not easily visible environmental inputs – a point that has been made repeatedly by developmental psychobiologists. However, it is important also to recognize the processes that ensure the reliability of this exposure to the required environmental inputs. It is this process of *developmental niche construction* that allows us to speak of the inheritance of certain aspects of experience by the offspring as a result of the activities of their caregivers. Again, here we are not stressing just the importance of gene-environment interaction, a point easily conceded by many. We are arguing against the dichotomy made between genetic and environmental causes on the grounds that one is *stable and inherited* and the other *spurious and acquired*. Song learning in many bird species is a case in point, as the research by Meredith West and Andrew King has shown. In some species of birds, such as the Brown-headed Cowbird, all birds belonging to a population sing the same song (while in many others the songs of individuals may differ substantially, such as in the Australian Lyre bird, or the Indian Common Mynah). While instances of uniform songs were once taken as support for the genetic determination of the song acquisition (Marler and Slabbekoorn 2004), it is now known that cowbirds have to be exposed to other members of their species in order to acquire their population-specific song. The story in cowbirds, which are nest parasites and are therefore not raised by their own parents or even a member of their own species, is even more complicated and intriguing than with birds which acquire the song from their parents, or learn to recognize their own species-typical sounds from exposure to themselves and siblings within the nest (Gottlieb 1981). The details of how they acquire their song need not be described here – it suffices to say that cowbirds nevertheless always learn to sing the particular dialect of the population they belong to because of the reliability with which they meet, recognize and flock with members of their own species and are therefore exposed to the right stimulating experiences during development (West, King, and Duff 1990; Freeberg et al. 2002).

Transgenerational stability need not rely on the faithful transmission of DNA. Natural selection selects for adaptive traits or phenotypes; that is, it selects for outcomes and not for developmental mechanisms. Outcomes always derive from nonlinear interactions among a range of diverse developmental resources. Their organization frequently exhibits phenotypic plasticity, a capacity that allows the organism to react adaptively to different environmental conditions (Pigliucci 2001; West-Eberhard 2003; Gilbert and Epel 2009). The stable inheritance of this adaptive phenotype depends on the reliable transmission of all the necessary developmental factors across generations. In other words, phenotypic plasticity relies on a dependable yet flexible inheritance of a 'developmental niche' which is faithfully constructed and reconstructed by the species, the parent and the organism itself

(West and King 1987). The subject of selection is the whole developmental system, not just its genetic endowment, as is generally stated in both developmental, evolutionary and psychology textbooks. Lip service to development and the importance of gene-environment interaction is just not the same as acknowledging the existence of epigenetic, behavioral, ecological and cultural inheritance.

5.2.4 *Extragenetic Inheritance and Developmental Niche Construction*

The construction of the developmental niche relies heavily on the extragenetic inheritance of developmental resources. This heterogeneous process includes maternal and paternal effects, which cannot be reduced to just the influence of parental genes or RNAs on their offspring, but includes all processes of care for the offspring. These are comprised of imprinting systems, cellular structures, gut organisms, differential provisioning of resources, preference induction (oviposition, imprinting on food, habitat, and mates), and social learning, to name just a few (Jablonka and Lamb 2005; Mousseau and Fox 1998; Maestripieri and Mateo 2009). Ontogenetic niche construction is one way to conceptualize 'extended inheritance'. Inheritance systems have evolved to allow for the transmission of crucial information from parents to offspring. A principled definition of inheritance must include whatever is reliably present in each generation due to the parental generation and necessary to reconstruct the life cycle. One should not single out a particular type of resource as the source of intergenerational stability. A reliably reproduced developmental system is the result of the reliable provision of a wide range of developmental resources necessary to reconstruct the organism's life cycle, of which DNA is just one element. Organisms place DNA into a developmental setting that is always highly characteristic of a lineage and commonly owes much of its structure to the activity of previous generations. Evolution has come up with a wide range of strategies to construct the ontogenetic niche to dependably guide the developmental process. Developmental systems are often "designed to be as open as ecologically possible and thus immediately sensitive to ecological change" (West and King 2008, 393). A reliable developmental niche allows for a high amount of developmental and phenotypic plasticity without compromising stability and reliability (Lamm and Jablonka 2008). Such a system is the antithesis to a canalized or closed developmental system.

The concept of the ontogenetic niche and its active construction by the organism is closely related to the concept of 'niche construction' developed by Odling-Smee, Laland and Feldman (2003). However, while their niche construction refers to the active construction of a 'selective environment' which in turn influences the selective pressure on a population, the development niche talks about the construction of a 'developmental environment' which in turn influences the development of both parental and offspring generations, and through non-genetic inheritance, the evolution of the lineage. The idea that organisms are not passive, but actively construct their environments which then influences the organism in turn, goes back to Lewontin (1983).

What all the above cases of inheritance through environment construction have in common is that they make the transmission of crucial information more reliable. Parental activity can facilitate, guide and entrench social learning. Some of the aforementioned mechanisms have at first sight not much in common with the construction of cognitive or epistemic structures. However, in the latter cases of behavioral, ecological and cultural inheritance the biological shades smoothly into the cognitive. For example, the emergence of cognitive capacities for tracking objects that are out of sight depends on the development of motor systems regulating embodied actions such as reaching (Smith and Breazeal 2007 show how cognition emerges out of non-cognitive processes).

West and King were among the first to "Ask not what's inside the genes you inherited, but what your genes are inside of" (West and King 1987, 552). A look at the enormous complexity of gene expression of eukaryotes reveals a very flexible and reactive genome open to many intra-and extra-organismal environmental influences which makes it necessary for organisms to manage aspects of their own ontogenetic environment. It is not which genes you have that has phenotypic consequences, but how they are expressed by the higher order network of gene regulation that controls the time- and tissue dependent expression of genes. There have been repeated attempts to reduce epigenetic mechanisms to the action of inherited or parent-of-origin genes, so that ultimately the real causes are all genetic (see for instance Rosenberg 1997). This special pleading fails in light of the discovery that the regulated expression of genes ultimately depends on a host of environmental factors.

5.2.5 Environmental Regulation of Gene Expression

Genetic activity is involved in all biological processes, but so are non-genetic factors. Explanations listing only interacting genes are biased at best and relatively vacuous at worst. More informative explanations give an account of why and how certain genes are expressed at a particular place and time, an account that necessarily includes a range of very specific additional factors, including environmental signals which can have influences on the short- and long-term regulation of gene expression. Many accounts of the importance of gene-environment interaction represent how different genotypes impact on how the organism reacts to environmental perturbations (e.g., Caspi et al. 2010). These accounts, however, often neglect how different environments lead to a wide range of phenotypic responses by the same genotype — especially the responses particularly studied by the field of (environmental) epigenetics and epigenomics which focuses on so-called environmentally affected epimutations or metastable epialleles (Szyf, McGowan, and Meaney 2008; Dolinoy and Jirtle 2008).

Postgenomic biology has brought with it a new conception from the *active* gene to the *reactive* genome that is regulated by cellular processes that include signals from the internal and external environment. This regulation not only includes the regulated *activation* but particularly the regulated *selection* and sometimes even the *creation* of genetic coding sequences, processes one of us has termed "molecular

epigenesis" (Stotz 2006). This is not the place to report the details now available on the mind-numbing complexities of the expression of genes during development; instead a few central ideas should suffice. The last decade of genome-sequencing has revealed the paradox that the complexity of an organism is not related to its number of genes (Claverie 2001). Instead, organism complexity seems to be related to the complexity of the expression of a limited number of coding sequences. These sequences become more and more modularized in more complex organisms, with a higher number of coding sequences, called exons, that are separated by intervening introns. This increased modularity allows for the creation of a wide array of gene products out of a limited number of genes because the units can be differentially cut and pasted, or otherwise processed, through processes such as alternative splicing or RNA editing. (These processes are described in more detail in section 5.4.4.)

These mechanisms do not just control when genes are switched on and off, but also which parts of the DNA sequence will be transcribed, spliced and edited in complicated ways, and translated at specific rates. Often the particular mixture of gene products and their interacting cellular signaling factors are referred to as the *cellular splice code*. The cytoplasmic chemical gradients plus the maternal gene products inherited with the mother's egg give this process a head start. But the mother's control over her progenies' genes and their environment does not stop there. Chemical processes in the womb and after birth, including those induced by rearing practices such as the differential licking of pups by rat mothers, continue to influence (neurological) development through gene expression levels (Moore 1984; Meaney 2001). Chemical modification of the DNA and the surrounding protein packaging of the DNA, either inherited through the parents or added by environmental factors, influences the expression of genes throughout the offspring's life. This imprinting and epigenetic system is often called the histone- or *epigenetic code*, in the sense that the exact combination of chemical modifications – of which there are at least a handful – specifies the expression status of the underlying DNA sequence.

In the last two decades development has become equated with differential gene expression, but what is often forgotten in this definition is the complex network of other molecules (such as proteins and metabolites), cellular structures, 3-dimensional cellular assemblages and other higher-level structures that control or are otherwise involved not only in this differential expression of genes but in a wide range of other developmental processes decoupled from the direct influence of DNA sequences. Genes have an important role in development, but their role can only be properly understood within the larger system that exerts a controlling influence over them. Robert summarizes this attitude thus:

> To take development seriously is to take development as our primary explanandum, to resist the substitution of genetic metaphors for developmental mechanisms … The translation of embryology's hard problem (how a specific organism arises from a single, relatively homogenous cell) into a problem about gene action and activation generates explanations at the level of genes; but these explanations solve (or, rather, begin to solve) the subsidiary problem of the role of genes in development, not the problem of development as such. … There is indeed good reason to believe that genetics reduces to development, and not the other way around. (Robert 2004, 22)

5.2.6 A New Epigenesis

What a new account of development really has to accomplish is not just to go beyond these vexed dichotomies such as innate and learned, but to provide a framework that integrates a complex set of heterogeneous factors into a system of developmental resources all of which are reliably reproduced in succeeding generations of a developmental system but none of which really belong alone to either 'gene', 'organism' or 'environment' (the famous "Triple Helix" of Richard Lewontin 2000). Its contextualization of genes should obviate "even naïve temptations toward gene/environment dichotomies, and … will open up a very rich area of empirical investigations to examination and conceptualization in developmental-system terms. … Ultimately, such a view should work towards overcoming inner/outer dichotomies in favor of self-organizing, causally reciprocal systems of interaction" (Moss 2001, 85). Developmental Systems Theory (DST), an alternative approach to integrating evolution, development and inheritance, provides just such a framework and its conception of development is basically the one promoted in this paper. DST's central tenets, as stated in what could be considered its manifesto, are a) joint determination by multiple causes and distributed control, b) context sensitivity and contingency, c) extended inheritance, and d) development and evolution as construction (Oyama, Griffiths, and Gray 2001b). What this section has contributed is on the one side a necessary background for the argument by introducing our understanding of the process of development. This will be necessary in order to apply DST's framework to a new pressing question, namely how should one conceptualize the relationship between development and learning, which is the central question in this paper. On the other side we introduce some new developments that support DST's general perspective and develop it further: 1. There have been a lot of discoveries in the regulation of gene expression in the last decade that give ammunition to DST's general framework. In particular, the field of environmental epigenetics and epigenomics, which was hardly known just 15 years ago, has taken off as a mainstream area of research. 2. The idea of ontogenetic niche construction, introduced some 20 years ago to formalize extragenetic inheritance, has not yet been taken up by DST because it was unknown to the authors when they formulated their manifesto, quite in contrast to (selective) niche construction which was immediately emphatically embraced (Oyama, Griffiths, and Gray 2001b). In addition, the idea of providing some prerequisites to overcoming the nature-nurture divide, while wholly compatible with DST's teaching, was originally developed in one of the present authors' other publications (Stotz 2008).

The important systems features of such a view include the rejection of dichotomous descriptions of behavior in favor of a full analysis in terms of continuing interaction among, and joint determination by, heterogeneous developmental resources. Learning may be involved but only as part of an overall concept of experience which includes less obvious contributions, such as self-stimulation. An important part of such an analysis implies seeing behavior as belonging to the organism's overall anatomical and physiological make-up. A dynamical systems

view of locomotor development exemplifies such an approach very well by including the growth of muscles and the infant's strength in an account of behavioral coordination of movement (e.g., Thelen 1995; Spencer et al. 2006). Other important features of a developmental systems account are: (i) context sensitivity and developmental contingency of all developmental factors; (ii) distributed control of development upon its heterogeneous resources, and the acknowledgement of the role played by the developmental system to control its further development; (iii) extending the idea of inheritance to include other factors than DNA, including factors formerly thought of as 'environmental' or 'experiential' if they are reliably reproduced or 'passed on' to succeeding generations; and, last but not least, (iv) reconceptualization of development (and evolution) as the interactive construction in a thoroughly epigenetic account of development that "never sidesteps the task of explaining how a developmental outcome is produced" (Oyama, Griffiths, and Gray 2001b, 4). This broad conception of 'epigenetics' is expressed succinctly by Eva Jablonka:

> Epigenetics … focuses on the general organizational principles of developmental systems, on the phenotypic accommodation processes underlying plasticity and canalization, on differentiation and cellular heredity, on learning and memory mechanisms. Epigenetics includes the study of the transmission of subsequent generations of developmentally-derived differences between individuals, thereby acknowledging the developmental aspect of heredity. (Jablonka, pers. comm., cited in Gottlieb 2001)

5.2.7 Reclaiming the Environment

Such an epigenetic view of development necessitates a new appreciation of the environment, which has been conspicuously absent from the last 100 years of developmental research. The rise of the new science of *Entwicklungsmechanik* (developmental mechanics) in the late 19th and beginning of the 20th century saw the demise of the anatomical tradition which, due to its evolutionary framework and its methods of observation of developing organisms in their natural context, came to be regarded as old-fashioned and unscientific. Mystical ideas of epigenesis were completely rejected. The new mantra of *experimentation* – with its new methodology of manipulating the animal in controlled laboratory settings – brought the discipline of embryology, now called developmental biology, from the sea shore to the indoors. It is necessary to understand the emerging 'model organism' approach against this background. To make the scientist independent from the dictates of seasonal availability and natural variability, laboratories started to breed their own animals with the goal of making them constantly available and as uniform as possible. This constrained the choice of organism which "must be selected for the inability of their development to be influenced by specific environmental cues". In other words, "the influence of … environmental sources of phenotypic diversity were progressively eliminated under the physiological context of embryology" (Gilbert 2003, 88f).

While the physiological tradition favored the whole organism at the expense of the environment, a newly emerging genetics focused on genes at the expense of the organism. Both research traditions discounted and dispensed with the environment, the former disregarding the external habitat of the organism and the latter disregarding the internal cellular environment of genes and their expression. This shows a parallelism with the contemporaneous and ironically named 'environmental determinism' movement in behavioristic psychology, which, by moving the study of animal behavior and learning into the laboratory, dispensed with both the variety of organisms and their natural habitat in favor of uniform organisms and controlled ('environmental') test conditions.

We contend that ecological validity will be an indispensable factor for studying development, experience and learning. This has long been acknowledged by ethologists and behavioral ecologists, as well as by developmental psychobiologists, and has become accepted by integrative approaches to animal behavior in the lab which ask for the ecological validity of experimental designs. For example, the "adaptive specialization approach" relates species differences in cognitive processes to the ecological problems faced by that species (Emery 2006). Other approaches that call for the investigation of organisms 'in the real world', such as *Ecological Developmental Biology* (Gilbert 2001; Gilbert and Epel 2009) and *Developmental Ecology* (West 2003), have inspired a flood of new observations and experiments cementing the influential role of ecology in the study of behavior.

5.3 Experience and Learning: from Subtle Influences to Obvious Connections

This section looks at the history and current accounts of research into mechanisms of learning in animals, with an emphasis on simple systems approaches in which the boundary between learning and other kinds of experience becomes fluid. This is an important step toward reconciling accounts of learning with our conception of epigenetic development that necessarily includes some form of experience in the construction of any physiological or behavioral trait.

5.3.1 Naked Behavior: the Loss of Internal Cognition and the Natural Environment

In the 19th century, studies of complex behaviors typically contrasted innate, instinctual behavior with the products of learning and intelligence (Crowley & Allen, 2008). But even some pre-Darwinian writers such as Henry Lewis Morgan argued that explanations in terms of instinct were vacuous because they merely attributed to an unknown material cause what would otherwise be regarded as the

product of intelligence (Johnston 2002). Nevertheless, the distinction persisted and by the late 19th century the concepts of instinct and intelligence were both understood within the general framework of evolutionary biology. Both notions remained controversial even within that framework. Comparative psychologists, exemplified by Conwy Lloyd Morgan, struggled with the question of how to deal with the subjective aspects of intelligence in a rigorous experimental fashion. At the same time, earlier experimental work on instinctive behavior was called into question. For instance, T. Mann Jones and Lloyd Morgan both repeated Douglas Spalding's experiments on feeding behaviors in chicks and found that, contrary to Spalding's conclusion, they involved a learned component (Boakes 1984).

Among the people influenced by Lloyd Morgan was Edward L. Thorndike who, at the beginning of the 20th century, demonstrated just how empirically tractable animal learning could be. Thorndike's experiments with animals escaping "puzzle boxes" showed how to quantify learning in terms of the decrease in time to escape with experience (Thorndike 1911). But Thorndike's methods also initiated a new trend in comparative psychology towards using laboratory setups that had little connection to evolutionary biology. Thorndike tested a range of different species and emphasized the comparative aspects of psychology, but his use of artificial situations and his formulation of general laws of learning such as his famous Law of Effect suggested that species differences were secondary.

Consequently, with the rise of behaviorism, there came a biologically uninformed environmentalism that regarded the main differences between species as the range of stimuli and reinforcers that could support classical Pavlovian stimulus-stimulus (S-S) conditioning and instrumental response-outcome (R-O) conditioning. Rats, pigeons, but few other species, were intensively studied, because it was assumed that, for the purposes of general learning theory, species differences were relatively unimportant. The behaviorists' categories of S-S and R-O conditioning, and their interpretations of animal behavior, were inseparably linked to, and ultimately defined by, their experimental methods. The terms 'associative' and 'nonassociative' learning are both theoretical abstractions. They are not the result of direct observation; their occurrences are merely inferred. Also, the distinction between single event learning, such as habituation and sensitization, and related event learning, i.e. classical (Pavlovian) and instrumental learning, can be seen as rather arbitrary, because it classifies types of learning according to a formal outcome (in a laboratory experiment, no less) rather than considering the underlying mechanisms, which might be quite similar at the neural or molecular levels (Grau and Joynes 2005). This operationalist approach to learning involved little or no regard for the animal's evolutionary or developmental history, its ecological habitat, and its cognitive processes. Or, at least, no explicit regard. For, as William Timberlake has argued, behaviorists' experimental apparatuses were 'tuned' to evolutionary, developmental, and ecological aspects of the organisms studied (Timberlake 2002). Rat learning, but not pigeon learning, was investigated in mazes, and the use of different operant responses, whether pecking or bar pressing, and even subtler aspects of equipment design, such as the size and positioning of

levers in a Skinner box[3], implicitly reflect the experimenters' adaptation of laboratory setups to biological features of the organisms under study.

Comparative psychologists have been paying explicit attention to ecological aspects of learning for well over a decade now (see, e.g., an early review by Shettleworth 1994). The last decade has seen much exciting work on varieties of social and observational learning, and even 'insight' learning, which do not fit the standard models for classical and instrumental conditioning. Nevertheless, the recent lively debate about these topics has been conducted largely in terms of operational definitions and experimental protocols, rather than underlying mechanisms. Insofar as there is discussion of mechanisms, it has been to pitch "associative" mechanisms (i.e. of the sort described in traditional learning approaches and discussed above) against "cognitive" mechanisms (such as "theory of mind" or knowledge of physical forces). But even among opponents in this debate there is growing recognition that it has been carried on at too abstract a level to be fruitful (Allen 2006; Papineau & Heyes 2006; Penn & Povinelli 2007a).

Questions about which species are capable of which forms of learning are typically treated as if organisms come to the task as fully-formed representatives of their species. Thus questions about, for example, the imitative capacities of primates rarely take individual development into account (Jones 2005). In fact, it is widely believed on the basis of non-developmental studies that monkeys aren't capable of genuine imitation or are very poor at it while apes are naturally more capable (Byrne 2004). But the importance of development is underscored by experimental findings with human-reared or enculturated apes (mother-raised in captivity with human interaction; nursery-raised; laboratory-trained; and raised within human culture) which gave rise to the strong "enculturation hypothesis", which was later modified into the weaker "socialization hypothesis" (Call and Tomasello 1996; Tomasello and Call 2004; see also Bering 2004; Furlong, Boose, and Boysen 2008). While some argue that the developmental evidence so far is still too scant to draw definite conclusions (e.g. Penn and Povinelli 2007b), these explanations provide an *epigenetic model* of the differential effects of enculturation in human socio-cultural environments on the development of a whole range of capacities in great apes. Among those are many which nativist theories assign to humans alone, such as mental representational capacities and a whole range of social cognitive capacities such as gaze following, joint attention, intentional understanding, empathy, and 'true imitation'. A recent report describing imitation by Japanese macaques points in the direction of a similar conclusion about the importance of the social context for development of imitative abilities. These macaques were raised in an environment where joint attention with human caregivers was emphasized through the use

[3]The "Skinner box", also known as an "operant conditioning chamber" is an apparatus still in wide use that was designed by the psychologist B.F. Skinner. In the chamber, stimuli, punishments, and rewards can be mechanically delivered on a predetermined schedule to the animal subject in the chamber, whose behaviors, e.g., bar pressing (used for rats) or key pecking (used for pigeons), are mechanically recorded and carefully logged in their temporal relation to the delivery of associated stimuli and reinforcers.

of gestures such as pointing and the communicative use of eye-gaze, and they subsequently performed well in imitation tasks that macaques typically fail (Kumashiro et al. 2003; but compare with Subiaul et al. 2004 whose monkeys mastered cognitive imitation without social facilitation). Likewise, Savage-Rumbaugh's investigation of the bonobos Kanzi and Panbanisha for such capacities as language comprehension, symbolic communication, and tool use, especially when these two bonobos are contrasted with unenculturated bonobos in the lab such as P-Suke, points to the need for systematic studies of development (but see Savage-Rumbaugh, Fields, and Spircu 2004 for a step in this direction). Lloyd (2004) argues convincingly that many of Savage-Rumbaugh's critics have seriously underestimated the importance of development. It is, for example, beside the point to argue that symbolic communication is outside the repertoire of mature, natural-born bonobos. As she puts it, "in order to draw conclusions about *potentialities*, we must investigate them" (Lloyd 2004, 587). As Sue Savage-Rumbaugh has said (pers.comm.), "If apes don't talk because they have nothing to say, then we must provide them with something to talk about." McGonigle & Chalmers have also criticized psychologists for underestimating the role of learning in cognitive development because their "investigations are rarely followed through from one learning episode to another to assess the cumulative benefits (if any) as a function of the agent's task and life history" (McGonigle and Chalmers 2002).

5.3.2 Simple Learning Systems

Neuroscientists and molecular geneticists interested in animal learning have generally adopted the behaviorists' classificatory scheme of S-S (classical Pavlovian) and R-O (Skinnerian operant) conditioning, but have also attended to 'simpler' forms of single-stimulus learning, such as habituation, dishabituation, and sensitization. Invertebrate organisms, especially leeches and sea slugs, have provided much of our basic understanding of the role of mechanisms of synaptic change in single-stimulus and associative learning (Castellucci et al. 1970; Burrell and Sahley 2001). In most such work, the basic classificatory scheme is methodological and not tied to individual life histories in any detailed way (but see Stopher et al. 1991 for a developmental approach to learning in *Aplysia*; see section 5.4 for more details).

Some behavioral neuroscientists have recognized the shortcomings of the operationalism underlying the traditional classification scheme. For instance, Grau & Joynes argue for a 'neurofunctionalist' approach which seeks to classify learning in terms of both neural mechanisms and adaptive function (Grau and Joynes 2005a, 2005b). The work done in Grau's lab has shown remarkable learning and plasticity in the rat spinal cord, detached from the rat brain. Their results include long-term effects of nociceptive experience on spinal learning and on its capacity to recover from spinal injury (reviewed in Grau et al. 2006; Allen, Grau, and Meagher 2009). These results suggest that even in the spinal cord, 'experience' has lasting effects on the capacity of neurons to respond adaptively to future environmental conditions.

Even better, spinal cords can 'learn to learn' and are susceptible to an analog of 'learned helplessness' in which adaptive responses and learning capacity are both impaired. Despite the obvious developmental significance of these results, organismic development is not an explicit component of their research program.

With hindsight, perhaps no one should have been surprised that the vertebrate spinal cord is a plastic, adaptive system in its own right. After all, invertebrates with fewer neurons than the typical rat spinal cord nevertheless show various forms of learning. The basic cellular mechanisms for learning and memory are highly conserved between invertebrates and vertebrates (Burrell and Sahley 2001) and may even go further back in evolutionary history. For example, the NMDA receptors involved in the synaptic plasticity of neurons use proteins for binding amino acids that are highly conserved from bacteria (Kuryatov et al. 1994).

Even the simplest organisms, bacteria, respond differently to similar configurations of cues in their surroundings on the basis of their specific life experiences. Some of the physical properties of the cellular boundary and the bacteria's complement of cell-surface and internal receptors can react during early growth to environmental factors such as kinds of nutrients, temperature, pH, or concentrations of toxins. Other processes formerly thought to be restricted to more complex organisms have now been described as the norm rather than the exception of prokaryotic behavior. These include the processes of morphogenesis (change in form), cellular differentiation (change in function), aging, communication, and a whole range of group-mediated, cooperative behavior, such as aggregation and sporulation (Lyon 2006; see also Shapiro 2007; Ben-Barak 2008; Zimmer 2008). Shapiro argues that sophisticated information processing capacities in prokaryotic cells warrants a more contemporary view of bacteria as cognitive entities acting in response to sensory inputs. He describes how smart even the smallest living cells can be due to their capacity for meaningful intercellular communication. "Here the term cognitive refers to processes of acquiring and organizing sensory inputs so that they can serve as guides to successful action. The cognitive approach emphasizes the role of information gathering in regulating cellular function" (Shapiro 2007, 812).

The *concept* of bacterial learning may thus be no mere philosophical abstraction. But, someone concerned with preserving old distinctions might press us, do bacteria really 'learn'? The answer one gives, of course, depends very much on one's definitions of learning and experience. The question gets a negative answer if learning is restricted to organisms with nervous systems that connect sensory to motor systems, and sensory systems are conservatively defined as specialized organs with specialized receptor cells that connect a specialized cognitive system that has specialized information-transmission cells to the outside world to extract information from the environment for action (or behavior, narrowly defined). However, the answer may well be in the affirmative if 'environment' is understood as the source of a "quite heterogeneous mix of resources called experience" (Moore 2003, 350) extracted by a wide variety of means, and if knowledge and means for behavior derive from more than what is known to the senses defined in the strictest sense. It is definitely 'yes' if cell-surface receptors are subsumed under the rubric of sensory organs (Baker and Stock 2007).

When scientists extend the application of concepts beyond their usual meanings it is right for philosophers and other scientists to be skeptical about whether the extension is warranted or helpful. This became obvious in the debate between the developmental psychobiologist Gilbert Gottlieb and the behavioral geneticists Gottesman and Turkheimer on the notion of 'development' in so-called 'developmental behavioral genetics' (Gottlieb 1995; Turkheimer, Goldsmith, and Gottesman 1995; see Griffiths and Tabery 2008 for some critical reflections on this debate). With the application of concepts such as learning (Tagkopoulos, Liu, and Tavazoie 2008), memory and anticipation (Saigusa et al. 2008), and cognition (Baker and Stock 2007) to single-celled organisms there will, no doubt, be missteps along the way. But not all of these extensions can or should be dismissed on the basis of definitions provided *a priori*. As biologists have become better and better attuned to the extensive network of interactions between genomes and environments, and have gained a greater appreciation of the plasticity of biological systems, the old distinction between development and learning looks increasingly untenable. Almost a century ago, Carmichael made an early attempt at such a synthesis. He wrote that "in all maturation there is learning; and in all learning there is hereditary maturation" (Carmichael 1925, 260; quoted in Johnston 2001). This was ill-conceived because of its poor concept of development as – predetermined – hereditary maturation, which kept the dichotomy between innate and acquired wholly intact. We believe that it is now possible to make good on the promise of treating learning as just one process of experience, and to bring it all under the general umbrella of development (in keeping with the quote by Midgley and Morris in the introduction).

5.4 Synthesizing Development and Learning

There are many ways to reorganize the relationship between two disciplines – here biology and psychology – and their concepts or processes – i.e., development and life versus learning and cognition. Greater appreciation for the biological underpinnings of cognition has led some to propose the equation of life and cognition. Most notably, the 'Santiago' theory of Maturana and Varela makes this explicit: "Living systems are cognitive systems, and living as a process is a process of cognition. This statement is valid for all organisms, with or without a nervous system" (Maturana and Varela 1980, 13). All living organisms, even individual cells, react adaptively to individual experience of external perturbations to maintain their identity. According to this view, cognition, understood as the very basic operation of making a distinction, defines the boundaries of the system and is therefore the activity involved in the self-production (autopoiesis) of living systems. In other words, cognition is not equivalent to, but an indispensable characteristic of, life.

We certainly don't want to propose the fusion of biology and psychology to the extent that both would lose their distinct identities. Rather we promote a biologically-informed psychology and a psychologically-informed biology. This would require the reciprocal reconciliation, integration and synthesis of their overlapping

areas of study, such as the study of behavior, and of their central concepts, among which are development and learning, and as we propose, experience. We should again mention here that the field of developmental psychobiology is explicitly committed to such an approach. Also the developmental psychology of Jean Piaget attempted such a biologically-informed psychology that compared and integrated the processes of biological and psychological development. For instance, Piaget understood behavior as one means for the organism to actively adapt to the environment. And he understood that intelligence, as an extension of biological organization, functions as the organizing activity that enables yet more satisfactory adaptation (Piaget 1971/1967; Stevenson 1962). The study of behavior looks at three interconnected time-scales: evolution, development, and situated behavior. This integration is based on an essential role for biology in a theory of behavior. Central to the project of synthesizing development and learning is to identify cases of epigenetic interaction, both narrowly and widely construed, the role of experience and learning in development, and the role of development in the phenomenon of learning. From a psychobiological perspective, learning appears as a category within an overall framework of development as the lifelong, adaptive construction of the phenotype out of the interaction between genes, the organism and its environment. Taking the idea of phenotypic plasticity seriously may lead to a conception of development as a lifelong process of 'learning' or 'acquiring' a mode of living in an environment that is partly constructed by the organism or the previous generation. The other way around, learning understood as the acquisition of novel behavior and gain of knowledge about the environment, becomes synonymous with developing.

5.4.1 The Role of Epigenetic Mechanisms in Development and Learning

Traditionally, behavior has been explained by dissociable influences of genes – producing hardwired, innate behavior – and environment – causing acquired, learned behavior. Today it is slowly becoming apparent that the picture is much more complex and intertwined, and that experience, or any kind of environmental inputs in general, achieve their effect on behavior at least in part through the regulation of gene expression in some or all cells, but particularly the nervous system. The environment achieves its influence on the organisms via three general mechanisms: One is signal transduction from the environment through the sensory system to the genome, mediated by the neuroendocrine system and their associated hormones that function as both transcription factors and neuro-transmitters. Secondly, during the process of direct induction, environmental factors interact directly with the cell where they can either activate or repress signal transduction cascades that activate gene expression. Thirdly, following environmental induction epigenetic molecular mechanisms alter gene expression by chemically modifying the DNA nucleotide bases or the DNA chromatin structure. DNA methylation and histone deacylation repress gene expression by blocking access to the DNA by transcription factors,

while the opposite mechanisms of demethylation and acetylation render the DNA active by allowing certain transcription factors with promoter binding capacity to recruit the transcriptional machinery to the DNA.

Chromatin is the chromosomal complex made up of DNA and histone proteins that enables DNA to be tightly packaged into the nucleus and helps to control transcriptional access to DNA. Originally thought to be a rather structurally static complex, chromatin has been shown to be part of a very flexible and dynamic mechanism of precise transcriptional regulation (Sweatt 2009). Accordingly, chromatin remodeling such as methylation (DNA and proteins) and acetylation (protein) does not only occur in early development but can happen later in life, thus providing a mechanism through which the environment sculpts the genome and affects the phenotype throughout the whole life cycle. Hence the response of the 'epigenome' to environmental influences is a biological mechanism that serves as a medium for the adaptability of the genome to altered environments during life (Jaenisch and Bird 2003).

There are at least four different aspects of how this epigenetic control of gene expression relates to development and learning: 1) In the course of morphogenesis and psychogenesis undifferentiated and totipotent stem cells divide into pluripotent cells that are able to react to environmental signals by remodeling chromatin to change the cell's gene expression; during this process these originally pluripotent cells develop into fully differentiated cells identified by their individual 'epigenetic code' and its associated differential gene-expression pattern. These relatively stable alterations of the chromatin structure are one of the cell's main memory mechanisms by which they inherit and maintain their differentiated phenotype.

2) Epigenetic changes are also the main mechanisms underlying the process called 'fetal programming', "the concept that epigenetic factors in the intrauterine environment have a profound effect on the trajectory of prenatal development" (Nathanielsz and Thornburg 2003) that can lead to lasting effects of neonatal experience on adult physical (cardiovascular, metabolic diseases) and psychological (stress reaction, neural plasticity, depression, schizophrenia) phenotype. There is a wealth of experimental evidence that relates maternal care in mammals to epigenetic changes of genes in the cells of selected neural and organ systems. For instance, increased NMDA receptor expression can influence hippocampal synaptic development and function, which then translates into differential spatial learning and memory abilities (Meaney 2001a). In other words, there exists a developmental need for epigenetic mechanisms to allow the formation of a normal nervous system.

3) Beyond these two developmental aspects, the roles of epigenetic mechanisms in cognitive processes throughout life, such as learning and memory formation, are becoming increasingly appreciated. These include influences on associative fear conditioning, extinction of conditioned fear, latent inhibition, spatial learning and memory, and memory recovery. Epigenetic mechanisms have also been implicated in the positive effects of environmental enrichment on memory capacity. It can be postulated that DNA methylation/histone modification–mediated gene regulation is not only important for neural cell differentiation but also crucial for synaptic plasticity and high-order cognitive functions such as learning and memory, especially the formation of long-term memories (Sweatt 2009).

Together these studies demonstrate that experience, activity, and neurotransmitter-dependent activity increases histone acetylation *and* DNA methylation and that both are required for learning and memory. Indeed, there seems to be a dynamic balance between inhibition of a memory suppressor gene (protein phosphatase-1) and induction of a memory enhancer gene (reelin), antagonistically driven by increased states of DNA methylation and histone acetylation, respectively (Miller, Campbell, and Sweatt 2008; Noh et al. 2005).

The finding of specific types of memories associated with specific patterns of histone modifications suggests the intriguing possibility for a type of epigenetic "code for memory formation" (Wood, Hawk, and Abel 2006). In general, the formation of long-term memory requires NMDA-receptor-dependent synaptic transmission. However, different types of long-term memory seem to be associated with distinctive kinds of epigenetically induced modifications of the genetic material: Acetylation of hippocampal histone H3 but not H4 is significantly increased in the hippocampus after an animal is trained with a contextual fear-conditioning paradigm. A different form of long-term memory, latent inhibition, was associated with altered acetylation of histone H4, whereas H3 acetylation was unaltered by this paradigm (Chwang et al. 2006).

4) If an environmentally produced sensory input induces a change in behavior that persists beyond the presence of the original stimulus, we are speaking of memory. Such a notion of memory applies not just to neural systems, but also to cells that use (among other mechanisms) chromatin modifications to maintain changes of gene expression through cell divisions for the remainder of a cell's life. Paradigmatically, learning is understood as a usually adaptive, neural response to an input (an external stimulus or the organism's own behavior) in which the input-response relation is memorized. The recall of these memorized relations can later be the basis of a more effective response. It has recently been suggested that this very general characterization of learning applies not only to neuronal systems but also to cellular responses that are based on epigenetic mechanisms of cell memory (Ginsburg and Jablonka 2009).

The extensive interplay between epigenetic mechanisms and learning is well enough established that it is no longer adequate to ignore it, pleading that it's too complex to consider or outside the scope of traditional learning experiments and theory. The problem is of course that most evidence comes from animal experimentation, and hence the legitimate question arises, "How does this translate to humans?" So far both practical and ethical limitations have precluded the possibility of human studies in a way that truly and sufficiently captures the dynamic interplay of factors (Miller 2010). But many researchers now believe that the wealth of animal studies may provide a solid rationale to direct the search for epigenetic alterations in humans. For instance, over the next five years the Canadian Institutes of Health Research is financing a $4 million study, called the MAVAN project (Maternal Adversity Vulnerability and Neurodevelopment), in which Michael Meaney and colleagues from across the country will examine the effects of parental care on child development.[4]

[4] http://www.douglasrecherche.qc.ca/news/1006

5.4.2 *Learning and the Provisioning of Experience as (part of) Development*

As we stated at the outset, our objective is to integrate the idea of learning into a wider concept of experience and development. Specifically, in a systems view of development "all bio-behavioral phenomena – "innate" and "acquired" – are the products of "a continuous developmental process from fertilization through birth to death" (Kuo, 1967/1976, p. 11, cited in Midgley and Morris 1992). In such a framework learning may appear as just one among many processes in which experience influences the phenotype in general and behavior in particular. Experience affects behavior on many time scales. Even the most fleeting behavioral effects involve gene regulation and expression, but there may be no lasting effects unless the experience is repeated or other conditions coincide to shift the system into a new, relatively stable region of its phase space. Some experiences or combinations of experience, however, produce long lasting changes in the systems' dynamics, and when such changes are (typically) adaptive, we may label them as either development or learning.

But our use of the phrase 'development or learning' is too ambiguous. It could mean that these are two fundamentally different kinds of processes that happen to have a similar temporal relationship to experience: relatively long term, typically adaptive effects resulting from interactions with contingent aspects of the extracellular environment. Or, it could mean that the development and learning are more tightly assimilated to one another. Not learning *and* development, but learning *as* part of development. It is this latter interpretation that we wish to defend, if only for the sake of forcing a reconceptualization of a crude dichotomy between these terms.

To begin, we start with a relatively uncontroversial description of development as the process of organismic transformation from a single cell to a differentiated, structured entity. Because this characterization of development tends to suggest a material or anatomical conception of the organ or organism, it can seem like a category mistake to force learning into the same mold. However, it is important to realize that learning is a specialized process of (typically neural) differentiation and structural change that supports (adaptive) modification of behavior by experience. From this it follows that learning is a kind of developmental process: i.e., learning as development. As we have already indicated in connection with bacterial development, we think that this assimilation of learning to development is no mere metaphor: The processes underlying bacterial development and neural modification during learning are evolutionarily conserved to a surprising degree.

We are also willing to go quite far in the other direction, assimilating development to learning. Many if not all biologically significant developmental processes produce lasting changes in behavior as a function of experience. Even something as directly anatomical as limb development has behavioral consequences. Given one standard conception of learning as *change of behavior as a function of experience*, one may conclude that development is a kind of learning process. It might be objected that this conception of learning is excessively behavioristic — better definitions involve acquisition of knowledge, or other mental structures. However, in

our view, such 'knowledge' is itself biologically insignificant unless it results in behavioral change.

Several investigators have used the new framework of the developmental niche as one way to go beyond both nativist and empiricist oversimplifications of ontogeny and to highlight how learning processes are part of species-typical and individual development. Jeff Alberts conceptualizes the development of the rat in terms of four consecutive ontogenetic niches through which the pup passes on the way to adulthood (Alberts 2008). Common to each niche are channels of sustenance for the developing organism, such as nutrients, warmth and insulation, behavioral and social stimuli. The ontogeny of species-typical rat behavior is directed by olfactory cues that are provided by the different stages of the ontogenetic niche. For example, in the second stage, immediately postnatal, olfactory cues on the dam's nipples guide the pup's attachment and suckling. The pup's developing sensoria *learn* to recognize the odor for the nipple through chemical cues in the amniotic fluid provided by the prenatal 'uterine niche'. The spread of amniotic fluid after birth over the dam's body bridges the pre- and post-natal niches of the pup.

Another example of how the rat's developmental niche affords the necessary experience for the developing pup is the 'huddle'. Huddling is an important, species-typical behavior of the rat exhibited from day 15. Filial huddling preferences are mediated by *learned* olfactory cues. The olfactory-guided species preference is induced by thermotactile stimulation provided by the 'natal niche'. Alberts notes:

> Again we find a stereotyped, species-typical, developmentally-fixed behavior is learned, with all of the key components […] existing as natural features of the ontogenetic niche. … Specific features of these [nurturant] niches elicit specific reactions and responses in the developing offspring. These reactions and responses constitute conditions sufficient for the formation of a learned association and, as a result, the differentiation of behavior. … The utter reliability of the ontogenetic niches and the affordances that exist in each *are inherited as surely as are genes*. (Alberts 2008, 300, emphasis added).

Meredith West and Andrew King have shown over many decades of painstaking research that a nest parasite, the Brown-headed Cowbird, is not a paradigm example of a 'hardwired' species, as normally assumed. If there is a 'safety net' it is not in a 'genetic program' but in the social structure of the flock. An individual cowbird's niche is defined by his or her position within the flock, which "gates' what is "bioavailable" to be culturally transmitted or learned throughout the lifespan. According to West and King the developmental system is designed to be as open as ecologically possible. To that effect evolution has trusted an exogenetic developmental niche to transmit information that is vital to cowbird reproduction from one generation to the next (West and King 2008). "It's the dependability of the niche in delivering certain resources to the young that makes it a legacy. They inherit the senses and the surrounding to find what they need" (West, King, and Arberg 1988, 46; West and King 1987). One of their important claims is that the ontogenetic niche gates what is available to be learned, in other words what really matters is the bioavailability of stimulation and experience rather than simple exposure.

Many more examples of the way in which developmental niches provide for the reliability of encountering experiences necessary for normal development could be provided were there space to do so. Such examples would include human language

learning, food and habitat imprinting in insects (oviposition); maternal care and stimulation for neural development (sexual behavior and fear reaction in rats; learning disposition in chickens).

5.4.3 The Development of Learning

There are some quite straightforward reasons why a more developmentally sensitive approach to animal learning is useful. It may help to uncover age-related behavior differences as well as age-related changes in learning that subjects bring to tasks, and to control for, or even exploit, the effects of earlier experiments with the same subjects. Further interesting questions concern whether mechanisms and content of learning change ontogenetically, and if so, what this can tell us about the generality of learning mechanisms in adults. How are experiential regulation of brain development and general learning mechanisms related? Do developing and mature brains share the same information storage mechanisms or does neural plasticity in early life interfere with the processes of learning and memory (Shair, Barr, and Myron 1991, chapters 1, 6, 14, and 15)?

A developmental approach to forms of learning and memory in *Aplysia* helped to differentiate a novel inhibitory process that is often masked by sensitization in the adult, and two different forms of response facilitation, which emerge at different developmental times. Dishabituation and sensitization differ in fundamental ways, such as their developmental timetable, their time of onset, and their stimulus requirements. The investigators therefore concluded that a formerly held, simple dual-process view is inadequate to the features of these two kinds of nonassociative learning processes (Stopher et al. 1991).

5.4.4 The Quest for New Distinctions

Our deliberate attempt to erase long-held dichotomies and boundaries doesn't deny the existence of distinctions. Drawing useful distinctions is an important part of the scientific process of categorization, but sometimes one has to let go of long-held beliefs in order to cast new light on an issue, in order to see general principles and continuities instead of clear cut distinctions. Understanding development as the contingent process of construction as outlined above helps to overcome the unscientific dichotomies of nature versus nurture, instinct versus learning or innate versus acquired, and replaces them with scientifically more meaningful and fruitful distinctions. As Lenny Moss puts it:

> What the sad endurance of that tired old dichotomy consisting of (conflated) genes and (ill-defined) environment has helped to obscure, are the many levels of biological ordering that mediate between individual molecules and whole developmental systems. To give up the preformationist umbilical cord is not to drop into an abyss of limitless complexity but rather to remain empirically open to discovering what levels of biological ordering is most relevant for one's explanatory purposes (Moss 2001, 91).

Such distinctions need to rest on a deep analysis of the causal roles played by the diverse developmental resources within the ontogenetic process. The causal role of 'genes', the coding sequences in the DNA sequence, is the 'causal specificity' of the linear sequence of gene products. However, there are cellular processes called alternative splicing that change the length and the linear order of the original coding sequences in reaction to the external environment and conditions within the cell. Other processes even alter the original sequence through the deletion, insertion or substitution of some of the nucleotides that make up the original sequence during the transcription process, a mechanism called RNA editing. Hence the molecules that provide causal specificity to these sequence-modifying processes, such as splicing and editing factors, regulatory sequences in the DNA sequence and environmental signaling factors, *share the causal role of genes*. These processes provide us with an argument against Ken Waters, who asserts that only genes are causally specific difference makers in the production of gene products such as RNAs and proteins, and that the parity thesis of DST is therefore wrong. The parity thesis merely maintains that some roles are shared by several, quite different cellular factors, not that all causal factors are on a par (which is how Waters 2007 misunderstands the thesis; see Stotz 2006 for a sustained critique). The case of causal or sequence specificity shows quite clearly how sometimes the same causal role can not only be shared by several developmental resources, but how these can also change their role depending on context. For example the causal role 'activator', namely binding to particular regulatory sequences in the DNA such as enhancers, is shared by a range of diverse DNA-binding proteins. These same proteins may fulfill other causal roles, such as 'inhibitor', depending on the context in which they find themselves, to which group of interacting molecules they are recruited, and their combinatorial action.

An analysis of the causal role of factors reveals a hierarchy of functions, and how many factors share a single causal role depends in part on the amount of detail used to describe this role. So while coding DNA sequences, splicing and editing factors share the (generic) causal role of enforcing sequence specificity, exactly how they fulfill this role within the complicated process of gene expression is quite diverse. Hence, if the causal role of coding sequences were to be specified in much more detail, one might find that indeed 'only' genes fulfill this role. But for most intents and purposes it is the causal role of enforcing sequence specificity that was attributed to genes in the Central Dogma of Molecular Genetics, and it is at this level of description that Waters made his argument.

Developmental resources play a range of causal roles. Many cases cited in the 'eco-devo'[5] literature show that what Scott Gilbert has called the 'instructive' role can be carried out by environmental factors, while the genes involved play merely a 'permissive' role, e.g. in cases of polyphenisms in response to different ecological conditions. Examples include the temperature- or context-dependent sex determination of many reptiles, fishes and worms. This context-dependency of the morphological

[5]Ecological developmental biology concerns itself with the interactions between developing organisms and their environmental contexts in the real world (Gilbert and Epel 2009).

and behavioral phenotype is a "necessary condition of integrating the developing organism into its habitat" (Gilbert 2003, 98). A developmental systems view promotes another distinction: between resources that are 'reliably reproduced' (therefore inherited) from one generation to the next and those that are 'novel' or 'contingent'. Within inherited resources, the mechanism of transmission may be 'sample-based' or 'informational', and may serve the role of increasing developmental 'plasticity' or 'canalization' of phenotype. For any causal role at the center of an investigation, there will be a range of factors that are 'causally specific' with respect to it, and others that are mere background conditions. But whatever role is being investigated, hardly ever will it divide developmental resources neatly in 'genes' and 'environment', innate and acquired, and last but not least nature and nurture (Griffiths and Gray 2005).

5.5 Conclusion

The last decade has witnessed enormous scientific advances in genomics, systems biology, social neuroscience, evolutionary, and ecological and developmental biology, introducing notions such as 'evo-devo', 'eco-devo', phenotypic plasticity, niche construction, extragenetic inheritance, and developmental systems theory. In light of these advances it is no longer reasonable to defend either strongly nativist, gene-centered and pre-deterministic explanations of behavior on the one hand or strongly environmentalist explanations on the other. Nature and nurture are not separable entities with nature as the a priori plan and nurture as the contingent experiences shaping the outcome of the plan's execution. Instead, what we have tried to argue here is that every trait develops out of the nonlinear interactions among a range of very diverse developmental resources. We maintain that, for the purpose of explaining the origin of behavioral and cognitive capacities of animals, their causes are not usefully divided into genetic and non-genetic factors. Behavioral development starts with the environmental regulation of gene expression, and depends upon a range of experiences beneath the skin and above the gene, to construct the stages of sensory and social learning in vertebrates, to the exquisitely sensitive learning capacities of the human brain. Given that people won't stop talking about nature or nurture, we should at least try to place them into a non-dichotomous context: 'Nurture' is this ongoing *process* of development, while 'nature' is the *product* of the organism-environment-system (Oyama 1999).

Our aim in this paper has been to argue that the separation of questions about learning in human and nonhuman animals from questions about their development is as untenable as the old distinction between nature as genetic causes and nurture as environmental causes. Too many experiments testing the cognitive abilities of animals are done without fully reporting the developmental backgrounds of the research subjects, let alone systematically investigating the role of a lifetime of experience in the construction of the animals' behavioral-cognitive phenotype. Our hypothesis in this paper has been that one of the causes of this neglect is the difficulty some research

areas have in taking development seriously as a contingent, rather than a predetermined process. We have warned that vacuous concepts such as instinct, innateness, or genetic program forestall deeper investigation into the real causes of behavior.

We have also argued for a more conceptual point about the assimilation of learning and development. That argument can be summarized as follows:

1. Development is the process encompassing the complete life cycle of organismic transformation from a single cell to a differentiated, structured entity, from birth to reproduction and death.
2. Learning is a specialized process of (typically neural) differentiation and structural change that supports (typically adaptive) modification of behavior by experience throughout the lifespan of the organism.
3. Therefore, learning is part of the developmental process: i.e., learning as development.

We advance more tentatively an argument for the identification in the opposite direction, viz.:

4. All developmental processes (that matter biologically) produce a change in behavior as a function of experience.
5. Therefore, development is a kind of learning process, broadly defined.

How might taking development seriously be important to matters of philosophical and scientific interest? In our view, comparative psychology (and the philosophy of animal cognition) engages rather too frequently in mapping the cognitive capacities of animals to the cognitive stages of humans, as if it is unquestionably meaningful to compare the cognitive abilities of chimpanzees to two or three year old children, for example. Such comparisons are explicit in the titles of many research articles (e.g. Collier-Baker & Suddendorf, 2006) not to mention the news headlines generated by such studies. Likewise, we think that philosophical and scientific energy would be better spent on trying to understand the experiences that are important for individuals of various species to develop a capacity for learning from experience, rather than assuming that basic learning mechanisms are themselves preprogrammed or innate.

Although we do not have the space to develop these points in detail, we also think that various philosophical attempts to naturalize intentional content embody preformationist assumptions. The major theories of the past couple of decades treat adult concepts as fixed points of meaning. The assumption was explicit in Jerry Fodor's nativism and it appears more subtly in Dretske's early attempts to ground meaning as the stable outcome of a discrete pre-semantic learning phase (Fodor 1975; Dretske 1981). While Dretske's account allowed an important role for experience in fixing meanings, it foundered on the impossibility of finding a sharp divide between the pre-semantic phase and the adult stage of fixed meanings. The systems view of development treats learning as one among many processes in which experience influences behavior in a lifelong process of adaptive construction of the phenotype in its environment. This perspective suggests that the existing strategies for naturalizing content are doomed to fail in face of the developmental facts. Any single-factor theory of content will, at best, be a low-dimensional abstraction of

what is a much richer set of interactions between organisms and their environments. To some this might suggest an eliminativist stance towards notions such as representation, meaning, and intentional content. Here, however, we do not want to take a stand on this issue, although we believe that low-dimensional abstractions are sometimes important tools for scientific modeling. However, more inclusive and detailed models are the ultimate goal.

A similar understanding of development to the one proposed here had been arrived at independently in one corner of the cognitive science community, specifically among those interested in situated and embodied cognition, most notably the branch called Dynamical Systems Theory. Samuelson and Smith advocate for the dynamical systems perspective when they write, "We believe that in the next century, coupling the dynamics of perceiving and remembering with the dynamics of development will lead us to a more complete theory of knowledge and its development" (Samuelson and Smith 2000, 98).

Our vision of how to integrate the concepts of learning and development is based on a wider understanding of the role of the heterogeneous mix of resources making up 'experience' or 'environment' (Moore 2003). To what extent does the experienced environment correspond to the environment of behavioral development? If experience is defined to involve only what is known through senses, then it is a subset of the latter. A wider conception of developmental environments may include non-obvious influences with no straightforward connection to their effects on the organism. These influences are the object of study in developmental psychobiology but rarely ever investigated by comparative psychologists. With the concept of experience playing a central role, a biologically-informed psychology would have as one of its consequences that there would be hardly any features whose development is outside the scope of the psychological sciences. It would not, then, be appropriate to take these features as given. Calling a feature 'innate' or announcing that it 'matures' is simply issuing a "promissory note against future developmental psychology and biology" (Griffiths and Stotz 2000, 38). An elucidation of the developmental cascade by which a behavioral-cognitive capacity develops will contain both biological and psychological factors. It is our contention that any adequately naturalized account of philosophically significant notions such as intentionality, meaning, and knowledge can ill afford to ignore the best going scientific account of an organism's nature.

Acknowledgements This paper has been a long time developing. It stems from our interactions while KS was a postdoctoral research associate in the Cognitive Science Program at Indiana University. Members of the Indiana University Biology Studies Research Group provided comments on an early version of this paper, and we are especially grateful to Lisa Lloyd for her written comments on that version. We would both like to thank Indiana University's New Frontiers program for supporting the symposium "Reconciling Nature and Nurture in the Study of Behavior" organized by KS in 2007. We benefitted from a presentation of these ideas at the 2007 meeting of the Society for Philosophy and Psychology, which included commentary by Luc Faucher. We are grateful to the editors Katie Plaisance and Thomas Reydon for their comments, as well as two anonymous referees for the press. We would also like to thank Ulrike Pompe for her careful reading of the penultimate draft. KS's research is funded by the Australian Research Council's Discovery Projects funding scheme (project number 0878650). CA was supported by the Alexander von Humboldt Foundation while visiting the Ruhr University, Bochum, during the final preparation of this manuscript.

References

Alberts, J. R. (2008): 'The nature of nurturant niches in ontogeny', *Philosophical Psychology* 21 (Special Issue, Reconciling Nature and Nurture in the study of Cognition and Behavior): 295–303.

Allen, C. (2006): 'Transitive Inference in Animals: Reasoning or Conditioned Associations?' In Hurley and Nudds (eds.) *Rational Animals?* Oxford: Oxford University Press, 175–185.

Allen, C., Grau, J. W. and Meagher, M. W. (2009): 'The Lower Bounds of Cognition: What Do Spinal Cords Reveal?' In: J. Bickle (ed.): *The Oxford Handbook of Philosophy and Neuroscience*, Oxford: Oxford University Press.

Baker, M. D. & Stock, J. B. (2007): 'Signal Transduction: Networks and Integrated Circuits in Bacterial Cognition', *Current Biology* 17: R1021–R1024.

Ben-Barak, I. (2008): *Small Wonders: How Microbes Rule Our World*, Carlton North (Vic.): Scribe Publication.

Bering, J. M. (2004): 'A critical review of the "enculturation hypothesis": the effects of human rearing on great ape social cognition', *Animal Cognition* 7: 201–212.

Blumberg, M. (2005): *Basic Instinct: The Genesis of Behavior*, New York: Thunder's Mouth Press.

Boakes, R. (1984): *From Darwinism to Behaviorism*, Cambridge: Cambridge University Press.

Burrell, B. D. & Sahley, C. L. (2001): 'Learning in simple systems', *Current Opinion in Neurobiology* 11: 757–764.

Byrne, R. W. (2004): 'Detecting, understanding, and explaining animal imitation', In: Hurley, S. and Chater, N. (eds.): *Perspectives on Imitation: From Mirror Neurons to Memes*, Cambridge (MA): MIT Press.

Cacioppo, J. T. & Berntson, G. G. (eds.) (2004): *Essays in Social Neuroscience (Social Neuroscience Series)*, Cambridge (MA): MIT Press.

Call, J. & Tomasello, M. (1996): 'The effects of humans on the cognitive development of apes', In A. E. Russon, K. A. Bard and S. T. Parker (eds.): *Reaching into Thought*, New York: Cambridge University Press.

Call, J. & Tomasello, M. (1998): 'Distinguishing intentional from accidental actions in orangutans (*Pongo pygmaeus*), chimpanzees (*Pan troglodytes*) and human children (*Homo sapiens*)', *Journal for Comparative Psychology* 112: 192–206.

Carmichael, L. (1925): 'Heredity and Environment: Are they antithetical?', *Journal of Abnormal and Social Psychology* 20: 245–260.

Caspi, A., Hariri, A., Holmes, A., Uher, R., & Moffitt, T. E. (2010): 'Genetic sensitivity to the environment: The case of the serotonin transporter gene (5-HTT) and its implications for studying complex diseases and traits', *American Journal of Psychiatry* 167: 509–527.

Castellucci, V., Pinsker, H., Kupfermann, I. & Kandel, E. R. (1970): 'Neuronal mechanisms of habituation and dishabituation of the gill-withdrawal reflex in Aplysia', *Science* 167: 1745–1748.

Chwang, W. B., O'Riordan, K. J., Levenson, J. M. & Sweatt, J. D. (2006): 'ERK/MAPK regulates hippocampal histone phosphorylation following contextual fear conditioning', *Learning and Memory* 13: 322–328.

Claverie, J. M. (2001): 'Gene number: what if there are only 30,000 human genes?' *Science* 291: 1255–1257.

Collier-Baker E. & Suddendorf, T. (2006): 'Do chimpanzees (Pan troglodytes) and 2-year-old children (Homo sapiens) understand double invisible displacement?' *Journal of Comparative Psychology* 120: 89–97.

Crowley, S. J. & Allen, C. (2008): 'Animal Behavior: E pluribus unum?' In: M. Ruse (ed.): *The Oxford Handbook of the Philosophy of Biology*, Oxford: Oxford University Press, 327–348.

Dolinoy, D. C. & Jirtle, R. L. (2008): 'Environmental epigenomics in human health and disease', *Environmental and Molecular Mutagenes* 49: 4–8.

Donald, M. (2000): 'The central role of culture in cognitive evolution: A reflection on the myth of the "isolated mind"', In: L. Nucci, G. B. Saxe and E. Turiel (eds.): *Culture, Thought and Development*, Mahwah (NJ): Lawrence Erlbaum Associates.

Dretske, F. (1981): *Knowledge and the Flow of Information*, Cambridge (MA): MIT Press.
Emery, N. J. (2006): 'Cognitive ornithology: the evolution of avian intelligence', *Philosophical Transactions of the Royal Society B* 361: 23–43.
Fodor, J. A. (1975): *The Language of Thought*, New York: Crowell.
Freeberg, T. M., West, M. J., King, A. P., Duncan, S. D. & Sengelaub, D. R. (2002): 'Cultures, genes, and neurons in the development of song and singing in brown-headed cowbirds (*Molothrus ater*)', *Journal of Comparative Physiology* 188: 993–1002.
Furlong, E. E., Boose, K. J. & Boysen, S. T. (2008): 'Raking it in: The impact of enculturation on chimpanzee tool use', *Animal Cognition* 11: 83–97.
Gácsi, M., Virányi, Z., Kubinyi, E., Belényi, B. & Miklósi, Á. (2009): 'Explaining dog wolf differences in utilizing human pointing gestures: Selection for synergistic shifts in the development of some social skills', *PLoSONE* 4: e6584.
Gilbert, S. & Epel, D. (2009): *Ecological Developmental Biology: Integrating Epigenetics, Medicine, and Evolution*, Sunderland (MA): Sinauer Associates.
Gilbert, S. F. (2001): 'Ecological developmental biology: Developmental biology meets the real world', *Developmental Biology* 233: 1–22.
Gilbert, S. F. (2003): 'The reactive genome', In: G. B. Müller and S. A. Newman (eds.): *Origination of Organismal Form: Beyond the Gene in Developmental and Evolutionary Biology*, Cambridge (MA): MIT Press.
Ginsburg, S. & Jablonka, E. (2009): 'Epigenetic learning in non-neural organisms', *Journal of Bioscience* 33: 633–646.
Gottlieb, G. (1981): 'Roles of early experience in species-specific perceptual development', In: R. N. Aslin, J. R. Alberts and M. P. Petersen (eds.): *Development of Perception*, New York: Academic Press.
Gottlieb, G. (1995): 'Some conceptual deficiencies in 'developmental' behavior genetics', *Human Development* 38: 131–141.
Gottlieb, G. (1997): *Synthesizing Nature-Nurture: Prenatal Roots of Instinctive Behavior*, Hillsdale (NJ): Lawrence Erlbaum Associates.
Gottlieb, G. (2001): 'A developmental psychobiological systems view: Early formulation and current status'. In: S. Oyama, P. E. Griffiths and R. D. Gray (eds.): *Cycles of Contingency: Developmental Systems and Evolution*, Cambridge (MA): MIT Press.
Grau, J. W., Crown, E. D., Ferguson, A. R., Washburn, S. N., Hook, M. A. & Miranda, R. C. (2006): 'Instrumental learning within the spinal cord: Underlying mechanisms and implications for recovery after injury', *Behavioral and Cognitive Neuroscience Reviews* 5: 191–239.
Grau, J. W. & Joynes, R. L. (2005a): 'A neural-functionalist approach to learning', *International Journal of Comparative Psychology* 18: 1–22.
Grau, J. W. & Joynes, R. L. (2005b): 'Neurofunctionalism revisited: Learning is more than you think it is', *International Journal of Comparative Psychology* 18: 46–59.
Griffiths, P. E. (2002): 'What is Innateness?' *The Monist* 85: 70–85.
Griffiths, P. E. (2004): 'Instinct in the '50s: The British reception of Konrad Lorenz's theory of instinctive behaviour', *Biology and Philosophy* 19: 609–631.
Griffiths, P. E. & Gray, R. D. (2005): 'Three ways to misunderstand Developmental Systems Theory', *Biology and Philosophy* 20: 417–425.
Griffiths, P. E. & Stotz, K. (2000): 'How the mind grows: A developmental perspective on the biology of cognition', *Synthese* 122: 29–51.
Griffiths, P. E. & Tabery, J. (2008): 'Behavioral genetics and development: Historical and conceptual causes of controversy', *New Ideas in Psychology* 26: 332–352.
Herrmann, E., Call, J., Hernandez-Lloreda, M. V., Hare, B. & Tomasello, M. (2007): 'Humans have evolved specialized skills of social cognition: The Cultural Intelligence Hypothesis', *Science* 317: 1360–1366.
Hinde, R. A. (1966): *Animal Behaviour: A Synthesis of Ethology and Comparative Psychology*, New York: McGraw Hill.
Inouea, S. & Matsuzawa, T. (2007): 'Working memory of numerals in chimpanzees', *Current Biology* 17: R1004–R1005.

Jablonka, E. & Lamb, M. J. (2005): *Evolution in Four Dimenesions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*, Cambridge (MA): MIT Press.
Jaenisch, R. & Bird, A. (2003): 'Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals', *Nature Genetics* 33 Suppl.: 245–254.
Johannsen, W. (1911): 'The genotype conception of heredity', *American Naturalist* 45: 129–159.
Johnston, T. (2002): 'An early manuscript in the history of American comparative psychology: Lewis Henry Morgan's Animal Psychology', *History of Psychology* 5: 323–355.
Johnston, T. D. (2001): 'Towards a systems view of development: An appraisal of Lehrman's critique of Lorenz'. In: S. Oyama, P. E. Griffiths and R. D. Gray (eds.): *Cycles of Contingency: Developmental Systems and Evolution*, Cambridge (MA): MIT Press.
Jones, S. S. (2005): 'Why don't apes ape more?', In: S. Hurley and N. Chater (eds.): *Perspectives on Imitation: From Cognitive Neuroscience to Social Science*, Cambridge (MA): MIT Press.
Keller, E. F. (2000): *The Century of the Gene*, Cambridge (MA): MIT Press.
Kumashiro, M., Ishibashi, H., Uchiyama, Y., Itakura, S., Murata, A. & Iriki, A. (2003): 'Natural imitation induced by joint attention in Japanese monkeys', *International Journal of Psychophysiology* 50: 81–99.
Kuryatov, A., Laube, B., Betz, H. & Kuhse, J. (1994): 'Mutational analysis of the glycine-binding site of the NMDA receptor: structural similarity with bacterial amino acid-binding proteins', *Neuron* 12: 1291–1300.
Lamm, E. & Jablonka, E. (2008): 'The nurture of nature: Hereditary plasticity in evolution', *Philosophical Psychology* 21: 305–319.
Lehrman, D. S. (1953): 'Critique of Konrad Lorenz's theory of instinctive behavior', *Quarterly Review of Biology* 28: 337–363.
Lehrman, D. S. (1970): 'Semantic & conceptual issues in the nature-nurture problem'. In: D. S. Lehrman (ed.): *Development & Evolution of Behaviour*, San Francisco: W. H. Freeman and Co.
Lewontin, R. C. (1983): 'The organism as the subject and object of evolution', *Scientia* 118: 65–82.
Lewontin, R. C. (2000): *The Triple Helix: Gene, Organism, and Environment*, Cambridge (MA): Harvard University Press.
Linquist, S., Machery, E., Griffiths, P. E. & Stotz, K. (2011): 'Exploring the folkbiological conception of human nature', *Philosophical Transactions of the Royal Society B* 366(1563): 444–453.
Levitis, D. A., Lidicker Jr., W. Z. & Freund, G. (2009): 'Behavioural biologists do not agree on what constitutes behaviour', *Animal Behaviour* 78: 103–110.
Lloyd, E. A. (2004): 'Kanzi, evolution, and language', *Biology & Philosophy* 19: 577–588.
Lyon, P. (2006): *The Agent in the Organism: Towards a Biogenic Theory of Cognition*. PhD dissertation, Australian National University.
Maestripieri, D. & Mateo J. M. (eds.) (2009): *Maternal Effects in Mammals*, Chicago: The University of Chicago Press.
Maienschein, J. (2005): 'Epigenesis and preformationism', In: Zalta, E. N. (ed.): *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/epigenesis/.
Marler, P. & Slabbekoorn, H. (eds.) (2004): *Nature's Music: The Science of Birdsong*. San Diego: Elsevier.
Maturana, H. R. & Varela, F. J. (1980): *Autopoiesis and Cognition: The Realization of the Living*, New York: Springer.
McGonigle, B. O. & Chalmers, M. (2002): 'The growth of cognitive structures in monkeys and men'. In: S. B. Fountain, M. D. Bunsey, J. H. Danks and M. K. McBeath (eds.): *Animal Cognition and Sequential Behaviour. Behavioural, Biological and Computational Perspectives*, Boston: Kluwer Academic Publishers.
McGonigle, B. O. & Chalmers, M. (2008): 'Putting Descartes before the horse (again!). Commentary on Penn, D., Povinelli, D.J and Holyoak, K.J.', *Behavioral and Brain Sciences* 31: 142–143.
Meaney, M. J. (2001a): 'Maternal care, gene expression, and the transmission of individual differences in stress reactivity across generations', *Annual Review Neuroscience* 24: 1161–92.
Meaney, M. J. (2001b): 'Nature, Nurture, and the Disunity of Knowledge', *Annals of the New York Academy of Sciences* 935: 50–61.

Michel, G. F. & Moore C. L. (1995): *Developmental Psychobiology: An interdisciplinary science*, Cambridge (MA): MIT Press.

Midgley, B. D. & Morris, E. K. (1992): 'Nature = f(nurture): A review of Oyama's "The Ontogeny of Information: Developmetal Systems and Evolution"', *Journal of the Experimental Analysis of Behavior* 58: 229–240.

Miller, C. A., Campbell, S. L. & Sweatt, J. D. (2008): 'DNA methylation and histone acetylation work in concert to regulate memory formation and synaptic plasticity', *Neurobiology of Learning and Memory* 89: 599–603.

Miller, G. (2010): 'The seductive allure of behavioral epigenetics', *Science* 329: 24–27.

Moore, C. L. (1984): 'Maternal contributions to the development of masculine sexual behavior in laboratory rats', *Developmental Psychobiology* 17: 347–356.

Moore, C. L. (2003): 'Differences between organism-environment systems conceived by Lehrman and Gibson: What's in the nest of reciprocities matters', *Developmental Psychobiology* 42: 349–356.

Moss, L. (2001): 'Deconstructing the gene and reconstructing molecular develomental systems'. In: S. Oyama, P. E. Griffiths and R. D. Gray (eds.): *Cycles of Contingency: Developmental Systems and Evolution*, Cambridge (MA): MIT Press.

Mousseau, T. A. & Fox, C. W. (eds.) (1998): *Maternal Effects as Adaptations*. Oxford: Oxford University Press.

Nathanielsz, P. W. & Thornburg, K. L. (2003): 'Fetal programming: from gene to functional systems–an overview', *Journal of Physiology* 547: 3–4.

Newman, S. A. (2003): 'From physics to development: the evolution of morphogenetic mechanisms', In: Müller, G. B. & Newman, S. A. (eds.): *Origination of Organismal Form: Beyond the Gene in Developmental and Evolutionary Biology*, Cambridge (MA): MIT Press.

Noh, J., Sharma, R. P. Veldic, M. Salvacion, A. A. Jia, X. & Chen, Y. (2005): 'DNA methyltransferase 1 regulates reelin mRNA expression in mouse primary cortical cultures', *Proceedings of the National Academy of Sciences of the U.S.A*. 102: 1749 –1754.

Odling-Smee, F. J., Laland, K. N. & Feldman, M. W. (2003): *Niche Construction: The Neglected Process in Evolution*, Princeton (NJ): Princeton University Press.

Oyama, S. (1985): *The Ontogeny of Information: Developmental systems and evolution*, Cambridge (MA): MIT Press.

Oyama, S. (1999): 'The nurturing of natures'. In: Grunwald, A., Gutmann M. & Neumann-Held E. M. (eds.): *On Human Nature. Anthropological, Biological and Philosophical Foundations*, New York: Springer.

Oyama, S. (2001): 'Term in tension: What do you do when all the good words are taken?' In: S. Oyama, P. E. Griffiths and R. D. Gray (eds.): *Cycles of Contingency: Developmental Systems and Evolution*, Cambridge (MA): MIT Press.

Oyama, S., Griffiths, P. E. & Gray, R. D. (2001b): 'Introduction: What is developmental systems theory?' In: S. Oyama, P. E. Griffiths and R. D. Gray (eds.): *Cycles of Contingency: Developmental Systems and Evolution*, Cambridge (MA): MIT Press.

Oyama, S., Griffiths, P. E. & Gray, R. D. (eds.) (2001a): *Cycles of Contingency: Developmental Systems and Evolution*. Cambridge (MA): MIT Press.

Papineau, D. & Heyes, C. (2006): 'Rational or Associative? Imitation in Japanese Quail.' In Hurley and Nudds (eds.) *Rational Animals?* Oxford: Oxford University Press, 198–216.

Penn, D. C., Holyoak, K. J. & Povinelli, D. J. (2008): 'Darwin's mistake: Explaining the discontinuity between human and nonhuman minds', *Behavioral and Brain Sciences* 31: 109–129.

Penn, D. & Povinelli, D. (2007a): 'Causal Cognition in Human and Nonhuman Animals: A Comparative, Critical Review.' *Annual Review of Psychology* 58: 97–118.

Penn, D. C. & Povinelli, D. J. (2007b): 'On the lack of evidence that chimpanzees possess anything remotely resembling a 'theory of mind'', *Philosophical Transactions of the Royal Society B* 362: 731–744.

Povinelli, D. J. (2000): *Folk Physics for Apes*. New York: Oxford University Press.

Piaget, J. (1971/1967): *Biology and Knowledge: An Essay on the Relations between Organic Regulations and Cognitive Processes*, Chicago: Chicago University Press.

Pigliucci, M. (2001): *Phenotypic Plasticity: Beyond Nature and Nurture*, *Syntheses in Ecology and Evolution*, Baltimore: The Johns Hopkins University Press.
Robert, J. S. (2004): *Embryology, Epigenesis and Evolution: Taking Development Seriously*, Cambridge: Cambridge University Press.
Roe, S. A. (1981): *Matter, Life, and Generation: Eighteenth-Century Embryology and the Haller-Wolff Debate*, New York: Cambridge University Press.
Rosenberg, A. (1997): 'Reductionism redux: computing the embryo', *Biology and Philosophy*, 12: 445–470.
Saigusa, T., Tero, A., Nakagaki, T. & Kuramoto, Y. (2008): 'Amoebae Anticipate Periodic Events', *Physical Review Letters* 100: 018101.
Samuelson, L. K. & Smith, L. B. (2000): 'Grounding development in cognitive processes', *Child Development* 71: 98–106.
Savage-Rumbaugh, S., Fields, W. M. & Spircu, T. (2004): 'The emergence of knapping and vocal expression embedded in a Pan/Homo culture', *Biology & Philosophy*, 19: 541–575.
Schneirla, T. C. (1957): 'The concept of development in comparative psychology'. In: Harris, D. B. (ed.): *The concept of development*, Minneapolis: University of Minnesota Press.
Schneirla, T. C. (1966): Behavioral Development and Comparative Psychology, *Quarterly Review of Biology* 41: 283–303.
Shair, H. N., Barr, G. A. & Myron, eds. H. A. (1991): *Developmental Psychobiology: New Methods and Changing Concepts*. Oxford: Oxford University Press.
Shapiro, J. A. (2007): 'Bacteria are small but not stupid: cognition, natural genetic engineering and socio-bacteriology', *Studies in History and Philosophy of Biological and Biomedical Sciences* 38: 807–819.
Shettleworth, S. J. (1994): 'Biological approaches to the study of learning'. In: Mackintosh, N. J. (ed.): *Handbook of Perception and Cognition*, London: Academic Press.
Smith, L. B. & Breazeal, C. (2007): 'The dynamic lift of developmental process', *Developmental Science* 10: 61–68.
Spencer, J. P., Corbetta, D., Buchanan, P., Clearfield, M., Ulrich, B. & Schöner, G. (2006): 'Moving toward a Grand Theory of Development: In memory of Esther Thelen', *Child Development* 77: 1521–1538.
Sterelny, K. (2003): *Thought in a Hostile World: The Evolution of Human Cognition*, Oxford: Blackwell.
Sterelny, K. & Griffiths, P. E. (1999): *Sex and Death: An Introduction to the Philosophy of Biology*, Chicago: University of Chicago Press.
Stevenson, H. W. (1962): 'Piaget, Behavior Theory, and Intelligence', *Monographs of the Society for Research in Child Development* 27: 113–126.
Stopher, M. A., Marcus, E. A., Nolen, T. C. Rankin, C. H. & Carew, T. J. (1991): 'Learning and memory in Aplysia: A combined developmental and simple systems approach'. In: Shair, H. N. Barr G. A. & Myron, H. A. (eds.): *Developmental Psychobiology: New Methods and Changing Concepts*, Oxford: Oxford University Press.
Stotz, K. (2006): 'Molecular epigenesis: distributed specificity as a break in the Central Dogma', *History and Philosophy of the Life Sciences* 28: 527–544.
Stotz, K. (2008): 'The ingredients for a postgenomic synthesis of nature and nurture', *Philosophical Psychology* 21: 359–381.
Stotz, K. (2010): 'Human nature and cognitive-developmental niche construction', *Phenomenology and the Cognitive Sciences* 9: 483.
Subiaul, F., Cantlon, J. F., Holloway, R. L. Terrace, H. S. (2004): 'Cognitive Imitation in Rhesus Macaques', *Science* 305: 407–410.
Sweatt, J. D. (2009): 'Experience-Dependent Epigenetic Modifications in the Central Nervous System', *Biological Psychiatry* 65: 191–197.
Szyf, M., McGowan, P. O. & Meaney, M. J. (2008): 'The Social Environment and the Epigenome', *Environmental and Molecular Mutagenesis* 49: 46–60.
Tagkopoulos, I., Liu, Y. Tavazoie, S. (2008): 'Predictive behavior within microbial genetic networks', *Science* 320: 1313–1317.

Thelen, E. (1995): 'Time-scale dynamics and the development of an embodied cognition'. In: Port R. F. & van Gelder T. (eds.): *Mind as Motion: Explorations in the Dynamics of Cognition*, Cambridge (MA): MIT Press.

Thorndike, E. L. (1911): *Animal Intelligence*, Darien (CT): Hafner.

Timberlake, W. (2002): 'Niche-related learning in laboratory paradigms: the case of maze behavior in Norway rats', *Behavioural Brain Research 134* 134: 355–374.

Tomasello, M. (2000): *The Cultural Origins of Human Cognition*, Cambridge (MA): Harvard University Press.

Tomasello, M. Call, J. (2004): 'The role of humans in the cognitive development of apes revisited', *Animal Cognition*, 7:213–215.

Turkheimer, E., Goldsmith, H. H. & Gottesman, I. I. (1995): 'Commentary'' *Human Development* 38: 142–153.

Turkheimer, E. & Gottesman, I. I. (1991): 'Individual Differences and the Canalization of Human Behavior', *Developmental Psychology* 27: 18–22.

Waddington, C. H. (1942): 'Canalisation of development and the inheritance of acquired characters', *Nature* 150: 563–565.

Waters, C. K. (2007): 'Causes that make a difference', *Journal of Philosophy* CIV: 551–579.

West, M. J. (2003): 'The case for developmental ecology', *Animal Behaviour* 66: 617–622.

West, M. J. & King, A. P. (1987): 'Settling Nature and Nurture into an Ontogenetic Niche', *Developmental Psychobiology* 20: 549–562.

West, M. J. & King, A. P. (2008): 'Deconstructing innate illusions: Reflections on nature-nurture-niche from an unlikely source', *Philosophical Psychology* 21: 383–395.

West, M. J., King, A. P. Arberg, A. A. (1988): 'The Inheritance of Niches'. In: Blass, E. M. (ed.): *Handbook of Behavioral Neurobiology*: Plenum Press.

West, M. J., King, A. P. & Duff, M. A. (1990): 'Communicating about Communicating: When Innate Is Not Enough', *Developmental Psychobiology* 23: 585–598.

West-Eberhard, M. J. (2003): *Developmental Plasticity and Evolution*, Oxford: Oxford University Press.

Wheeler, M. & Clark, A. (2008): 'Culture, embodiment and genes: unravelling the triple helix', *Philosophical Transactions of the Royal Society B* 363: 3563–3575.

Wood, M. A., Hawk, J. D. & Abel, T. (2006): 'Combinatorial chromatin modifications and memory storage: A code for memory?' *Learning and Memory* 13: 221–244.

Zimmer, C. (2008): *Microcosm: E. coli and the New Science of Life*, New York: Pantheon Books.

Chapter 6
Re-Conceiving Nonhuman Animal Knowledge Through Contemporary Primate Cognitive Studies

Andrew Fenton

6.1 Introduction

A substantive epistemic subject has the capacity to (a) engage in activities of an epistemic nature (b) that are governed by rules or standards (henceforth referred to as epistemic activities), (c) adopted, or learned, by the individual in question and (d) held in common with her social group. The concept of a substantive epistemic subject arises from two distinct considerations jointly considered. On the one hand, the concept captures what it is to be a human epistemic subject, whose engagement in epistemic activities provides much of the material used in theorizing about knowledge, justified or rational belief. On the other hand, the concept allows that at least some animals[1] other than humans could also engage in epistemic activities relevant to theorizing about knowledge, justified or rational belief.

My conception of a substantive epistemic subject reflects the influence of a distinction, found in the literature on animal cognition, between active and passive knowing or active and passive cognition (see Gould and Gould 1994/99: 8, 87, 114, 120, 126). Gould and Gould describe this distinction as follows:

> Cognition can be innate – passive knowledge encoded in an animal's genes and used as instructions for wiring a nervous system to generate particular inborn abilities and specializations. Active cognition – the ongoing process of gathering, analyzing, and using knowledge – can incorporate several stages of mental processing beginning with sensation, which is the detection of stimuli by a sensory receptor organ and the subsequent processing of that sensory information by the brain. … It is the processing and analysis of sensory information that engenders knowledge, which can then be stored, recalled, and used in decision-making (Gould and Gould 1994/99: 8).

[1] In the discussion that follows I will adopt the locution 'animals' instead of the more cumbersome 'nonhuman animals.'

A. Fenton (✉)
Department of Philosophy, California State University-Fresno,
2380 East Keats Ave. M/S MB105, Fresno, CA 93740-8024
e-mail: andrew.fenton@gmail.com

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*,
Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_6,

The notion of active knowing present in the relevant literature amounts to the following. To be an active knower, an organism plays an important role in the acquisition of knowledge (it learns by manipulating/experimenting with its environment), and decides, though perhaps not consciously,[2] what information, among the knowledge already possessed, will be used in future behaviour (Gould and Gould 1994/99: 8, 114). To be a substantive epistemic subject on my account is to qualify as an active knower on Gould's and Gould's account.

The aforementioned analysis of what it is to be a substantive epistemic subject implicates, among other things, a capacity to gather (and use) evidence and the ability to achieve a degree of epistemic success. In this chapter I defend two claims that support the thesis that chimpanzees are substantive epistemic subjects. First, I defend the claim that chimpanzees are evidence gatherers (broadly construed to include the capacity to gather and use evidence). In the course of showing that this claim is probably true I will also show that, in being evidence gatherers, chimpanzees engage in a recognizable epistemic activity. Second, I defend the claim that chimpanzees achieve a degree of epistemic success while engaging in epistemic activity.

The *prima facie* implications of my claims that chimpanzees are evidence gatherers and enjoy a degree of epistemic success are modest—just as human knowledge plays an integral role in intentional human behaviour, so chimpanzee knowledge also plays an integral role in intentional chimpanzee behaviour. However, this way of seeing chimpanzees reveals a path for re-examining animal knowledge. Treatments of animal knowledge in the philosophical literature tend to go in one of two directions: They (i) embrace reliabilism and so construe animal knowledge as reliably produced true beliefs (or, if not beliefs, the relevant analogue for non- or pre-linguistic animals)[3] (see Goldman 1976; Kornblith 1999; Sosa 1991a; Steup 2003), or (ii) embrace an anthropocentric stance that treats animals as knowers only when they find themselves behaving in circumstances that, were it true of humans, would imply the presence of causally efficacious knowledge (see Davidson 1982; Russell 1948). Though reliabilism applied to animal knowledge comes in several forms (see Dretske 1989; Goldman 1988; 1989; Kornblith 2002; Sosa 1991a; 1991b), they share the view that knowledge need not involve metacognition, where a metacognition condition requires that an epistemic subject's reasons for believing something to be true are accessible to her as objects of thought to be explicitly related to her belief (or the relevant analogue for non-linguistic animals) as justifiers. Knowledge, on a reliabilist account, can result from reliable belief forming mechanisms as long

[2]I think it is safe to interpret Gould and Gould as not requiring any accompanying phenomenal consciousness when ascribing active cognition, though I may be wrong here—see Gould and Gould (1994/99), p. 70.

[3]To avoid using the caveat "or, if not beliefs, the relevant analogue for non- or pre-linguistic animals" whenever I use 'beliefs' to describe a sub-class of mental states possessed by non-linguistic animals, I will use 'beliefs*' in what follows to refer to either beliefs or, where appropriate, their analogues for non- or pre-linguistic animals.

as said mechanisms are sensitive to relevant negative feedback from the environment that indicates the inaccuracy of the relevant beliefs* (see Dretske (1989)). What I am calling here the anthropocentric stance requires a longer explanation.

Again, according to the anthropocentric stance to animal knowledge, animals are knowers only when they find themselves behaving in circumstances that, were it true of humans, would imply the presence of causally efficacious knowledge. In other words, in certain circumstances, animals act 'as if' from causally efficacious knowledge, where the exemplar is properly functioning, human adults. This anthropocentric stance can be understood as either realist or non-realist. I understand Bertrand Russell to be offering a realist approach in *Human Knowledge: Its Scope and Limits*. His discussion of a dog's knowledge (1948: 182-183, 428-429) resembles what I have in mind. Russell argues that

> [t]he expectations of animals, and of men except in rare scientific moments, are caused by experiences which a logician might take as premises for an induction. My dog, when I take out her leash, becomes excited in expectation of a walk. She behaves as if she reasoned: "Taking out the leash (A) has invariably, in my experience, been followed by a walk (B); therefore probably it will be so followed on this occasion." The dog, of course, goes through no such process of reasoning. But the dog is so constituted that if A has been frequently followed by B in her experience, and B is emotionally interesting, A causes her to expect B. (428-429)

Importantly, for Russell, any analysis of knowledge must recognize that the capacities that facilitate its emergence in human activities predate our species. This amounts to an appeal to evolutionary continuity as a constraint in theorizing the nature of knowledge, and commits Russell to seeing causally efficacious knowledge in the behaviour of animals other than humans (1948: 421).[4]

In contrast, Donald Davidson writes,

> [a]gainst the dependence of thought on language is the plain observation that we succeed in explaining and sometimes predicting, the behavior of languageless animals by attributing beliefs and desires and intentions to them. This method works for dogs and frogs much as it does for people. And, it may be added, we have no general and practical alternative framework for explaining animal behavior. (1982: 323)

Davidson goes on:

> But there would be a clear sense in which it would be wrong to conclude that dumb … animals have propositional attitudes. To see this it is only necessary to reflect that someone might easily have no better or alternative way of explaining the movements of a heat-seeking missile than to suppose the missile wanted to destroy an airplane and believed it could by moving in the way it was observed to move. This uninformed observer might be justified in attributing a desire and beliefs to the missile; but he would be wrong. (ibid.)

Davidson, then, allows that we ascribe knowledge to other animals analogically without granting the realist position.

It should be noted that neither of these accounts imply that the relevant nonhuman animals are substantive epistemic subjects as characterized above. It is difficult

[4] Russell's discussion of animal belief on pages 94-95, 99 of Russell (1948) is also relevant here.

to determine the view of knowledge informing its ascription to nonhuman animals in comparative psychology, ethology, and primatology. Kornblith has suggested that a reliabilist account of knowledge will capture the sense of knowledge assumed in these animal sciences (2002: 53-62). I suspect, however, that a more active cognitive account of nonhuman knowledge, one that presents many nonhuman animals as knowers on their own terms, as it were, better accords with what many comparative psychologists, ethologists, and primatologists are ascribing to their nonhuman subjects.[5] What I offer here is a way of understanding non-linguistic animals, in this case chimpanzees, as knowers in this more active sense.[6]

6.2 Terms of the Discussion

Before proceeding further I should clarify what I mean by evidence gathering and epistemic activity. For the purposes of my discussion, to be an evidence gatherer is to engage in, or be capable of engaging in, the collection (and use) of information about one's physical, social or phenomenological environment in ways that tend to produce representational states in one's noetic structure (or, though perhaps only for linguistic animals, one's belief system) that can then be used to assess the epistemic value (e.g. the truth or probable truth) of beliefs* that are already in one's noetic structure, or are at least being considered for inclusion (though not necessarily consciously considered). Minimally then, *evidence* is information *both* relevant to assessing the epistemic value (e.g. the truth, probable truth, or falsity) of beliefs* already, or potentially, in an individual's noetic structure *and* available to be so used by an evidence gatherer. My account of evidence is broad enough to include experience(s) and does not require meta-cognitive capacities (i.e., using new information to order, revise or reject beliefs one already holds need not involve meta-cognition).

[5] Kornblith uses some of Carolyn Ristau's work on the piping plover to try and show the applicability of his account (2002: 53-55). However, it is clear from Ristau's comments on the significance of her choice of cognitive vocabulary when explaining and describing the behaviour of her nonhuman animal subjects that (a) her subjects possess knowledge, and (b) it is reasonable to think this because they seem to be cognitively engaged with their environment. In other words, for Ristau, her subjects – understood as cognizers – are sufficiently sensitive and responsive to their environment to be, in some important sense, epistemic subjects (see Ristau 1991a: 93, 124; Ristau 1991b: 309-310).

[6] In using the word 'non-linguistic' it is not my intent to dismiss human language research using chimpanzees. Even given the successes in communicating with nonhuman great apes using symbol systems or American Sign Language, however, the majority of chimpanzees remain non-linguistic in that they lack a comprehension of, and ability to communicate using, a natural language or symbol system. Also, and more importantly, my account of being a substantive epistemic subject can be applied to animals who are even more clearly non-linguistic than chimpanzees. It is important, then, not to lose sight of my view that there are non-linguistic animals, chimpanzees among them, who can be appropriately regarded as substantive epistemic subjects.

My treatment of evidence, or by implication evidence gathering, may seem too liberal but we should pause and reflect upon what qualifies as the possession of evidence, or evidence gathering capacity, among *human* conspecifics who are quintessential evidence gatherers and users. Of course, we want examples of quintessential human evidence gathers whose requisite cognitive capacities are reasonably ascribed to such animals as chimpanzees. Consequently, consider the evidence gathering capacity of young human children.[7] As human children play with objects in their environments (e.g. striking two toys together or fitting them into various boxes/containers or dropping them in water), they are in effect gathering information about the objects. This information serves as grounds for future responses to, or inferences about, relevantly similar objects in their environment, even affecting what future *information* is taken to be relevant in responding, or making inferences useful, to a task at hand ((Langer 1996; Santrock 2001: 257-260). Note that this kind of behaviour, though reasonably regarded as evidence gathering, does not require a degree of cognitive sophistication that it is unreasonable to ascribe to chimpanzees (see Chapters 3 and 4 of Gómez (2004)).

This sense of evidence and evidence gathering resonates, though to different degrees, with how Laurence Bonjour and Matthias Steup, to name just two examples, seem to understand them (see Bonjour 2002: 39-43 and Steup 2003: 313-314). Steup is clear, however, that evidence gathering and use involves metacognition (as I described metacognition earlier), at least if it is to be epistemically significant (Steup 2003: 314). Though less explicit on this point (see Bonjour 2002: 41, 224-226), Bonjour probably differs with Steup on the importance of metacognition. By his own admission, (i) it is reasonable to suppose that many humans, including children, possess knowledge or justified beliefs and (ii) this is acquired without engaging in metacognition (Bonjour 2002: 225, 226).[8] Robert Audi also does not think it is plausible to hold that metacognition is necessary for evidence gathering or use. Interestingly, Audi's rejection of what he calls second-order internalism[9] – nicely exemplified by Steup – is at least partially based upon the plausibility of talking of the justified beliefs of young humans who have as yet to develop extensive conceptual

[7] I am not suggesting that chimpanzee cognition compares with the developmental level of properly functioning human children. For example, it is obvious that many adult chimpanzees enjoy a degree of independence or self-sufficiency absent in many children. Rather, I wish to find examples among humans of behaviour and cognitive capacities that would not be regarded as 'too sophisticated' to be ascribed to chimpanzees.

[8] Nicholas Rescher is another epistemologist whose understanding of evidence gathering clearly *requires* metacognition (2001: 14-16, 19-20).

[9] Basically, epistemological internalism requires that justifiers for an epistemic subject's belief are accessible to her and can be explicitly related by the epistemic subject in such a way as to ground the judgment that the belief is true or probably true (Steup 2003: 310).

frameworks (see Audi 1989: 309, 311). As even internalist epistemologists,[10] who tend to be the more conservative of contemporary epistemologists, are not in total agreement about whether evidence use *requires* meta-cognition, my treatment here does not require it.

Epistemic activity, on my account, is any cognitive activity (e.g., evidence gathering) that results in beliefs* that, due to this activity, have varying degrees of positive epistemic status. *Minimally*, this involves the processing of information, ranking the resulting beliefs* using values of an epistemic nature relative to the individual's continuing environmental feedback, and manipulating these resulting beliefs* in ways that affect the individual's future behaviour. On my account, epistemic activity neither requires metacognitive capacity nor does it implicate phenomenally conscious states, though it does implicate a to-be-specified degree of sensitivity and responsiveness to environmental feedback.[11]

6.3 On Chimpanzee Hunters (of Knowledge) and (Evidence) Gatherers

The claim that chimpanzees engage, with *some* degree of sensitivity and responsiveness, in activities which can be appropriately described as gathering evidence has a degree of *prima facie* plausibility, and for the following reasons. First, chimpanzees begin life lacking many of those skills that will, as they mature, be needed to find nourishment, protect themselves from the aggressive behaviour of conspecifics, find mates, and so on.[12] Young chimpanzees will acquire some of these skills while observing the behaviour of older conspecifics, including their mothers (Gómez 2004: 18-19, Hauser 2000: 35, 135-136; Russon 1997: 175, 184-185). To accomplish this in the context of tool use, these young apes attend to the activities of others around them, and not only respond to the relevant stimuli, which itself will probably reflect innate dispositions to find certain stimuli attractive, but combine certain objects in ways that resemble what they have just observed (Hauser 2000: 135; Hirata 2009: 5; Matsuzawa 1996: 201-203; Matsuzawa and Yamakoshi 1996: 215, 217, 226-229; Parker 1996: 351, 352-355). Think here of very young chimpanzees who will re-insert a discarded probe into a termite nest after the mother has finished

[10] Steup, Bonjour and Audi are all properly regarded as epistemological internalists. The judgment that epistemological internalists are the more conservative of contemporary epistemologist is, of course, a comparative claim.

[11] In the philosophical literature, the sensitivity and responsiveness of animals to environmental feedback figures in contexts related to this one. See Allen (1999) concerning responsiveness to error; Kornblith (2004) concerning responsiveness to counterevidence; Saidel (1998) concerning responsiveness to a failure to achieve a goal.

[12] This is generally true of nonhuman primates (Strier 2000: 255-256, 263, 266-271).

feeding at that particular site.[13] To acquire some of these skills in the context of social interactions, these young apes learn, among other things, which behaviours precede, or tend to precede, aggressive activity and which do not, which chimpanzees are more dominant than others, which male chimpanzee is the most dominant, and which individuals are a part of the 'range community' and which are not (de Waal 1987: 421-429; de Waal and Aureli 1996: 86-87, 88-89; Fruth et al. 1999: 66-67, 69; McGrew 2004: 131, 157-159; Nishida and Hiraiwa-Hasegawa 1987: 167-172, 174-176).[14] These features of their social environment are not fixed, and so a degree of sensitivity and responsiveness to, say, changes in the social hierarchy are required if they are to successfully navigate this environment.

Second, chimpanzees, as well as bonobos, have demonstrated a remarkable ability to acquire proto-linguistic, or *perhaps* weak linguistic, skills within artificial settings (Fouts and Fouts 1999: 252-255; Gómez 2004: 277-291; Greenfield and Savage-Rumbaugh 1990/94: 541-574). As examples consider two chimpanzees in 'language' research: Loulis' ability to sign to other chimpanzees or human attendants (Fouts and Fouts 1999: 253-254, 255) or Ai's ability to reliably respond (i.e., consistently respond above the level of chance) to various lexigrams (symbols) or Japanese *kanji* (Matsuzawa 2002: 191-195). Loulis's case is interesting, not only because of his communicative skills, but because he developed these skills primarily through his relationship with one or more conspecifics. For five years (beginning when Loulis joined the study), human researchers and caregivers were restricted to seven signs in American Sign Language (ASL) when signing in the presence of Loulis. This restriction was to test the hypothesis that chimpanzees trained in ASL could transmit their knowledge of ASL to a conspecific. Four other chimpanzees (including the well known 'language ape' Washoe), all trained in ASL, interacted with Loulis during this time. Over a period of 73 months, Loulis acquired a vocabulary of 51 signs that he could reliably use to communicate (Fouts, Jensvold and Fouts 2002: 288).[15] Ai is a part of a 14 member chimpanzee group in the Primate Research Institute at Inuyama, Japan (Matsuzawa 2002: 191). Born in 1976, Ai joined the Primate Research Institute in late 1977.[16] By the age of five, Ai had been

[13] There are videos associated with Sanz et al. (2004) that can be viewed when accessing it through *The American Naturalist* online. Video 1, titled "Chimpanzees Approaching Nest", appears to show a young chimpanzee copying the behaviour of his mother as she forages for termites (see http://www.journals.uchicago.edu/doi/full/10.1086/424803).

[14] A very general description of the kinds of social knowledge developed by individual nonhuman primates can be found in Ray (1999) or Chapter 7 of Tomasello and Call (1997).

[15] The implication of this study is that Loulis acquired these additional signs from his chimpanzee companions. Video recordings of these chimpanzees suggest that they use their knowledge of ASL in interactions with each other. They reliably use signs to initiate play (e.g., the sign for chase would reliably precede bouts of chasing behaviour), request objects or seek bodily contact (e.g., request grooming) (Fouts and Fouts 1999: 254; Fouts, Jensvold and Fouts 2002: 286-288).

[16] For a limited biography of Ai at the Primate Research Institute see http://www.pri.kyoto-u.ac.jp/ai/friends/indexE.html (accessed on May 8, 2010).

trained to match lexigrams to 11 colors as well as 14 objects (Matsuzawa 1985: 57). In a study to test Ai's numerical competence she was trained to count from 1 to 5 through trials that displayed colored objects with which she had been previously trained. By the final trials Ai was able to reliably identify the color, object and number of 125 sample items (Matsuzawa 1985). When these skills have not been moulded (as with Loulis), the relevant animals seem to have acquired the skills through observation and *perhaps* imitation (Savage-Rumbaugh and Lewin 1994: 135-142; Matsuzawa 2002: 192, 194).

Taken together, these facts about chimpanzees suggest that they are evidence gatherers. A closer examination of these facts about chimpanzees, then, is warranted. Several points bear mention before delving deeper, however. (i) A sensitivity and responsiveness to environmental feedback is an important part of efficient learning (Saidel 1998: 1-8). (ii) The learning that is of interest to me here need not involve imitation, or what psychologists call 'insight' (Byrne 1995: 45-48). Even instrumental learning can be epistemically significant, though perhaps only if the relevant organism remains sensitive or responsive to their environment after having learned certain behaviour (Byrne 1995: 56-62). (iii) When information from environmental feedback positively or negatively affects the status of information *already* stored in an animal's central nervous system (i.e., the information states already possessed by the relevant animal), this *newly acquired* information arguably qualifies as evidence (or plays an evidentiary role). This may seem to be too loose a sense of evidence, or by implication evidence gathering, but think back to the earlier example of children playing with objects (e.g. striking two toys together, fitting them into various boxes/containers, or dropping them in water). As I suggested earlier, children playing with objects are in effect gathering information about them, or their relations with other objects (Crain 1992: 173-174, 322-323; Tomasello and Call 1997: 59, 68-71, 97). It is evidence gathering, so observed in children, that informs my analysis here.

Let us now return to some of the facts about chimpanzees I listed earlier. Consider a common tool-using activity among wild chimpanzees—termite fishing. (1) Chimpanzees who forage for termites in termite nests typically do not do so year round, their foraging behaviour is correlated with the seasonal activities of termites (see, for example, Goodall 1988/97: 74-75). Here we see a *hint* of selective behaviour, though it is not sufficient to suggest that this behaviour is not driven by environmental contingencies. (2) That this foraging behaviour is not simply an expression of a set behavioural pattern or a predisposed response to a particular stimulus is strongly suggested by the facts that (i) not all chimpanzees – even from the same sub-species in similar ecological conditions – will hunt termites and (ii) not all chimpanzees – even from the same sub-species in similar ecological conditions – hunt the *same species* of termite (Matsuzawa and Yamakoshi 1996: 219; McGrew 1994/96: 30-31; McGrew 2004: 113; Sanz et al. 2004: 567-568). (3) Importantly, before beginning to forage at a nest, a chimpanzee will first *investigate* the level of its activity. She does this by disturbing the nest structure and *observing* the reaction of the resident termites. Enough activity will incline her to dip a grass blade or thin twig – denuded of protruding leaves – into the nest (Sanz et al. 2004: 574). (4) What

community this chimpanzee belongs to is a relatively reliable indicator of what material substrate she will use for termite fishing (McGrew 2004: 111-113) and how she removes the termites from the probe is a weak indicator of how conspecifics around her have done this in the past (McGrew 1994/96: 31-32). (5) The chimpanzee infant typically spends a significant part of the waking day clinging to the body of her mother. Often attentive to what is happening around her, the infant seems to at least sometimes watch the mother foraging for termites, including her preparation of the probe and how she removes the termites upon extracting the probe from the nest. As the infant matures, becoming physically mobile and moving about in the vicinity of the mother, she will probably pick up a discarded probe and, with enough time taken in the past to exploring such an object's features, begin to insert it into holes left by the mother's foraging (Lonsdorf 2006: 36-37, 42-43).[17]

As the infant learns the termite fishing technique, either by watching conspecifics or exploring the nest with a discarded probe, she processes a good deal of information about her own body, the termite nest structure, termites, probes, how to extract a probe without losing a lot of termites and how to extract the termites without getting bitten (Byrne 2004: 36; Yamakoshi 2004: 163-164). This information processing, it is reasonable to suppose, yields, among other things, a to-be-specified number of information, affective and conative states that will have an effect on the future behaviour of this maturing ape. It is also reasonable to suppose that, as the infant matures, new information obtained in play or 'practice' will inform the direction the infant takes in manipulating objects in her environment, even inclining her to adopt new ways of accomplishing old tasks (e.g. new ways of holding twigs, better ways to prepare the probe for insertion into a termite nest, how to insert the probe into a nest and so on). Here evidence gathering and use, as I characterized it above, seems to be at work early on in a chimpanzee's life.

Consider further some chimpanzee stone tool use. In certain parts of West Africa, some of the members of *Pan troglodytes verus* will forage for nuts using hammers and anvils to break open the casing of oil palm, coula or panda nuts (Matsuzawa 1994/96: 353; McGrew 2004: 118-120). Anvils will be any hard surface (e.g. rock, tree root or tree stump) that can both hold the nut and provide resistance to the force of the hammer used by the chimpanzee. Hammers are typically rocks used to strike, and break open, the nut casing (Matsuzawa 1994/96: 356-360; McGrew 1994/96: 35; McGrew 2004: 118). To explain this behaviour we need to posit causally efficacious information, affective and conative states—as I will illustrate shortly, no other explanations seem adequate to the task. Young chimpanzees learn to successfully use stone tools between the ages of three and five, but it takes "almost ten years to acquire the refined level of skill shown by adults" (Matsuzawa 1994/96: 367). Clearly, this is a case of learned behaviour, rather than the result of a fixed action pattern or even the combination of fixed actions as a conditioned response to the right physical stimulus. Not all chimpanzees use stone (or wood) tools in this way,

[17] Again see the videos associated with Sanz et al. (2004) which can be viewed when accessing it through *The American Naturalist* online.

only the subspecies *Pan troglodytes verus* (in West Africa) (McGrew 1994/96: 33), and not all members of the subspecies *Pan troglodytes verus* engage in nut cracking behaviour (McGrew 1994/96: 30). This behaviour is not ecologically determined. The rocks (or wood) and nuts are available in habitats frequented by at least one of the other subspecies of chimpanzee (e.g. *Pan troglodytes troglodytes*) (McGrew 1994/96: 35). It *seems* to be a pattern of behaviour that chimpanzees can learn to apply through the example of others. A female chimpanzee (named Yo), in a community of chimpanzees who did not break open coula nuts,[18] immediately did so when a study area watched by a group of primatologists was seeded with coula and oil-palm nuts (Matsuzawa 1994/96: 364). The other adults of this community, who witnessed Yo crack open the coula nuts and eat the kernels, showed little interest in doing the same (Matsuzawa 1994/96: 364; Matsuzawa 1996: 202). Some of the younger chimpanzees, however, gathered around to watch Yo break the coula nuts' casing and consume the kernels. In the days that followed two of these juveniles copied Yo's behaviour, cracking open the coula nuts, obtaining the nut's kernel and tasting it (Matsuzawa 1994/96: 364-365, 367; Matsuzawa 1996: 202). Note that the adults in the group did not begin to mimic the female in question (Matsuzawa 1994/96: 364, 367; Matsuzawa 1996: 203). So, whatever the source of this behaviour, it does not arise as a result of mere stimulus enhancement.

Also take note that Yo did not learn this behaviour in the group of which she was now a member, nor was she disposed to break open any nut or nut-like object encountered in a feeding area. A year after the aforementioned experiment was conducted, an area frequented by this group of chimpanzees was seeded with wooden balls that resembled coula nuts in both shape and size. Yo, though not the aforementioned curious juveniles, ignored these wooden balls (Matsuzawa 1996: 202). It would appear, then, that this chimpanzee possessed information about particular nuts that were not normally in her environment and, when the opportunity arose, used this information to obtain some food. Just in these two incidents alone we have the presence of causally efficacious information, affective and conative states that contribute to Yo's foraging and which are selectively used to accomplish this.

Once more, evidence gathering is evident in this type of behaviour. In Yo's case, she is sensitive to certain features of various small nut-like objects in her surrounding environment. Before using a stone to break a small nut-like object, that object must relevantly resemble nuts she has broken open in the past. Arguably, Yo is using already stored information (i.e. memories of some past experience), comparing it to information recently received from her senses and then using a positive correlation as evidence that an edible object is in her field of vision. None of this need happen at the level of awareness, nor need it be realized as a syllogism, to qualify as evidence gathering or use. It is this kind of evidence gathering and use that is surely the more prevalent form at work in human daily affairs.

[18] Members of the community in Bossou of which she was a part did crack open nuts, but only oil-palm nuts (Matsuzawa 1994/96: 364).

As indicated above, it takes chimpanzees almost ten years to acquire the nut-cracking skill of experienced adults (Matsuzawa 1994/96: 367). Matsuzawa has noted that there are at least three developmental stages in a young chimpanzee's ability to forage for nuts using stone tools. He writes,

> First is the action manipulating a single object, such as a nut or a stone ….Second is the action of relating two objects; a nut and a stone, or a stone and another stone. Third is coordinating the multiple actions of manipulated objects. (1996: 201)

As the young chimpanzee matures, she can be observed first playing with individual nuts or stones, or taking a kernel for consumption from off of her mother's anvil after her mother has broken open a nut's casing. After a time, she begins rolling a nut off of her mother's anvil or pushing one stone against another. She might even try hitting the nut with her hand while the nut is either on the ground or is sitting on a stone, clearly emulating the behaviour of older chimpanzees around her. She might, instead, strike a nut against a root, trunk or stone. After a time, she will begin to strike the nut with a stone, and learn to place the nut onto a stone or other hard substrate before she strikes it (Matsuzawa 1994/96: 356-359).[19] Again, all of this behaviour requires a to-be-specified amount of information processing, including the integration of new information over time about individual objects, relations between objects, and her own body relevant to developing the skills required for breaking open nut casings. This all seems to relevantly resemble what I described earlier when talking about the evidence gathering activities of young humans. Young chimpanzees appear to be evidence gatherers. Coupled with the reasonable suspicion that these young apes also possess a to-be-specified number of information states which inform, in conjunction with various affective or conative states, their interactions with nuts, stones or other material substrates, we can reasonably hold that these young chimpanzees already resemble epistemic subjects.

I mentioned earlier that chimpanzees must learn various social skills if they are to successfully navigate their social environments. Within the context of their social interactions there are suggestions of evidence gathering. One common 'practice' among chimpanzees who have been victims of recent aggression is to insert a finger into the mouth of the one who behaved aggressively, typically the more dominant chimpanzee (de Waal 1990/96: 80). This is a risky behaviour. Chimpanzees have been known to bite off digits, or worse, in moments of aggression (de Waal 1990/96: 60, 80). How is the behaviour to be construed? It seems to play an evidentiary role in revealing the present disposition of the relevant conspecific. A positive response

[19] The reader should not be misled by the play behaviour through which the aforementioned young chimpanzees develop their increasingly complex interaction with stones and nuts. Play can be an important way in which young animals acquire information and skills that are needed as they mature to adulthood (Manning and Dawkins 1998: 84-88). This is not to argue that play behaviour has this primary role, nor is such a primary role necessary for my discussion. The play of these chimpanzees, as described by experienced primatologists like Matsuzawa, clearly involves increasingly complex relations between the chimpanzee, nuts and stones. Allen and Bekoff provide an interesting discussion of the possible roles of play behaviour (Allen and Bekoff 1997: 108-112).

to the finger insertion leads to a relaxing of the victim, with grooming often ensuing (de Waal 1990/96: 40-41, 43, 80). Arguably, the positive response is taken as evidence that the aggressor is not going to behave aggressively for the time being, or something to that effect.[20]

A second area, rich in suggestive examples of evidence gathering in a social context, concerns the acquisition and use of information about chimpanzee social hierarchy. As I mentioned earlier, the social hierarchy within chimpanzee groups is flexible—something that is not uncommon among primates (including, of course, humans) (de Waal 1994/96: 248; McGrew 2004: 157-159). Among the males, one chimpanzee enjoys alpha status, typically giving him, among other things, first access to common food, a good deal of uninterrupted access to sexually receptive females, and a certain 'license' to express himself aggressively to conspecifics within the group (i.e. aggressive behaviour will not typically result in *retaliation* from others within the group) (McGrew 2004: 157). This status is not achieved or maintained on brute strength alone, so it is not always the strongest or biggest chimpanzee male that 'ascends' to alpha status. It is not uncommon to find (more longer term) alliances or (shorter term) coalitions[21] that maintain a male's dominance over the group (de Waal 1990/96: 49, 50-51; McGrew 2004: 157-159). Evidence of a male's dominance resides, at least in part, in the periodic repetition of submissive behaviour of others within the group. A male who approaches a more dominant male will typically exhibit submissive behaviour. This consists of rather stereotyped behaviour, including a relatively low approach to the more dominant male and the vocalization of certain sounds christened "submissive grunts" (de Waal 1990/96: 44-45, 52-53). Such behaviour reveals the relative status of two interacting chimpanzees, and other chimpanzees observing this behaviour seem attuned to its significance. Changes in the social hierarchy (e.g. the fall in status of one male and the rise of another) can be evidenced by the change in the frequency of submissive behaviour between previously dominant and subordinate chimpanzees and the rise of behaviour among conspecifics that is uncharacteristic of the past hierarchy—e.g., approaching sexually receptive females despite the agitation,

[20] This is risky behaviour (and the interpretation might elicit scepticism in my readers), but it is not uncommon. De Waal puts it this way: "Chimpanzees have a habit of putting their fingers or the back of one hand between the teeth of dominant group members. A friendly gesture, it is *also a test* of the dominant's state of arousal and often is used in ambiguous situations. … [I]n the Arnhem colony I have seen quite a few instances when fingers were not treated … gently during appeasement attempts. Young chimpanzees of three years or less, who may have *lacked the experience to judge* whether the gesture was safe or not, were almost always the victims of … bites" (de Waal 1990/96: 80 [emphasis mine]). I have highlighted de Waal's choice of words where they seem to enjoy epistemic significance.

[21] Coalitions are described as "two or more individuals joining forces against one or more conspecific rivals" (Nishida and Hosaka 1996: 114). Alliances are coalitions that survive for a lengthy period of time within a given community (though the amount of time required for a coalition to qualify as an alliance is, as far as I know, unspecified) (Nishida and Hosaka 1996: 114). Coalitions seem to be contrasted with alliances both because of their brevity of existence and opportunistic character (Nishida and Hosaka 1996: 114).

or aggressive responses, of the 'current' alpha male, or more straightforward aggressive behaviour directed towards the 'current' alpha male (see de Waal 1990/96: 50, 52, 57-61, 63-69). Young and old alike, in order to avoid becoming victims of aggression, must learn the social significance of such behavioural changes or expressions of submission.

It is reasonable to suppose that a chimpanzee who observes such behavioural changes, or expressions of submission, is storing information about the social hierarchy of the group that can be used in future behaviour. This stored information will consist of a to-be-specified number of information states which, in conjunction with various affective or conative states, can incline an individual to behave submissively or aggressively when approaching a particular conspecific in possession of some food or pursuing a sexually receptive female. The pay-off will be the avoidance of personal injury – or the continuation of a relatively peaceful day – or the continued possession of, or access to, various resources or conspecifics (Tomasello and Call 1997: 194-195, 196-197, 202-203).

What is more, the relevant information states concerning the dominance ranking within the relevant group will have to change over time, and sometimes very quickly, to keep up with the changes in social hierarchy. A chimpanzee that is too inattentive may find himself on the 'wrong side' of a fight over, say, a common food source. Past experience being the victim of aggressive behaviour by an 'up and coming' male no doubt 'teaches' chimpanzees to stay attuned to such changing interactions within the group (Tomasello and Call 1997: 194, 205, 207, 208-209). Once again, there is good reason to think that chimpanzees are evidence gatherers and with a, not insignificant, degree of sensitivity or responsiveness to changing circumstances in their environment.

The other examples with which I began this section can all receive the kind of analysis I just gave, but I do not think that this is necessary to defend the claim that chimpanzees are evidence gatherers. When all is said and done, there are good grounds for believing it to be true.

6.4 Knowing Success

Arguably, the most fertile ground for finding clear and strong evidence of epistemic success is skilled behaviour. It is reasonable to think that skilled behaviour consists of (i) *coordinated* (ii) *goal-directed* behaviour that an organism has (iii) *learned* during its ontogeny, that (iv) requires a *non-haphazard application of past experience* in (v) successfully *achieving a desired end*, and (vi) involves *ends that are themselves selected by the organism*[22] in question (vii) based upon its past

[22] Once again, these do not have to be consciously chosen nor do the ends need to be non-species specific or in some important sense idiosyncratic. That is to say, even ends that arise out of what an animal is predisposed to find salient will qualify as ends selected by this animal in the relevant way.

experience and preferences.[23] This analysis of skilled behaviour distinguishes it from the mere expression of genetic predispositions of the kind encountered in the behaviour of digger wasps (Gould and Gould 1994/99: 39-43) or sphex wasps (Dennett 1984: 11) *without excluding* associative or instrumental learning as a component of skilled behaviour—learning that we even see in some of the skilled behaviour of humans (Crain 1992: 165).

For the sake of brevity I will focus on the example of chimpanzee *stone tool* use discussed in the previous section (though what I have already discussed in that section implies both skilled behaviour and epistemic success). Several features of this activity are worth highlighting. (1) Chimpanzee nut-cracking behaviour is learned (Matsuzawa 1994/96: 356-359). (2) It requires the presence of causally efficacious information states about the relevant species of nut, the utility of the relevant tools for the task at hand, and the desirability of a certain end (e.g. the acquisition of the relevant nut kernel) (see Matsuzawa 1996: 202-203). (3) These information states enjoy a certain prominence in the individual's noetic structure in the relevant foraging context (after all, they, rather than competing information states about other sources of nourishment, inform the behaviour of the foraging chimpanzee in a 'nut-cracking context'). (4) These information states enjoy their aforementioned prominence in the relevant chimpanzee's noetic structure in the face of ongoing feedback from that chimpanzee's physical environment.

The behaviour of Yo and some of the juveniles in her group, mentioned in the previous section, seem to clinch the matter. Remember, of the adults in her group, only Yo immediately placed a seeded coula nut on an anvil, broke open its shell, retrieved the kernel and consumed it. Two juveniles watched her behaviour, and in the days that followed were observed successfully retrieving a coula nut kernel from each of the nuts they cracked, though they initially spat them out after only briefly tasting them (Matsuzawa 1996: 202).

What does this set of observations show? First, Yo seems to have possessed information states with content identifying coula nuts as food that contain an edible core. This is suggested not just by her apparently lone appreciation of a coula nut as something that could be broken open, but her eagerness to eat the kernel—something the younger chimpanzees were not initially willing to do (presumably because of the difference in taste from the oil-palm nuts) (Matsuzawa 1996: 202). Second, knowledge, or something akin to it, can be '*transmitted*' from one generation to the next.[24] This is not only relevant to the chimpanzee culture debate (see de Waal 2001: 227-229;

[23] Arguably something like this notion of skilled behaviour underlies James and Carol Gould's discussions of learning and insight (see Gould and Gould 1999: 65-67, 68-87, 100-113).

[24] Note that I need no other learning mechanisms at work here than stimulus enhancement and instrumental learning. Even if these, and not more social learning, mechanisms best explain how the juveniles began to acquire the skills associated with cracking open coula nuts, they still acquired knowledge (or something akin to it) of the edibility of coula nuts similar to the knowledge (or something akin to it) possessed by Yo, and only learned of this property of coula nuts from observing Yo's foraging behaviour.

Gómez 2004: 249-265; McGrew 2001: 248 for examples), but is relevant to analytic epistemologists interested in the history or scope of social knowledge (see Longino (2002); Schmitt (1994)). Third, it suggests that at least some chimpanzees are sensitive to the information possessed by others.[25] Here, then, we seem to see acquired information affecting the behaviour of chimpanzees, and within a context of action requiring skilled behaviour.

Did Yo also engage in epistemic – and not 'merely' evaluative – activity using epistemic standards she had adopted? As I stated in (4) above, these information states enjoy prominence in the relevant chimpanzee's noetic structure in the face of ongoing feedback from that chimpanzee's physical environment. Each time Yo engages in nut-cracking behaviour she receives further reinforcement from her success. In other words, the relevant, causally efficacious information states receive ongoing positive feedback when Yo succeeds in obtaining an edible kernel from breaking open the relevant nut. Presumably, this means that Yo is more inclined to use these information states in relevantly similar circumstances in the future. These facts about Yo's nut-cracking behaviour, and the continuing prominence of certain information states conducive to this behaviour, speaks to the accuracy of the relevant information states. As accuracy is a straightforwardly epistemic value, there is an epistemic value at work in the cognitive activity required for Yo to break open nut casings.

We can see evidence of a contrary instance of information states that lack this degree of accuracy in the behaviour of the juveniles who had copied Yo in breaking the coula nut casings. As I briefly mentioned in the previous section, a year after the aforementioned experiment was conducted, an area frequented by this group of chimpanzees was seeded with wooden balls that resembled coula nuts in both shape and size. Yo, though not the aforementioned curious juveniles, ignored these wooden balls (Matsuzawa 1996: 202). Matsuzawa writes,

> The youngsters … seemed ready to crack any objects resembling edible nuts even if the objects were unfamiliar. Their attempts to crack open wooden balls may reveal an abiding tendency to try to crack open unfamiliar nut-like objects which was facilitated by their observing Yo's cracking new nuts in the last year. (1996: 202)

Interestingly, these juveniles appeared to possess causally efficacious information states that, unlike Yo's, lacked a certain accuracy. Perhaps better yet, these youngsters possessed rules of action that allowed information states with a degree of inaccuracy to enjoy a prominence in their respective noetic structures while engaging in nut-cracking behaviour. Presumably, this was registered by the juveniles upon receiving negative feedback from their attempts to break open the wooden balls.

Important to my point here is that accuracy of the relevant, causally efficacious information states is important to the success of these chimpanzees, and that at least

[25] Call and Tomasello (2008) provide a brief but useful overview of available evidence that chimpanzees track the knowledge of conspecifics.

some of these animals favour accurate information states over time and through various circumstances. In effect, these chimpanzees are tracking the truth or falsity of said information states. At any rate, accuracy is a value clearly at work in this kind of behaviour, at least some of the time. Since it is clearly an epistemic value, the importance of accuracy to the ongoing activities of chimpanzees evinces (i) the existence of chimpanzee epistemic activities and (ii) information states that meet the epistemic standards (at least concerning accuracy) adopted by these chimpanzees themselves. Consequently, this example of chimpanzee skilled behaviour supports the claim that chimpanzees can, and sometimes do, achieve a degree of epistemic success while engaging in epistemic activity.

To sum up this section, I have provided an example of skilled chimpanzee behaviour that (a) suggests or implies that these animals engage in epistemic activities, and (b) these activities track the accuracy of the relevant information states that inform the subsequent skilled behaviour. If this is right, I have shown not only that chimpanzees are evidence gatherers, but that they can achieve a degree of epistemic success while engaging in epistemic activity.

6.5 On Why this Matters

The importance of these observations partially resides in their implications for both future work in chimpanzee cognitive studies and naturalized epistemology. There are enough data on chimpanzee cognition and behaviour for naturalized epistemologists to now begin to develop analyses of knowledge geared toward primatologists. More importantly, these analyses can reflect the active cognitive activities of chimpanzees. This offers primatologists a way of moving beyond metaphor or perhaps even analogy, and ascribing knowledge to chimpanzees that is, in many ways, relevantly similar to what we ascribe to ourselves. By recognizing chimpanzees as substantive epistemic subjects, and recognizing in at least some of their behaviour epistemic activities, we deepen the picture of what it means for animals to be actively cognitively engaged with their physical or social environments. This also deepens our shared understanding of epistemic subjectivity and offers a way of exploring its evolutionary history.

What I offer here is a corrective to both contemporary reliabilism and the anthropocentric stance mentioned earlier in understanding animals like chimpanzees as knowers in a philosophically significant sense.[26] I have argued that these animals

[26] I do not mean to imply that regarding chimpanzees as substantive epistemic subjects will take us far afield from epistemological reliabilsm. I would agree that knowledge, and positive epistemic status more generally (e.g., justified, rational or warranted belief [or their analogues for non- and pre-linguistic animals]), is intimately connected with reliably produced true belief (or its analogue for non- and pre-linguistic animals). As I suggest in this section, reliabilists must take greater care to provide epistemological analyses that accommodate and, in some important sense (and at some level of description), reflect the epistemic standards of all substantive epistemic subjects. This means working harder than we have to understand and then incorporate the epistemic activities and perspectives of animals like chimpanzees into universal analyses of knowledge (and positive status more generally).

engage in epistemic activities: that is, goal-directed activities governed by rules evincing values (and goals) of an epistemic nature. These activities, and the relevant values, ought to figure in future naturalistic analyses of knowledge or, perhaps, other forms of positive epistemic status. This claim largely arises from considerations of method in analytic epistemology.

In developing a theory of knowledge, epistemologists adopt one of three approaches: a top-down, bottom-up or hybrid approach. A top-down approach consists of positing an analysis (including conditions) of knowledge, and by implication of what it is to be a substantive epistemic subject, that is then tested against *prima facie* cases or instances of knowledge (and the capacities of epistemic subjectivity required to acquire such knowledge). In contrast, a bottom-up approach involves gathering together an extensive pool of *prima facie* cases or instances of knowledge, as well as a contrasting set of non-knowledge, from which an analysis of knowledge (and the capacities of epistemic subjectivity required to acquire such knowledge) can be gleaned. A hybrid approach, exemplified by seeking a reflective equilibrium between conditions and particular cases or instances of knowledge, possesses elements of both a top-down and bottom-up approach (see Chisholm 1973, pp. 12-15). Neither the top-down nor the bottom-up approach can be purely top or bottom. Top-down theorists have their intuitions about what cases of putative knowledge are clearly knowledge, and these intuitions inform the analysis they proffer. Bottom-up theorists have their intuitions about what conditions must be met for a case of putative knowledge to qualify as knowledge, and these intuitions inform the cases they pick out as paradigmatic (Chisholm 1973, pp. 9-11, 12-21). Of course, those in the middle (e.g. advocates of reflective equilibrium) are even more sensitive to the dynamic between epistemological theory and, for want of a better term, epistemic data (Cohen 1991, pp. 185-88). *If* we are seeking a universal theory of knowledge, quite irrespective of whether we are top-down theorists, bottom-up theorists, or advocates of reflective equilibrium, we will want to attend to those cases of putative knowledge taken to be knowledge by others than ourselves (or our belief communities). We will want to attend to their conceptions of knowledge or the epistemic standards *they* use in their epistemic activities. Ensuring that the instances of knowledge we use in theorizing about it reflect a diversity of activities and standards minimizes the mistake of highlighting capacities that are, upon reflection, unnecessary for knowledge (e.g., meta-cognitive capacities). If we should attend to diversity of cognitive practice, epistemic standards and even conceptions of knowledge when developing adequate universal theories of knowledge, then we ought to avail ourselves of the epistemic perspectives of a representative sample of epistemic subjects. If, as I have argued, chimpanzees are substantive epistemic subjects, epistemologists should not ignore their epistemic perspectives.

Anthropocentric approaches to animal knowledge ignore the epistemic activities and implicit epistemic values of nonhuman substantive epistemic subjects by virtue of justifying ascriptions of animal knowledge through analogy to instances of human knowledge in relevantly similar circumstances. As Russell illustrates, realists adopting this approach do not deny that many other animals are epistemic subjects, but their 'epistemic citizenry' is of a secondary nature (or a poorer cousin to what we find

among humans). The data, including epistemic values, that inform epistemological analyses are drawn from human epistemic activities.[27]

As mentioned previously, however, reliabilism arose out of concerns that traditional approaches to knowledge favoured cognitive capacities absent in very young humans or animals (other than humans). This seems to evince sensitivity to the epistemic perspective of animals as advocated above. However, we should take care to notice that contemporary reliabilist epistemologies tend to prioritize human epistemic activities and values when developing or defending their analyses (see Goldman 1976; Goldman 1988; Kornblith 1999). Reliability of beliefs, or belief forming mechanisms, is a recognizable *epistemic value* to human epistemic subjects. Arguably, this explains the persuasiveness of reliablist epistemologies. To ignore, or not properly appreciate, that reliability may not be a recognizable epistemic value to other animals – perhaps because they do not track environmental feedback over time in a fashion that could under-write an appreciation of a belief*'s (or its underlying mechanism[s]'s) reliability, or they lack the capacity to think in terms of belief*-forming mechanisms – prioritizes the perspectives of substantive epistemic subjects for whom it is. It is true that epistemological reliabilism does not require that all epistemic subjects who possess knowledge are capable of analyzing the reliability of their relevant belief-forming mechanisms.[28] However, a universal analysis of knowledge that purports to offer conditions of knowledge that resonate with the relevant epistemic judgments of successful epistemic subjects should take great care to ensure that these epistemic subjects include more than properly functioning adolescent or adult humans. Contemporary work in chimpanzee cognitive studies offers naturalized epistemologists a chance to correct this oversight.

6.6 Conclusions

I have provided examples of chimpanzee evidence gathering and, what might be reasonably described as, epistemic success. This strongly implies that chimpanzees engage in epistemic activities, identifying them as substantive epistemic subjects markedly similar to ourselves. If chimpanzees are properly regarded as substantive

[27] Up until now, with few exceptions, the epistemic activities and values informing the development and defence of analytic theories of positive epistemic status, or epistemic subjectivity, have been drawn from human behaviour (typically, the activities and values of mature, properly functioning, adult humans).This has *tended* to yield analyses of positive epistemic status or epistemic subjectivity that require sophisticated cognitive capacities (see Bonjour 2002; Rescher 2001; Steup 2003).

[28] As a form of epistemological externalism, reliabilist epistemology does not require that the justifiers which confer positive epistemic status are accessible to the relevant epistemic subject, nor that she be capable of understanding her belief's justifiers as such. As, however, Goldman has rightly recognized, a to-be-specified sensitivity and responsiveness to defeaters (e.g., counter-evidence to a belief's truth or probable truth) is required for epistemic success (see Goldman 1988).

epistemic subjects, this has some significant consequences for both contemporary research in chimpanzee cognitive studies and naturalized epistemology. Naturalized epistemologists now have the data needed to begin to develop analyses of positive epistemic status, and even epistemic subjectivity, that are sensitive to the epistemic activities of, and implicit epistemic values held by, chimpanzees. This will be of use in tracking bona fide examples of chimpanzee epistemic activity in free-living or captive chimpanzee populations, and understanding how knowledge, understood philosophically, affects the behaviour of some animals other than humans. This also deepens our shared understanding of epistemic subjectivity and offers a way of exploring its evolutionary history. It may also enable naturalized epistemologists to effectively move beyond lingering anthropocentricities in their epistemic frameworks, properly putting nature back into naturalized epistemology.

References

Allen, C. (1999): 'Animal concepts revisited: the use of self-monitoring as an empirical approach'. *Erkenntnis* 51: 33–40.

Allen, C. and Bekoff, M. (1997): *Species of Mind: The Philosophy and Biology of Cognitive Ethology*. Cambridge (MA): The MIT Press.

Audi, R. (1989): 'Causalist Internalism'. *American Philosophical Quarterly* 26: 309–320.

Bonjour, L. (2002): *Epistemology: Classic Problems and Contemporary Responses*. Lanham: Rowman and Littlefield.

Byrne, R. W. (1995): *The Thinking Ape: Evolutionary Origins of Intelligence*. New York: Oxford University Press.

Byrne, R. W. (2004): 'The manual skills and cognition that lie behind hominid tool use'. In Russon, A. E. and Begun, D. R. (eds.): *The Evolution of Thought: Evolutionary Origins of Great Ape Intelligence*.. New York: Cambridge University Press, pp. 31–44.

Call, J. and Tomasello, M. (2008): 'Does the chimpanzee have a theory of mind? 30 years later'. *Trends in Cognitive Sciences* 12: 187–192.

Chisholm, R. M. (1973): *The Problem of the Criterion*. Milwaukee: Marquette University Press.

Cohen, L. J. (1991): 'Stephen P. Stich, *The Fragmentation of Reason*'. *Philosophy and Phenomenological Research* 51: 185–188.

Crain, W. (1992): *Theories of Development: Concepts and Applications*. Third Edition. New Jersey: Prentice Hall.

Davidson, D. (1982): 'Rational Animals'. *Dialectica* 36: 317–327.

de Waal, F. B. M. (1987): 'Dynamics of Social Relationships'. In: Smuts, B., Cheney, D. L., Seyfarth, R. M., Wrangham, R. W., and Struhsaker, T. T. (eds.): *Primate Societies*. Chicago: The University of Chicago Press, pp. 421–429.

de Waal, F. B. M. (1990/96): *Peacemaking among Primates*. Cambridge (MA): Harvard University Press.

de Waal, F. B. M. (1994/96): 'Chimpanzee's Adaptive Potential: A Comparison of Social Life under Captive and Wild Conditions'. In: Wrangham, R. W., McGrew, W. C., de Waal, F. B. M., and Heltne, P. G. (eds.): *Chimpanzee Cultures*. Cambridge (MA): Harvard University Press, pp. 243–260.

de Waal, F. B. M. (2001): *The Ape and the Sushi Master: Cultural Reflections of a Primatologist*. New York: Basic Books.

de Waal, F. B. M. and Aureli, F. (1996): 'Consolation, reconciliation, and a possible cognitive difference between macaques and chimpanzees'. In Russon, A. E., Bard, K. A. and Parker, S. T.

(eds.): *Reaching into Thought: The minds of the great apes*. Cambridge: Cambridge University Press, pp. 80–110.
Dretske, F. (1989): 'The Need to Know'. In: Clay, M. and Lehrer, K. (eds.): *Knowledge and Skepticism*. Boulder (CO): Westview Press, pp. 89–100.
Fouts, R. S. and Fouts, D. H. (1999): 'Chimpanzee Sign Language Research'. In: Dolhinow, P. and Fuentes, A. (eds.): *The Nonhuman Primates*. Mountain View, California: Mayfield Publishing Company, pp. 252–256.
Fouts, R. S., Jensvold, M. L. A. and Fouts, D. H. (2002): 'Chimpanzee Signing: Darwinian Realities and Cartesian Delusions'. In: Bekoff, M., Allen, C. and Burghardt, G. M. (eds.): *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*. Cambridge (MA): The Massachusetts Institute of Technology Press, pp. 285–291.
Fruth, B., G. Hohmann and W. C. McGrew. (1999): "The Pan Species. In: Dolhinow, P. and Fuentes, A. (eds.): *The Nonhuman Primates*. Mountain View, California: Mayfield Publishing Company, pp. 64–72.
Goldman, A. I. (1976): 'Discrimination and Perceptual Knowledge' *Journal of Philosophy* 73: 771–791.
Goldman, A. I. (1988): 'Strong and Weak Justification'. In: *Philosophical Perspectives, 2, Epistemology, 1988*, pp. 51–69.
Goldman, A. I. (1989): 'Précis and Update of *Epistemology and Cognition*'. In: Clay, M. and Lehrer, K. (eds.): *Knowledge and Skepticism*. Boulder (CO): Westview Press, pp.69–87.
Gómez, J. C. (2004): *Apes, Monkeys, Children, and the Growth of Mind*. Cambridge (MA): Harvard University Press.
Goodall, J. (1988/97): *In the Shadow of Man*. London: Orion Books Limited.
Gould, J. L. and Gould, C. G. (1994/99): *The Animal Mind*. New York: Scientific American Library.
Greenfield, P. M. and Savage-Rumbaugh, E. S. (eds.) (1990/94): 'Grammatical combination in *Pan paniscus*: Processes of learning and invention in the evolution and development of language'. In: Parker, S. T. and Gibson, K. R. (eds.): *"Language" and Intelligence in Monkeys and Apes: Comparative Developmental Perspectives*. New York: Cambridge University Press, pp. 540–578.
Hauser, M. D. (2000): *Wild Minds: What Animals Really think*. New York: Henry Holt and Company.
Hirata, S. (2009): "Chimpanzee social intelligence: selfishness, altruism, and the mother-infant bond." *Primate* 50: 3–11.
Kornblith, H. (1999): "Knowledge in Humans and Other Animals." In *Nous, Vol. 33 Supplement: Epistemology 1999*, pp. 327–346.
———. (2002): *Knowledge and its Place in Nature*. New York: Oxford University Press.
———. (2004): "Sosa on Human and Animal Knowledge." In *Ernest Sosa and His Critics*. Edited by John Greco. Malden: Blackwell Publishing Limited, pp. 126–134.
Langer, J. (1996): "Heterochrony and the evolution of primate cognitive development." In *Reaching into Thought: The minds of the great apes*. Edited by Anne E. Russon, Kim A. Bard and Sue Taylor Parker. Cambridge: Cambridge University Press, pp. 257–277.
Longino, H. E. (2002): *The Fate of Knowledge*. Princeton: Princeton University Press.
Lonsdorf, E. V. (2006): "What is the role of mothers in the acquisition of termite-fishing behaviours in wild chimpanzees (*Pan troglodytes schweinfurthii*)." In *Animal Cognition* 9: 36–46.
Manning, A. and M. S. Dawkins. (1998): *An Introduction to Animal Behaviour*. Fifth Edition. New York: Cambridge University Press.
Matsuzawa, T. (1985): "Use of numbers by a chimpanzee." In *Nature* 315: 57–59.
———. (1994/96): "Field Experiments on Use of Stone Tools by Chimpanzees in the Wild." In *Chimpanzee Cultures*. Edited by Richard W. Wrangham, W.C. McGrew, Frans B.M. de Waal, and Paul G. Heltne. Cambridge (Mass.): Harvard University Press, pp. 351–370.
———. (1996): "Chimpanzee intelligence in nature and in captivity: isomorphism of symbol use and tool use." In *Great Ape Societies*. Edited by William C. McGrew, Linda F. Marchant, and Toshisada Nishida. Cambridge: Cambridge University Press, pp. 196–209.

———. (2002): "Chimpanzee Ai and Her Son Ayumu: An Episode of Education by Master-Apprenticeship." In *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*. Edited by Marc Bekoff, Colin Allen, and Gordon M. Burghardt. Cambridge (Mass.): The MIT Press, pp. 189–195.

Matsuzawa, T. and G. Yamakoshi. (1996): "Comparison of chimpanzee material culture between Bossou and Nimba, West Africa." In *Reaching into thought: The minds of the great apes*. Edited by Anne E. Russon, Kim A. Bard, and Sue Taylor Parker. Cambridge: Cambridge University Press, pp. 211–232.

McGrew, W. C. (1994/96): "Tools Compared: The Material of Culture." In *Chimpanzee Cultures*. Edited by Richard W. Wrangham, W.C. McGrew, Frans B. M. de Waal, and Paul G. Heltne. Cambridge (Mass.): Harvard University Press, pp. 25–39.

———. (2001): "The Nature of Culture: Prospects and Pitfalls of Cultural Primatology." In *Tree of Origin: What Primate Behavior Can Tell Us about Human Social Evolution*. Edited by Frans B.M. de Waal. Cambridge (Mass.): Harvard University Press, pp. 231–254.

———. (2004): *The Cultured Chimpanzee: Reflections on Cultural Primatology*. New York: Cambridge University Press.

Nishida, T., and M. Hiraiwa-Hasegawa. (1987): "Chimpanzees and Bonobos: Cooperative Relationships among Males." In *Primate Societies*. Edited by Barbara Smuts, Dorothy L. Cheney, Robert M. Seyfarth, Richard W. Wrangham, and Thomas T. Struhsaker. Chicago: The University of Chicago Press, pp. 165–177.

Nishida, T. and K. Hosaka. (1996): "Coalition strategies among adult male chimpanzees of the Mahale Mountains." In *Great Ape Societies*. Edited by William C. McGrew, Linda F. Marchant, and Toshisada Nishida. Cambridge: Cambridge University Press, pp. 114–134.

Parker, S. T. (1996): "Apprenticeship in tool-mediated extractive foraging: The origins of imitation, teaching, and self-awareness in great apes." In *Reaching into thought: The minds of the great apes*. Edited by Anne E. Russon, Kim A. Bard, and Sue Taylor Parker. Cambridge: Cambridge University Press, pp. 348–370.

Ray, E. (1999): "Hierarchy in Primate Social Organization." In *The Nonhuman Primates*. Edited by Phyllis Dolhinow and Agustín Fuentes. Mountain View, California: Mayfield Publishing Company, pp. 211–217.

Rescher, N. (2001): "Philosophical Methodology." In *Two Roads to Wisdom? Chinese and Analytic Philosophical Traditions*. Edited by Bo Mou. Chicago: Open Court Publishing Company, pp. 3–25.

Ristau, C. (1991a): "Aspects of the Cognitive Ethology of an Injury-Feigning Bird, the Piping Plover." In *Cognitive Ethology: The Minds of Other Animals*. Edited by Carolyn A. Ristau. Hillsdale: Lawrence Erlbaum Associates, Publishers, pp. 91–126.

———. (1991b): "Cognitive Ethology: An Overview." In *Cognitive Ethology: The Minds of Other Animals*. Edited by Carolyn A. Ristau. Hillsdale: Lawrence Erlbaum Associates, Publishers, pp. 291–313.

Russell, B. (1948): *Human Knowledge: Its Scope and Limits*. New York: Simon and Schuster.

Russon, A. E. (1997): "Exploiting the expertise of others." In *Machiavellian Intelligence II: Extensions and Evaluations*. Edited by Andrew Whiten and Richard Byrne. New York: Cambridge University Press, pp. 174–206.

Saidel, E. (1998): "Beliefs, desires, and the ability to learn." In *American Philosophical Quarterly* 35 (1): 1–12.

Santrock, J. W. (2001): *Child Development*. Ninth Edition. New York: McGraw-Hill Ryerson.

Sanz, C., D. Morgan, and S. Gulick. (2004): "New Insights into Chimpanzees, Tools, and Termites from the Congo Basin." In *The American Naturalist* 164 (5): 567–581.

Savage-Rumbaugh, S. and R. Lewin. (1994): *Kanzi: The Ape at the Brink of the Human Mind*. New York: John Wiley and Sons, Inc.

Schmitt, F. F., Editor. (1994): *Socializing Epistemology: The Social Dimensions of Knowledge*. Lanham: Rowman and Littlefield Publishers, Incorporated.

Sosa, E. (1991a): "Knowledge and intellectual virtue." In *Knowledge in Perspective: Selected Essays in Epistemology*. Edited by Ernest Sosa. New York: Cambridge University Press, pp. 225–244.

———. (1991b): “Reliabilism and intellectual virtue.” In *Knowledge in Perspective: Selected Essays in Epistemology*. Edited by Ernest Sosa. New York: Cambridge University Press, pp. 131-145.

Steup, M. (2003): “A Defense of Internalism.” In *The Theory of Knowledge: Classical and Contemporary Readings*. Third Edition. Edited by Louis P. Pojman. Belmont: Wadsworth/ Thomson Learning, pp. 310–321.

Strier, K. B. (2000): *Primate Behavioral Ecology*. Boston: Allyn and Bacon.

Tomasello, M. and J. Call. (1997): *Primate Cognition*. New York: Oxford University Press.

Yamakoshi, G. (2004): “Evolution of complex feeding techniques in primates: is this the origin of great ape intelligence”? In *The Evolution of Thought: Evolutionary Origins of Great Ape Intellgience*. Edited by Anne E. Russon and David R. Begun. New York: Cambridge University Press, pp. 140–171.

Part IV
Evolutionary Explanations of Behavior

Chapter 7
Evolving the Future: Sketching a Science of Intentional Change

David Sloan Wilson

Evolution is a process of change that adapts organisms to their environments. It is therefore ironic that evolution is often thought to result in an incapacity for change when it comes to human affairs. This is the specter of genetic determinism, which has haunted discussions of evolution and human behavior for decades (Ehrenreich & McIntosh 1997, Wilson 2005). According to the reasoning of genetic determinism, if behaviors are coded by genes and genes only change over the timescale of hundreds and thousands of generations, then we are stuck with the behaviors that we would like to change over much shorter timescales. This reasoning has led generations of thinkers to acknowledge the importance of evolution for all other species, for human physical traits and a few instincts such as our urge to eat and have sex, but to regard our rich behavioral and cultural diversity as somehow outside the orbit of evolutionary theory.

This essay describes a seismic shift in our thinking about evolution and human behavior. My use of the term "seismic shift" is carefully chosen. A geological seismic shift occurs when pressures that have been accumulating for a long time suddenly overcome the forces of friction. The intellectual seismic shift that I am describing reflects gradual scientific developments that have been taking place, especially over the last two decades, which now need to overcome resistance based on previous configurations of ideas that no longer make sense. In both cases, the suddenness of the seismic shift is based on gradual changes suddenly overcoming longstanding resistance, not a major event immediately precipitating the change.

D.S. Wilson (✉)
Departments of Biology and Anthropology, Binghamton University,
Binghamton, NY 13902, USA
e-mail: dwilson@binghamton.edu

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*,
Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_7,

The purpose of this essay is to sketch the new configuration of ideas that deserves to replace the configuration associated with genetic determinism. It will help to briefly list the elements of the sketch before fleshing them out in more detail.

1) All organisms are capable of changing in response to their environments, which is called *phenotypic plasticity*. Understanding phenotypic plasticity in other species is an important prerequisite for understanding human phenotypic plasticity.
2) Some kinds of phenotypic plasticity can be described by the paradoxical phrase "rigidly flexible". Your tax preparation software or Big Blue, the chess-playing computer, are examples of rigid flexibility. They are amazingly flexible at executing the task for which they have been designed but can't do anything else. Their flexibility requires rigidly specified environmental information that is rigidly processed in exactly the right way.
3) Other kinds of phenotypic plasticity are based on more open-ended processes that count as evolutionary in their own right. An example is the capacity of our immune system to produce roughly 100 million different antibodies and to select the ones that successfully bind to antigens. This open-ended capacity has been aptly termed a Darwin Machine: a fast-paced process of evolution built by the slow-paced process of genetic evolution (Calvin 1987, Plotkin 1994).
4) Both kinds of phenotypic plasticity are found in most species but humans have a capacity for open-ended behavioral change that is transmitted across generations, therefore becoming cultural change, surpassing all other species (Deacon 1998, Jablonka & Lamb 2006). That makes us highly distinctive but does not remove us from the orbit of evolutionary theory. On the contrary, we need to tell two evolutionary stories for every Darwin Machine: how it evolved by genetic evolution and how it employs open-ended variation-and-selection processes in its own right.
5) All evolutionary processes, fast or slow, lead to outcomes that can be either good or bad for long-term human welfare. It is not the case that evolution automatically makes everything nice. Neither is it the case that evolution makes everything nasty. Rather, evolution can result in the full spectrum of outcomes associated with human welfare, from the best to the worst. To produce desired outcomes, we must become wise managers of evolutionary processes.
6) The prospect of using evolutionary theory to manage cultural change raises the specter of Social Darwinism, the use of evolutionary theory in the past to justify policies such as eugenics, genocide, and lack of welfare support for the poor. Social Darwinism is one form of social engineering, a term with a bad reputation no matter what its theoretical underpinning. The horrifying prospect of social engineering is that it will be used as a tool of exploitation. The solution is to be vigilant against exploitation in all its forms and to decide by consensus how to use knowledge to improve the human condition. Evolutionary knowledge is no different than any other kind of knowledge in this respect. Despite the sorry history of Social Darwinism, contemporary evolutionary theory provides a powerful argument for egalitarianism, since human cooperation can only be achieved by suppressing the potential for exploitation within groups (Boehm 1999, Sober & Wilson 1998, Wilson 2002).

7) These points are so basic (at least in retrospect) that they are unlikely to be wrong. However, they are also abstract and need to be made more concrete to manage behavioral and cultural change in a practical sense. Fortunately, the applied human-related sciences offer many successful case studies, ranging from therapeutic methods for individuals to changing the cultural practices of large populations. When viewed through the lens of evolutionary theory, these case studies can be seen as Darwin Machines in action, intelligently designed to use variation-and-selection processes to produce benign outcomes.

I will now briefly elaborate on each of these points. My main goal is to help the reader conceptualize behavioral and cultural change as firmly inside the orbit of evolutionary science, which not only transforms intellectual understanding but also provides an essential toolkit for managing change in a practical sense.

7.1 All organisms are capable of changing in response to their environments

Some terms that should be part of everyone's vocabulary are *phenotype*, *genotype*, *norm of reaction*, and *phenotypic plasticity* (Pigliucci 2001, West-Eberhard 2003). A phenotype is any trait that can be observed in an organism, behavioral or otherwise. A genotype is the organism's genetic composition. A norm of reaction describes the relationship between the phenotype of an organism with a particular genotype and the organism's environment. A norm of reaction is often displayed as a graph with an environmental variable (such as temperature) on the x-axis and a phenotypic trait of the organism (such as body size) on the y-axis. If the line is flat, then the organism is not phenotypically plastic with respect to that trait. If the line departs from flatness in any way, then the organism is phenotypically plastic with respect to that trait. This graphical portrayal makes it clear that there are many ways to be phenotypically plastic. Every genotype has a norm of reaction and genetic evolution winnows among genotypes, resulting in norms of reaction that cause organisms to change (or not change) in response to their environments in just the right way.

Examples of phenotypic plasticity include but go far beyond behavioral change. Sex is determined by the presence or absence of a Y-chromosome in our species (except in extremely atypical environments), but in other species it is phenotypically plastic. In some reptile species, any individual can become either a male or a female depending upon the temperature experienced during egg development (Crews et al. 1994). Sex is socially determined in some fish species; every individual begins life as a female and physiologically changes into a male when it becomes the largest member of its group (Devlin & Nagahama 2002).

Some species undergo extreme makeovers in response to chemicals indicating the presence of predators in their environment (e.g., Relyea 2002). They change their morphological form (such as growing more muscular tails), behaviors

(such as moving less), and life histories (such as maturing earlier). Some species of caterpillars resemble the flowers of their host plant when they hatch during the spring, when the flowers are present, but resemble twigs when they hatch later in the season, when flowers are absent. The environmental cue is their diet (Greene 1996).

Why are some traits more phenotypically plastic than others? It depends largely upon the patterns of environmental variation experienced during genetic evolution. In the caterpillar example, flowers are reliably present in spring and absent in summer, favoring a particular matching of phenotype to environment. In the predation example, some species inhabit environments where predators might or might not be present, favoring the capacity to switch between a predator-absent suite of traits and a predator-present suite of traits. This capacity does not evolve when predators are always absent. On some oceanic islands where large mammalian predators have never existed, the birds confuse people for trees and do not have the capacity to change their response on the basis of their experience. Those species are now largely extinct (MacPhee & Sues 1999). All of these examples illustrate the general concept that the existence and specific pattern of phenotypic plasticity in a given species reflects the existence and specific pattern of environmental variation during the genetic evolution of the species.

Human skin color provides an outstanding example of both the presence and absence of phenotypic plasticity in our own species (Tadokoro et al. 2005). It reflects a tradeoff between the harmful effects of the sun and the need for the skin to receive sunlight to manufacture vitamin D. Too much and too little sunlight are both harmful. In open tropical environments, sunlight is always present and human skin color evolved to be permanently dark. In the temperate zones, sunlight is variable and human skin color evolved to be phenotypically plastic, darkening in response to exposure to the sun. The capacity to suntan is just as much a genetically evolved adaptation as permanently dark skin.

7.2 Some kinds of phenotypic plasticity can be described by the paradoxical phrase “rigidly flexible”

The examples listed above are similar to conditional and unconditional statements in a computer program. If you were writing a computer program, you would assign values to some parameters that don’t change during the execution of the program (e.g., let x = 1) but you would allow other parameters to have different values depending upon certain conditions (e.g., if y = 2 then let z = 3). Genetic evolution has endowed organisms with “let” statements for some traits (such as permanently dark skin color in some people) and “if-then” statements for other traits (such as the capacity for tanning in other people).

The computer programming analogy nicely illustrates the concept of *rigid flexibility*. The conditional statement “if y = 2 then let z = 3” specifies a particular phenotypic response (z = 3) to a particular environmental parameter (y = 2).

The right environmental information (the value of y) must be provided for the phenotypic response to occur and any other response ($z \neq 3$) is prohibited. The seemingly opposite terms *rigid* and *flexible* are joined at the hip, like the opposites of a Zen Koan.

The example of human skin color can be used to illustrate an important implication of rigid flexibility. People capable of tanning are also vulnerable to *burning* when their skin is suddenly exposed to sun after a long period of low exposure. Why can't they tan faster? During most of our genetic evolutionary history, our ancestors spent most of their time outdoors and never experienced the situation of being suddenly exposed to the sun after a long period of low exposure. Gradual tanning was always sufficient and rapid tanning was never required. When people changed their lifestyle during very recent times by doing such things as flying to Florida during winter for a week's vacation, they encountered a pattern of environmental change that had no counterpart to anything experienced during their genetic evolution. We are stuck with genes that are only capable of gradual tanning and there is nothing we can do about it—except by wearing clothing, smearing sun blocking lotions on our skin, or staying indoors.

The good news about rigid flexibility is that it can magnificently adapt organisms to the particular patterns of environmental change experienced during its evolution. The bad news is that rigid flexibility can go horribly wrong when the pattern of environmental change itself changes, a problem that only be solved by subsequent genetic evolution or a behavioral and cultural intervention. Might behavioral and cultural interventions also count as evolutionary?

7.3 Other kinds of phenotypic plasticity are based on more open-ended processes that count as evolutionary in their own right

The vertebrate immune system includes many components that are rigidly flexible but it also includes another kind of phenotypic plasticity that is more open-ended in its flexibility. The immune system can produce approximately 100 million different kinds of antibodies. Each is like a hand that can grasp a particular organic surface and collectively they can grasp almost any conceivable organic surface. When a particular antibody latches onto an invading disease organism, it summons other components of the immune system to attack the invader and triggers the cells that produce the antibody to reproduce. In this fashion, antibodies that *vary* are *selected* based on their ability to bind to antigens (Sompayrac 2008).

This is not a happy accident. Every part of the process, from the mechanisms that create different antibodies to the mechanisms that amplify the ones that successfully bind to antigens, is a sophisticated product of genetic evolution. Yet, the variation-and-selection process built by genetic evolution results in a new kind of phenotypic plasticity that can rapidly adapt to new environments, rather than merely following if-then statements winnowed by past environments. If a new disease organism

invaded from outer space that never before existed on earth, our immune systems could probably take care of it.

Learning of the sort that B.F. Skinner made famous is an open-ended process similar to the immune system. In operant conditioning, an organism behaves in different ways and is capable of detecting which behaviors work better than others, for example by resulting in a food reward. The most successful behaviors are adopted, enabling the organism to rapidly adapt to new environments, just like the immune system can adapt to new disease organisms.

Skinner (1981) explicitly described operant conditioning as a rapid evolutionary process in its own right, built by the slow-paced process of genetic evolution. He grasped the basic concept of a Darwin Machine but erred in other respects. For example, he tried to explain too much with his principle of operant conditioning and perversely insisted that the study of behavioral change should be restricted to input-output relationships without actually opening the black box of the mind and directly studying the mechanisms that accomplish the transformation. Thinking of the human capacity for behavioral change as comparable to the immune system enables us to keep the "baby" of the Skinnerian tradition without the bathwater, as I have elaborated elsewhere in an essay titled "Learning from the Immune System about Evolutionary Psychology" (Wilson 2010a).

Two points need to be stressed for the purpose of this essay. First, the variation-and-selection process of a Darwin Machine results in a different kind of phenotypic plasticity than rigid flexibility, one that is capable of producing genuinely new adaptations to new environments. Second, Darwin machines do not replace rigidly flexible mechanisms but complement them and are utterly dependent upon them. In his lucid book on how the immune system works, Sompayrac (2008) compares the open-ended component to a quarterback who cannot possibly function without other members of the football team, all of whom are relying upon if-then statements winnowed by genetic evolution.

An example from the immune system will show why these two points matter for our understanding of human behavioral/social/cultural change. Throughout our evolutionary history, the bodies of our ancestors were inhabited by a diverse community of species living in our guts. They weren't necessarily *welcome*, but they were always *there* and the immune system evolved to rely upon their presence to develop antibodies against them. With the advent of modern medicine and public health measures such as sanitary water supplies, it became possible for the first time in human history to largely eliminate elements of our gut biota such as intestinal worms. This might seem like an unambiguous blessing but instead it results in the same kind of problem that we encounter when we fly to Florida for a winter vacation. In the absence of intestinal worms, our immune system can react inappropriately and unleash a storm of friendly fire against our own bodies (Yazdanbakhsh et al. 2002). We call these immune system disorders but in most cases they are examples of normal immune systems malfunctioning in modern environments. Our immune system cannot solve this problem any more than our skin can speed up its tanning capacity. There must be solutions comparable to clothing, sunscreen, and staying indoors or there will be no solutions at all.

How many human behavioral/social/cultural disorders are comparable to sunburns and immune system disorders? We'll never know until we begin to understand the human capacity for change from a sophisticated evolutionary perspective.

7.4 Both kinds of phenotypic plasticity are found in most species but humans have a capacity for open-ended behavioral change that is transmitted across generations, therefore becoming cultural change, surpassing all other species

Even pigeons have the capacity for open-ended learning that Skinner made famous by putting them in his boxes. To get from pigeons to humans, we must tell a story about human evolution per se. Three distinctive features of our species that we need to explain are a) our distinctive *cognition*, including our capacity for symbolic thought; b) our distinctive ability to transmit learned information across generations, resulting in cumulative *culture*; and c) our distinctive ability to *cooperate* with individuals who are not our close genetic relatives or narrow reciprocators. A consensus is emerging that of these three C's, cooperation came first and the other two C's are themselves forms of cooperation (Wilson 2007, Wilson et al. 2008, Tomasello 2009, Tomasello et al. 2005).

In all group-living species, natural selection can occur among individuals within groups or among the groups in the total population (Wilson & Wilson 2007). The balance between levels of selection is not static but can itself evolve. When between-group selection becomes sufficiently strong compared to within-group selection, groups become so functionally organized that they qualify as organisms in their own right (Maynard Smith & Szathmary 2005, 2009). All of the entities that we currently recognize as organisms, including multicellular organisms such as ourselves, are tightly regulated social groups whose members led a more autonomous and conflictive existence in past ages. Social insect colonies also qualify as organisms by virtue of their group-level functional organization, even though their members are not physically connected to each other (Seeley 1995, Holldöbler & Wilson 2008).

Human evolution represents a major transition, similar to these previous transitions (Boehm 1999, Wilson 2006, 2007, Wilson et al. 2008). Our ancestors became the primate equivalent of a social insect colony. The key event was the ability to suppress competition and deviance within groups, so that the driving force of evolution became how well groups succeeded relative to other groups. Achieving a balance of power within groups need not have been a cognitive event—it could have been based on the ability to throw projectiles with deadly force, for example, which originally evolved to deter predators and competitors on the savannah but then could be used to deter would-be alpha males (Bingham 1999). However it happened, this kind of guarded egalitarianism allowed our cognitive and cultural abilities to evolve in a direction predicated on trust and cooperation within groups.

The sharing of learned information takes place to a limited degree in the absence of trust but can take place to a much greater degree in its presence. Symbolic thought is not a private cognitive process but requires an inventory of symbols with meanings that are shared across individuals (Deacon 1998). In this fashion, the major transition that took place in our ancestors was like crossing a watershed, enabling primate intelligence to flow in a cooperative rather than a competitive direction. Our capacity for open-ended behavioral change became so great we spread over the globe, adapting to all climatic zones and hundreds of ecological niches. We remained a single biological species but our cultural diversity was like an entire phylum (Pagel and Mace 2004). Then the invention of agriculture enabled population size to increase many orders of magnitude in only a few thousand years (Diamond 1997).

7.5 All evolutionary processes, fast or slow, lead to outcomes that can be either good or bad for long-term human welfare

Everything that counts as functionally organized is either directly or indirectly a product of evolution (Campbell 1960). Yet, many products of evolution count as pathological from the standpoint of long-term human welfare. It is essential to understand the basic relationships between evolution, adaptation, and long-term human welfare to become wise managers of evolutionary processes.

In the first place, many outcomes of evolution aren't adaptive in any sense. Examples include traits that evolve by genetic drift, traits that were adaptive to past environments but not the present environment, traits that are costly byproducts of adaptations, and costly traits that "hitchhike" on adaptations by being located close to them on the same chromosome. Adaptations evolve by natural selection, which is opposed by many forces, as the late evolutionist Stephen Jay Gould tirelessly argued (Gould 2007). It is theoretically possible for a non-adaptation to benefit long-term human welfare, but only as a happy coincidence.

Even when a trait does count as an adaptation, it can be selfish and short sighted, benefiting some individuals and groups at the expense of others or providing immediate benefits despite long-term costs. Long-term human welfare is inherently about benefiting the common good and restraining ourselves in the present for the sake of the future. Thus, many adaptations are highly functionally organized in their own way but become part of the problem as far as long-term human welfare is concerned.

A good example concerns the "problems" of early pregnancy in women and violent behavior in men. In a landmark study, evolutionary psychologists Margo Wilson and Martin Daly (1997) related these problems to average life expectancy in the city of Chicago. The neighborhoods of Chicago vary greatly in their quality of life, which is reflected in average life expectancy, from the high 70s in the best neighborhoods to the 50s in the worst. There is a very strong positive relationship between age of first reproduction in women and average life expectancy of the neighborhood.

When women in the worst neighborhoods are asked why they begin having babies so young, they give a response that can only evoke sympathy: they want to see their grandchildren and want their mothers to see their children. They observe people "weathering" all around them and have calibrated their reproductive schedule accordingly, consciously or unconsciously. It makes no sense to postpone one's reproduction in such an environment.

There is a 100-fold difference between the worst and best neighborhoods in the rate of homicide among men. Homicides are removed from average life expectancy for this comparison, so this is not a matter of correlating something with itself. This enormous range of variation means that when there are very few opportunities for success, especially reproductive success, many men are willing to "get rich or die tryin'" as the album and movie by the rapper 50 Cent puts it. In safe and secure environments, when survival and reproduction can be achieved non-violently, men are no more likely to commit homicide than women.

The "problems" of early reproduction in women and violent behavior in men are clearly adaptations to highly insecure environments, in the evolutionary sense of the word "adaptation". They remain important problems to solve, but understanding them from an evolutionary perspective points to solutions that might not occur to us otherwise. It is both impractical and morally questionable to counsel women in the worst neighborhoods to delay their reproduction and even men to refrain from violence when these are their best options for their own reproductive success in their current environment. On the other hand, if the kind of environment that leads to a high average life expectancy can be created, then women are likely to delay their reproduction and men are likely to become less violent on their own.

More generally, the traits associated with long-term human welfare *can* win the Darwinian contest, but only under the right environmental conditions, where "environment" is interpreted broadly to include much that is socially constructed by humans. Provide the right conditions and the world can become a better place seemingly by itself. Provide the wrong conditions and even the most heroic efforts to make the world a better place can fail miserably. A sophisticated knowledge of evolution, including genetic evolution and all the Darwin Machines produced by genetic evolution, is required to engineer the right environments.

7.6 The prospect of using evolutionary theory to manage cultural change raises the specter of Social Darwinism

Using evolution to inform social policy is not new. Consider Julian Huxley, one of the pre-eminent evolutionists of the 20th century and grandson of Thomas Huxley, "Darwin's bulldog". Julian Huxley was a passionate humanist who felt that mankind must take charge of is own destiny. In addition to his book *Evolution: The Modern Synthesis* (1942), which literally defined the field of evolutionary biology for the ensuing decades, his humanistic books include *Religion without*

Revelation (1927, 1957), *Evolutionary Ethics* (1943), *Essays of a Humanist* (1964), and *The Future of Man* (1966). Here is an example of his humanistic side:

> There is no separate supernatural realm: all phenomena are part of one natural process of evolution. There is no basic cleavage between science and religion…I believe that [a] drastic reorganization of our pattern of religious thought is now becoming necessary, from a god-centered to an evolutionary-centered pattern.
>
> Many people assert that this abandonment of the god hypothesis means the abandonment of all religion and all moral sanctions. This is simply not true. But it does mean, once our relief at jettisoning an outdated piece of ideological furniture is over, that we must construct something to take its place (Huxley 1969).

This could have been written by Richard Dawkins or even by myself, although as a thoroughgoing atheist I am more respectful of religion than either Huxley or Dawkins (Wilson 2010b). Here is another passage:

> The lowest strata are reproducing too fast. Therefore…they must not have too easy access to relief or hospital treatment lest the removal of the last check on natural selection should make it too easy for children to be produced or to survive; long unemployment should be a ground for sterilization (Huxley 1947).

This passage sounds horrifying to most of us today, certainly to myself. Even more horrifying is the fact that Huxley had lots of company. It was acceptable at that time for social planners to argue that mankind should take charge of its destiny in this particular way. More horrifying still, their talk was not idle and led to social policies on both sides of the Atlantic that can only be looked back upon with shame. Yet, I would argue that the culprit is not evolutionary thinking but a worldview that regarded it as acceptable for the privileged to impose life and death decisions on the unprivileged without their consent. Rebecca M. Lemov's book *World as Laboratory: Experiments with Mice, Mazes, and Men* (2005) chronicles shameful public policies during the same period inspired by the "blank slate" tradition of behaviorism.

Given the history of Social Darwinism, it is important to address the question of whether evolutionary theory inherently lends itself to policies that favor social inequality. Social policies are most likely to become problematic when they involve some people imposing their will on others without their consent. Social policies are most likely to remain benign when they are agreed upon by all who will be affected by the policies. These statements are true regardless of the theoretical perspective that informs social policy.

If anything, modern evolutionary theory is biased in favor of egalitarian social policies. People are horrified by the prospect of other people determining their fate without their consent for the best of reasons—it provides no safeguards against exploitation within groups. Cooperative human life requires these safeguards and always has—suppressing selection within groups is what major evolutionary transitions are all about. In addition to these basic theoretical considerations, there is compelling empirical evidence that inequality is toxic for human social life at all scales, including nations and states within the Unites States (e.g., Wilkinson & Pickett 2009).

It is common for political ideologies to claim the support of *any* authoritative idea, religious, scientific, or otherwise. The solution to this problem is to challenge

the association between the ideology and the idea, not to accept the association and shun the idea. Moreover, it's not as if the world was a nice place before Darwin and then became mean on the basis of his theory. Before Darwin, the religious concept of divine right was used to commit genocide, dispossess people of their land, enslave them, and so on.

The nature of ideological thinking, exploitation and cooperation within groups, and exploitation and cooperation among groups, are all subjects that urgently need to be understood from a genetic and cultural evolutionary perspective, leading to knowledge that can be used to formulate humane social policies agreed upon by consensus. In this sense, knowledge derived from evolutionary theory is no different than knowledge derived from any other source. All knowledge is a form of power that can be used for good or ill. It is up to us to use it responsibly. For better or worse, we live in a world of our own making and must use our knowledge to manage our affairs. It is time to make use of the knowledge provided by evolutionary theory.

7.7 The applied behavioral sciences offer many successful case studies

It might seem that an enlightened Social Darwinism only exists in the future, perhaps the far future. On the contrary, outstanding examples of intentional change can be found in the applied human-related sciences (Biglan & Hinds 2009, Luyben 2009). When these examples are viewed through an evolutionary lens, they can be seen as variation-and-selection processes that are carefully managed to achieve desired outcomes. I will briefly describe three examples of changes at very different scales—individuals, small groups, and large populations.

Changing individuals: Hundreds of psychotherapeutic methods exist to help individuals who are functioning poorly and earnestly want to change. Some of these methods actually work and have been rigorously validated in randomized trials. One method called Acceptance and Commitment Therapy (ACT) builds upon previous behavioral and cognitive therapies (which are successful in their own right) by adding a component of mindfulness, which is drawn from meditative religious practices (Hayes 2004).

Stated in evolutionary terms, people who have need to seek therapy have two problems. First, their behavioral repertoire has become limited to avoid exacerbating their problems. Second, their criteria for adopting behaviors does not correspond to their true goals in life. The goal of therapy is to help the client increase the range of behavioral variation and select the behaviors according to the right criteria. This is partially a matter of conscious choice (the rationale of cognitive therapy) but also a matter of managing the psychological machinery of learning that takes place beneath conscious awareness (the rationale of behavioral therapy).

The mindfulness component of ACT encourages the client to distance oneself from one's problems and accept the fact that some problems might not go away, but that this need not prevent the achievement of one's most important goals.

One metaphor employed in ACT therapy asks the client to imagine being a bus driver, stopping to let people on and off on the way to a final destination. You might not like the people who get on the bus. In fact some might be downright scary. However, your challenge is to manage the people on the bus as best you can on your way to your final destination.

Metaphors such as these and other elements of ACT therapy have been proven to be highly effective in randomized trials, even on the basis of a single therapeutic session. The efficacy of ACT is based in part on the capacity of the human mind for symbolic thought and the power of symbolic systems to govern behavior (Hayes et al. 2001, Hayes 2004). Space does not permit a fuller account but I hope that I have described ACT just enough to show how it can be viewed as a managed variation-and-selection process that is informed by a detailed understanding of the human mind as a product of genetic evolution.

Changing small groups: Everyone wants to improve American public school education but no one is entirely sure of the best way to do so. Perhaps surprisingly, then, there is an intervention program called the Good Behavior Game, which has been shown to have transformative effects even in the toughest of inner city public schools (Embry 2002).

Invented by a teacher and perfected over a period of decades by researchers, the GBG begins by having the teacher ask the students what *they* think counts as good and bad classroom behavior. Even first graders are capable of coming up with the same dos and don'ts that the teacher might impose, but the fact that *they* decided upon the rules makes a big difference.

After the dos and don'ts are discussed and conspicuously displayed, the class is divided into groups that compete to be good. At first the competition is for a short period, such as doing schoolwork for a ten-minute period. Any group that manages to avoid committing a certain number of don'ts receives a small prize, such as picking from a prize bowl or even an opportunity to let loose and commit a don't – armpit farts are a popular reward for winning! Competing as a member of a group is highly motivating and causes peer pressure to promote normative rather than deviant behaviors.

Gradually the game is played more often and for longer periods. Sometimes it is played unannounced. The reward for winning is gradually deferred to the end of the day or week. In this fashion, the norms of good behavior become the culture of the class. The benefits of the GBG are astonishing. In one comprehensive study conducted in the inner city public schools of Baltimore, Maryland, the GBG was implemented in some 1st and 2nd grade classrooms but not others in a randomized design. The progress of the children was then carefully followed as they matured. At the end of the 6th grade, the GBG kids were less likely to be diagnosed with conduct disorder, to have been suspended from school, or to be judged in need of mental health services. During grades 6-8, they were less likely to use tobacco or hard drugs such as heroin, crack, and cocaine powder. In high school, the GBG kids scored higher on standardized achievement tests, had a greater chance of graduating, of attending college, and a reduced need for special education services. In college, the GBG kids had a reduced risk for suicide ideation, lower rates of anti-social personality

disorder, and lower rates of violent and criminal behavior. The GBG was especially effective at improving the lives of boys. All of the above-cited results are statistically significant and can be attributed to the effect of the GBG, played in the 1st and 2nd grades only, because the students were randomly assigned to the two treatment groups. The detailed results are reported in a 2008 supplement of the *Journal of Drug and Alcohol Dependence* (Volume 95, Supplement 1, pp. S1-S104).

These lifelong benefits might seem too good to be true, until we realize that the classes that didn't play the GBG were so disruptive that almost no learning was taking place. Like money in a bank earning interest, learning the habits of cooperative behavior and harvesting their benefits over a two year period can indeed accrue benefits that last a lifetime.

When the GBG is viewed through an evolutionary lens, it can be seen to provide the conditions that favor cooperative behavior in any human group, not just a group of children. People of all ages hate being bossed around but will conscientiously abide by rules that are established by consensus. Most people are strongly motivated to become respected members of groups and even more motivated when groups are competing with each other. These motivations can be stronger than earning rewards merely for oneself. The same motivations can lead to destructive outcomes, of course, but the whole point of managing the evolutionary process is to intelligently steer them toward productive outcomes. The success of the GBG also enables us to revisit the specter of Social Darwinism. Not only is the GBG a benign social policy informed by evolutionary theory, but it also illustrates the essential role of egalitarianism for cooperative social interactions at any age.

Changing large populations: A program that successfully reduced cigarette sales to minors in the states of Wyoming and Wisconsin shows that change can be accomplished at the scale of large populations, if one knows what to do (Embry et al. 2010). Federal agencies regulating tobacco sales employ underage kids as secret agents who enter retail stores and attempt to purchase cigarettes. When they are successful more than 20% of the time in a given state, the state is put on notice that it stands to lose millions of dollars provided by the federal government in the form of block grants. Wyoming and Wisconsin were in this dilemma, with cigarette sales to minors hovering above 30%, and sought the help of two prevention scientists, Dennis Embry and Anthony Biglan, to do something about it. Biglan and Embry accomplished their mission. How did they do it?

Their first step was to build a meaningful consensus against illegal sales. Biglan and Embry made the rounds among key legislators, state department heads, and other important people to stress the need for action. Even though most of these people had a genuine interest in the long-term welfare of their constituents, the immediate danger of losing millions of dollars in federal support was a more powerful incentive. Anti-tobacco organizations and other stakeholders were also brought into the process, resulting in a declaration endorsed by leaders at the state level that could then be endorsed by leaders at each locality within the state.

The declaration was publicized by an advertising campaign using the same techniques that are effective at marketing cigarettes—social branding, rather than product branding. TV and radio commercials portrayed a convenience store clerk being

rewarded for doing the right thing. Slogans were invented such as "Wyoming Wins!" Political figures and celebrities endorsed the cause. Owners of retail outlets were informed of the consensus and provided with materials to distribute to their clerks.

All of this was required to establish the criteria for selecting behaviors, much like ACT at the individual level and the GBG in a single classroom. Much more effort was required to meaningfully establish a consensus at the scale of an entire state but it could still be done, as Biglan and Embry were able to demonstrate.

Now that "the right thing" was clear in everyone's mind, the next task was to reinforce the right thing by making our psychological mechanisms for learning and copying work for us rather than against us. Biglan and Embry created task forces with their own underage secret agents who attempted to buy cigarettes. Clerks who turned them away were richly rewarded with coupons from local businesses, articles in the local newspaper, and their picture on the wall of the store. Clerks who obliged were mildly punished with a reminder to uphold the law. Biglan and Embry also held a contest among the Wisconsin clerks for the most clever thing to say when faced with a minor trying to buy cigarettes. The winning entries were printed in the form of cards that could be handed to the underage customers, which were provided to all the clerks—an exceptionally clever use of a variation-and-selection process to discover and spread best practices.

The program was rigorously assessed and highly effective at reducing cigarette sales to minors. Baseline information gathered before the intervention reported average rates of illegal sales of tobacco of 43% in Wyoming and 35% in Wisconsin. After the intervention, those numbers declined to 10.8% and 8.1%, where they have remained stable to the present day. Even better, reducing illegal sale of tobacco directly to minors was effective at reducing their smoking rate; they did not entirely make up for it by obtaining tobacco from other sources.

What Biglan and Embry accomplished at a statewide scale takes place naturally at a small scale. For our hunter-gatherer ancestors, most challenges to survival were obvious, a consensus was established around the campfire, and social rewards and punishment took place through the spontaneous expression of emotions. What comes naturally at a small scale does not happen automatically at a large scale. Something must be constructed at a large scale that interfaces with our genetically evolved instincts for learning and copying. If that "something" isn't added, then large-scale society cannot be expected to function well. Biglan and Embry had a clear idea of what to do to make a large society function like a small group, preventing thousands of smoking-related deaths over the long term. How many other problems faced by large-scale society might be solved in the same way?

7.8 Summary

The idea that evolution accounts for our physical bodies and a few basic impulses but has nothing to say about our rich behavioral and cultural diversity is bizarre in retrospect. Once our capacity for change is seen as a sophisticated product of genetic

evolution and a collection of fast-paced evolutionary processes in their own right, every branch of knowledge about humans is brought inside the orbit of evolutionary science.

How radical a transformation will this be? The study of every human-related subject is sophisticated in its own right and has resulted in the accumulation of durable knowledge. Perhaps this knowledge is consistent with evolutionary theory, even if evolutionary theory was not explicitly invoked. If so, then approaching a given subject from an evolutionary perspective will merely result in reinventing the wheel.

This will sometimes be the case. As we have seen, the applied human sciences offer outstanding examples of intentional behavioral and cultural change that were developed without explicit reference to evolution. However, it will not *always* be the case. Anyone familiar with the human-related disciplines knows that they are a kaleidoscope of perspectives that are not consistent with each other, much less an overarching evolutionary perspective. The implicit assumption that "what I think is consistent with evolution without requiring much knowledge about evolution" will often prove to be false. Adopting an explicit evolutionary perspective will therefore result in new insights for each discipline and a unification of disciplines that has not occurred otherwise.

The situation is similar to biological knowledge in Darwin's day. A great deal of information had accumulated and much of it was accurate, but it wasn't organized so that every branch of knowledge could be interrelated with every other branch. Darwin provided the organizing framework, whereby all aspects of life could be understood in terms of "the same laws acting around us", as he put it at the end of the *Origin of Species*. The integration that took place in the biological sciences during the 20th century (and continuing) is now in progress for our knowledge of humanity. Not only is this an exciting intellectual prospect, but it provides tools for improving the quality of human life in a practical sense. I hope that this sketch will encourage the reader to become involved in the integration that is already in progress.

Acknowledgements This essay is a sketch of a more comprehensive article that will be coauthored with Tony Biglan, Dennis Embry, and Steve Hayes, whose work is featured in Section 7.7. The same themes are presented in trade book form in Wilson (2011).

References

Biglan, A., & Hinds, E. (2009): Evolving Prosocial and Sustainable Neighborhoods and Communities. *Annual Review of Clinical Psychology* 5: 169–196.

Bingham, P. M. (1999): Human Uniqueness: A general theory. *Quarterly Review of Biology* 74: 133–169.

Boehm, C. (1999): *Hierarchy in the Forest: Egalitarianism and the Evolution of Human Altruism*. Cambridge, Mass: Harvard University Press.

Calvin, W. H. (1987): The brain as a Darwin machine. *Nature* 330: 33–34.

Campbell, T. D. (1960): Blind variation and selective retention in creative thought and other knowledge processes. *Psychological Review* 67: 380–400.

Crews, D., Bergeron, J. M., Bull, J. J., Flores, D., Tousignant, A., Skipper, J. K., et al. (1994): Temperature-dependent sex determination in reptiles: Proximate mechanisms, ultimate outcomes, and practical applications. *Developmental Genetics* 15: 297–312.

Deacon, T. W. (1998): *The Symbolic Species*. New York: Norton.

Devlin, R. H., & Nagahama, Y. (2002): Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. *Aquaculture* 208: 191–364.

Diamond, J. (1997): *Guns, Germs, and Steel*. New York: Norton.

Ehrenreich, B., and J. McIntosh (1997): The New Creationism: Biology under Attack. *The Nation*, June 9, 1997: 11–16.

Embry, D. D. (2002): The good behavior game: a best practice candidate as a universal behavioral vaccine. *Clinical Child and Family Psychology Review* 5: 273–297.

Embry, D. D., Biglan, A., Galloway, D., McDaniels, R., Nunez, N., Dahl, M. J., et al. (2010): Reward and Reminder Visits to Reduce Tobacco Sales to Young People: A Multiple-baseline across two states. Unpublished manuscript.

Gould, S. J. (2007): *The Essential Stephen Jay Gould*. New York: Norton.

Greene, E. (1996): Effect of light quality and larval diet on morph induction in the polymorphic caterpillar Nemoria arizonaria (Lepidoptera, Geometridae). *Biological Journal of the Linnean Society* 58: 277–285.

Hayes, S. C. (2004): Acceptance and commitment therapy, relational frame theory and the third wave of behavioral and cognitive therapies. *Behavior Therapy* 35: 639–665.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001): *Relational Frame Theory: A Post-Skinnerian Account of Human Language and Cognition*. New York: Springer.

Holldöbler, B., & Wilson, E. O. (2008): *The Superorganisms*. New York: Norton.

Huxley, J. S. (1947): *Man in the Modern World*. London: Chatto & Windus.

Huxley, J. S. (1969): *Essays of a Humanist*. New York: Penguin.

Jablonka, E., & Lamb, M. J. (2006): *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Cambridge, MA: MIT Press.

Lemov, R. (2005): *World as Laboratory: Experiments With Mice, Mazes, and Men*. New York: Hill and Wang.

Luyben, P. D. (2009): Applied Behavior Analysis: Understanding and Changing Behavior in the Community–A Representative Review. *Journal of Prevention and Intervention in the Community* 37: 230–253.

MacPhee, R. D. E., & Sues, H.-D. (Eds.). (1999): *Extinctions in Near Time: Causes, Contexts, and Consequences*. New York: Springer.

Maynard Smith, J., & Szathmáry, E. (1995): *The Major Transitions in Evolution*. New York: W.H. Freeman.

Maynard Smith, J., & Szathmáry, E. (1999): *The Origins of Life: From the Birth of Life to the Origin of Language*. Oxford: Oxford University Press.

Pagel, M., & Mace, R. (2004): The cultural wealth of nations. *Nature* 428: 275–278.

Pigliucci, M. (2001): *Phenotypic Plasticity: Beyond Nature and Nature*. Baltimore, MD: Johns Hopkins University Press.

Plotkin, H. (1994): *Darwin Machines and the Nature of Knowledge*. Cambridge, MA:
Harvard University Press.

Relyea, R. A. (2002): Local population differences in phenotypic plasticity: predator-induced changes in wood tadpoles. *Ecological Monographs* 72: 77–93.

Seeley, T. (1995): *The Wisdom of the Hive*. Cambridge, Mass.: Harvard University Press.

Skinner, B. F. (1981): Selection by Consequences. *Science* 213: 501–504.

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

Sompayrac, L. M. (2008): *How the Immune System Works (Third Edition)*. Hoboken, NJ: Wiley-Blackwell.

Tadokoro, T., Yamaguchi, Y., Batzer, J., Coelho, S. G., Zmudzka, B. Z., Miller, S. A., Wolber, R., Beer, J. Z. and Hearing, V. J. (2005): Mechanisms of Skin Tanning in Different Racial/Ethnic Groups in Response to Ultraviolet Radiation. *Journal of Investigative Dermatology* 124: 1326–1332.

Tomasello, M. (2009): *Why We Cooperate*. Boston: MIT Press.
Tomasello, M., Carpenter, J., Call, J., Behne, T., & Moll, H. (2005): Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and Brain Sciences* 28: 675–735.
West-Eberhard, M. J. (2003): *Developmental Plasticity and Evolution*. Oxford: Oxford University Press.
Wilkinson, R., & Pickett, K. (2009): *The Spirit Level: Why Greater Equality Makes Societies Stronger*. London: Bloomsbury Press.
Wilson, D. S. (2002): *Darwin's Cathedral: Evolution, Religion and the Nature of Society*. Chicago, Ill.: University of Chicago Press.
Wilson, D. S. (2005): Evolutionary Social Constructivism. In J. Gottschall and D.S. Wilson (Eds): *The Literary Animal: Evolution and the Nature of Narrative*. Evanston, Ill.: Northwestern University Press, pp. 20–37.
Wilson, D. S. (2006): Human groups as adaptive units: toward a permanent consensus. In P. Carruthers, S. Laurence & S. Stich (Eds.): *The Innate Mind: Culture and Cognition*. Oxford: Oxford University Press, pp. 78–90.
Wilson, D. S. (2007): *Evolution for Everyone: How Darwin's Theory Can Change the Way We Think About Our Lives*. New York: Delacorte.
Wilson, D. S. (2010a): Learning from the Immune System about Evolutionary Psychology. *Evolutionary Review* 1: 13–17.
Wilson, D. S. (2010b): The Truth is Sacred. In G. Levine (Ed.), *The Joy of Secularism: Eleven Essays For How We Live Now*. Princeton: Princeton University Press.
Wilson, D. S. (2011): *The Neighborhood Project: Using Evolution to Improve My City, One Block at a Time*. New York: Little, Brown
Wilson, D. S., Van Vugt, M., & O'Gorman, R. (2008): Multilevel selection and major evolutionary transitions: implications for psychological science. *Current Directions in Psychological Science* 17: 6–9.
Wilson, D. S., & Wilson, E. O. (2007): Rethinking the theoretical foundation of sociobiology. *Quarterly Review of Biology* 82: 327–348.
Wilson, M. & Daly, M. (1997): Life expectancy, economic inequality, homicide, and reproductive timing in Chicago neighborhoods. *British Medical Journal* 314: 1271–1274.
Yazdanbakhsh, M., Kremsner, P. G. & van Ree, R. (2002): Allergy, Parasites, and the Hygiene Hypothesis. *Science* 296: 490–494.

Chapter 8
Human Artistic Behaviour: Adaptation, Byproduct, or Cultural Group Selection?

Johan De Smedt and Helen De Cruz

8.1 Art as a Human Universal

One morning, when writer Elizabeth Gibson was on her way for coffee as usual, she spotted a conspicuous and colourful canvas in a pile of rubbish. Although she knew nothing of modern art, she felt compelled to take the painting to her crammed Manhattan apartment because, as she put it, "it had a strange power". The canvas hung for several years in her flat until she discovered that it was actually the famed work *Tres Personajes* by the Mexican painter Rufino Tamayo, stolen some twenty years before. After realizing its value, Gibson returned the picture to its rightful owners. This anecdote illustrates that we have an intuitive concept of art—even without any formal training in aesthetics or art history, we recognize art when we see it. Indeed, experimental studies (e.g., Seifert, 1992) reveal that Western college students without any formal training in art display and freely express aesthetic sensitivities to works of visual art, even if they are unfamiliar with them, like African sculpture.

What is it that we see and intuit in works that we denote as 'art'? This is one of the most outstanding problems in contemporary philosophy of art, and attempting a solution to this problem falls outside the scope of this paper. Objects and performances that we routinely classify as art share features like skill, strikingness and

J. De Smedt (✉)
Department of Philosophy and Ethics, Ghent University,
Blandijnberg 2, 9000 Ghent, Belgium
e-mail: johan.desmedt@ugent.be

H. De Cruz
Centre for Logic and Analytical Philosophy, Katholieke Universiteit Leuven,
Kardinaal Mercierplein 2, 3000 Leuven, Belgium
e-mail: Helen.DeCruz@hiw.kuleuven.be

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*,
Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_8,

beauty, but it is easy to come up with counterexamples for each of these features; for example, ready-mades do not clearly exhibit artistic skill. Some philosophers of art (e.g., Dutton, 2006) have therefore proposed to take only unproblematic cases to guide any definition of art. Others (e.g., Gaut, 2005) propose a cluster concept of art, where an art object can have several characterizing features, but where none of these is necessary, while some suggest a plurality of art concepts (Mag Uidhir and Magnus, 2011). All these approaches have in common that they focus on the objects, rather than on the causes of these objects.

Instead of taking the art objects as a starting point, we examine the human cognitive faculties and behaviours that are responsible for the creation and enjoyment of these objects. This shift in focus allows us to include objects and performances from distant places and cultures. It is an oft-stated truism that other cultures do not have a term equivalent to our western notion of art for art's sake. Yet although Hellenistic sculptors, Gothic architects and Melanesian wood carvers did not possess the modern western concept of art, we readily appreciate and appropriate their work. And just as sculptures from sub-Saharan Africa and Oceania adorn western homes, artists from these cultures have eagerly adopted western styles and media. In his inventory of human universal characteristics, Brown (1991) cites art, including music, dance, oral or written literature, visual art and performance. It occurs in complex societies as well as in societies with very little material culture, where it often appears in the shape of beads or other forms of body decoration. As will be expounded later, forms of body decoration are also found in great quantities in prehistoric hunter-gatherer living sites. Interestingly, once we move away from the western concept of art for art's sake, and focus on human behaviour, the similarities between western art production and the production of objects and performances in other cultures become apparent.

The universality of artistic behaviour across cultures seems to warrant an explanation in biological terms (Carroll, 2004). This view is strengthened by the fact that both the ability to create and to appreciate art arise remarkably early in development. From the age of about two years onwards, young children spontaneously engage in singing, dancing and drawing, and they move and vocalize to music even before their first birthday. Although they are not skilled artists, toddlers nevertheless name their drawings using the same names as the real-world objects that capture their interest, such as 'cat' or 'daddy'. As Bloom (2000) has remarked, these early works are similar to those of adult artists in that both the artist and the child take an intentional perspective towards categorizing and naming the artwork. Also, like adults, children as young as two years take the intention of the maker when they name a drawing. For example, when they witness an adult drawing a circle that could be either of two unfamiliar disc-shaped objects, they take the gaze direction of the artist as a cue for which of the items was depicted. The toddlers reliably point at the object that the adult was looking at when asked which object was being depicted (Preissler and Bloom, 2008). Slightly older children also assume this stance for their own work: when one asks four-year-olds to draw a picture of a lollipop and a balloon, the drawings look virtually identical. Yet the children will consistently refer to the pictures according to what they intended to depict when they produced the drawings (Bloom and Markson, 1998).

Although representational visual art is not produced in all cultures, several systematic studies have shown that people unfamiliar with fairly abstract, pictorial representations can recognize these images spontaneously. An early study (Hochberg and Brooks, 1962) focused on a western child, brought up without exposure to any pictorial representations, such as picture books, television or figurative wallpaper. At 19 months, the boy was able to recognize and reliably name drawings made by others of his toys and common objects. Deregowski et al. (1972) showed line drawings of fairly complex scenes, such as a hunter stalking a goat, to people from an Ethiopian culture without pictures or drawings. Again, these people recognized and named the drawings correctly. Martlew and Connolly (1996) asked children from a Papua New Guinean culture without figurative art or access to photography to draw a man. Although the children had never produced drawings before, they drew recognizable anthropomorphic figures, often resembling the stick figures made by western three-year-olds.

8.2 Is Art an Adaptation?

8.2.1 Adaptationist Explanations of Art

The universality of art across cultures, our ability to recognize and appreciate it and its early emergence in development seem to suggest that producing and enjoying art may be a stable part of human cognition. There are two possible evolutionary explanations for this: either it is an adaptation, which has evolved in direct response to one or more selective pressures in our ancestral past, or it is a byproduct of other adaptations without being adaptive in itself. Its complexity makes it implausible that artistic behaviour would have evolved through random genetic drift, which is the only other explanation in evolutionary terms at the level of the individual organism.

Those who favour the view that art is an adaptation invoke its universality across cultures, its costliness, and its early and spontaneous development in children. Miller (2000) argues that art and other forms of human creative behaviour evolved as the result of sexual selection: their costliness in terms of time and energy provided ancestral hominid females with an honest signal of the fitness of the art-producing male. Just like a lush but burdensome tail in peacocks or birds of paradise are good signals of their owners' qualities to live with such a handicap, the artist's works are honest signals of his qualities as a mate. Tooby and Cosmides (2001) point out that pretend play emerges universally in toddlers. They argue that this ability provides us with the imagined worlds of (oral) literature and visual art, risk-free environments where learning can take place through vicarious experience. Dissanayake (2000) proposes that art is the intentional act of making everyday behaviour special through exaggeration, formalization, or manipulation of expectations: dance exaggerates and formalizes normal bodily movements; songs distort normal speech and prosody. Performing these actions in groups relieves tension

and anxiety, thus improving social bonds within the community. She traces the evolutionary precursor to these behaviours to mother-infant dyadic interactions, where mothers and infants spontaneously modify their vocalizations, facial expressions and gestures.

8.2.2 Problems With Adaptationist Explanations of Art

Clearly, it is not difficult to imagine adaptive functions for art, but that is exactly the problem of such adaptationist accounts—theorizing about them remains fairly unconstrained. Also, the category of objects that is being explained is wider than what we conceive of as art. Miller explains not only art, but also humour and even conspicuous consumption. Tooby and Cosmides themselves point out that their adaptive account is about fiction, the broad human ability to imagine counterfactual worlds and situations, rather than about art specifically. Dissanayake provides an explanation not only for art but also for ritual and even ritualized behaviour, which is not restricted to humans, but can be observed in many animals living in captivity.

Another potential problem with the adaptationist view of art is that the neural structures responsible for artistic behaviour would have to be modularly organized. If artistic behaviour is directly targeted by natural selection, we expect its organization in the brain to be modular. The evolvability argument, developed by biologists like Lewontin (1978) and philosophers like Wimsatt (2001) and Sterelny (2004), holds that unless cognition is to some important degree modular, it is incapable of evolving away from its current organization. In a nonmodular brain, a change in one component will be connected to many other changes, thus the slightest modification might have disastrous effects for the organism. Only modularly organized cognitive capacities can evolve without affecting the rest of the brain. Although the extent to which the human brain is modularly organized is subject to debate, most evolutionary psychologists endorse a modular conception of the human mind (see e.g., Cosmides and Tooby (1994) for a theoretical discussion of the central position of modularity in the evolutionary psychological research programme). Moreover, if a given capacity is modular, evolutionary psychologists often take this to be a strong indication of its adaptive value. From the perspective of evolutionary psychology, one would therefore expect that cognitive faculties that evolved through natural or sexual selection are modularly organized.

The most straightforward way to find out if a given cognitive faculty is modularly organized is to examine whether it consistently activates the same network of neural circuits. Tasks that probe our theory of mind, for example, consistently activate the same network of neural circuits, including the medial prefrontal cortex, superior temporal sulcus, and temporal poles across a wide diversity of mentalizing tasks, such as hearing stories, seeing objects move intentionally across a screen and interpreting cartoons (Gallagher and Frith, 2003). However, a series of independent neuroimaging studies indicates that perceiving art or engaging in artistic

behaviour does not yield a consistent activation of the same neural network. Instead, different forms of art recruit different neural pathways. A PET study of subjects who tango (Brown et al., 2006) revealed that dance involves a network of neural circuits normally involved in ordinary bipedal locomotion and the organization of complex sequences of movements. In contrast, music exploits modules normally involved in auditory processing. Remarkably, New World monkeys that do not produce music themselves can distinguish between atonal and tonal melodies, and can recognize a melody played in different keys (Hauser and McDermott, 2003). Thus, music likely exploits auditory sensitivities that are phylogenetically ancient and that did not evolve for music appreciation. Within visual art, different kinds of works elicit different sensory responses. While the pure forms and shapes of Piet Mondriaan and Kazimir Malevich activate orientation-selective cells in the primary visual system that respond selectively to straight lines (Zeki, 1999), kinetic art, such as Jean Tinguely's mobiles, targets the motion-sensitive cells of area V5 (Zeki and Lamb, 1994).

A comparison of the regions of interest (i.e., regions where most neural activity takes place) reveals that there is no area common to all forms of art perception, hence current cognitive neuroscience has not detected a specialized art centre in the brain. Rather, art hijacks the properties of the normal perceptual neural circuits. Lesion studies of visual artists provide an equally compelling case: art production seems to continue irrespective of the location or extent of the lesions in the artists' brain (see Zaidel (2005) for a comprehensive overview). Remarkably, some cases of brain damage even lead to the emergence of artistic skills in individuals who previously did not engage in artistic behaviour: patients with fronto-temporal dementia, who as a result of this have impaired linguistic and social skills but spared manual and visual capacities, sometimes start painting obsessively and produce impressive works of art (Miller et al., 1998). Some cases of brain damage can lead to changes in style in artistic production: an Asian-American artist, for instance, who suffered from fronto-temporal dementia, evolved from conventional Chinese-style paintings to expressionist and fauvist-like works as her illness progressed (Mell et al., 2003). If artistic behaviour just exploits brain circuits that fulfil normal functions, and if it persists despite various forms of brain damage, it seems rather doubtful that it would be a biological adaptation.

8.3 Is Art a Byproduct?

8.3.1 Byproduct Explanations of Art

Some evolutionary psychologists propose that art is not an adaptation, but a byproduct. One of the most influential proponents of this view is Pinker (1997, 524–525), who argues that art's primary purpose is "to press our pleasure buttons". Art exploits aesthetic preferences that were adaptive in other contexts, just like cheesecake gratifies our ancestral craving for sugar and fat. Indeed, Blood and Zatorre (2001) have

shown that aesthetically pleasing stimuli activate reward-based emotional circuits in the brain: participants who listened to their favourite music showed stronger activation in reward and motivation-related brain areas compared to control compositions. Subjects who look at paintings they deem beautiful, activate reward-based emotional circuits compared to duller paintings (Vartanian and Goel, 2004). A plausible reason why art should press our pleasure buttons is that artworks exploit evolved tendencies of the human brain, such as sensitivity to clear lines and marked colour contrasts that is characteristic of the primate visual system. Because of their importance to the survival and reproduction of the organism, some cues are given priority by the early perceptual systems. Ramachandran and Hirstein (1999) propose that visual artists maximally exploit these tendencies, thereby eliciting strong emotional responses. This would explain why works of art typically tap into several normal perceptual input systems and why they activate reward-based neural circuits. The cross-cultural prevalence of some art forms can be explained by their efficient exploitation of our cognitive predispositions. Newborns, for example, can already discriminate faces from other objects by detecting the shadowy patches created by the eye sockets and the mouth. Masks across the world exaggerate these facial features. Thus, they act as a superstimulus for our face recognition module, a compelling explanation for the use of masks in many cultures past and present across the globe (Sperber and Hirschfeld, 2004).

8.3.2 Problems With Byproduct Explanations of Art

At first blush, the byproduct explanation for art seems more cogent than the adaptationist approach. However, it faces two important problems. First, it predicts that we would invariably prefer works of art that maximally conform to evolved aesthetic preferences. Yet academic art by painters like William Adolphe Bougereau and Jean-Antoine Watteau which generally responds to our evolved tastes in depicting attractive people in lush landscapes, is nowadays derisively referred to as overpolished and clichéd. Experimental studies (Martindale, 1998) indicate that the lay public prefer academic art, and this is in line with byproduct explanations. But this does not explain the enduring appeal of works of visual art that are hardly eye candy, such as Francisco Goya's gloomy political canvases or Francis Bacon's haunting papal portraits.

Second, the costs in terms of time, material and energy that art requires seem difficult to reconcile with byproduct accounts. Pinker's view might seem plausible in the light of modern society, where we are constantly immersed in music and visual and narrative art, but art emerged within Palaeolithic hunter-gatherer societies where artists could not afford to live exclusively from their work but were hunting, gathering, building camps and crafting tools like everyone else. Among the oldest examples of representational art are mammoth ivory figurines from Swabia, Germany dated at about 35,000 years ago (Conard, 2003). Due to the growth structure of mammoth tusks, this material is notably difficult to work with and each of

these tiny figures probably took several days to make. Likewise, in contemporary small-scale societies, such as hunter-gatherer and horticulturalist groups, people put a disproportionate amount of time and energy in the production of art. Why do these costly behaviours persist over such long stretches of time, when we would expect strong selective forces operating against such wastes of time and energy? To date, byproduct explanations of art have not addressed this question.

8.4 A Cultural Group-Selectionist Alternative

8.4.1 What is Cultural Group Selection?

At present, neither the adaptationist nor the byproduct account can adequately explain the data. While this by itself is not a reason to reject such explanations out of hand, it does provide room for an alternative explanation that will be explored here. This explanation, we will argue, accords well with empirical observations of art use in contemporary small-scale societies and with the archaeological record of Palaeolithic art. According to this explanation, some forms of art evolved through cultural group selection, in particular as a means to emphasize within-group identity. In the remainder of the paper, we will examine the proliferation of some forms of art in the Late Pleistocene (in particular, the mobiliary art from the Magdalenian) by appeal to cultural group selection. Group selection was originally proposed as a mechanism to explain altruism (e.g., Wynne-Edwards 1962). In this view, groups composed of altruists do better as a whole than groups composed of selfish individuals, favouring the retention of altruistic behaviour. In the second half of the 20th century, prominent evolutionary theorists like John Maynard Smith (1964) and George Williams (1966) argued that the assumptions on which group selection relies are very implausible. For one thing, altruistic groups are vulnerable to subversion from within—given that a single cheater within a group of altruists has higher fitness, this individual's genetic success will far outstrip the success of the altruists, as the latter have costs as well as benefits. Moreover, the replicators in biological evolution are genes, and it turns out that most instances of altruism can be explained in terms of the propagation of these genes (kin selection). As a result, group selectionist ideas fell on hard times in mainstream evolutionary theory. While no one claimed that group selection was inherently impossible, it was argued that special conditions need to be fulfilled before it could work. Group selectionist ideas are making a comeback, both in theories of altruism in the natural world (e.g., Wilson and Hölldobler, 2005) and in models of the cultural evolution of human altruistic behaviour (e.g., Henrich, 2004). Cumulative culture, which gives humans the capacity to transmit complex behavioural traits at a fast rate, indeed creates a set of special circumstances that might allow for group selection to occur.

A sensible way to interpret group selection is to see it as claiming that groups can fulfil the same role as organisms. In mainstream evolutionary theory, a distinction is made between *replicators* (genes) and *vehicles* (entities that interact with

the environment). Genes can interact as cohesive wholes with their environment through their vehicles, typically organisms. Thus the behaviour of a given vehicle has direct consequences for its replicators: the vehicles' differential reproductive success ultimately causes the reproductive success of their replicators, thereby making them important units of selection (Sterelny, 1996). From this, it already becomes intuitively clear that groups must be distinct from each other and form cohesive wholes for group selection to occur.

Group selection requires that the fitness benefits of altruistic groups over selfish groups must outweigh the fitness benefits of selfish individuals over altruistic individuals within mixed groups. This condition can be mathematically described using the Price equation (Price, 1972), which provides a formal way to study changes in the frequency of heritable traits at two levels. In this case, we are interested to find out whether the benefits of art for the group (i.e., all members of the group together, including nonproducers) is greater than the fitness costs of the production of art by individual members of the group. The Price equation is a statistical statement that relates the expected change in the frequency of a gene or cultural trait ($\Delta\overline{x}$) per generation, the absolute fitness W_j, and the current frequency of the trait x_j. We start with a population of N individuals subdivided into groups indexed by j, each with n_j members. There are no restrictions on how the groups are composed, except that all groups must contain at least one individual.

$$\overline{w}\Delta\overline{x} = \overbrace{Cov(w_j, x_j)}^{between\ groups} + \overbrace{E(W_j \Delta x_j)}^{within\ groups} \tag{1}$$

The first term on the right side of equation 1 represents the relationship between the fitness of the groups and the initial frequency of the culturally transmitted trait within them, i.e., what is the effect of having this trait in the group as a whole as compared to other groups. The second part represents the expected changes in this trait, based on its impact on the fitness of individual members of the group. Given that covariance expresses the product of a variance and a regression coefficient (β), we can rewrite the Price equation as follows (simplifying by ignoring factors like mutation and recombination):

$$\overline{w}\Delta\overline{x} = \beta_{w_j, x_j} Var(x_j) + E\left(\beta_{w_{ij}, x_{ij}} Var(x_{ij})\right) \tag{2}$$

The two terms on the right side of the equations 1 and 2 oppose each other, since altruism increases group fitness but decreases individual fitness to a certain extent. If most of the variance in the population is within the group, but all groups have nearly the same frequency of the culturally transmitted traits, then the variation between groups $Var\ (x_j)$ will be very small, whereas the expectation of the variation within groups $Var\ (x_{ij})$ will be nearly the entire variance of the population. In this case, cultural traits that favour altruism will not be maintained. If groups can be isolated from each other, the variance between groups can become larger than the variance within groups due to cultural drift, which provides an ideal basis for the development of altruistic behaviour. Cultural drift is the emergence and spread

of cultural elements that arise by chance within a given group and that are copied randomly by members of that group. This is a mechanism that results in between-group differences when groups are sufficiently isolated. However, frequent contact between groups and migration can quickly undermine this: behavioural traits from one group can percolate into another, which increases variation within groups *Var* (x_{ij}) at the expense of variation between groups *Var* (x_j). Cultural mechanisms that enable humans to mark group identity and to maintain between-group differences can counter these effects, giving rise to within-group altruistic behaviour. Subversion from within is routinely countered by social rules, such as altruistic punishment (Fehr and Gächter, 2002), that discourage selfish behaviour and non-conformism. Furthermore, the presence of conformists dramatically increases the group size for which cooperation can be sustained (Guzmán et al., 2007).

Henrich's (2004) derivation of the Price equation (see equation 3) also shows that group selection only works if the benefits of being in an altruistic group outweigh the costs of bestowing benefits to other members of the group:

$$\beta_{w_i x_i} + \beta_{w_i x_j} \beta_{x_j x_i} > 0 \tag{3}$$

The first term of equation 3 is always positive, as it models the benefits of being in an altruistic group. The second term is always negative, because it represents the costs of bestowing benefits to others. Obviously, the sum of both terms needs to be larger than 0 for cultural group selection to occur.

Cultural group selection is one type of group selection in which the group is defined through cultural markers, such as distinct language or dialect, religious beliefs, dress code, food taboos, or other cultural norms. Cultural groups are fairly stable because people have a conformist bias: they tend to follow the norms of the culture in which they were raised. This conformist tendency is well attested ethnographically (Richerson and Boyd, 2005; Tehrani and Collard, 2002) and archaeologically (Collard et al., 2006), in the way material culture tends to evolve together with a particular ethnic group. As groups are culturally, rather than genetically, defined, and given that such culturally defined groups are fairly stable, cultural group selection can be invoked to explain human prosociality, i.e., the exceptional degree of cooperation and altruism found within most human societies. Rather than explaining this through genetic changes, one could argue that human culture, with its ability to differentiate groups from each other, allowed for the formation of distinct groups that each have their own norms and cultural practices (Henrich, 2004). As we have seen above, once stable groups have been formed, altruistic behaviour can be favoured within such groups, and the individuals within such groups will have higher reproductive success compared to members of other groups.

During the Late Pleistocene (126,000-10,000 years ago) members of *Homo sapiens* began to create various forms of material culture that, because of their aesthetic properties and putative symbolic value, are often referred to as visual art. Unfortunately, the archaeological record does not provide reliable evidence for music until much later, namely the recovery of flutes made of bird bone and mammoth ivory of Aurignacian sites in southwest Germany, dated at about 36,000 years

ago (Conard et al., 2009). The evidence for dance can be indirectly inferred, from Magdalenian representations of dancers, for instance in the Grotte des trois frères. Because of the poor archaeological preservation of music and dance, we will here concentrate on visual material culture. The earliest convincing examples of visual art are in the form of body decoration, in particular shell beads from Israel and Algeria, dated to 135,000-100,000 years ago (Vanhaeren et al., 2006), shell beads from Blombos cave, South Africa, dated to 75,000 years ago (Henshilwood et al., 2004) and ostrich eggshell beads from Kenya, dated to 50,000 years ago (Ambrose, 1998). As we will explain in more detail later, there are good reasons why the earliest art is in the form of body decoration. For reasons of space and clarity, we leave aside the engraved ochre artefacts from Blombos and other South African sites, as their status as art objects is still contested within the archaeological community. Figurative art, such as figurines, paintings and engravings, appear somewhat later still.

Although it remains unclear whether cognitive or cultural changes lie at the basis of this transition, theoretical models (e.g., Powell et al., 2009) indicate that cultural changes brought about by different patterns of interaction and population density can explain the emergence of art without the need to invoke new cognitive capacities due to genetic mutations. A cultural account of art is also supported by the fact that different art forms (e.g., musical instruments, beads, rock paintings, engravings) emerged independently at different time periods across the world, a pattern that cannot be explained by gaps in the archaeological record alone. To give but one example, figurative paintings on rock surfaces appear significantly earlier in Europe (about 33,000 years ago, in Chauvet cave, France) than in Africa (about 27–25,000 years ago, Apollo 11 cave in Namibia (Conard, 2003)), or in Australia (about 17,500 years ago, Kimberley region, northern Australia (Roberts et al., 1997)). We here propose that some forms of Palaeolithic art, in particular mobiliary art and body decoration, could have been invented as a way to signal group identity which allows for a differentiation between groups, an essential condition for cultural group selection to occur. We will now consider two theoretical models to explain in detail this signalling function of art: green beards and ethnic markers. We will pit these models against the archaeological record to determine how useful they are for explaining the emergence of some forms of art.

8.4.2 Green Beards

Art may have been used as a conspicuous tag to signal altruism directly. In theoretical models such tags are often referred to as *green beards*: if green-bearded creatures bestow their altruism exclusively on fellow green beards, natural selection will promote the presence of the tag as well as the altruism. This theoretical framework can be easily extended to cultural evolution. Simulations (e.g., Riolo et al., 2001) indicate that cooperation can evolve easily in a population of agents who follow the simple rule "cooperate with others who bear the same tag as you". But as Dawkins (1989) already recognized, green beard altruism can be undermined by cheaters, who show the tag

Table 8.1 Different fitness outcomes of signallers and non-signallers, adapted from McElreath and Boyd 2007, p. 202

Genotype	Phenotype	Fitness
NN	Non-altruist, no green beard	w_0
NG	Non-altruist, green beard	$pb + w_0$
AN	Altruist, no green beard	$q(-c) + w_0$
AG	Altruist, green beard	$pb - cq + w_0$

but are not altruistic. The inherent instability of green beards has been demonstrated extensively in models of biological forms of green beard, where the linkage between the allele that signals the altruism (*A*) and the allele that codes for the display of the altruistic trait (*G*) gets disrupted (McElreath and Boyd, 2007). The possible combinations of such genes in a haploid organism are summarized in table 1.

Here, fitness outcomes are calculated as follows: genotype *NN* represents baseline fitness w_0, p is the frequency of altruists in the population, b is the benefit one receives from an altruistic donor. Given that signalling nonaltruists *NG* can always expect to get b, but that they do not incur costs (c), their benefit is $pb + W_0$. Altruists without signal (*AN*) are worst off, because they only suffer costs c whenever they encounter a potential recipient who signals, the frequency of which is given by q. Finally, altruistic signallers (*AG*) get benefit pb but have to pay cost qc. It is easy to see that genotype *NG* always has higher fitness than any other type as long as $c > 0$ and $b > 0$. Therefore, any process that breaks up the association between the tag G and the altruism A can result in an invasion of *NG*s thus eroding the value of the signal. Selective forces work against linkage between A and G. This association can be expressed as D (linkage disequilibrium). If A is associated with G, D is positive, if A and G are assorted at random, $D = 0$, and if A is paired with N, D will be negative. Because *NG* has the higher fitness, D will decline, until selection no longer favours the A allele. (Note that the selective force that breaks the linkage between the alleles coding for green beards and altruism does not play when green beards are rare alleles that are good proxies of relatedness. Due to the dynamics of kin selection, if two organisms that are reasonably closely related have the same rare marker, they can use this as a reliable indicator of relatedness.)

In cultural evolution, to counter this effect, one can change the tag regularly. Once a tag becomes too common, the chance increases that one encounters an organism with the tag but not the altruistic intentions. This can be mathematically expressed in equation 4:

$$\beta(p_j, q_j) = \frac{\frac{pq + D}{q} - p}{1 - q} = \frac{D}{q(1-q)} \tag{4}$$

Here p_j is the frequency of the altruism trait in the donor given the frequency of the green beard characteristic in the recipient q_j, p is the frequency of altruists in the population, and q is the frequency of green beards. As mentioned earlier, D expresses the association between green beard and altruism. One can see that the strength of cultural group selection through tags is proportional to the amount of D, but inversely

proportional to the variance of the green beard trait, in other words, rare markers work best. Additionally, green beards can repel cheaters if the tag is costlier to produce for cheaters than for cooperators. As simulations (e.g., van Baalen and Jansen, 2003) show, a population of agents that signal their altruism through green beards can withstand cheaters when the temptation to cheat is very low, i.e., when the costs of adopting the tag are very high.

When we pit these criteria against the archaeological record, it seems unlikely that green beards can be a good model for the evolution of Palaeolithic art. As art is not a genetic characteristic, *D* will not be high—there is no intrinsic reason why those who make and/or display art would be more altruistic than those who do not. Therefore, populations using art as a signal for altruism can be easily invaded by cheaters. Furthermore, as rare markers work best, green beard models predict that the signal for mutual altruism should be rare and subject to frequent stylistic turnovers. However, taphonomic and other destructive processes in consideration, Palaeolithic art is found in abundance. Next to this, art styles in the Upper Palaeolithic are remarkably stable in space and time: they are typically in use for several thousands of years with few stylistic changes over large areas (we refer to section 8.4.4. for an example). Although mobiliary art requires much effort to produce, it can be displayed by anyone. Several Upper Palaeolithic child burials have been found, where the individuals were covered with hundreds, or sometimes thousands of beads, each of which took considerable skill, time and energy to make—it seems unlikely that the children would have produced these beads themselves. The positioning of the beads suggests that they were attached to clothing, such as shoes, trousers or parkas, suggesting that the children did not receive them as exceptional grave gifts, but that they were part of their attire (Vanhaeren and d'Errico, 2005). Clearly, the person who made the beads and bore the costs of its production was not always the one who displayed the tag, and this association is a necessary condition for green beards to work. It is also not clear how mobiliary art could be less costly to produce for people who behave altruistically than for those who do not. In sum, green beard dynamics are an unlikely explanation for the emergence of Palaeolithic art.

8.4.3 Ethnic Markers

Like green beards, ethnic markers are easily recognizable tags that mutual altruists can use to exhibit or infer altruistic intentions. The crucial difference is that ethnic markers do not signal altruism per se, but provide information on an agent's behaviour during social interactions—they are a proxy for social norms and conventions, such as marriage rules, religion, or moral practices. As social norms and conventions are not readily observable, arbitrary characteristics, like hairstyle or dress code, can provide good indications for them. Meeting an individual with similar ethnic markers facilitates social interactions which can be conducive to cooperation. An influential illustration of how ethnic markers can work is Nettle and Dunbar's (1997) model of languages and dialects. Their simulation indicates that individuals with similar languages or dialects can cooperate better and as a result of this achieve higher

fitness. Given a limited memory-span, these individuals can withstand invasion from cheaters who speak the same language, especially given that cheaters need to relearn another language each time they are found out and have to move to another group where they are not known as cheaters. We will now examine how art could have been used as an ethnic marker. Cross-culturally, artistic ethnic markers are widely observed. Examples include decorated functional artefacts, where the style gives information about the ethnic group the owner belongs to, such as arrow point style as social information in Kalahari San (Wiessner, 1983). Artistic style enables individuals to distinguish people who belong to the ingroup from those who do not. Like dialects, artistic styles are difficult to imitate—it typically takes years for an artist to master a particular style.

Anthropological studies show that hunter-gatherers typically live in small bands of about 25 individuals; they are highly mobile within a large territory, moving on when resources are depleted. During parts of the year when resources are concentrated and abundant, these small groups aggregate with other bands that share their language, customs and beliefs. Group size is then between 200 and 800 individuals, depending on the capacity of the environment. During such seasonal aggregations, information, gifts and sexual partners are exchanged (Stein Mandryk, 1993). In the Upper Palaeolithic, we see the alternation between these group sizes in two types of sites: smaller residential sites with relatively little material culture, and larger sites with high concentrations of material culture. Altruism within small bands is widely attested in the ethnographic record in the form of food sharing (Hill, 2002) or alloparenting (Ivey, 2000). It can be easily explained by two well-established evolutionary mechanisms: kin selection (since most members of these small bands are related) and reciprocal altruism (since all members have social contact on a daily basis). There is also anthropological evidence that members of maximum bands help each other in times of hardship (Whallon, 2006). This type of altruism is much more difficult to explain through biological evolutionary mechanisms, since most people within the maximum band are not that closely related, and social contact between them typically takes place sporadically. Thus, kin selection and reciprocal altruism alone cannot explain why people from different small groups would help those of other groups.

From a behavioural ecological point of view, it is easy to understand why hunter-gatherers who live under marginal or unpredictable climatological circumstances, such as the present-day Inuit or the Kalahari !Kung, help each other to lessen the risk of local scarcity. When resources are unevenly spread in the landscape, small bands will sometimes starve before they find food. Under very difficult circumstances that are both cold and dry (the environment typical for Late Pleistocene Europe) it is not uncommon that 10% of the population dies of starvation each year (Stein Mandryk, 1993). This is a situation that is characteristic for Late Pleistocene Europe (126,000-10,000 years ago), where people mainly subsisted on herds of large animals, like reindeer, horse, mammoth and bison. Under these circumstances, where the main sources of food are unpredictable and patchy in distribution, inter-group contact and movement will become increasingly advantageous and necessary. Fruitless (wrong) moves can be lethal, leading to starvation and population decline. Not only do groups need information on where to find resources, they must also get access to them. These conditions set the stage for alliance networks between minimum bands, who

can through visits, gift-giving and other regular contacts exchange valuable information on resources, and help each other in times of need. This help can take the form of passive tolerance, for instance, allowing another group to trespass on their territory, or can consist of active food sharing (Whallon, 1989). Social security networks come with a set of defined rights and obligations that people can exercise when they are in need or that they must fulfil when others are in distress (Gamble, 1982; Whallon, 1989). Such mutualistic ties are widely attested in ethnographically documented hunter-gatherers from tundra and arctic environments, such as the Tareumiut and the Nunamiut Inuit in northwest Alaska (Minc, 1986), and desert and arid environments, such as the well-known *hxaro* network of the Kalahari hunter-gatherers.

How could such networks be maintained? Although face-to-face contacts can play an important role, they are limited to adjacent local groups, and cannot be used to establish relationships between individuals from groups that have little or no previous face-to-face contacts. The use of a tag turns out to be a stable strategy to signal social security network membership. The *hxaro* network of the Kalahari !Kung uses ostrich eggshell beads as gifts to keep their social security network up to date. Ostrich eggshell is difficult to obtain, because the eggs are jealously guarded by both parents who ferociously defend their brood. The shell is also notably difficult to work: it has to be fresh but nevertheless fractures easily. Interestingly, ostrich eggshell beads from the Kenyan Middle Stone Age site of Enkapune Ya Muto are among the oldest examples of uncontested body decoration, dated to about 50,000 years ago (Ambrose, 1998). Many of the beads broke prematurely and were discarded as waste, which shows how difficult it is to produce them. Other anthropological parallels of long-distance exchange networks include the Trobriand exchange of shell necklaces and bracelets in the Kula ring, or the exchange of woven mats by women from Tonga and surrounding archipelagos. Upper Palaeolithic Europe saw a prolific production of beads from mammoth ivory, tooth and shell. Interestingly, although some beads were found in burial contexts, most of them were found in living sites (White, 1982). These findings suggest that beads were part of the everyday attire of European Ice Age hunter-gatherers. The production of the beads and the acquisition of the raw materials required effort and time. Experimental archaeological studies (e.g., White, 1997) indicate that fashioning one mammoth ivory bead, as is found in Aurignacian western European sites, takes one to two hours. Some beads were made of shells that are found in sites up to 600 kilometres removed from the Atlantic or Mediterranean coasts (Whallon, 2006). Such high investments of time and energy can be explained when one interprets these objects as ethnic markers.

8.4.4 The Case of the Magdalenian

We will focus now on the Magdalenian, a European cultural complex, which presents a pertinent illustration of how art may have played an important role in maintaining social security networks. Although the Magdalenian spanned Europe from the Pyrenees to Poland and Ukraine, its material culture was remarkably invariant.

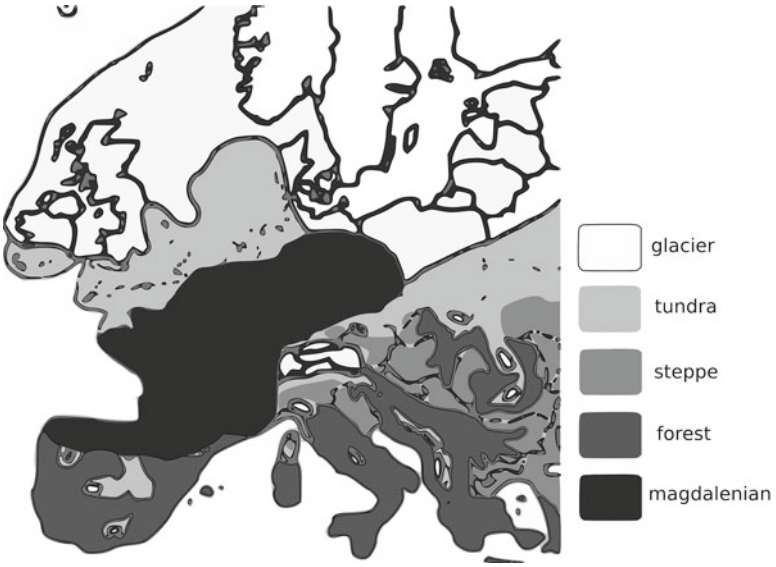

Fig. 8.1 Extent of vegetation types at the end of the Last Glacial Maximum and range of distribution of the Magdalenian, adapted from Jochim et al. (1999)

During the Last Glacial Maximum, which lasted from about 25,000 to 18,000 years ago, temperatures had plunged and ice sheets had expanded from Scandinavia and the Alps. Most of Europe was depopulated, because conditions were too harsh for human subsistence. Only southern France and northern Iberia were hospitable enough to maintain high population densities. From these regions, humans gradually recolonized Europe between 18,000 and 11,000 years ago. The recolonization is supported by archaeological data, which show the spread of the Magdalenian, a markedly uniform material culture from south of the Loire to the rest of Europe (Jochim et al., 1999). It is also confirmed by analysis of mtDNA sequence variations in extant European populations which indicate that a population originating from southern France and northern Iberia spread to central and eastern Europe about 15,000 years ago (e.g., Torroni et al., 1998). Due to the severe population bottleneck that took place during the Last Glacial Maximum, about 60% of the European mitochondrial DNA lineages (Richards et al., 2002) and even a higher proportion of Y chromosome lineages (Semino et al., 2000) can be traced back to the Magdalenian recolonization. Figure 8.1 shows the area of distribution of the Magdalenian, as well as the vegetation types at the end of the Last Glacial Maximum.

Since the Magdalenian spans an enormous geographic area with a low population density, we would expect human groups to become isolated and their artistic production and other forms of material culture to diverge. Also, the climate, the geography of the areas and types of prey show considerable variability across Europe, which again leads to the prediction that these groups would diverge. For example, settlements closer to water relied to an important extent on aquatic food

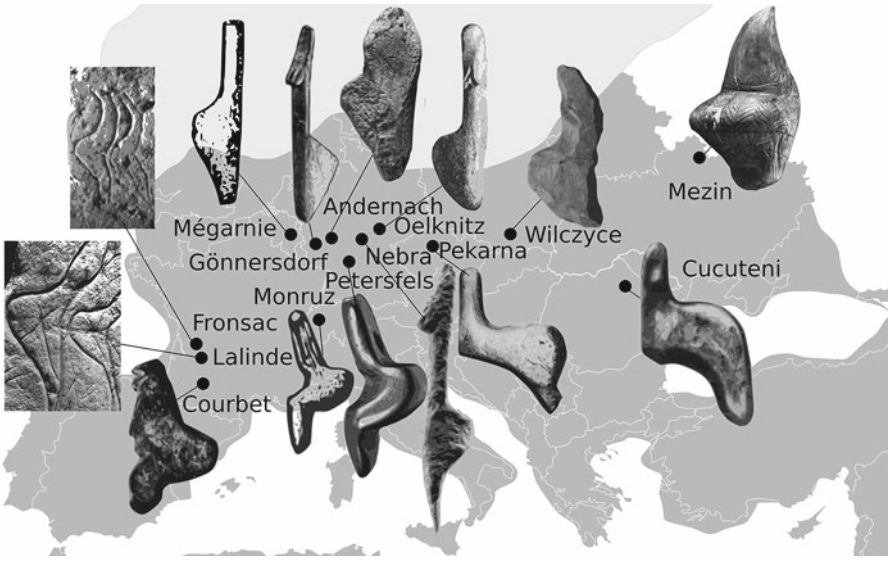

Fig. 8.2 A selection of Magdalenian so-called Gönnersdorf-Lalinde type Venus figurines and their locations

resources, whereas groups living inland subsisted mainly on reindeer and other large terrestrial mammals, reflected in a larger size of the settlements as preying upon large herds requires many hunters and can sustain higher population densities. However, the striking uniformity of the Magdalenian material culture suggests that groups maintained extensive contacts. Cultural innovations such as harpoons and spear-throwers (the latter already invented during the preceding Solutrean) were ubiquitous. Also, the frequent occurrence of exotic shells, amber and nonlocal stones found hundreds of kilometres away from their place of origin suggests the maintenance of long-distance exchange networks (Dolukhanov, 1997). The Magdalenian expansion was also characterized by a significant increase in population density. During the Last Glacial Maximum, the density of sites across the southwest European landscape remained low, suggesting a population size of about 4400 to 5900 individuals. The Magdalenian recolonization led to a marked increase in site density across western and central Europe, suggesting a population of up to 28,800 individuals (Bocquet-Appel et al., 2005).

Colonizing marginal territory requires extensive social security networks, since environmental conditions are unpredictable. Similar mobiliary art and body decoration in the form of beads and pendants enabled these small bands to maintain contact and to signal membership of large aggregation bands. Over thousands of kilometres, Magdalenian art shows striking stylistic similarities, including perforated bone discs with zoomorphic figures, antler spear-throwers with zoomorphic sculpture, and hundreds of stylized female figures in profile. These figurines have been found in a wide geographical area from the Dordogne to Ukraine, as can be seen in Fig. 8.2.

These objects were sculpted from a wide variety of materials, including flint, bone, ivory and steatite, which all have specific properties in terms of workability, fracturing and density. Despite this diversity in raw materials, they are stylistically markedly homogeneous, representing stylized women in profile with large buttocks, elongated headless torsos, small or absent breasts, without arms or feet. None of the figurines, including those made of flint, show traces of wear so they were not used as tools, but often they exhibit traces of extensive polishing, which firmly establishes that the artisans were concerned with their aesthetic properties. The statuettes fall within the Late Magdalenian, between 16,000 and 14,000 years ago (Fiedorczuk et al., 2007), a period characterized by population expansion and settlement of humans in large open-air and rock shelter sites. Long-distance contacts are documented in the transfers of exotic materials such as Mediterranean shells and Baltic amber found more than 600 kilometres from their places of origin (Gamble et al., 2005). We propose that the abundance of these figurines within living sites, e.g., more than 20 in Wilczyce, Poland (Fiedorczuk et al., 2007), the continent-wide adherence to a canon, and the care with which the objects were made suggest their use as ethnic markers. The fact that some of the objects (e.g., in Monruz, Switzerland, and Petersfels, southern Germany) have holes for suspension (Braun, 2005) strengthens this interpretation, as they were probably worn by individuals, as necklaces or other types of body decoration signaling group identity. Importantly, none were found in burial sites, which indicates they were not associated with particular individuals but rather with groups. As the climate became milder due to the start of an interglacial period, Magdalenian visual art in all its forms disappeared. Large animals became extinct or rare, and were replaced by smaller game such as deer, birds and hares, which are more evenly spread across the landscape. Although we still find evidence of long-distance contact in the form of exchange of seashells, which were probably valued for their exotic character, the risk of starvation became smaller and social security networks were less essential for survival in this richer environment. The lack of material manifestations of social safety nets in the archaeological record during this period supports our hypothesis.

8.5 Conclusion

Based on converging lines of evidence, we have sketched a cultural group selectionist model in which Palaeolithic mobiliary art and body decoration were used as a signal of membership of mutual altruistic groups. Archaeological and genetic evidence show that anatomically modern humans migrated out of Africa during the Last Ice Age. Around 50,000 years ago, they colonized Australia, including the arid inland with its inhospitable and unpredictable climate. At around 45,000 years ago they expanded into arctic Siberia. As ethnographic parallels and our case study of the Magdalenian show, risky and marginal environments can only be colonized by hunter-gatherer groups if they form social security networks. These networks require recognizable ethnic markers in the form of portable art and body decoration. It is no

coincidence that mobiliary art and pierced shell beads were first made during the last two Ice Ages, as soon as population density allowed it (first in Africa and later in Eurasia), as at least some forms of art can be explained as an adaptive cultural response to harsh and unpredictable environmental conditions. Mutual altruism was necessary for Upper Palaeolithic hunter-gatherers, since they lived in uncertain and marginal environments, where the risk of starvation was always considerable.

It is important to note that our model was not designed to provide an all encompassing explanation for artistic behaviour, in the sense that traditional adaptationist approaches have attempted. Indeed, the fact that art spontaneously arises as a byproduct of normal perceptual and motivational processes leads us to suspect that no silver bullet theory will be able to successfully explain all forms of art production. Art objects have a diversity of roles and meanings in present and past human societies, and each of these roles and meanings might require different explanatory frameworks. The purpose of this paper was to examine how some forms of art in a particular context (such as the mobiliary art from the Magdalenian) could proliferate and be maintained through cultural group selection.

Acknowledgements We would like to thank Katie Plaisance, Thomas Reydon, an anonymous reviewer and members of the Human Evolution and Behavior Network (HEBEN) for their comments on an earlier version of this paper. This research was funded by grant 3H070815 from the Research Foundation Flanders and grant COM07/PWM/001 from Ghent University.

References

Ambrose, S.H. (1998): Chronology of the Later Stone Age and food production in East Africa. *Journal of Archaeological Science* 25: 377–392.

Blood, A.J., and Zatorre, R.J. (2001): Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences of the USA* 98: 11818–11823.

Bloom, P. (2000): *How children learn the meanings of words*. Cambridge, Mass.: MIT Press.

Bloom, P., and Markson, L. (1998): Intention and analogy in children's naming of pictorial representations. *Psychological Science* 9: 200–204.

Bocquet-Appel, J.P., Demars, P.Y, Noiret, L., and Dobrowsky, D. (2005): Estimates of Upper Palaeolithic meta-population size in Europe from archaeological data. *Journal of Archaeological Science* 32: 1656–1668.

Braun, I.M. (2005): Art mobilier magdalénien en Suisse. *Préhistoire, Art et Sociétés* 60: 25–44.

Brown, D.E. (1991): *Human universals*. New York: McGraw-Hill.

Brown, S., Martinez, M.J., and Parsons, L.M. (2006): The neural basis of human dance. *Cerebral Cortex* 16: 1157–1167.

Carroll, N. (2004): Art and human nature. *Journal of Aesthetics and Art Criticism* 62: 95–107.

Collard, M., Shennan, S.J., and Tehrani, J.J. (2006): Branching, blending, and the evolution of cultural similarities and differences among human populations. *Evolution and Human Behavior* 27: 169–184.

Conard, N.J. (2003): Palaeolithic ivory sculptures from southwestern Germany and the origins of figurative art. *Nature* 426: 830–832.

Conard, N.J., Malina, M., and Münzel, S.C. (2009): New flutes document the earliest musical tradition in southwestern Germany. *Nature* 460: 737–740.

Cosmides, L., and Tooby, J. (1994): Origins of domain specificity: The evolution of functional organization. In L. Hirschfeld and S.A. Gelman (eds.), *Mapping the mind. Domain specificity in cognition and culture* (pp. 85–116). Cambridge: Cambridge University Press.
Dawkins, R. (1989): *The selfish gene* (2nd ed.). Oxford: Oxford University Press.
Deregowski, J.B., Muldrow, E.S., and Muldrow, W.F. (1972): Pictorial recognition in a remote Ethiopian population. *Perception* 1: 417–425.
Dissanayake, E. (2000): *Art and intimacy: How the arts began*. Washington, D.C.: University of Washington Press.
Dolukhanov, P. (1997): The Pleistocene-Holocene transition in northern Eurasia: Environmental changes and human adaptations. *Quaternary International* 41: 181–191.
Dutton, D. (2006): A naturalist definition of art. *Journal of Aesthetics and Art Criticism* 64: 367–377.
Fehr, E., and Gächter, S. (2002): Altruistic punishment in humans. *Nature* 415: 137–140.
Fiedorczuk, J., Bratlund, B., Kolstrup, E., and Schild, R. (2007): Late Magdalenian feminine flint plaquettes from Poland. *Antiquity* 81: 97–105.
Gallagher, H.L., and Frith, C. (2003): Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences* 7: 77–83.
Gamble, C. (1982): Interaction and alliance in Palaeolithic society. *Man* 17: 92–107.
Gamble, C., Davies, W., Pettitt, P., Hazelwood, L., and Richards, M. (2005): The archaeological and genetic foundations of the European population during the Late Glacial: Implications for 'agricultural thinking'. *Cambridge Archaeological Journal* 15: 193–223.
Gaut, B. (2005): The cluster account of art defended. *British Journal of Aesthetics* 45: 273–288.
Guzmán, R.A., Rodríguez-Sickert, C.Rowthorne, R. (2007): When in Rome, do as the Romans do: The coevolution of altruistic punishment, conformist learning, and cooperation. *Evolution and Human Behavior* 28: 112–117.
Hauser, M.D., and McDermott, J. (2003): The evolution of the music faculty: A comparative perspective. *Nature Neuroscience* 6: 663–668.
Henrich, J. (2004): Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior and Organization* 53: 3–35.
Henshilwood, C., d'Errico, F., Vanhaeren, M., van Niekerk, K., and Jacobs, Z. (2004): Middle Stone Age shell beads from South Africa. *Science* 304: 404.
Hill, K. (2002): Altruistic cooperation during foraging by the Ache, and the evolved human predisposition to cooperate. *Human Nature* 13: 105–128.
Hochberg, J., and Brooks, V. (1962): Pictorial recognition as an unlearned ability: A study of one child's performance. *American Journal of Psychology* 75: 624–628.
Ivey, P.K. (2000): Cooperative reproduction in Ituri forest hunter-gatherers: Who cares for Efe infants? *Current Anthropology* 41: 856–866.
Jochim, M., Herhahn, C.Starr, H. (1999): The Magdalenian colonization of southern Germany. *American Anthropologist* 101: 129–142.
Lewontin, R.C. (1978): Adaptation. *Scientific American* 239: 156–169.
Mag Uidhir, C. Magnus, P.D. (2011): Art concept pluralism. *Metaphilosophy* 42: 83–97.
Martindale, C. (1998): *Bouguereau is back*.http://www.science-of-aesthetics.org/proceedings/abwed.html
Martlew, M., and Connolly, K.J. (1996): Human figure drawings by schooled and unschooled children in Papua New Guinea. *Child development* 67: 2743–2762.
Maynard Smith, J. (1964): Group selection and kin selection. *Nature* 201: 1145–1147.
McElreath, R., and Boyd, R. (2007): *Modeling the evolution of social behavior. A guide for the perplexed*. Chicago and London: University of Chicago Press.
Mell, J.C., Howard, S.M., and Miller, B.L. (2003): Art and the brain. The influence of frontotemporal dementia on an accomplished artist. *Neurology* 60: 1707–1710.
Miller, B.L., Cummings, J., Mishkin, F., Boone, K.Prince, F., Ponton, M., and Cotman, C. (1998): Emergence of artistic talent in frontotemporal dementia. *Neurology* 51: 978–982.
Miller, G. (2000): *The mating mind. How sexual choice shaped the evolution of human nature*. London: William Heineman.

Minc, L.D. (1986): Scarcity and survival: The role of oral tradition in mediating subsistence crises. *Journal of Anthropological Archaeology* 5: 39–113.
Nettle, D., and Dunbar, R.I.M. (1997): Social markers and the evolution of reciprocal exchange. *Current Anthropology* 38: 93–99.
Pinker, S. (1997): *How the mind works*. London: Allen Lane.
Powell, A., Shennan, S., and Thomas, M. (2009): Late Pleistocene demography and the appearance of modern human behavior. *Science* 324: 1298–1301.
Preissler, M.A., and Bloom, P. (2008): Two-year-olds use artist intention to understand drawings. *Cognition* 106: 512–518.
Price, G. (1972): Extension of covariance selection mathematics. *Annals of Human Genetics* 35: 485–490.
Ramachandran, V.S., and Hirstein, W. (1999): The science of art. *Journal of Consciousness Studies* 6: 15–51.
Richards, M., Macaulay, V., Torroni, A., and Bandelt, H. (2002): In search of geographical patterns in European mitochondrial DNA. *American Journal of Human Genetics* 71: 1168–1174.
Richerson, P.J., and Boyd, R. (2005): *Not by genes alone. How culture transformed human evolution*. Chicago: University of Chicago Press.
Riolo, R.L., Cohen, M.D., and Axelrod, R. (2001): Evolution of cooperation without reciprocity. *Nature* 414: 441–442.
Roberts, R, Walsh, G., Murray, A., Olley, J., Jones, R, Morwood, M., Tuniz, C., Lawson, E., Macphailk, M., Bowderyk, D., and Naumann, I. (1997): Luminescence dating of rock art and past environments using mud-wasp nests in northern Australia. *Nature* 387: 696–699.
Seifert, L.S. (1992): Experimental aesthetics: Implications for aesthetic education of naive art observers. *Journal of Psychology* 126: 73–78.
Semino, O., Passarino, G., Oefner, P.J., Lin, A.A., Arbuzova, S., Beckman, L.E., De Benedictis, G., Francalacci, P.., Kouvatsi, A., Limborska, S., Marcikiæ, M., Mika, A., Mika, B., Primorac, D., Santachiara-Benerecetti, A.S., Cavalli-Sforza, L.L. and Underhill, P.A. (2000): The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: A Y chromosome perspective. *Science* 290: 1155–1159.
Sperber, D., and Hirschfeld, L.A. (2004): The cognitive foundations of cultural stability and diversity. *Trends in Cognitive Sciences* 8: 40–46.
Stein Mandryk, C. (1993): Hunter-gatherer social costs and the nonviability of submarginal environments. *Journal of Anthropological Research* 49: 39–71.
Sterelny, K. (1996): The return of the group. *Philosophy of Science* 63: 562–584.
Sterelny, K. (2004): Symbiosis, evolvability, and modularity. In G. Schlosser and G.P. Wagner (eds.), *Modularity in development and evolution* (pp. 490–516). Chicago: University of Chicago Press.
Tehrani, J., and Collard, M. (2002): Investigating cultural evolution through biological phylogenetic analyses of Turkmen textiles. *Journal of Anthropological Archaeology* 21: 443–463.
Tooby, J., and Cosmides, L. (2001): Does beauty build adapted minds? Toward an evolutionary theory of aesthetics, fiction and the arts. *SubStance* 30: 6–27.
Torroni, A., Bandelt, H., D'Urbano, L., Lahermo, P., Moral, P., Sellitto, D., Rengo, C., Forster, P., Savontaus, M., Bonné-Tamir, B. and Scozzari, R. (1998): mtDNA analysis reveals a major Late Paleolithic population expansion from southwestern to northeastern Europe. *American Journal of Human Genetics* 62: 1137–1152.
van Baalen, M., and Jansen, V.A. (2003): Common language or Tower of Babel? On the evolutionary dynamics of signals and their meanings. *Proceedings of the Royal Society of London B* 270: 69–76.
Vanhaeren, M., and d'Errico, F. (2005): Grave goods from the Saint-Germain-la-Rivière burial: Evidence for social inequality in the Upper Palaeolithic. *Journal of Anthropological Archaeology* 24: 117–134.
Vanhaeren, M., d'Errico, F., Stringer, C., James, S., Todd, J., and Mienis, H. (2006): Middle Paleolithic shell beads in Israel and Algeria. *Science* 312: 1785–1788.

Vartanian, O., and Goel, V. (2004): Neuroanatomical correlates of aesthetic preference for paintings. *NeuroReport* 15: 893–897.

Whallon, R. (1989): Elements of cultural change in the Later Palaeolithic. In P. Mellars and C. Stringer (eds.), *The human revolution. Behavioural and biological perspectives on the origin of modern humans* (pp. 433–454). Edinburgh: University of Edinburgh Press.

Whallon, R. (2006): Social networks and information: Non-utilitarian mobility among hunter-gatherers. *Journal of Anthropological Archaeology* 25: 259–270.

White, R. (1982): Rethinking the Middle/Upper Paleolithic transition. *Current Anthropology* 23: 169–176.

White, R. (1997): Substantial acts: From materials to meaning in Upper Paleolithic representation. In M.W. Conkey, O. Soffer, D. Stratmann, and N.G. Jablonski (eds.), *Beyond art: Pleistocene image and symbol* (pp. 93–121). San Francisco: California Academy of Sciences.

Wiessner, P. (1983): Style and social information in Kalahari San projectile points. *American Antiquity* 48: 253–276.

Williams, G.C. (1966): *Adaptation and natural selection*. Princeton: Princeton University Press.

Wilson, E.O., and Hölldobler, B. (2005): Eusociality: Origin and consequences. *Proceedings of the National Academy of Sciences of the USA* 102: 13367–13371.

Wimsatt, W.C. (2001): Generative entrenchment and the developmental systems approach to evolutionary processes. In S.Oyama, P.E. Griffiths, and R.D. Gray (eds.), *Cycles of contingency. Developmental systems and evolution* (pp. 219–237). Cambridge, MA: MIT Press.

Wynne-Edwards, V.C. (1962): *Animal dispersion in relation to social behavior*. Edinburgh: Liver and Boyd.

Zaidel, D.W. (2005): *Neuropsychology of art. Neurological, cognitive and evolutionary perspectives*. Hove and New York: Psychology Press.

Zeki, S. (1999): *Inner vision*. New York: Oxford University Press.

Zeki, S., and Lamb, M. (1994): The neurology of kinetic art. *Brain* 117: 607–636.

Chapter 9
Sensory Exploitation: Underestimated in the Evolution of Art As Once in Sexual Selection Theory?

Jan Verpooten and Mark Nelissen

9.1 Introduction

Before addressing the question of the evolution of art it may be useful to consider another question first: what is art? This question has no agreed-upon answer. Some philosophers of art even claim that art is intrinsically indefinable (e.g., Gaut 2005). Others devote their careers trying to define art (see for a summary: Adajian 2007). Definitions or rather descriptions of art seem to be extremely dependent on the perspective of the (sub)discipline from which they are undertaken, and the works of art that are considered relevant by researchers; for example, video games are seldom considered art today, but probably will be by a new generation. Maybe it is because the term "art" traditionally denotes something of value or significance (comparable to the impact of the label "scientific") that people never seem to stop discussing what is art and what is not. Some – especially artists – will claim art to be indefinable, thus contributing to its charm and appeal.

However, when considering art from an evolutionary perspective we *need* some sort of a description of art to work with, and a rather general one, since evolutionary theory — as a scientific theory — is about general processes. In most approaches of natural scientists art is described as "aesthetically pleasing" (e.g., Dissanayake 1992; Miller 2000, 2001; Ramachandran and Hirstein 1999; Pinker 1997, 2002),

J. Verpooten (✉)
Konrad Lorenz Institute for Evolution and Cognition Research, Adolf Lorenz Gasse 2, 3422 Altenberg, Austria

Department of Biology, University of Antwerp, Universiteitsplein1, 2610 Antwerpen, Belgium
e-mail: jan.verpooten@ua.ac.be

M. Nelissen
Department of Biology, University of Antwerp, CGB – Groenenborgerlaan 171, 2020 Antwerpen, Belgium
e-mail: mark.nelissen@ua.ac.be

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*, Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_9,

but this is arguably a too narrow description of art. Meaning (symbolic, in the sense of referring to something outside the work of art) is also important in art, and is usually not reducible to aesthetic appeal, if the work of art is meant to be aesthetically pleasing at all. So, our general description should ideally cover such disparate examples as placing a *urinoir* entitled "Fountain" in an exhibition space, the extremely popular and extremely violent video game Grand Theft Auto, and a tradition of weaving ornamental baskets. Van Damme (2008, p. 30) writes: "Numerous contemporary definitions of the term "art" mention in one way or another both "aesthetics" (denoting say, high quality or captivating visual appearance) and "meaning" (referring to some high quality or captivating referential content) as diagnostic features, although any clear-cut distinction between the two appears unwarranted, if only since there is no signified without a signifier." Furthermore, we will consider art as a signaling *behavior*, following Dissanayake's (1992, p. 8) ethological approach: "a 'behavior of art' should comprise both making and experiencing art, just as aggressive behavior presupposes both offense and defense." Thus, here we view "artistic behavior" as producing and experiencing "signals" (or a perceivable object emitting signals) with captivating meaning and/or form (design) to group members.[1]

The concept outlined in this chapter takes all this into account and is based on a biological model of signal evolution, namely Sensory Exploitation (SE). SE is a fairly recent model that is currently gaining field in sexual selection theory, where it offers a refreshing alternative to the classic perspective on the evolution of signal sending and receiving in courtship behavior. We argue that it should do the same for the evolution of human artistic behavior. SE deserves more attention in evolutionary thinking about art than it has received until now. To avoid any misunderstandings we would like to stress that using a model from sexual selection to address questions about the evolution of human artistic behavior does not in any way imply (or exclude) that art evolved as a sexual display. How this works will be explained below.

Many proposals about the evolution of art have been based on or linked to sexual selection in one way or another (e.g., Low 1979; Eibl-Eibesfeldt 1989a, b). The first ideas in this direction came, as so often in evolutionary biology, from Darwin himself. They can be found in his second book on evolution in which he covered both sexual selection and "the descent of man" (Darwin 1871). For example, Darwin suggested that bird song and human proto-song, which he thought would have been especially exerted during the courtship of the sexes, were evolutionary analogues. He even posited that some animals possessed a "sense of beauty" quite similar to ours and that this capacity had significant evolutionary consequences (Darwin 1871, p. 301): "When we behold a male bird elaborately displaying his graceful plumes or splendid colors before the female, whilst other birds, not thus decorated, make no such display, it is impossible to doubt that she admires the beauty of her male partner." Put differently, Darwin was the first to postulate that elaborate male display traits

[1] Although art may also be "captivating" to other groups of the same species or even to other species on earth or elsewhere, this is not necessarily so. Moreover we will argue art evolved *because* it is captivating to group members (and to artists themselves).

(such as ornament, song, and dance)[2] have evolved by appealing to choosy females' senses. The idea that a sense of beauty would have evolutionary consequences is obviously inspiring in relation to questions about the evolution of aesthetic signals and art. (The above-mentioned concern that art is not only about beauty does not devalue the general principle of Darwin's hypothesis, provided that sexual selection is perceived from the SE perspective.)

We will review and evaluate two existing applications of sexual selection to the evolution of art, borrowing ideas and contrasting our view with them. In order to do this, a preliminary discussion of current models of sexual selection is required. In section 2 we discuss two types of sexual selection models that address the evolution of male display traits and female preferences. There is the indirect benefit model in which females develop preferences for certain male traits that are adaptive (or indicators thereof). These preferences are indirectly selected for in the course of evolution, because the good choices (for males with adaptive traits) are rewarded with fitter offspring (since they inherited both the genes for good choice and the adaptive traits, which they pass on to their sons and daughters). This circular process can run out of hand. Since genes for good choice and genes for adaptive traits become genetically correlated (meaning they are passed on together to the next generations), they can be caught in a potentially maladaptive runaway process. It is basically this indirect benefit model that has been used by both Miller (1998, 1999, 2000, 2001) and Boyd and Richerson (1985, ch. 8) to address the evolution of aesthetic displays and art in humans. Miller proposes that art may in fact quite literally have evolved as a sexual display through indirect benefit processes on the genetic level. Boyd and Richerson (1985, ch. 8) focus specifically on the explanatory possibilities of the runaway process. They apply the model to cultural level processes, thus using a sexual selection model to postulate a non-sexual,[3] cultural runaway process that leads to the spread of cultural aesthetic traits. These two hypotheses are reviewed and discussed in the first part of section 3.

The other sexual selection model discussed in section 2 is SE. From the SE perspective, female preferences are sensory biases that have originated in another context than the current mating context and that may be maintained by the utility they have in that context (e.g., finding food). A male evolves display traits that exploit these female sensory biases, since captivating the female's attention or just plainly misleading her (e.g., by mimicking food) increases his reproductive success. We conclude section 2 with summarizing why this alternative (or at least addition) to the classic indirect benefit model is important in sexual selection theory. In the second part of section 3, SE is applied to human artistic behavior as an addition or even alternative to the existing hypotheses. So here we argue that art evolved

[2]Often a distinction is useful in mating behavior between intersexual signaling and intrasexual competition for mates. While peacocks use their tails to court peahens, antlers and other "weapons" are used to fight same-sex rivals. Here we focus on the former.

[3]Cultural variants as analogues to genes are also passed on through reproduction, but not through sexual reproduction; however, they are reproduced through imitation and other forms of social learning.

by exploiting human biases for certain meanings as well as design or formal aspects. Animal biases that are exploited can be quite complex, determined not only by innate dispositions or engineering details of the sensory system of the signal receiver but also by psychological factors such as emotions and (social) learning (e.g., Guilford and Dawkins 1991) and we can expect the same for human biases. To the person who experiences a work of art there might be no direct utility involved, just as the female that is misled by the male mimicking food may not benefit from being sensorily fooled. SE is typically applied to sexual selection cases in which the traits or signals exploiting biases are genetically encoded male display traits (e.g., orange spots resembling food in guppies). However, borrowing from Boyd and Richerson's (1985, ch. 8) model, sensory exploitation also applies to non-sexual contexts, and exploiting signals may be culturally transmitted as well. So, SE does not need to imply that art evolved through courtship. Here we are not specifically interested in the reproductive success of the artists, but in the reproductive success of artistic signals themselves that spread through cultural transmission regardless of beneficial effects to individuals that transmit them, just as male ornaments evolve through sensory exploitation without the need of any benefits to females. This possibility of non-functional evolution of art will be a theme throughout this chapter. We will mainly focus on iconic representations and also briefly discuss "self-exploitation" and make a sketchy comparison of art and religion in relation to human mental biases. In section 4, we summarize our evaluation and articulation of existing hypotheses based on the SE view on art.

9.2 Sexual selection theory

To make our argument it is not necessary to provide a full overview of sexual selection theory. We will only focus on those models applicable to the evolution of art. These are the indirect benefit or "Fisher-Zahavi model" (Eshel et al. 2000; Kokko et al. 2003) and SE (e.g. Ryan 1990, 1998). Both Boyd and Richerson and Miller use the former; our concept is based on the latter.

Mate choice is an important evolutionary process that imposes sexual selection on the other sex and accounts for spectacular traits and behaviors that would otherwise remain unexplained by natural selection (Darwin 1871; Andersson 1994). Both the indirect-benefit model and SE describe the relation between mate choice and these traits and behaviors. For an insightful review of sexual selection models in general — much in this section is based on it — see Kokko et al. (2003).

9.2.1 Indirect-benefit model

The Fisher-Zahavi model is an indirect-benefit model of mate choice. Both the so-called good genes selection hypothesis (or fitness indicator theory) and Fisher's runaway process fall within this category. The good genes selection hypothesis

simply states that females choose partners based on indicators of genetic quality. The evolutionary logic behind this behavior is that they as such provide their offspring with good genes. Choosing good genes positively influences the viability of the offspring and increases the chances that the female's offspring reaches reproductive age. So female choice for indicator traits is indirectly selected by piggybacking on the directly naturally selected good genes (Fisher 1930, formally demonstrated by Lande 1981). Closely related to the good genes hypothesis is the handicap principle. It predicts the game-theoretic constraint that indicators must be costly to be reliable because if not they can be faked too easily (Zahavi 1975, 1991; Zahavi and Zahavi 1997).

Thus, fitter males, and the females who preferentially mate with them, will have offspring that inherit the genes for both fitness and the mating preference. The resulting linkage disequilibrium[4] between preference genes and male fitness favors the spread and elaboration of the preference by indirect selection. Fisher's insight, that the increased importance of attractiveness as a component of male fitness can drive the exaggeration of a male trait signaling fitness beyond its otherwise naturally selected optimum, is known as the "Fisherian runaway" process. So long as the process is unchecked by severe counterselection (i.e., survival costs), it will advance with ever-increasing speed (Fisher 1930).

9.2.2 Sensory Exploitation

Selection operating directly on the psychosensory system in contexts other than mate choice may either maintain or drive changes in mating biases (Williams 1966; Sober 1984; West-Eberhard 1984, 1992; Ryan 1990, 1995, 1998; Ryan and Rand 1990, 1993; Ryan and Keddy-Hector 1992; Endler 1992; Arak and Enquist 1993, 1995; Shaw 1995; Dawkins and Guilford 1996; Endler and Basolo 1998; Autumn et al. 2002). To some extent mate choice may thus evolve by a process variously known as SE (e.g., Ryan 1990, 1998), sensory drive (e.g., Endler 1992), pre-existing bias, or sensory trap (e.g., Christy 1995). For example, across some populations of guppies the strength of attraction to orange objects in a non-mating context explains 94% of the inter-population variation in female mating preferences for orange male ornaments (Rodd et al. 2002). This means that in populations where females are strongly attracted to orange food items, they will also tend to choose males mimicking these orange food items; hence, the reproductive success of males that happen to have orange spots in these populations increases and over a certain number of

[4]In population genetics, linkage disequilibrium is the non-random association of genes at two or more loci. In this specific case it means that the "gene" for preference for certain male display traits becomes correlated to the "gene" for the male display trait itself, since both genes are inherited by offspring. In sons the gene for the preference trait is not expressed, but it is in the sons' daughters, and vice versa, the gene for the display trait is not expressed in the daughters but it is in the daughters' sons.

generations these orange spots may become ever more accurate mimics[5] of orange food items. Thus female sensitivity to orange-colored food items may be at least as important to the evolution of female mating preferences for males with large orange spots as any direct and indirect benefits that more-orange males deliver to their mates. SE may do more than offer a quirky exaptive[6] alternative for how mating biases and male display traits evolve. Whenever studying a biological trait within the Darwinian framework it is important to distinguish between the selective forces that led to its origin, its evolution, and the processes that maintain it (Fisher 1930). The origin of mating biases and displays are relatively hard to explain with the indirect-benefit model (Arnqvist 2006). SE, however, may provide the initial "nudge" often required initiating choice-display coevolution (Arak and Enquist 1995; Payne and Pagel 2000). Recent empirical research and theoretical models suggest that origin by SE has been widespread (Rodriguez and Snedden, 2004; Arnqvist 2006). And maybe choice-display coevolution is not even required to explain the evolution of male ornaments, as we will discuss below.

Arnqvist (2006) distinguishes two classes of origins of sensory biases. Firstly, females are adapted to respond in particular ways to a range of stimuli in order to, for example, successfully find food, avoid becoming food for predators and breed at optimal rates, times, and places. Such multi-dimensional response repertoires form a virtually infinite number of pre-existing sensory biases that are potential targets for novel male traits. These he names "adaptive sensory biases." Notice that male traits that result from exploiting these adaptive sensory biases are in fact mimics. Secondly, pre-existing sensory biases need not be the direct result of selection. In theory, they can simply be incidental and selectively neutral consequences of how organisms are built (Ryan 1990; Endler and Basolo 1998). For example, artificial neural network models have shown that networks trained to recognize certain stimuli seem to generally produce various sensory biases for novel stimuli as a byproduct (Enquist and Arak 1993, 1994; Arak and Enquist 1993; Johnstone 1994). Similarly, research in "receiver psychology" (e.g. Guilford and Dawkins 1991; Ghirlanda and Enquist 2003) has also suggested that higher brain processes may incidentally produce pre-existing sensory biases for particular male traits. Following Arak and Enquist (1993), Arnqvist (2006) refers to such sensory biases as "hidden preferences". These, then, can be seen as side effects or contingencies of how the sensory system, defined in its widest sense, of the receiver is constructed. Usually it results in abstract biases, e.g., for symmetrical or exaggerated traits (Ryan 1998). Arnqvist's (2006) distinction is quite similar to the one mentioned above between "aesthetics" and "meaning", which is made in most contemporary definitions of art. In the next section we will exploit this similarity for constructing our SE concept of art.

[5]The term "mimic" usually refers to a whole, mimicking organism (e.g., Pasteur 1982), but as Maran (2007, p. 237) usefully points out, from a semioticist viewpoint "neither the mimic nor the model needs to be a whole organism but can be just a part of an organism both in spatial or temporal terms or just a perceptible feature." So here we use mimic in the latter sense.

[6]An exaptation is a pre-existing trait that acquires a new beneficial effect without modification to the phenotype by selection (Gould, 1991).

All sensory systems have biases, and mating biases are therefore inevitable (Kirkpatrick and Ryan 1991; Arak and Enquist 1995). Of course, not all possible sensory biases are exploited in a mating context, although theoretically they could be. For example, Burley (1988) showed that female zebra finches prefer males whose legs have been experimentally decorated with red or black plastic bands, while males with blue and green bands were rejected. Basolo (1990) showed that female platyfish prefer males with colorful plastic "swords" glued on the ends of their tails, suggesting that this preference also pre-dated the evolution of such ornaments in their close relatives the swordtails. These could be called "latent" preferences (Miller 1998, 2000), preferences resulting from biases that are present but not exploited in a sexual context.

9.2.3 Sensory Exploitation versus indirect-benefit model?

The preceding discussion shows us how SE and indirect benefits are generally considered intimately intertwined in determining the evolution of female biases and male display traits. Thus Kokko et al. (2003) write: "Even when a male trait has evolved to exploit a pre-existing sensory bias, indirect selection on the female preference may occur owing to the benefits accruing from the production of more-attractive sons. Such a signal may potentially then become secondarily genetically correlated with other fitness-enhancing traits." So, Kokko et al. (2003) state here that even if SE happens, indirect selection will likely influence female mating preferences, which would in turn influence male display traits and so on, hence a runaway process. However, there is no theoretic reason to assume this would be a necessary outcome. Consider the example of the female preference for orange spots in male guppies again. The female preference for orange spots is in fact a preference for orange food and the preference for orange food is maintained by the fact that it is useful in food gathering. As a result, the mating preference for orange-spotted males can't be altered without selecting against something highly useful for food gathering. SE happens because of stabilizing selection[7] against changes to the preferences, which would have to be mediated by changes to the perceptual system that would be detrimental to the guppies in other ways (given the limited number of ways to get guppies to do what they need to do). In that sense, then, SE is sensitive to the problem of the evolution of female preferences, it's just that the guppies have the orange spot preferences they do because any other genuinely biologically possible preferences would be detrimental, not because orange spot preferences are linked to fitness in some further way. Moreover, Kokko et al. (2003)'s use of the

[7]Stabilizing selection, also referred to as purifying selection or ambidirectional selection, is a type of natural selection in which genetic diversity decreases as the population stabilizes on a particular trait value. Put another way, extreme values of the character are selected against. It is probably the most common mechanism of action for natural selection.

concepts “fitness” and “indirect benefit” are misleading. It can mean: good genes for survival *and/or* good genes for acquiring mates (attractiveness). Kokko et al. (2003) suggest the evolution of male display traits such as orange spots could be mediated by indirect benefits. But do they supply good genes for survival or are they just indicative of sexy son genes? It is quite possible that having orange spots does not correlate at all with genetic quality for viability. In that case, orange spots cannot and will not be selected as indicators of good genes for survival. These are important observations because they imply the possibility that evolution of male display traits may have more to do with the mechanism of SE than with indirect selected traits such as female preferences for indicators of genetic quality for viability (see Fuller et al. 2005). The strong version of SE can thus be perceived as an alternative to the indirect-benefit model in sexual selection and some researchers have offered it as such. At least some of the sensory biases and displays we find in nature might be the result of SE alone (West-Eberhard 1984; Ryan 1990, 1998). We stress this possibility because it will be central in our argument in the next section that the strong version of the SE concept might offer an alternative model for the evolution of art.

9.2.4 Biological mimicry

In some cases it is clear that good genes selection and runaway processes can never happen, but that nevertheless impressive ornaments evolve through signal evolution — that is in situations where benefits for the exploiter cannot in any way imply benefits for the signal-receiver. Some cases of biological mimicry fall within this category. For instance, in the genus *Ophrys*, plants evolved to attract male bees as pollinators by mimicking female mating signals. Here evolution by SE — the plants don’t give any rewards in return — seems to be the only possible explanation (Schiestl and Cozzolino 2008; Jersakova et al. 2006). Of course, in this example indirect genetic benefits don’t apply because sensory biases of another species are exploited. But even intra-species SE in a sexual context may occur without good genes for viability selection, as the following example illustrates. Many cichlid fish species independently have evolved mouthbreeding as a highly specialized brood care behavior. Egg dummies, resembling the ova of the corresponding species, formed of various parts of the body can be found in different lineages of mouthbreeding cichlids. Most abundant are egg spots, which are conspicuously yellow spots on the anal fin of males. Females of mouthbreeding cichlids undoubtedly evolved sensory capabilities to detect eggs and are supposed to have a strong affinity for them, because they pick them up immediately after spawning. In fact, the ability to detect the eggs directly affects the female’s fertility. Every missed egg results in a reduction in fitness. Consequently, a pre-existing sensory bias might have occurred in early mouthbreeders and might still occur in mouthbreeding species without egg dummies. As a consequence, males would have evolved egg spots in response to this sensory bias (Tobler 2006).

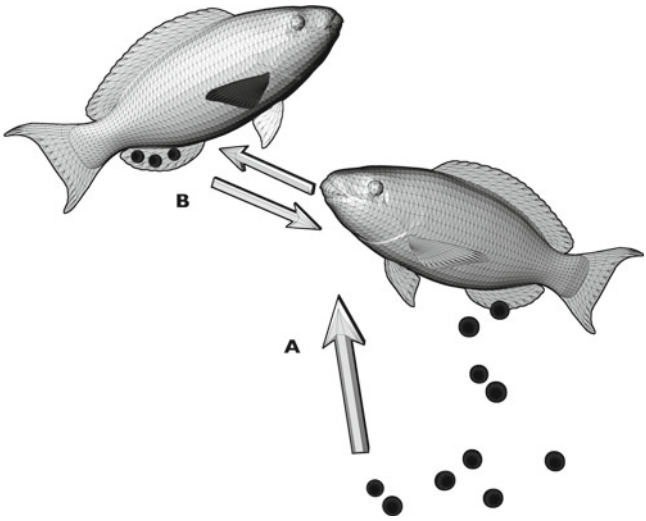

Fig. 9.1 The mating system of mouthbreeding cichlids. (A) After laying her eggs the female (right) sucks them up in her mouth. Her ability to detect the eggs is strongly selected for, since every missed egg results in a reduction of fitness. This ability depends on a hair trigger response to "egg signals." (B) Subsequently, males (left) evolved egg spots, accurate two-dimensional mimics of the eggs, to exploit this female response. Choice-display coevolution is inhibited by the fact that the female's bias for eggs is vital for detecting the real eggs, and there is no reason to a priori state that the effectiveness of the male egg spots are linked to genetic quality. So, this may well be an example of the strong version of sensory exploitation. (artwork: Alexandra Crouwers and Jan Verpooten.)

After the female (receiver) has picked up her eggs (model), the male displays in front of her, showing the egg spots on his anal fin (mimic). The female responds to the life-like egg illusion by a sucking reaction – and obtains a mouthful of sperm from the canny male in the process. One of us (Nelissen) has performed quite some research on cichlids and has described the system of the egg spots (in *Tropheus* and *Simochromis*). During courtship males vibrate their body while showing the egg spots to the female. It could well be that by doing this they enhance the egg illusion, giving it a more three-dimensional effect in combination with the light-dark grading in color and the colorless outer ring the egg spots exhibit (e.g., Wickler 1962). It may be that the female's mating preference for a male with well-elaborated egg spots does not yield in any direct benefits for the female, nor any good genes for viability of the female's offspring. Runaway selection is also limited by the mimicking function of the egg spots: they may need to remain life-like in order to mislead the female. As explained above, female preference for egg-like signals cannot be altered because of the functional importance of this preference outside the courtship context. Thus this might well be an example of the strong version of SE. The female's mating preference may be solely maintained by exploiting the benefit of the detection of eggs after spawning (Tobler 2006) (Fig. 1). Interesting to the problem of the evolution of human representational art is that cases of mimicry, such

as this one, show how SE can produce two-dimensional representations (the egg spots) on a surface (the anal fin of the male) of three-dimensional objects (the eggs). In section 3.3.1. we will use this case as an example of SE in non-human animals and compare it to visual art in humans from a semiotic viewpoint.

9.2.5 Summary of section 9.2

SE is a crucial addition to or possibly even an alternative — at least under certain conditions — to the indirect-benefit model to explain the evolution of signals used in sexual contexts. Likewise, as we will argue in the next section, it also applies to the evolution of art. Here is a short summary:

- SE may provide the initial nudge for the evolution of male displays.
- SE may either maintain or drive changes in mating biases. As a result, male display traits may not necessarily be indicators of good genes for viability (i.e., survival).
- Cases of mimicry are clear-cut examples of the influence of SE as a mimic evolves to exploit sensory biases. Moreover, stabilizing selection on the female's sensory system inhibits changing its adaptive sensory biases by choice-display coevolution.

In section 3 we will show that a substantial portion of the discussion about the evolution of art is situated around the same questions as the ones covered in this section. We will thus use these summarized insights from this section to address them.

9.3 Hypotheses about art

Both Miller and Boyd and Richerson built their hypotheses upon the indirect-benefit model, although they do so in quite different ways. In particular, the framework in which they apply the indirect-benefit model differs. Both their hypotheses are Darwinian, but Boyd and Richerson formalize the influence of culture into their models while Miller's model focuses on genes. Both approach art from a signal evolution perspective: there is a signaler (the producer of art), and a set of receivers (who perceive or experience the work of art).

9.3.1 Miller's proposal

Being an evolutionary psychologist, Miller (2000, 2001) considers the capacity to produce and appreciate art as a "psychological adaptation": an evolved domain-specific mental capacity. Art as such serves a sexual function, as an extension, as

Miller argues, of the human mind that itself evolved as a seducing device or an "entertainment system" by sexual selection (Miller 2000). In Miller's view human art making is exactly like bower building by male bowerbirds as follows. Females prefer to mate with males who construct larger, better quality, and more highly ornamented bowers (e.g., Borgia 1995). The bower can be considered as the "extended phenotype" of the male bowerbird (Dawkins 1982): a genetically evolved, species-specific artifact constructed outside the individual's body, but very much in the service of the individual's genes. Just like a bower, art is an aesthetic display that coevolved with aesthetic preferences (Miller 1998, 1999, 2000, 2001). It is an indicator of fitness. This means it is an indicator of reproductively important traits such as health, fertility, and genetic quality. "Perhaps beauty boils down to fitness" and "an art-work's beauty reveals an artist's virtuosity", Miller (2001) states. Virtuosity, indicative of creative application of high skill and high intelligence, is such a fitness indicator (Miller 2001).

As Darwin (1871) noted, female animals are often choosier about their mates, and males often display more intensely than females. Accordingly, Miller (1999) identified a significant sexual dimorphism in cultural production (public paintings, books, music albums and plays). Miller explains this dimorphism with a "cultural courtship model": human cultural production (i.e., art) functions largely as a courtship display, and the persistent sex difference in public cultural production rates reflects an evolved sex difference in courtship strategies (Miller 1999).

Criticism of Miller's proposal mainly focuses on the last two points: the implied competitiveness for mates that drives art and the claim that the sexual dimorphism[8] of art production that Miller identified in recent western society can be universalized. Critics stress the importance of tradition, which constrains individual competition and promotes cooperation among group members in traditional societies (Dissanayake 2001; Coe 2003). They argue that the bulk of human visual art has been traditional and our perception is biased by an overemphasis on certain short periods where individual creativity and competitiveness were important, such as the Renaissance (Coe 2003). The western non-traditional individualistic society of today is not representative but rather an exception. Moreover, if artists today are driven by competition, it is perhaps for media attention, not for mates. Another problem with Miller's proposal is that in traditional societies, females are sometimes the main producers of art (Dissanayake 2001; Coe 2003).

9.3.2 Boyd and Richerson's proposal

If traditions are capable of consistently influencing the human phenotype, meanwhile significantly constraining individual competition in favor of the genes of that

[8] Sexual dimorphism is a measure of differences between the sexes (e.g., height, color, etc.), mostly due to the operation of sexual selection.

individual,[9] it may arguably be necessary to incorporate culture into the Darwinian framework as an inheritance system that is partly independent from the genetic inheritance system. This is what Boyd and Richerson (1985) dubbed "Dual Inheritance Theory". They pointed out that Darwin's theory does not explicitly distinguish cultural inheritance from genetic inheritance. Darwin was a self-declared Lamarckian who believed that acquired variation (through social learning, e.g., a mechanism that transmits cultural information) played an important role in evolution (Richerson and Boyd 2001). So, Darwin's assumptions about beauty and evolution, which we mentioned in the introduction, should be viewed within a gene-culture coevolutionary framework.

Thus, within this framework, Darwinian selectionism is not exclusively applied to the genetic level but to both the genetic and cultural levels. Also, how both inheritance systems interact in human evolution (i.e., gene-culture coevolution) is investigated in a formalized manner (Boyd and Richerson 1985, 2005). Analogous to how population geneticists model the way different forces change gene frequencies in a population, they model how forces interact to bias cultural transmission in a population — that is, how culture[10] evolves. In Dual Inheritance Theory, the evolution and maintenance of culture is described by several mechanisms including transmission bias. One of these mechanisms or forces is "indirect" or "model" bias (Henrich and McElreath 2003; McElreath and Henrich 2007). Boyd and Richerson (1985, ch. 8) postulated that this force might cause a "cultural runaway process" that in turn offers an explanation for the evolution of aesthetic traits and art. In short, individuals imitate successful people because they provide the highest chance of acquiring adaptive information (Flinn and Alexander 1982). They prefer a certain value of an indicator of success (e.g., number of children or acres of land). This system of indicator trait and preference trait can, under certain conditions, be caught in a runaway process. A self-enforcing feedback loop between indicator and preference can cause the indicator trait, which was initially an adaptive sign of success, to become exaggerated following its own internal logic. "Much as peacock tails and bowerbird houses are thought to result from runaway sexual selection, the indirect bias runaway process will generate traits with an exaggerated, interrelated, aesthetically pleasing but afunctional form" (Boyd and Richerson 1985, p. 278).

As we suggested before, the fact that women clearly also engage in art production, especially in traditional societies, which are the rule in human evolution, but also fairly recently in the emancipated west, poses a problem for Miller's argument that art making is a sexual adaptation since it strongest support is the apparent sexual dimorphism in art making, with men showing off artistically and women choosing. In his contributing chapter to the book "The evolution of culture", Miller (1999) uses data on human sexual dimorphism in "cultural output" (i.e., art making) as evidence for the operation of sexual selection. Sexual dimorphism is one of the most

[9] Thus reducing the genes' relative importance in determining human behavior.

[10] The term culture refers here not to a specific culture, but to "information" (ideas, beliefs, etc.) which is transmitted in a population through social learning.

convincing proofs one can find for sexual selection operating, since sexual selection is the main cause of sexual dimorphism in organisms. As Darwin (1871) noted, since female animals are often choosier about their mates (because they usually invest more in less offspring than males), males may evolve quite elaborate displays as a response to female choosiness. The conspicuous sexual dimorphism in the peafowl is a clear-cut example: peacocks have large and costly tails, peahens are drab in color, differences that are obvious consequences of sexual selection. So Miller states that a work of art is like a peacock's tail: very costly, but compensated by reproductive success and thus adaptive.[11] There are at least two problems with this "empirical support" for Miller's proposal that art making evolved as a male sexual adaptation. Firstly, mating succes is a poor proxy for reproductive succes in post-birth-control cultures (also see Fitch 2006). Secondly, the sample of artists Miller (1999) uses (jazz musicians in the west prior to female emancipation) is not representative for humans in general. In many traditional societies women also engage in elaborate artistic behavior. Miller (2000) may have realized the shortcomings of his sexual dimorphism argument when he subsequently suggested in his book "The mating mind" that art making may be the result of a special kind of sexual selection, namely, *mutual* sexual selection. Under mutual sexual selection both males and females evolve sexual ornaments, consequently dissolving the sexual dimorphism. In the case of art, both men and women would have evolved to make art in order to attract mates and appreciate art to assess mates. However, by abandoning the sexual dimorphism argument, which is a strong one for sexual selection, the case for art as a sexual adaptation is severely weakened. All other aspects of art (its costliness, its captivating capacity, etc.) can easily be explained by other processes. Furthermore, if art evolved under mutual sexual selection it would predict that men are specifically interested in female art and women in male art. However, at first sight, the reverse might be the case, people especially being interested in art from same-sex peers. In fact, this would be highly consistent with SE, since the more the maker and the experiencer of art are similar, the more their pre-existing biases will be (also see 3.3.2.).

Boyd and Richerson offers another possible way out of this problem as in their cultural model the sex of the individuals do not play a role:

> Notice that in the case of the cultural runaway process colorful displays are not as likely to be limited to the male sex as they are with the genetic analog. A prestigious male or female

[11] The peacock's tail could only have evolved if the survival costs of having one are compensated by its reproductive benefits. In other words, there is an evolutionary tradeoff between investing in survival and in reproduction. Imagine there are 2 types of peacocks in a population. There are 20 type 1 peacocks with less attractive but also less risky tails, half of which reach reproductive age. Type 2 peacocks have enormous, conspicuous tails, and there are also 20 of them in the population. As a result, 19 type 2 peacocks are eaten by tigers and only one of them survives to reproductive age. If, however, this one male is so attractive in comparison to the others of group 1 so that he acquires, say, 90 % of the matings, the trait of the enormously large tail will spread over the population and persist at the expense of smaller tails, regardless of the high fatality it causes among males, because its mean evolutionary payoff is higher.

> can have an unlimited number of cultural offspring by non-parental transmission, whereas in the genetic case only males can take advantage of multiple matings to increase their fitness enough to compensate for costly displays. The fact that women as well as men participate in elaborate symbolic behaviors is more consistent with a cultural than with a genetic runaway explanation. (Boyd and Richerson 1985, pp. 278-279)

This cultural hypothesis about art illustrates that application of sexual selection models to the evolution of art doesn't imply that art needs to have a sexual function. The model, in this case Fisher's runaway, is assumed to apply to non-sexual cultural transmission as well. However, we will argue that the concept of SE applied to art implies a runaway process (which is a secondary force resulting from indirect benefits as we have mentioned above) is not even required for aesthetics and art to evolve. Exploitation of sensory biases — a primary force — can do the trick just as well.

9.3.3 The concept of Sensory Exploitation

Our proposition is based on the observation that both existing proposals show how sexual selection theory applied to artistic behavior offers valuable mechanistic insights into its evolution, but that they may underestimate the importance of SE in sexual selection and as such in the evolution of art. We will argue that SE may need to play a more substantial role in the evolutionary approach to art just like it does today in sexual selection theory. Art is believed to lie at the heart of culture, so if any behavior should be considered from a gene-culture coevolutionary perspective it must be artistic behavior. Thus, we will not a priori exclude the influence of cultural transmission from our model.[12]

As stated, we view "artistic behavior" as producing and experiencing signals (or a perceivable object emitting signals) with captivating meaning and/or form (design) to group members. The distinction between aesthetics and meaning made in most contemporary definitions of art roughly corresponds to the distinction made by Arnqvist (2006) between hidden preferences influencing the design of signals and adaptive sensory biases influencing the content of signals, resulting in mimicking signals, respectively. Thus, from a broad signal evolution perspective we can state that what Van Damme (2008, p. 30) has called aesthetics, corresponds to design and results from the exploitation of hidden preferences, and what he has called "meaning" corresponds to content and results from exploitation of adaptive sensory biases by mimicking signals or traits.

Elaborating on the discussion in section 2, let us first consider the origin of artistic behavior. Pre-existing biases of the psychosensory system are the most plausible

[12]Notice, however, that Dual Inheritance Theory does not exclude that art could have been sexually selected; e.g., Boyd and Richerson (1985, p 277): "Cultural traits which affect mating preference could similarly affect genetic evolution through the action of sexually selection."

candidate for many of the origins of female mate preferences, influencing which male display traits will evolve (e.g. Arnqvist 2006). Analogously, human pre-existing psychosensory biases may influence the direction in which art evolves. Our argument is that by focusing upon an indirect-benefit model this influence may be underestimated. For example, Miller (1998, p. 107) argues against the sensory bias evidence that "latent preferences are not necessary, according to R. A. Fisher's (1930) runaway theory. Even chance fluctuations in mate preferences, combined with a strange kind of evolutionary positive-feedback loop, could produce quite extreme mate preferences and quite exaggerated courtship traits." However, this argument can be easily reversed: Why do you need to postulate a combination of chance fluctuations and a secondary process such as Fisher's runaway when "latent preferences" are inevitably present anyway (see Kirkpatrick and Ryan 1991, Arak and Enquist 1995)? As mentioned, this critique also applies to Boyd and Richerson's runaway model. SE delivers a more parsimonious explanation for the origin and evolution of aesthetics — although it does not exclude secondary processes such as runaway. Miller (1998, 2000) also tends to minimize the sensory bias model by limiting it to preferences that are mere side-effects due to engineering details of the sensory system (i.e., *hidden preferences)*, ignoring *adaptive sensory biases*. That adaptive sensory biases influence the evolution of male traits is evidenced by clear-cut cases of mimics as sexual displays (Fuller et al. 2005). Consider the classic example used to explain Fisher's runaway process, the peacock's tail. Ridley (1981) suggested that tails with multiple eyespots, such as those of the peacock and the Argus pheasant, play upon a widespread responsiveness to eye-like stimuli in animal perception. In certain cases runaway is definitely limited by the need to maintain mimicking function. Miller (2000, p.142ff.) also voices the concern that a sensory bias model ignores the importance of an organisms' avoiding having sexual preferences for any ornaments that offer no fitness benefit or negative fitness benefit to them (surely there would be selection against this?). This concern is again easily addressed with the argument of stabilizing selection mentioned before: selection against adaptive sensory biases is unwarranted since they serve crucial functions in other, non-mating contexts. Another concern of Miller (2000, p. 146) is that: "For highly social animals like most primates, finding potential mates is not the problem. Many primates already live in large groups, and interact regularly with other groups. They are spoiled for choice. When mate choice depends more on comparing mates than locating mates, the sensory engineering argument seems weaker." It may be that in animals living in social groups sensory exploitation is less important than in solitary animals. However, we would like to stress that although the argument is contra sensory exploitation it is not necessary pro good genes selection. In social animals intra-sexual selection becomes more important, resulting in the development of weapons (such as antlers) rather than appealing ornaments (Andersson 1994). Moreover, the assumption that social animals *compare* mates already implies they are looking for good genes. Finally, Miller reduces sensory exploitation again here to engineering details. When males evolve mimics to mislead females, competition between males is guided by the success of the mimic in eliciting a response and not by comparison between mates.

Another important criticism of Miller's proposal is that he does not really grasp what Fisherian runaway and costly signaling means (Haufe 2008). Miller (2000, p. 147) employs the following reasoning against SE, arguing that sensory biases will always be entrained by good genes selection: "[i]f sensory biases led animals to choose lower-fitness animals over higher-fitness animals, I suspect that the biases would be eliminated rather quickly." However, as Haufe (2008, p. 124) explains:

> Genetic modeling of sexual selection does not confirm Miller's suspicions. In fact, it directly contradicts them. …, it follows analytically from the most basic Fisherian runaway model (as well as from other kinds of models) that a preference which causes (say) females to prefer "lower-fitness" (i.e., lower viability) animals over "higher-fitness" (i.e., higher viability) animals can spread and persist in a population, even when a preference for "optimal" (in terms of viability) males is introduced. Not only that, according to the basic model the preference which initiated runaway will itself become exaggerated, causing males to have even lower viability. Miller presumably is aware of this feature of runaway. However all of this gets tossed aside in pursuit of "hidden adaptive logic."

So, the strong version of our concept predicts that SE not only exerts a substantial influence on the direction in which art evolves, but that it may also maintain artistic behavior. In section 2 we explained how this is theoretically possible in the evolution of male display traits. Analogously, this possibility applies to the evolution of art making. It is clear from the evidence in sexual selection that the primary force of SE will always be present. The same applies to art. Secondary forces, such as indirect benefits may be operating but are in principle not required for art to evolve. So here we explore how far we can get without a priori invoking these secondary processes.

9.3.3.1 Iconic representation

The role of perceptual biases in the evolution of art has already been extensively investigated by several researchers (e.g., Hodgson 2006; Kohn and Mithen 1999; Ramachandran and Hirstein 1999). Essentially, they all have focused on the abstract, geometric aspect of visual art. They state that art emerged because its geometric patterns are supernormal stimuli to the neural areas of the early visual cortex. As such (exaggerated) symmetry, contrast, repetition, and so on, in visual art hyperstimulate these early neural areas. Thus, they have focused on what we have called hidden preferences. We agree with these authors that hidden preferences probably play an important role in the design aspects of human visual representations as they do in the design of male display traits.

However, as indicated by Van Damme's definition, design is only one aspect of human visual art – content, or meaning (mimics/iconic representations as the result of adaptive sensory biases) is at least as important in most cases. We will make this clear by way of an example — a comparison between egg spots in cichlids and visual art in humans from a semiotic viewpoint. This is followed by an introduction to some of the human adaptive sensory biases exploitable by iconic representations.

Semioticists generally agree that biological mimicry is a semiotic phenomenon (Maran 2007). In his essay, "Iconicity," Sebeok (1989) demonstrates that mimicry is a case of iconicity in nature. "A sign is said to be iconic when the modeling process employed in its creation involves some form of simulation" (Sebeok and Danesi 2000), and this is exactly what happens when adaptive sensory biases are exploited. We suggest that this also works the other way around: not only are mimics icons, visual art, or more specifically iconic representations (i.e., realistic art, figurative imagery) can be usefully perceived as mimics resulting from exploitation of human adaptive sensory biases.

Van Damme (2008, p. 38) defines iconic representations as: "The two- or three-dimensional rendering of humans and other animals, or to be more precise, the representation of things resembling those in the external world, or indeed imaginary worlds, fauna and flora especially, but also topographical features, built environments, and other human-made objects." This definition is equally applicable to mimics. We have discussed the case of the egg spots in section 2. What is interesting for the problem of the evolution of human representational art, is that cases of mimicry like this one show how ordinary selection via SE can produce two-dimensional representations (the egg spots) on a surface (the anal fin of the male) of three-dimensional objects (the eggs). To a female cichlid both the signal from the egg and the signal from the egg spot mean "egg", in the sense that she responds indiscriminately towards both those signals with a sucking reaction. In the same way, humans react towards iconic representations — even though we might "know" we are dealing with an illusion — as we react to the real thing. However, there is a difference between humans looking at art and the female cichlid looking at the egg spots: she really is deceived, whereas we know we are looking at a painting of a landscape and not at the real thing. But does this distinction really matter? Not materially. For even though we know that, say, the movie or novel is not real, we still become deeply emotionally involved. Even though we know it is fiction, we react as if it is not. Art exploits our visual system in the case of iconic representations and our emotional and cognitive biases in general, regardless of our consciousness of the distinction between fiction and reality. Human iconic representations are mimics and as such also result from SE. Of course the female reacts toward formal features, design in other words, but this design is not *just* design but design designated to evoke meaning in order to exploit her.

So instead of focusing on geometrical patterns resulting from exploiting activation of early visual areas of the cortex, we focus on the exploitation of perceptual and mental biases for iconic images, that is, on a higher level of visual processing, say, face recognition. Humans have a hair-trigger response to faces. Everywhere we look, we see faces. In cloud formations, in Rorschach inkblots, and so on. The "fusiform face area" is a part of the human visual system, which may be specialized for facial recognition (first described by Sergent et al. 1992). It has recently been suggested that non-face objects may have certain features that weakly trigger the face cells. In the same way objects like rocky outcroppings and cloud formations may set off face radar if they bear enough resemblance to actual faces (Tsao and Livingstone 2008). Whether the hair-trigger response to faces is innate or learned,

it represents a critical evolutionary adaptation, one that dwarfs side effects. The information faces convey is so rich — not just regarding another person's identity, but also their mental state, health, and other factors. It's extremely beneficial for the brain to become good at the task of face recognition and not to be very strict in its inclusion criteria. The cost of missing a face is higher than the cost of declaring a non-face to be a face. So, face recognition is an adaptive sensory bias, which is highly susceptible to exploitation by a depiction of a face as a side effect. If our brain had been less sensitive to faces and had stricter inclusion criteria, perhaps many fewer portraits would have been painted throughout art history.

However strong the bias for faces is, it is not always exploited. In fact, in many prehistoric iconic representations, the face is not extensively elaborated. This is probably due to the specific context in which the depiction is produced and experienced (analogously, it might be that female cichlids are much less sensitive to "egg-like signals" a long time before spawning or after spawning). In many representations of the human figure much more attention is given to specific parts of the body. For instance, in the well known upper paleolithic "Venus" figurines, the head is rather schematic whereas breasts, buttocks, and belly are sculpted in great detail and disproportionately exaggerated. Many different hypotheses have been proposed to explain these distorted female representations (for an overview see McDermott 1996). While speculative, McDermott's (1996) interpretation is particularly interesting for our approach. He proposes that these disproportions resulted from egocentric or autogenous (self-generated) visual information obtained from a self-viewing perspective. In other words, the disproportions in Venus figurines result from the position of the female creators' eyes relative to their own bodies. Self-exploitation of perceptual biases[13] may have been the first step in the emergence of iconic art (Verpooten and Nelissen 2010). Whether these Venus figurines were created as self-representations, as fertility symbols or as erotic items, and whether they were created by men and/or women, they may constitute material evidence of strong adaptive sensory biases for above-mentioned parts of the female body.

Another frequently recurring theme in art history and even more so in art prehistory is the depiction of animals (large wild animals are among the most common themes in cave paintings). Again, a set of adaptive sensory biases might be one of the underlying causes of the tendency to depict animals. In particular, some have speculated that this could well be drawn back to the shared human capacity for "biophilia" (Wilson 1984). Biophilia is defined as a biologically based or innate predisposition to attend to, or affiliate with, natural-like elements or processes (Kellert and Wilson 1993). This set of tendencies is claimed to be the result of human evolution in a natural world in which human survival significantly depended on interactions with natural elements and entities, such as animals (animals could be, for example, predator or prey). Leading biophilia theorists have characterized it as

[13]In this case the adaptive attention toward vital, reproductively functional parts of her own body.

including both positive and negative affective states towards natural-like elements.[14] These affective states may be exploitable by artificial natural-like signals, such as iconic representations of natural elements. For instance, the depictions of large cats in the Grotte Chauvet (believed to be one of the oldest two-dimensional iconic representations) might have elicited a fear response, drawing attention to the depiction. What art needs to be maintained, improved, and reproduced over different generations, in other words to become a "tradition", is to have attention drawn to it by exhibiting captivating or even gripping aesthetics and/or meaning.

9.3.3.2 Self-exploitation

Visual art is extra-corporal. A consequence of its extra-corporal aspect is that it is equally perceivable by its producers as by its receivers. When producers are also perceivers and possess more or less the same sensory system with comparable psychosensory biases, SE would predict they are equally prone to exploitation as any other receivers. In other words, same species SE via extra-corporal traits implies the possibility of self-exploitation. Such a self-exploitation would be evidence that traits can be exploitative without any direct or indirect benefits. And it exists. Courting male fiddler crabs sometimes build mounds of sand called hoods at the entrances to their burrows. It has been shown that burrows with hoods are more attractive to females and that females visually orient to these structures. Interestingly, a recent study showed that males themselves were also attracted towards their own hoods as a consequence of SE or sensory trap (Ribeiro et al. 2006). Hence, hood building causes self-exploitation. The same may apply to human visual art. As artists are always the first ones to perceive their artworks, they are most likely the first ones to be exploited by the signals they produce. Miller (2000) likes to use Picasso as an example of a successful artist, who produced a lot of paintings and had a lot of mistresses, to support his hypothesis that art evolved as a sexual display of good genes. But maybe Van Gogh, who hardly sold any paintings during his lifetime nor had a lot of success with women, to say the least, and locked himself in an attic so to speak to devote himself to his art — to self-exploit his psychosensory biases, is more exemplary of artistic behavior?

9.3.3.3 Art as a spandrel

In Boyd and Richerson's (1985, ch. 8) cultural runaway model aesthetic traits are maintained as non-functional byproducts of the otherwise adaptive indirectly biased cultural transmission. In our SE concept, we entertain the possibility as well that art,

[14] Some also make a distinction between biophilia and biophobia: the former refers to positive, while the latter to negative affective states towards natural-like processes and elements (see Ulrich, 1993). This however seems largely a terminological discussion. The crux of the matter is that there are some biologically-based affective responses to biological categories.

resulting from exploitation of sensory biases, is non-functional. At least, we argue art does not *need* to be functional to have evolved in humans. At certain times and places throughout human evolution, producing and experiencing iconic representations may have been neutral or even maladaptive, depending on specific conditions. The question as to whether visual art such as iconic representations is or has been adaptive or not is thus a tricky one, and hard to answer. Illustrative of this are the divided opinions on adaptiveness of visual art (e.g., Pinker 2002). Moreover, under the proponents of art as adaptive there is no consensus in what way it actually is. To some it is a sexual adaptation (e.g., Miller 1998, 1999, 2000, 2001), to others it is a group bonding adaptation (Coe 2003; Dissanayake 1992, 2001). We conclude that if it can be shown that iconic representations evolve even when they are maladaptive, they definitely will do so when they induce some kind of benefits on any kind of unit of selection. It is a well-known fact in evolutionary biology that the evolutionary function(s) of a particular trait often change substantially over time (cf. Reeve and Sherman, 1993). As stressed by Williams (1966) in his foundational work, adaptation is an "onerous concept" to be demonstrated, not assumed. So, instead of a priori assuming adaptiveness, parsimony demands that we first explore whether art could have evolved even without any adaptive function at all. On our view art can evolve without any adaptiveness assumptions, as a mere consequence of SE. As stated, to the experiencer of a work of art there might be no direct utility involved, just as the female that is mislead by the male mimicking food may not benefit from being sensorily fooled. Here we are not interested in the reproductive success of the artists, but in the (reproductive) success of artistic signals themselves, that spread through cultural transmission[15] regardless of beneficial effects to individuals that transmit them, just as male ornaments evolve through sensory exploitation without the need of any benefits to the females. In this sense, it follows from the SE perspective that iconic art making could have evolved as a culturally transmitted spandrel. Spandrels are byproducts of adaptive capacities but not specifically adaptive themselves, borrowing an architectural term for a necessary but non-functional concomitant of primary load-bearing functions (Gould and Lewontin 1979). In this view, art evolved as a byproduct of sensory biases on the part of experiencing art. (On the part of art making it may have evolved as a byproduct of adaptive skills in

[15]There are some indications from the archaeological record that iconic art production is a mainly culturally transmitted behavior, while the ability to experience and interpret art is not and does in fact predate art production, just as the origin of female sensory biases leading to mate preferences sometimes predates exploitation (e.g., Ryan 1998). One of these indications is provided by Hodgson (2006). He remarks that the "first art", both (pre)historical and developmental (children's first drawings are abstract patterns), is geometric. So what he calls "geometric primitives" predates iconic art. Hodgson further notices that no culture has ever been shown to have an iconic art tradition without a geometric tradition, but vice versa, some cultures only have a geometric tradition. He draws from this that the making of geometrics may be a more accessible process than the making of representational motifs and that knowledge of geometrics may be innate whereas, we could add, making representations is not and requires individual learning and social transmission of skills to be evolutionary maintained (Fig. 2).

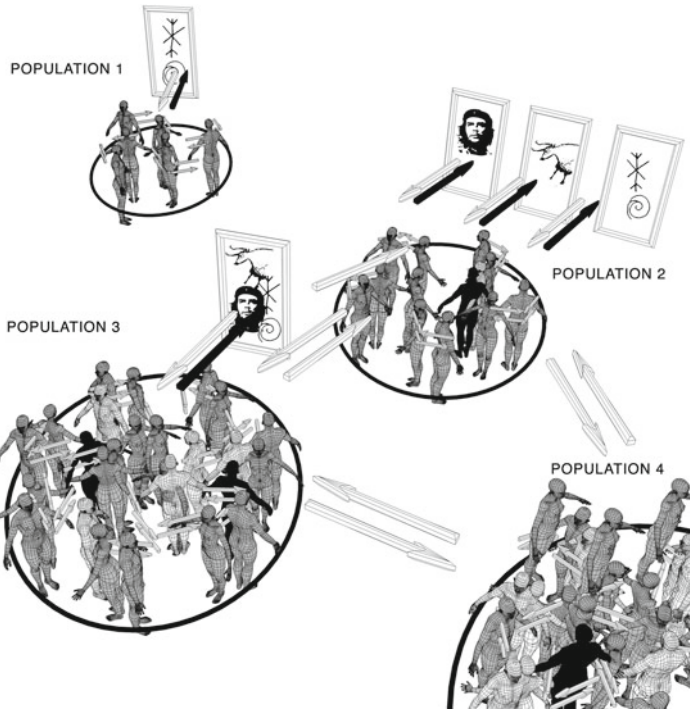

Fig. 9.2 Sensory exploitation, cultural transmission and the influence of the size of the interacting pool of social learners on art. 4 hypothetical populations of social learners and the artworks that they produce are shown. Arrows stand for the direction in which "information" is transmitted. In addition, when the arrow is black, that information directly determines the outward appearance of an artwork. This kind of information will come from the artist that created the work, which are also represented in black. Driven by the process of sensory exploitation, artists will create artworks that exploit their own and others' pre-existing biases. Portraits result from exploitation of biases caused by face recognition and animal depictions from biases caused by biophilia (or biophobia). Population 1 is a small and isolated population of social learners. As a result, the innovations required for its members to produce iconic art will not accumulate. They will however produce abstract art that does not require (much) social learning (Hodgson 2006). In populations 2-4 iconic art traditions will naturally and necessarily occur because they are large and interconnected, creating an interacting pool of social learners that is large enough for innovations required for production of iconic art to spontaneously accumulate and persist regardless any beneficial effects of the resulting artworks. (artwork: Alexandra Crouwers and Jan Verpooten.)

tool use, among other things.) If this artistic behavior does not impose too much costs upon its practicioners in an initial phase, art may have emerged spontaneously, exploiting their biases, without any utility (Fig. 2). It may, however, subsequently be exapted by delivering benefits to art producers and/or experiencers. For a detailed discussion of the relation between SE, cultural transmission and the emergence of visual artistic traditions, see Verpooten and Nelissen (2010).

9.3.3.4 A comparison with religion

Recently there has been a surge of interest in the biology and evolution of religion (e.g., Atran 2002; Culotta 2009; Dawkins 2006, ch. 5; Wilson 2002). Research results in this more intensely studied area may be useful to the study of art. From an evolutionary perspective, religion and art seem to have a lot in common. For one thing, both are complex human behaviors that cannot be explained easily in evolutionary terms. An adaptive explanation based on one selective pressure does not suffice for neither. Religion has maladaptive aspects, probably some functional aspects as well; however, just as in the case of art, depending upon specific conditions and as such varying across populations and cultures in human evolution (for examples, see Atran 2002).

Another interesting similarity between art and religion is that they are both based on some form of primary non-functional deception or illusion (and, as said, possible beneficial "after"- effects only crop up on a secondary level). We have typified art as such from the SE perspective, and in evolutionary religious studies too it is stressed that "[a]ll known human societies, past and present, bear the very substantial costs of religion's material, emotional, and cognitive commitments to factually impossible worlds" (Atran 2002, p. 4). This has two, closely linked, interesting consequences for our discussion. Firstly, the SE perspective may be a useful conceptual tool for evolutionary religious studies too; perhaps some form of SE plays a role in the creation of religious deceptions as it does in art. Secondly, maybe some perceptual or mental biases known to play a role in the creation of religious deception play a role in artistic creation as well. In fact, there is at least one possible candidate for this, similar to the tendency to see faces where there aren't any as a result of a strong bias for face recognition, mentioned above. It is the trip-wired tendency to attribute random events or natural phenomena to the agency of another being, which has been described as a "hypertrophy of social cognition." According to the emerging cognitive model of religion, we are so keenly attuned to the designs and desires of other people that we are hypersensitive to signs of "agents": thinking minds like our own.[16] These findings suggest we all have a bias from childhood to see the natural world as purposefully designed. It's a small step to suppose that the design has a designer. This predisposition to "creationist" explanations has resonance with another tendency in the human mind, the "hypersensitive agency detection device": looking for a thinking "being" even in nonliving things. In classic experiments in the 1940s, psychologists found that people watching animations of circles, triangles, and squares darting about could identify various shapes as characters and infer a narrative (this passage about agents and religion is taken from Culotta 2009). So, exploiting the strong tendency to attribute agency to nonliving things, may have played an important role in the evolution of art as well (and in addition, the experiments also showed evidence of our tendency to make *narratives* with these agents, likely this is also an important tendency exploited in many different arts).

[16]For instance, in an experiment in which undergraduates had to respond under time pressure, they were likely to agree with nonscientific statements such as "The sun radiates heat because warmth nurtures life" (Culotta 2009).

In fact, biophilia, which we discussed earlier as a human bias exploited by depicted animals in cave art, might result from a combination of an hypersensitive agency detection device and the capacity to feel empathy for agents. This possibility should be further explored. Maybe it explains the intense emotions of connectedness with "something larger" that "tree huggers" report to experience.

On this note, this might explain people's disinterest for (post)modern art (especially "concept art"): this kind of art is not developed to captivate our attention through exploiting our agency detection device nor our empathic faculty, rather it is designed to investigate and analyze these responses to art (or to "deconstruct" them as contemporary art theorists would say). It is as if artists switched from the animistic method to the scientific method. Indeed as follows from the studies cited in Culotta (2009, p.785) "scientific literacy" requires "an uphill battle", so too seems to be the case with most modern art.

9.4 Conclusion

Darwin's theory of sexual selection provides a mechanistic basis to explain the evolution of male sexual display traits. This mechanistic approach has proven useful to developing hypotheses about the evolution of human art. Both Boyd and Richerson (1985, ch. 8) and Miller (1998, 1999, 2000, 2001) have applied an indirect-benefit model to the evolution of artistic behavior. We have argued that the mechanistic possibilities SE has to offer have remained underexplored so far, so we have proposed a concept based upon it and we have used it to evaluate these hypotheses.

Central to SE, being closely related to biological mimicry, is that it is in principle a non-functional or even counterfunctional (maladaptive) evolutionary process with regard to the receiver of signals, merely being driven by exploitation of the receiver's sensory biases. Applied to the evolution of human art, we considered these signals as being culturally transmitted spandrels, non-functional evolutionary byproducts of other traits, namely human perceptual and mental biases such as face recognition and agency detection device. This non-functional view on art has some interesting consequences.

Firstly, in both Miller's and Boyd and Richerson's model, "aesthetic preferences" and "aesthetic traits" (i.e., art) coevolved as a result of an indirect-benefit process that may derail into the Fisherian Runaway Process. We have shown, however, that it follows from the SE perspective that at least some of these aesthetic preferences already should exist *before* any aesthetic traits have evolved. The fact that the aesthetic preferences that are exploited in art are also elicited by non-art, like a natural phenomenon such as a tree, may be an indication of this. Moreover, art is not just about pleasing aesthetics. Meaning — pleasing or not — is also important in art. Analogously, meaning is important in SE of which the exploiting traits are mimics, such as egg spots that represent eggs. So, SE also covers the important characteristic of art that it represents something outside the art context.

Secondly, on this non-functional view it follows that art emerged spontaneously in human evolution by exploiting pre-existing biases and not because it was

selected for. As we have hoped to show, benefits are not prerequisite for art to evolve. It would be strange if they were, since on the one hand art today imposes costs without convincing evidence of compensation on any level (cf. Fitch 2006 for music) and since one would expect adaptiveness to differ considerably in populations across time and place (cf. Reeve and Sherman 1993), while nevertheless art is and has been universal for a long time. So, if the costs art usually imposes are not detrimental to the survival of individuals of a population engaging in artistic behavior, it may be borne by the carrying capacity[17] of this population. In fact it follows from our model that it is this carrying capacity of the population that limits the proliferation of culturally transmitted spandrels. If carrying capacity is high we expect high cost art and a lot of it, if it is low we expect the opposite, at equilibrium. As said, all cultures exhibit lower cost abstract art but not all cultures exhibit representational art, which imposes higher costs, for example in terms of time and energy invested in learning and passing on skills (Hodgson 2006, Verpooten and Nelissen 2010). It would be interesting to see whether there is a correlation between the occurrence of representational art and carrying capacity across populations. Hollywood, video games, and virtual reality are the cave art of today and in absolute terms they are obviously much more costly than cave art; maybe they are the direct result of the exceptionally high joint carrying capacity of current industrialized populations in combination with being culturally transmitted spandrels emerging naturally from exploiting our biases.

Thirdly, compensating for the costs or not, beneficial effects might influence the evolution of art on a secondary level. There are at least two types of possible benefits which may exert selective pressures on the evolution of art. One is transmission of valuable (functional) information through art. Some art may have evolved adaptively as a means of storing and transmitting valuable information. This is an appealing proposition; however, its role may not be so important. Why use art if you have language, which may plausibly be a far more efficient instrument to transmit and maintain information? Art may, however, instead of transmitting information itself be useful in *facilitating* transmission of information through language (such as the use of rhyme for better memorizing). Anyway, this possibility should be somehow taken into account in the above-suggested test, because it would mean some sort of compensation for art's costs. The second possible benefit was discussed in great detail in this chapter: the individual (male) benefit of increased reproductive success. When exactly this kind of secondary process will operate, should be further explored. Fuller et al. (2005) have suggested a number of tests to distinguish SE from other preference models in sexual selection in practice. These tests may be used for the

[17] According to Boyd and Richerson (1985, p. 278) each culture may contain a number of non-functional or counterfunctional traits at equilibrium. By carrying capacity we mean the number of non-functional or counterfunctional cultural traits a population of social learners can maintain. We suggest it depends on the utility of other traits in the population that compensate for the costs of counterfunctional traits, such as technological skills and on the size of the population (a larger population can sustain more costly traits), among other things (cf. Shennan 2001; Henrich 2004).

same purpose regarding the relative role of SE and indirect-benefit processes in the evolution of artistic behavior. However, even if indirect benefits prove to play some role under certain conditions, it would not disconfirm the SE view on the evolution of art. If art were a sexual adaptation, it would not lower the costs for the population as a whole. So it does not undermine our prediction of a relation between carrying capacity and abundance of costly art in a population.

Even if art proves to have been adaptive most of the time in human evolution, to individuals as a mating display, to groups as a container of valuable information or as a facilitator of bonding, it will draw upon existing perceptual and mental biases. As a consequence, all of the major hypotheses about art will need to make use of the SE concept, which will need to play a central role in articulating all of them.

Acknowledgements We thank two anonymous referees, the editors of this volume, Johan Braeckman, Tijs Goldschmidt, Yannick Joye, and David Sloan Wilson for sharing useful suggestions and/or commenting on earlier drafts. Also many thanks to Lokaal01_Antwerpen (www.lokaal01.org) for the opportunity to discuss our ideas among scientists and artists. We thank the Konrad Lorenz Institute for Cognition and Evolution Research for the support. Last but not least we thank Thomas Reydon and Katie Plaisance for having us at the conference.

References

Adajian, T. (2007): The Definition of Art. In: E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, http://plato.stanford.edu/archives/fall2008/entries/art-definition/

Andersson, M. (1994): *Sexual Selection*. Princeton: Princeton University Press.

Arak, A., and Enquist, M. (1993): Hidden preferences and the evolution of signals. *Philosophical Transactions of the Royal Society of London*, Series B 340: 207–213.

Arak, A., and Enquist, M. (1995): Conflict, receiver bias and the evolution of signal form. *Philosophical Transactions of the Royal Society of London*, Series B 349: 337–344.

Arnqvist, G. (2006): Sensory exploitation and sexual conflict. *Philosophical Transactions of the Royal Society of London*, Series B 361: 375–386.

Atran, S. (2002): *In Gods we Trust: The Evolutionary Landscape of Religion*. New York: Oxford University Press.

Autumn, K., Ryan, M.J., and Wake, D.B. (2002): Integrating historical and mechanistic biology enhances the study of adaptation. *Quarterly Review of Biology* 77: 383–408.

Basolo, A.L. (1990): Female preference predates the evolution of the sword in swordfish. *Science* 250: 808–810.

Borgia, G. (1995): Complex male display and female choice in the spotted bowerbird: Specialized functions for different bower decorations. *Animal Behavior* 49: 1291–1301.

Boyd, R. and Richerson, P.J. (1985): *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Boyd, R. and Richerson, P.J. (2005): *The Origin and Evolution of Cultures*. Chicago: University of Chicago Press.

Burley, N. (1988): Wild zebra finches have band-color preferences. *Animal Behavior* 36: 1235–1237.

Christy, J.H. (1995): Mimicry, mate choice, and the sensory trap hypothesis. *American Naturalist* 146: 171–181.

Coe, K. (2003): *The Ancestress Hypothesis: Visual Art as Adaptation*. New Brunswick: Rutgers University Press.

Culotta, E. (2009): On the origin of religion. *Science* 326: 784–787.

Darwin, C. (1871): *The Descent of Man, and Selection in Relation to Sex*. 2 vols. London: John Murray.
Dawkins, M.S. and Guilford, T. (1996): Sensory bias and the adaptiveness of female choice. *American Naturalist* 148: 937–942.
Dawkins, R. (1982): *The Extended Phenotype: The Gene as the Unit of Selection*. Oxford: W. H. Freeman.
Dawkins, R. (2006): *The God Delusion*. Boston: Houghton Mifflin.
Dissanayake, E. (1992): *Homo Aestheticus: Where Art Comes From and Why*. Seattle: University of Washington Press.
Dissanayake, E. (2001): Birth of the arts. Natural History 109: 84–91.
Driscoll, C. (2006): The bowerbirds and the bees: Miller on art, altruism, and sexual selection. *Philosophical Psychology* 19: 507–526.
Eibl-Eibesfeldt, I. (1989a): *Human Ethology*. Translated by Pauline Wiessner-Larsen and Annette Heunemann. New York: Aldine Degruyter.
Eibl-Eibesfeldt, I. (1989b): The biological foundations of aesthetics. In: I. Rentschler, B. Hertzberger and D. Epstein (Eds.) *Beauty and the brain: Biological Aspects of Aesthetics* (pp. 29–68). Basel: Birkhauser.
Endler, J.A. (1992): Signals, signal conditions, and the direction of evolution. *American Naturalist* 139: 125–153.
Endler, J.A., and Basolo, A.L. (1998): Sensory ecology, receiver biases and sexual selection. *Trends in Ecology and Evolution* 13: 415–420.
Enquist, M., and Arak, A. (1993): Selection of exaggerated male traits by female aesthetic senses. *Nature* 361: 446–448.
Enquist, M., and Arak, A. (1994): Symmetry, beauty and evolution. *Nature* 372: 169–172.
Eshel, I., Volovik, I., and Sansone, E. (2000): On Fisher-Zahavi's handicapped sexy son. *Evolutionary Ecology Research* 2: 509–523.
Fisher, R.A. (1930): *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press.
Fitch, W.T. (2006): The biology and evolution of music: a comparative perspective. *Cognition* 100: 173–215.
Flinn, M.V., and Alexander, R.D. (1982): Culture theory: The developing synthesis from biology. *Human Ecology* 10: 383–400.
Fuller R.C., Houle D., and Travis J. (2005): Sensory Bias as an Explanation for the Evolution of Mate Preferences. *American Naturalist* 166: 437–446.
Gaut, B. (2005): The Cluster Account of Art Defended. *British Journal of Aesthetics* 45: 273–288.
Ghirlanda, S., and Enquist, M.A. (2003): A century of generalization. *Animal Behaviour* 66: 15–36.
Gould, S.J. (1991): Exaptation: A crucial tool for evolutionary psychology. *Journal of Social Issues* 47: 43–65.
Gould, S.J., and Lewontin, R.C. (1979): The spandrels of San Marco and the panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society* B 205: 581–598.
Guilford, T., and Dawkins, M.S. (1991): Receiver psychology and the evolution of animal signals. *Animal Behaviour* 42: 1–14.
Haufe, C. (2008): Sexual selection and mate choice in evolutionary psychology. *Biology and Philosophy* 23: 115–128.
Henrich, J., and McElreath, R. (2003): The evolution of cultural evolution. *Evolutionary Anthropology* 12: 123–135.
Henrich, J. (2004): Demography and Cultural Evolution: How Adaptive Cultural Processes can Produce Maladaptive Losses: The Tasmanian Case. *American Antiquity* 69: 197–214.
Hodgson, D. (2006): Understanding the origins of Paleoart: The neurovisual resonance theory and brain functioning. *PaleoAnthropology* 2006: 54–67.
Jersakova, J., Johnson, S. D., and Kindlmann, P. (2006): Mechanisms and evolution of deceptive pollination in orchids. *Biological Reviews of the Cambridge Philosophical Society* 81: 219–235.
Johnstone, R. A. (1994): Female preference for symmetrical males as a byproduct of selection for mate recognition. *Nature* 372: 172–175.

Kellert, S.R., and Wilson, E.O. (1993): *The Biophilia Hypothesis*. Washington: Island Press.
Kirkpatrick, M. and Ryan, M.J. (1991): The evolution of mating preferences and the paradox of the lek. *Nature* 350: 33–38.
Kohn, M., and Mithen S. (1999): Handaxes: Products of sexual selection? *Antiquity* 73: 518–26.
Kokko, H., Brooks, R., Jennions, M.D., and Morley, J. (2003): The evolution of mate choice and mating biases. *Proceedings of the Royal Society* B 270: 653–664.
Lande, R. (1981): Models of speciation by sexual selection on polygenic traits. *Proceedings of the National Academy of Sciences of the United States of America* 78: 3721–3725.
Low, B.S. (1979): Sexual selection and human ornamentation. In: N. A. Chagnon and W. Irons (Eds.) *Evolutionary biology and human social behavior* (pp. 462–487). Boston: Duxbury Press.
Maran, T. (2007): Semiotic interpretations of biological mimicry. *Semiotica* 167–1/4: 223–248.
McElreath, R., and Henrich, J. (2007): Dual inheritance theory: the evolution of human cultural capacities and cultural evolution. In: R. Dunbar and L. Barrett (Eds.) *Oxford Handbook of Evolutionary Psychology*. Oxford: Oxford University Press.
McDermott, L. (1996): Self-representation in upper paleolithic female figurines. *Current Anthropology* 37: 227–275
Miller, G.F. (1998): How mate choice shaped human nature: A review of sexual selection and human evolution. In: C. Crawford and D. Krebs (Eds.) *Handbook of evolutionary psychology: Ideas, issues, and applications* (pp. 87–129). Mahwah: Lawrence Erlbaum.
Miller, G.F. (1999): Sexual selection for cultural displays. In: R. Dunbar, C. Knight, and C. Power (Eds.) *The evolution of culture* (pp. 71–91). Edinburgh University Press,.
Miller, G.F. (2000): *The Mating Mind*. London: Heinemann.
Miller, G.F. (2001): Aesthetic fitness: How sexual selection shaped artistic virtuosity as a fitness indicator and aesthetic preferences as mate choice criteria. *Bulletin of Psychology and the Arts* 2: 20–25.
Pasteur, G. (1982): A classificatory review of mimicry systems. *Annual Review of Ecology, Evolution, and Systematics* 13: 169–199.
Payne, R.J.H. and Pagel, M. (2000): Inferring the origins of state-dependent courtship traits. *American Naturalist* 157: 42–50.
Pinker, S. (1997): *How the mind works*. New York: Norton.
Pinker, S. (2002): *The blank slate. The modern denial of human nature*. New York: Viking.
Ramachandran, V.S., and Hirstein, W. (1999): The science of art: A neurological theory of aesthetic experience. *Journal of Consciousness Studies* 6: 15–51.
Reeve, H.K. and Sherman, P. (1993): Adaptation and the goals of evolutionary research. *Quarterly Review of Biology* 68: 1–32.
Ribeiro, P.D., Christy J.H., Rissanen R.J. and Kim T.W. (2006): Males are attracted by their own courtship signals. *Behavioral Ecology and Sociobiology* 61: 81–89.
Richerson, P:J. and Boyd, R. (2001): Built for Speed, Not for Comfort: Darwinian Theory and Human Culture. *History and Philosophy of the Life Sciences* 23: 425–465.
Ridley, M. (1981): How the peacock got his tail. *New Scientist* 91: 398–401.
Rodd, F.H., Hughes, K.A., Grether, G.F., and Baril, C.T. (2002): A possible non-sexual origin of a mate preference: are male guppies mimicking fruit? *Proceedings of the Royal Society* B 269: 475–481.
Rodriguez, R.L. and Snedden, W. (2004): On the functional design of mate preferences and receiver biases. *Animal Behaviour* 68: 427–432.
Ryan, M.J. (1990): Sexual selection, sensory systems and sensory exploitation. *Oxford Surveys in Evolutionary Biology* 7: 157–195.
Ryan, M.J. (1995): Female responses to ancestral advertisement calls in tungara frogs. *Science* 269: 390–392.
Ryan, M.J. (1998, review 1999): Sexual Selection, Receiver Biases, and the Evolution of Sex Differences. *Science* 281: 1999–2003.
Ryan, M.J., and Keddy-Hector, A. (1992): Directional patterns of female mate choice and the role of sensory biases. *American Naturalist* Supplement 139: 4–35.

Ryan, M.J., and Rand, A.S. (1990): The sensory basis of sexual selection for complex calls in the tungara frog, *Physalaemus pustulosus* (sexual selection for sensory exploitation). *Evolution* 44: 305–314.

Ryan, M.J., and Rand, A.S. (1993): Sexual selection and signal evolution: the ghost of biases past. *Philosophical Transactions of the Royal Society of London* B 340: 187–195.

Schiestl, F.P., and Cozzolino, S. (2008): Evolution of sexual mimicry in the orchid subtribe orchidinae: the role of preadaptations in the attraction of male bees as pollinators. *BMC Evolutionary Biology* 8: 27.

Sebeok, T.A. (1989): Iconicity. In: T.A. Sebeok (Ed.) *The Sign and Its Masters* (pp 107–127). Lanham: University Press of America.

Sebeok, T.A., and Danesi, M. (2000): *The Forms of Meaning: Modeling Systems Theory and Semiotic Analysis*. Berlin: Mouton de Gruyter.

Sergent, J., Ohta, S., and MacDonald, B. (1992): Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain* 115: 15–3.

Shaw, K.L. (1995): Phylogenetic tests of the sensory exploitation model of sexual selection. *Trends in Ecology and Evolution* 10: 117–120.

Shennan, S. (2001): Demography and cultural innovation: a model and its implication for the emergence of modern human culture. *Cambridge Archeological Journal* 11: 5–16.

Sober, E. (1984): The Nature of Selection. Evolutionary Theory in Philosophical Focus. Cambridge (MA): MIT Press.

Tobler, M. (2006): Die Eiflecken bei Cichliden: Evolution durch Nutzung der Sinne? (The eggspots of cichlids: Evolution through sensory exploitation?). *Zeitschrift für Fischkunde* 8: 39–46.

Tsao, D.Y. and Livingstone, M.S. (2008): Mechanisms of face perception. *Annual Review of Neurosciences* 31: 411–437.

Ulrich, R.S. (1993): Biophilia, biophobia, and natural landscapes. In: R.S. Kellert, E.O. Wilson (Eds.) *The Biophilia Hypothesis* (pp. 73–137). Washington: Island Press.

Van Damme, W. (2008): Introducing world art studies. In: W. Van Damme and K. Zijlmans (Eds.) *World Art Studies: Exploring Concepts and Approaches*. Amsterdam: Valiz.

Verpooten, J. and Nelissen, M. (2010): Sensory exploitation and cultural transmission: the late emergence of iconic representations in human evolution. *Theory in Biosciences* 129: 211–221.

West-Eberhard, M.J. (1992): Adaptation: current usages. In: E.F. Keller and E.A. Lloyd (Eds.) *Keywords in Evolutionary Biology* (pp. 13–18). Cambridge (MA): Harvard University Press.

West-Eberhard, M.J. (1984): Sexual selection, competitive communication, and species-specific signals in insects. In: T. Lewis (Ed.) *Insect Communication* (pp. 283–324). London: Academic Press.

Wickler, W. (1962): "Egg-dummies" as natural releasers in mouth-breeding cichlids. *Nature* 194: 1092–1093.

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

Wilson, E.O. (1984): *Biophilia*. Cambridge: Harvard University Press.

Wilson, D.S. (2002): *Darwin's Cathedral: Evolution, Religion, and the Nature of Society*. Chicago: University of Chicago Press.

Zahavi, A. (1975): Mate selection: A selection for a handicap. *Journal of Theoretical Biology* 53: 205–214.

Zahavi, A. (1991): On the definition of sexual selection, Fisher's model, and the evolution of waste and of signals in general. *Animal Behaviour* 42: 501–503.

Zahavi, A., and Zahavi, A. (1997): *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford: Oxford University Press.

Chapter 10
Heuristic Evolutionary Psychology*

Armin W. Schulz

10.1 Introduction

As is widely known, evolutionary psychologists claim that appealing to the mind as an evolved, biological organ is immensely useful for bringing the science of psychology forward. In particular, they think that important discoveries about how our minds work can be especially easily made once we consider the issue from an evolutionary biological point of view: this perspective is said to bring considerations into view that psychologists would otherwise have missed (see e.g. Pinker, 1997; Cosmides & Tooby, 1992; Carruthers, 2006).

However, from the moment of its inception, this kind of approach towards doing psychology has also not been without its critics. In particular, evolutionary psychology has frequently been accused of resting on nothing but (adaptationist) just-so story telling. More specifically, many critics of the program have claimed that the evolutionary hypotheses considered by these researchers are completely evidentially ungrounded, and therefore amount to nothing more than unconvincing speculation. For this reason, the scientific credentials of the program are often put into doubt: far from widening and systematising debates about the structure of our minds, evolutionary psychology seems rather to narrow and confuse them (see e.g. Richardson, 2007; Buller, 2005; Dupré, 2001; see also Kitcher, 1985, pp. 9-10).

In order to respond to this criticism, evolutionary psychologists have two major options open to them. Firstly, they can claim that the criticism rests on a false presupposition. Specifically, they can argue that, by and large, they *do* have the required evidence for the hypotheses they are considering. For this reason, they should not be

*I would like to thank Elliott Sober, Dan Hausman, Larry Shapiro, the editors of this volume, and two anonymous referees for many useful remarks about previous versions of this paper.

A.W. Schulz (✉)
Department of Philosophy, Logic, and Scientific Method, London School of Economics and Political Science, Houghton St., London WC2A 2AE, UK
e-mail: A.W.Schulz@lse.ac.uk

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*, Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_10,

accused of providing nothing but unscientific speculations – their approach does not differ substantially from other (reasonable) applications of evolutionary theory (see e.g. Pinker & Bloom, 1990; Cosmides & Tooby, 1992; 2005). Secondly, they can claim that the above criticism is subject to a fundamental misunderstanding of their program: they use the evolutionary perspective merely as a *heuristic device* (see e.g. Machery, forthcoming; Samuels et al. 2004; Shapiro & Epstein, 1998; Buss et al., 1998, p. 545; Andrews et al., 2002, p. 538). For this reason, their use of evolutionary theory is not in need of evidential backing – heuristic devices have the goal of *leading to* evidence for some theory; they themselves, though, do not need to be evidentially supported.

It is this second response that I want to consider further here. The main reason for this is that, as it stands, it is insufficiently well supported. In particular, no concrete cases have been presented that clearly bring out the ways in which evolutionary theory has been used in a purely heuristic way in psychology (see also Davies, 2002). Presenting such cases, though, is necessary, since the accusation that evolutionary psychology is evidentially ungrounded concerns *currently practiced*, *actual* evolutionary psychology – not some *merely possible*, *fictional* evolutionary psychology.

For this reason, I here present arguments for two conclusions. Firstly, I try to show that the typical ('high church') examples of evolutionary psychological research in fact do *not* fit to a heuristic reading of the program. Secondly, though, I also aim to show that there are cases that *do* fit such a reading – however, there are not very many of them, and it is far from straightforward to find them. In this way, I hope to make clear that the heuristic defence of evolutionary psychology is not entirely unconvincing – but also that it is far less often applicable than is supposed by many philosophers and psychologists (see e.g. Machery, forthcoming; Samuels et al. 2004; Andrews et al., 2002).

Before presenting these arguments in more detail, it is useful to make a brief remark about how the term 'evolutionary psychology' is to be understood here. In general, there are two different ways of using this term: a narrow and a wide one.[1] According to the narrow usage, 'evolutionary psychology' refers just to the 'Santa Barbara' school of evolutionary psychologists – comprising primarily Leda Cosmides, John Tooby, David Buss, Robert Trivers, Martin Daly, and Margo Wilson (see e.g. Buller, 2005; see also Richardson, 2007; Sterelny, 2003, chap. 6). According to the wide usage, 'evolutionary psychology' refers to evolutionary approaches to the mind generally, independently of any specific doctrines that particular evolutionary psychologists might choose to defend. As will also become clearer below, I here always use the term in the latter, wide sense: the issue is whether the introduction of evolutionary theory into psychology *in general* can be defended from a heuristic point of view, not whether specific theories of specific evolutionary psychologists can be defended in this way (see also Carruthers, 2006, p. 36; Machery, forthcoming).

[1] Buller (2005) calls these two understandings of evolutionary psychology 'EP' and 'ep' respectively.

The paper is structured as follows. In section 10.2, I briefly make clear how to determine when evolutionary theory is applied in a genuinely heuristic way, and when not. In section 10.3, I use the results of the previous section to show that most of the standard examples of evolutionary psychology do *not* employ evolutionary theory in a heuristic manner. In section 10.4, I similarly show that Gergely & Csibra's work on the psychology of human pedagogy *does* exemplify a heuristic form of evolutionary psychology. I conclude in section 10.5.

10.2 Evolutionary Theory as an Explanatory and Heuristic Tool

In order to determine whether evolutionary psychology can really be defended from a heuristic point of view, it is necessary to begin by making clearer what it means, more generally, to use evolutionary theory in a heuristic way. In turn, this requires us to get clearer on what the relevant non-heuristic uses of evolutionary theory are: what is the contrast class to which heuristic applications of the evolutionary perspective are meant to be compared?

Now, in the present context, it seems clear that the major alternative to a heuristic use of evolutionary theory is an *explanatory* (or evidential) use.[2] At least on the face of it, when evolutionary theory is not used in a heuristic way, it is used to give an account of why certain things happened in the way they did – i.e. it is meant to *explain* a set of facts. In more detail, this explanatory use of the theory can be described as follows.

Explanatory applications of evolutionary theory (as of any other theory) aim to help us account for phenomena that are already known to exist: they try to determine what caused some phenomenon to come about, or what led to it having the particular features it actually has, or some such. Of course, for this to be possible, the phenomenon at issue needs to be (somewhat) well understood to begin with: in particular, we at least need to know *that* it exists and what (some of) its *features* are – for it is this existence and these features that are at the heart of the explanatory project. In this kind of case, therefore, knowledge of the phenomenon comes first, and the appeal to evolutionary theory comes later: here, the theory tracks the data, and not the other way around.[3]

This is very different when it comes to *heuristic* uses of evolutionary theory, however. There, the theory aims to make helpful suggestions about which issues are worth exploring further – i.e. it points to some overlooked phenomena that it would be good to know more about. Trivially, for this to be at all compelling, these

[2]Note that there may also be uses of evolutionary theory that are not well classified as being either of the explanatory or the heuristic sort (e.g. when it comes to the testing of the truth of evolutionary theory itself). However, for present purposes, maintaining the dichotomy in the text is sufficient.

[3]Note that the reason the theory cites for why the phenomenon of interest came about need not be the true reason – for all we know, the application of the theory might be mistaken in various ways. The point here is just that this kind of application at least *aims* at truth.

phenomena must neither be already known, nor must they be inherently uninteresting: for a heuristic application of evolutionary theory to be truly fruitful, it needs to suggest phenomena that we had no idea existed, and which are of major theoretical concern to us. In this kind of case, therefore, the application of the theory comes first, and the knowledge of the phenomenon comes later – here, the data track the theory, and not the other way around.

In order to understand this heuristic application of evolutionary theory better, it is further important to note that there are two very different interpretations of it. Firstly, this kind of application could be read in an *expressive* way: on this reading, the claim that evolutionary theory suggests interesting phenomena to investigate is to be seen to refer to the way in which evolutionary psychologists express themselves in their work – it is a claim about what these evolutionary psychologists point to when they describe the origins of their studies. Secondly, this kind of application could be read in a *structural* way: on this reading, the claim that evolutionary theory suggests interesting phenomena to investigate is to be seen to refer to the most compelling way in which the relationship between evolutionary theory and the phenomenon at issue can be characterised – it is a claim about how the receipt of the relevant data is *best* accounted for, independently of whether this agrees with the evolutionary psychologists' own assessment of the situation.[4]

Now, for present purposes, it is only this second, structural reading that is relevant. Primarily, this is because the question at stake is whether evolutionary psychologists are *justified* in claiming that evolutionary theory can be used in a heuristic way in psychology – and not just whether they *do*, in fact, claim this. This is important, as it immediately makes clear that finding out exactly what various evolutionary psychologists are saying about their research is not sufficient to determine whether a plausible heuristic form of evolutionary psychology exists: for all we know, these evolutionary psychologists may be *wrong* about the role that evolutionary theory plays in their theory – after all, their expertise is in the study of the mind, not in the analysis of research programs. For this reason, the expressed opinions of evolutionary psychologists can, at best, make for *evidence* about whether a plausible heuristic form evolutionary psychology exists – by themselves, though, these opinions cannot *answer* this question. Hence, the expressive reading of heuristic uses of evolutionary theory can be left aside in what follows – only the *structural* reading matters here.

With this clarification in the background, the distinction between heuristic and non-heuristic (explanatory) applications of evolutionary theory can be summarised as

[4]Note that this structural understanding of heuristic evolutionary psychology should be distinguished from a purely psychological one: in the latter case, the goal is to uncover *the exact psychological processes* that led particular researchers to engage in the kinds of activities they did engage in. However, this purely psychological project is not so interesting here, since, for a general defence of the plausibility of heuristic evolutionary psychology, it is not necessary to determine *exactly* how the consideration of evolutionary theory has led some particular researcher to do one experiment rather than another. All that needs to be shown here is that it is plausible that evolutionary theory *somehow* played a crucial role in this – however, exactly, it did so.

follows. Assume that there is some set of empirical findings *E* (e.g. an experimentally or naturally occurring phenomenon), and some application of evolutionary theory *A*. Then *A* is a *heuristic* application of evolutionary theory vis-a-vis *E* if

(H) *A* gives rise to the discovery of *E*.

This is in contrast to *A* being an *explanatory* application of evolutionary theory vis-a-vis *E*, in which case

(X) *A* specifies a reason for the occurrence of *E*.

A few aspects of this distinction between (H) and (X) are usefully clarified here. Firstly, both (H) and (X) are very abstract, and leave a number of questions open. In particular, they do not specify in detail what it takes for the application of some theory to be a cause for the discovery of some set of experimental findings, as opposed to specifying a reason for the occurrence of the latter. Fortunately though, for present purposes, these kinds of questions can be left open: even though it might not be entirely clear how the distinction between causes for the discovery of E and reasons for the occurrence of E can be characterised in general, it seems clear that there is some such distinction, and that we can often recognise it fairly easily. Nothing else is needed here.

Secondly, it is important to note that (H) and (X) relativise the heuristic uses of evolutionary theory to a specific area – namely the set of empirical findings *E* in question. This is important to note, as otherwise, the two criteria would be trivial: it seems clear that for almost any theory – evolutionary theory included – there will be fruitful heuristic applications for *some* E (i.e. in *some* experimental context). The point at stake, though, is to determine whether evolutionary theory has heuristic applications for a *given* E – i.e. in a *given, fixed experimental context*.

Thirdly and relatedly, note that (H) and (X) are not necessarily mutually exclusive. In particular, it is plausible that many explanatory uses of a theory have heuristic effects; equally, it is plausible that many heuristic applications of a theory *also* turn out to be explanations of the phenomena they help to discover. However, this does not mean that the two criteria collapse into each other – just because some application of a theory can satisfy both (H) and (X) (for the same or different *E*'s), this does not mean that, in general, there is no difference in how the theory is used in the two situations.

Fourthly and finally, it is important not to conflate the *testing* of the explanation offered by (X) with a heuristic application of evolutionary theory in the vein of (H). *Any* testable explanation suggests further issues to investigate – namely, all those that help determine whether the explanation is true. However, this is an extremely weak and uninteresting sense of being a heuristic device, which ultimately reduces to being a compelling explanation. In the present context, more is looked for than that: what is at stake is whether evolutionary theory can plausibly be said to be used as a heuristic device *over and above* its providing possible testable explanations of some psychological phenomenon – for only this would make for a cogent heuristic-based defence against the criticism that evolutionary psychology is evidentially unconvincing. This will become important again below.

Given (H) and (X), it is now possible to consider whether there really are heuristic applications of evolutionary theory in psychology – and thus, whether the heuristic interpretation of evolutionary psychology can be made plausible. In order to do this, I lay out two (representative) examples of this kind of research – Cosmides & Tooby's work on *cheater detection* and Gergely & Csibra's work on *natural pedagogy* – and assess the extent to which they exemplify applications of evolutionary theory of type (H). Note that the goal in discussing these research programs is not to determine whether they are to be seen as successful or as yielding true conclusions; instead, the aim is merely to assess whether the they give clear support to a *heuristic reading* of evolutionary psychology. Accordingly, I shall not present or discuss in detail any criticisms that have been or could be made of these projects, and simply consider them as they stand.

10.3 Cheater Detection and Heuristic Evolutionary Psychology

In order to determine whether Cosmides & Tooby's work can be used as a basis for a defence of a heuristic form of evolutionary psychology, I proceed in two steps. Firstly, I lay out their research in as neutral and faithful a manner as possible. Secondly, I assess this research using the tools developed in the previous section. Consider these two steps in turn.

10.3.1 Cosmides and Tooby on Cheater Detection

Cosmides & Tooby begin their research by drawing attention to two sets of social psychological findings, established using the classic Wason Selection Task (see Wason, 1966).[5] Firstly, human subjects often do not do well when it comes to assessing the truth of various conditional statements (see e.g., Wason, 1983; Cosmides, 1985). For example, when trying to assess whether the statement 'If a card has a vowel on one side, it has an even number on the other' is true (concerning a particular set of cards), people tend to want to ascertain whether cards that have a vowel on one side have an even number on the other *and* whether cards that have an even number on one side have a vowel on the other – even though the latter conjunct could not possibly falsify the above conditional (see e.g. Cosmides & Tooby, 1989).

Secondly and in contrast to the above, though, other studies have shown that people can also be quite *good* at assessing the truth of a conditional statement (see e.g. Johnson-Laird, 1982; Cosmides, 1985). For example, when asked to

[5]The Wason Selection Task consists in presenting subjects with a set of two-sided cards (typically four) and then asking them to point to the cards they think *must* be turned over in order to evaluate the truth of some statement (typically a conditional) concerning these cards.

assess whether the statement 'If a person is drinking beer, then they must be over 21 years old' is true (concerning a set of people in a bar), people quickly and correctly seek to ascertain how old the beer drinkers are and what the *under* 21-years olds – not the *over* 21-year olds – are drinking (see e.g. Cosmides & Tooby, 1989).

Cosmides & Tooby have found this difference in reasoning ability to persist under many varieties of the above two conditionals. In fact, they (claim to have) noticed that the only aspect of the situation that reliably predicted subjects' success at solving the Wason Selection Task was whether the content of the conditional concerned the violation of a convention of social exchange (see e.g. Cosmides & Tooby, 2005, 1992b). That is, Cosmides & Tooby found that people tended to do well when their task was to assess whether someone cheated in a social exchange, but badly when their task was to assess conditionals about other topics (for more on this, see e.g. Cosmides & Tooby, 1992b; for a critical view, see e.g. Buller et al., 2005; for some replies, see Cosmides & Tooby, 2008).

Crucially, Cosmides & Tooby then went on to claim that this improved performance is not surprising when looked at from an evolutionary point of view. In particular, they argue that, given the great importance of the social environment in our evolutionary history, we are likely to have evolved adaptations for dealing with other people. Specifically, Cosmides & Tooby argue that we needed to find a way to solve the *free-rider problem*: since individuals that take advantage of a social arrangement without paying the cost for maintaining it can make this kind of arrangement unstable, a way needs to be found to prevent cheating. To do *that*, though, it needs to be possible to *identify* the cheaters – for only then can they be prevented from or punished for any possible free-riding (Cosmides & Tooby, 1992b).

Accordingly, Cosmides & Tooby further argue that it is plausible to think that we have evolved cognitive adaptations that make exactly this possible. Specifically, they claim that we are likely to possess a 'cheater detection module': a mental mechanism that is attuned to the occurrence of social exchanges, and which allows us to determine the circumstances in which the conventions governing these are violated (see e.g. Cosmides & Tooby, 1992b). In contrast to this, they think that we did *not* have to evolve adaptations for reasoning with conditionals *in general*: since solving general logic problems was not part of our 'environment of evolutionary adaptedness' (EEA), there was no need to evolve a general 'logical reasoning module'. Finally, Cosmides & Tooby claim that, together, these facts account for the above effect difference in our ability to evaluate the truth of conditional statements: this difference is the result of the existence of specific adaptations for dealing with situations of social exchange, but none for dealing with conditionals in general (see e.g. Cosmides & Tooby, 1989).

10.3.2 The Place of Evolutionary Theory in Cosmides and Tooby's Research

For present purposes, what is most important about this summary of Cosmides & Tooby's research is that it quite clearly shows that evolutionary theory is here

applied mostly in an *explanatory*, and not in a *heuristic* way. To see this, note that the key social psychological effect difference to be accounted for had *already been known* when Cosmides & Tooby put their evolutionary hypotheses forward: the difference in the success rates in evaluating the two kinds of conditionals was the *starting point* of their evolutionary investigation – and not an end state (this comes out particularly clearly in Cosmides, 1985, but any of their other publications supports this reading, too). Given (H) and (X) from section 10.2, this therefore makes clear that evolutionary theory is here being used in an explanatory way: it is best understood as putting forward a possible reason for the occurrence of these differences – not as a tool that led to their discovery.

In slightly more detail, the above analysis of Cosmides & Tooby's research shows that they should not be seen as having started by defending the proposition that humans in the EEA needed to have specialised cognitive tools for the detection of cheaters, and then using the Wason selection task to determine whether there really are traits of this sort. Instead, they should be seen as having *started* with the puzzling results of the Wason Selection Task, and then seeking to find an evolutionary reason that might *explain* these results (see e.g. Cosmides, 1985). Because of this, it seems clear that this case does not support a heuristic interpretation of evolutionary psychology – it quite simply does not exemplify any heuristic application of evolutionary theory at all.

Now, at this point, the following three objections to this conclusion might come to mind. Firstly, one might think that this conclusion underestimates the importance of Cosmides & Tooby's evolutionary perspective for *organising and sorting* the findings from the Wason Selection Task (see e.g. Samuels et al., 2004). Before Cosmides & Tooby's evolutionary work, these findings were hard to interpret and were generally seen to present a major psychological conundrum (see e.g. Cosmides, 1985, and the references therein). What Cosmides & Tooby did was to clarify how these findings hang together, and show how they can consistently be made sense of. For this reason, it may seem that there is a legitimate and defensible heuristic use of the evolutionary perspective here after all: the application of evolutionary theory guided us in understanding the relevant empirical findings better (see also Samuels et al., 2004).

However, while plausible on the surface, this objection does not in fact address issues of relevance in the present context. In the main, this is because the clarificatory use of evolutionary theory it appeals to is actually an instance of (X), and not of (H). This comes out most clearly from noting that Cosmides & Tooby's evolutionary hypothesis clarifies the interpretation of the divergent findings of the Wason Selection Task only to the extent that it is *true*. In particular, if it were to turn out that some other factor determines why people do better at evaluating certain conditionals than others (as has been claimed, e.g., by Buller, 2005, pp. 173-177), Cosmides & Tooby's way of grouping the above findings would actually be *misleading*. By criterion (X), therefore, this truth-focus makes clear that the evolutionary perspective here purports to present an *explanation* of how the findings of the Wason Selection Task are to be organised, and does not aim to *suggest* phenomena that we might otherwise have overlooked.

Secondly, one might object to the above argument by suggesting that the evolutionary perspective points to *further phenomena* that surround the detection of cheaters – and that it therefore *is* used in a heuristic way here (see also Cosmides & Tooby, 1992). For example, it might be argued that it is only because of Cosmides & Tooby's appeal to the evolutionary perspective that we found out about how well people can reason about unfamiliar situations involving social arrangements and about 'switched' social exchange conditionals.[6] Since the results of these findings were unsuspected, we might thus be said to have gained a better understanding of our minds – something that we would otherwise have been missing out on. In this way, it might seem like the evolutionary perspective is in fact used in a heuristic manner here.

However, this is not a compelling response to the above argument either. In the main, this is because the heuristic use of evolutionary theory it identifies is, at best, highly limited: it only concerns various *subsidiary findings*, but leaves all the main results of Cosmides & Tooby's work out of the picture. This comes out particularly clearly from noting that these findings are not particularly groundbreaking in and of themselves: for example, while certainly somewhat interesting, finding out how people reason about 'switched' social exchange conditional is not something we are interested in for its own sake – especially when compared to Cosmides & Tooby's main result (namely, that we can explain the puzzling findings of the Wason Selection Task by positing the existence of a cheater detection module). For this reason, it is better to see these findings as interesting mostly for their use as possible *tests* of Cosmides & Tooby's evolutionary hypotheses.[7] However, if this is granted, the above objection loses most of its force, since, as noted in section 10.2, this sort of application of evolutionary theory is not sufficiently strong to mount a compelling defence of heuristic evolutionary psychology: it is then better seen as an extension of an application of type (X), and not as an instance of type (H).

Thirdly, one might argue that the appeal to evolutionary theory was instrumental in helping Cosmides & Tooby think of *hypotheses* that might explain the data. That is, it might be claimed that the importance of the evolutionary considerations in the present context comes precisely from the fact that they suggest hypotheses that *could* account for the phenomena being made: these considerations make clear that a possible explanation of the above data can be found in the (supposed) fact that humans have evolved a mental module for detecting cheaters. Since the suggestion of hypotheses that *might* explain a phenomenon is not the same as *actually*

[6]Switched social exchange conditionals are conditionals where antecedent and consequent are switched – thus altering their truth conditions – but which are still meant to express the same social arrangement (e.g. 'If you give me your watch, I give you $20' is switched to 'If I give you $20, you give me your watch'). Interpreting these switched conditionals is very difficult, though, and not so relevant for present purposes (for more on this, see Cosmides & Tooby, 1992, 2005; Buller, 2005, pp. 183-188).

[7]Note also that this is precisely how Cosmides & Tooby themselves seem to understand the relevance of these findings – see e.g. Cosmides & Tooby (1992b). See also Buller (2005, pp. 183-185).

explaining that phenomenon, this might be seen to point to a defensible heuristic use of the evolutionary perspective after all.[8]

However, this objection, too, fails to be compelling. Virtually every theory will generate *possible* explanations for virtually every phenomenon one might care to mention. For example, a quantum field theoretic perspective suggests that the findings from the Wason Selection Task *may* be explicable using the interactions of fluctuating numbers of electrons and protons; a chemical perspective suggests that the findings *may* be explicable using the reactive properties of various kinds of molecules; and a Marxist perspective suggests that the findings *may* be explicable as showing that the bourgeoisie has found yet another tool for suppressing the workers. The trouble with this collection of hypotheses is that we are not interested in assembling it *for its own sake*: normally at least, we are interested in generating a set of possible explanations for a phenomenon only to the extent that this helps us to *actually* explain this phenomenon. What this means in the present context is that the interest of the evolutionary perspective cannot merely be seen in its presenting a *possible* explanation of the Wason Selection Task data – it must be seen in its presenting an *actual* (though possibly false) one. In turn, this truth-focus immediately marks this use of evolutionary theory as explanatory in the sense of (X) above, and not as heuristic in the sense of (H) above. Hence, this objection does not interfere with my argument either.[9]

For all of these reasons, it becomes clear that the best interpretation of Cosmides & Tooby's work remains an explanatory one. Moreover, it is easy to see that this conclusion generalises to many other evolutionary psychological research projects.

For example, most of David Buss's work on 'Sexual Strategies Theory' must also be seen as trying to *explain* the differences and similarities in the way in which human females and males choose mates (see e.g. Buss & Schmitt, 1993). This comes out clearly from the fact that Buss *begins* his research by empirically substantiating the widespread supposition that males tend to want different things from the things that females want (at least in some cases), and then uses Trivers's theory of minimal parental investment to *account* for these differences (see e.g. Buss, 2003; Buss & Schmitt, 1993; for some critical remarks concerning this theory, see Schulz, 2010). Much the same holds for Gigerenzer et al.'s work on simple heuristics (see e.g. Gigerenzer & Selten, 2001): Gigerenzer et al. use evolutionary theory only to explain various *known* social psychological findings about how we make decisions (see e.g. Simon, 1957) – they do not use evolutionary theory to contribute to these findings *being made*. Similar remarks can be made about much of Pinker's, Daly & Wilson's, and Symons's work, and that of many other researchers in this area (for more on this work, see e.g. Barkow et al., 1992).

Overall, therefore, it becomes clear that the case for the heuristic interpretation of evolutionary psychology has not yet been made: most of the classic examples

[8] I thank an anonymous referee for suggesting this objection to me.

[9] Note also that the existence of a cheating detection module cannot be taken for the 'phenomenon' suggested by the evolutionary perspective, as this would beg the question (it would build the theory into the observations). See also Sober, 2008, chap. 2.

of the research program – i.e. those associated with the Santa Barbara School (Buller's 'EP') – do not support this interpretation particularly well. However, as the next section aims to make clear, it *is* possible to find an instance of evolutionary psychological research that does so – it is just that it takes some work to do so.

10.4 Natural Pedagogy and Heuristic Evolutionary Psychology

Gergely and Csibra's work on 'natural pedagogy' stands in many ways in direct contrast to the typical research that goes on under the heading of 'evolutionary psychology'. For example, instead of embracing the nativism that frequently characterises the latter (see e.g. Sterelny, 2003; Carruthers, 2006), Gergely & Csibra emphasise the importance of *learning* and *development* for the way humans think and act. However, apart from this, their research remains very clearly within the confines of evolutionary psychology – in particular, they still use evolutionary theory as a key tool with which to study the features of our minds.[10] This last point is especially important here, for it is primarily through considering Gergely & Csibra's research that a limited defence of heuristic evolutionary psychology becomes possible after all.[11] To make this clearer, I again proceed in two steps: firstly, I present Gergely & Csibra's work as carefully as possible, and secondly, I assess it in light of the distinctions made in section 10.2.

10.4.1 Gergely and Csibra on Natural Pedagogy

Gergely & Csibra begin their research by noting that various kinds of imitation studies have thrown up three remarkable facts.[12] Firstly, it has turned out that, while all infants will tend to imitate adults *sometimes*, they will not do so with equal frequency in all circumstances. In particular, infants are much more likely to imitate an adult's action after the adult has made eye contact with the child, has raised her eyebrows when facing it, or has clearly and directly addressed it verbally (see e.g. Gergely & Csibra, 2009; Csibra & Gergely, 2006). Gergely & Csibra

[10] As made clearer in section 10.1 above, this is all it takes for research to be 'evolutionary psychological' in the sense relevant here.

[11] Andrews et al. (2002, p. 538) and Buss et al. (1998, p. 545) claim that Thornhill & Gangstead's work on female preferences for symmetric men (see e.g. Gangstead & Thornhill, 1997) provides another example of a heuristic form of evolutionary psychology. Whether they are right in this is not something I shall discuss here (for some critical remarks concerning this, see e.g. Fuentes, 2002); what matters for present purposes is just that *most* instances of evolutionary psychological research are *not* heuristic in structure, and that finding exceptions to this requires hard work. See also below in section 10.5.

[12] For more on these studies, see e.g. Meltzoff (1988), Tomasello (1999), Csibra & Gergely (2006), and Gergely & Csibra (2009).

interpret this finding as showing that infants need to be informed that an important teaching episode is about to begin: the infant needs to be told that the present is an instance where imitation is called for (see e.g. Csibra & Gergely, 2006).

Secondly, Gergely & Csibra also note that when infants are imitating actions that an adult has previously performed, they tend to ignore elements of the actions that do not seem necessary to achieving the *goal* of the action. For example, when shown an adult that presses a button with her head *when her hands are occupied* (e.g. due to her holding a blanket), infants are much more likely to press the button with their hands than with their heads – thus ignoring the manner in which the model outcome was achieved (see e.g. Csibra & Gergely, 2006; Gergely & Csibra, 2009). Gergely & Csibra interpret this finding as showing that human infants have a natural proclivity towards choosing the most 'rational' means towards some particular end (see e.g. Csibra & Gergely, 2006).

Thirdly, Gergely & Csibra note that infants seem to operate with a 'best explanation' heuristic when determining what the content of a learning episode is. That is, infants *will* imitate the manner with which the action was performed if there is no good reason for why the adult would teach the infant the *goal* of the action.[13] For example, in the above study, infants will imitate pressing the button with their head when there is no apparent reason for the manner in which the adult acted – e.g. when the adult does *not* hold a blanket that occupies their hands (see e.g. Csibra & Gergely, 2006). Equally, infants will imitate the manner in which a character (e.g. a mouse) arrives at its proper location (its house) if the fact that this is its proper location had already been made salient (see e.g. Gergely & Csibra, 2009). Gergely & Csibra interpret this finding as showing that infants presume that the adult teacher is rational, and that she would not engage in unnecessary behaviour – hence, the infants infer that there must be a reason for why the button ought to be pressed with one's head, or for why the mouse ought to arrive at its house in a particular way (see e.g. Gergely & Csibra, 2009).

Given these three findings, Gergely & Csibra draw the following two conclusions. Firstly, they claim that humans are born with an innate capacity for natural pedagogy: as infants they are attuned to changing their behaviour in the light of the lessons conveyed to them in designated teaching episodes; as adults, they are innately aware of how to signal when they are about to initiate a teaching episode. For what follows below, it is important to note that this conclusion, on its own, is perfectly in line with the results of many other researchers (see e.g. Tomasello, 1999; Premack & Premack, 2003). Where Gergely & Csibra differ from the latter is in the *details* of the capacity for natural pedagogy that they posit.

Specifically, in their second conclusion, Gergely & Csibra argue that this capacity for natural pedagogy is a psychological *adaptation* that allows humans to acquire generalisable local knowledge which it would be difficult to code for genetically

[13] Alternatively, it might be said that infants determine whether the goal of a model action includes the manner in which it was performed by considering whether there is an obvious reason for how the teacher has performed it. For present purposes, either of these interpretations is acceptable.

(see e.g. Csibra & Gergely, 2006). In particular, they claim that since environmental conditions vary across different locales, it was more efficient for humans to be equipped with mechanisms for the rapid acquisition of the appropriate knowledge than to be born with a large store of knowledge for all eventualities. This made it possible for humans to avoid having to be burdened with a vast set of facts, most of which will be irrelevant to any situation they will ever find themselves in.

In more detail, Gergely & Csibra claim that we have evolved the capacity for natural pedagogy when we reached a point where complex tool use became crucially important to deal successfully with our environment, and when the workings of particular tools were very difficult to learn just by trial and error (see e.g. Csibra & Gergely, 2006). When these tools furthermore turned out to be useful only in specific local environments – so that it was not practical to code for the understanding of the tools genetically – natural pedagogy evolved. In this way, they come to argue that, given the conditions in which we evolved (and some general facts concerning the relative benefits of genetically coded versus culturally coded knowledge in different circumstances), the capacity for natural pedagogy is an adaptation for acquiring *a specific kind of knowledge*: namely, knowledge that is *generalisable* (i.e. that is important to more situations than the learning episode) and *local* to the particular environments we develop in (Csibra & Gergely, 2006, pp. 252-254; Gergely & Csibra, 2009).

This evolutionarily derived idea led Gergely & Csibra to perform several novel experiments (in what follows, I shall call these 'taste / teaching experiments'). The main idea behind these experiments is that, if the function of the adaptation for natural pedagogy truly is the acquisition of generalisable local knowledge, then infants should distinguish *teaching episodes* – which concern the features of various *objects* – from the *personal tastes* of the teachers (see e.g. Csibra & Gergely, 2006, p. 256). That is, if the function of natural pedagogy is the acquisition of *objective* information, infants should not be expected to learn anything about the *subjective* features of the teacher during a teaching episode – and that is so even if these subjective features are an integral part of the teaching episode (see also Gergely et al., 2007, p. 144).

This is exactly what we do find (see e.g. Gergely et al., 2007; Gergely & Csibra, 2009): if infants are taught that some object has 'positive valence' (i.e. is 'good' for human beings), then they expect this object to be chosen over other available objects – and this is independent of whether the adult doing the choosing has previously rejected this object during a teaching episode.[14] Note that this does not mean that infants cannot attribute subjective tastes to adults – in fact, this is quite within their

[14]The experimental design here is somewhat complex. The general gist behind it is the following: learning episodes are made to be incompletely uniform – some teachers are made to teach that some object A is 'better than' some other object B, and some the reverse. Given this, Gergely & Csibra hypothesise that if enough teachers teach that A is better than B, the infant will take A to have an 'objective' positive valence. Crucially, however, this positive valence will be kept separate from the 'tastes' exhibited during the teaching episodes by the individual teachers. For more on this, see Gergely et al. (2007).

powers (see e.g. Gergely & Csibra, 2009; Gergely et al., 2007). What this means is just that infants distinguish what adults are doing during teaching episodes – namely, expressing general facts about the local environment – from what they are doing otherwise – namely, acting based on their beliefs and desires.

10.4.2 The Place of Evolutionary Theory in Gergely and Csibra's Research

Taking a step back, the above analysis thus makes clear that Gergely & Csibra use evolutionary theory in two ways in their research. On the one hand, they use it in an *explanatory* way: they put forward the hypothesis that our capacity for natural pedagogy is an adaptation in order to explain various findings concerning children's behaviour. For example, they use this hypothesis to account for the fact that the capacity for human pedagogy seems to be a human universal, that it is present from birth, and that it provides fitness benefits to an infant (see e.g. Csibra & Gergely, 2006; Gergely & Csibra, 2009). This is an explanatory use, as the relevant findings were *already known*, before Gergely & Csibra started appealing to evolutionary theory. Indeed, this use of evolutionary theory seems to be exactly parallel to Cosmides & Tooby's in the case of cheater detection: known phenomena are placed in a novel theoretical setting, which helps explain why they came about in the way that they did (or so it is claimed).

On the other hand, though, the above analysis also shows that Gergely & Csibra use evolutionary theory in a *heuristic* manner here. This use centres on the evolutionary hypotheses about the *particular nature* of our capacity for natural pedagogy that they put forward; it works in two steps. Firstly, Gergely & Csibra derive the specific nature of this learning mechanism – i.e. the fact that it concerns generalisable local knowledge – directly from the evolutionary considerations they put forward (see e.g. Csibra & Gergely, 2006; Gergely & Csibra, 2009). That is, they do not arrive at this hypothesis by considering vast amounts of empirical data (or the like), but by the careful consideration of their evolutionary arguments: they derive it only from what would be adaptive in a certain set of circumstances.

Secondly, it is then this specific nature of the capacity for natural pedagogy that must be seen to suggest to them the taste / teaching experiments described above. In particular, it is very plausible that it is only because of their consideration of what would have been adaptive in the EEA that they are led to inquire into whether infants can distinguish the tastes of the teacher from the content of a learning episode. Since the taste / teaching experiments confirm that infants in fact have this ability, evolutionary theory is thus shown to have been instrumental in our discovering features of our minds that we would otherwise have been ignorant about. In other words, it seems clear that it is only due to the consideration of the evolutionary perspective that we have become aware of the existence of a dedicated mental mechanism for teaching and learning (i.e. one that is separate from our mindreading skills in general). By (H), this therefore marks the use of evolutionary theory here as heuristic.

In this context, it is also worthwhile to note that, in so far as these evolutionary hypotheses are used in a heuristic manner, they themselves are not part of the tests that are being performed (see also Machery, forthcoming; Csibra & Gergely, 2006). In using evolutionary theory to *derive* the taste / teaching experiments, Gergely & Csibra are not seriously defending the above hypothesis about our cognitive evolution (in Csibra & Gergely, 2006, they call the derivation of this hypothesis a 'just-so' story). Of course, this hypothesis still *might* be true – however, establishing this is not the aim of this part of their inquiry. All that they seek to do there is find out more about how our minds work: evolutionary theory is relevant for this only to the extent that it helps us find out about phenomena that we would otherwise be ignorant about, and which are very revealing about the nature of our minds. To see this more clearly, it is useful to note two further aspects of the taste / teaching experiments.

Firstly, these experiments are not *obviously* interesting or suggestive about our minds. That is, comparing how infants react to teaching episodes with how they react to exhibitions of differing preferences among different people is not something that *straightforwardly* seems an interesting comparison to make. In fact, when first faced with the hypothesis of a capacity for teaching and learning, there seems to be little of interest in making such a comparison at all. For this reason, it seems clear that doing these experiments is not something that immediately suggests itself – their importance needs to be *discovered*. Hence, the value of the present heuristic use of evolutionary theory cannot be belittled by claiming that the phenomena it suggested were trivial or obvious to begin with.

Secondly, these experiments – or rather, their results – expand our understanding of our minds significantly. Finding out that, from an extremely young age onwards, we seem to be able to distinguish among differences in personal taste and the contents of learning episodes is a stunning result that greatly deepens our knowledge of human cognition. In particular, this result reveals a lot about the different psychological mechanisms that make up our minds – and thus, about the basic structure of our cognitive architecture. This matters, as it makes clear that, unlike in the case of Cosmides & Tooby's work, the findings suggested by the heuristic use of evolutionary theory in Gergely & Csibra's case are not subsidiary results, but the key elements of their account – it is primarily these experiments that suggest that humans have a capacity for natural pedagogy that is distinct from their abilities to imitate or mindread.[15] In this way, the phenomena revealed by Gergely & Csibra's evolutionarily-derived theory are shown to make for *new and deep* insights into human psychology, and thus to present issues whose further investigation is of great importance for a better understanding of our minds.

In short: since it is primarily due to the evolutionary perspective that the taste / teaching experiments have been performed in the first place, and since these

[15]For similar reasons, these experiments cannot be seen merely as *tests* of Gergely & Csibra's evolutionary hypothesis: the experiments have pointed to phenomena that are greatly interesting in and of themselves – whatever the best explanation for these phenomena will turn out to be.

experiments have pointed to phenomena whose investigation has greatly expanded our knowledge of our minds, this thus makes clear that Gergely & Csibra use their evolutionary hypotheses in a fruitful heuristic manner. In turn, this means that attacking these hypotheses for being evidentially ungrounded misses the point: they are not meant to explain why our mind has certain features – they are meant to *suggest* features that our minds *might* have, and which we should explore further in order to deepen our understanding of our psychological nature. For this reason, it becomes clear that Gergely & Csibra's work shows that a compelling heuristic form of evolutionary psychology really does exist.[16]

10.5 Conclusion

I have tried to argue that it *is* possible to defend the existence of a heuristic form of evolutionary psychology. More specifically, I have tried to show that the fact that the evolutionary hypotheses considered by evolutionary psychologists often lack evidential support *need not* mean that this makes the program scientifically dubious: in some cases, these hypotheses might merely be used as heuristic devices that point to issues that are usefully investigated further.

However, I have also tried to argue that this point must not be overemphasised – in fact, far from being a common occurrence, heuristic applications of evolutionary theory in psychology are actually quite a rarity. While such occurrences do exist, as yet, they are still in a minority: *most* cases of evolutionary psychological research – and, in fact, virtually all of the work of the Santa Barbara ('EP') School of evolutionary psychologists – employ evolutionary theory only to *explain* a known set of phenomena, not to lead us to *discover* these phenomena. Of course, this does not mean that these uses of evolutionary theory are necessarily unconvincing; however, it does mean that they cannot be defended by claiming that empirical support for them is not needed.

Looking forward, what this implies is that a compelling heuristic-based defence of any particular evolutionary psychological research project can only be done by carefully analysing the *details* of such a project. Only this can reveal whether a heuristic reading of this project is plausible or not: in particular, only this can show that, in the case in question, evolutionary theory was in fact instrumental in pointing to

[16]In this context, it is also worthwhile to note that Gergely & Csibra's work does not point to any specific features of evolutionary theory as being responsible for its heuristic usefulness. In particular, nothing in the above shows that it is specifically the fact that evolutionary theory is a backwards looking, population-level theory (or some such) that makes it a useful heuristic device. In fact, everything said here is perfectly consistent with the fact that theories from other sciences could play similar roles in psychology – as made clear in note 4 above, I here leave it open precisely *why* evolutionary theory can be used to suggest interesting phenomena to investigate further. Of course, as a matter of fact, no other theory has been given the prominence that evolutionary theory has when it comes to psychology. Why that is so, though, is an interesting question that has as yet not been convincingly answered.

novel and interesting phenomena about the way our minds work. Overall, therefore, it becomes clear that the heuristic defence of evolutionary psychology, while not fully implausible, must be treated with a lot of care.[17]

References

Andrews, P., Gangestad, S. & Matthews, D. (2002): 'Adaptationism – How to Carry out an Exaptationist Program'. *Behavioral and Brain Sciences* 25: 489–553.

Barkow, J., Cosmides, L. & Tooby, J. (Eds): *The Adapted Mind*. Oxford: Oxford University Press.

Buller, D. (2005): *Adapting Minds*. Cambridge, MA: MIT Press.

Buller, D., Fodor, J. & Crume, T. (2005): 'The Emperor is Still Under-Dressed'. *Trends in Cognitive Sciences* 9: 508–510.

Buss, D. (2003): 'Sexual Strategies: A Journey into Controversy'. *Psychological Inquiry* 14: 219–226.

Buss, D., Haselton, M., Shackelford, T., Bleske, A. & Wakefield, J. (1998): 'Adaptations, Exaptations, and Spandrels.' *American Psychologist* 53: 533–548.

Buss, D. & Schmitt, D. (1993): 'Sexual Strategies Theory: An Evolutionary Perspective on Human Mating'. *Psychological Review*, 100: 204–232.

Carruthers, P. (2006): *The Architecture of the Mind*. Oxford: Oxford University Press.

Cosmides, L. (1985): *Deduction or Darwinian Algorithms? An Explanation of the "Elusive" Content Effect on the Wason Selection Task*. Doctoral Dissertation, Harvard University.

Cosmides, L. & Tooby, J. (1989): 'Evolutionary Psychology and the Generation of Culture, Part II: A Computational Theory of Social Exchange' *Ethology and Sociobiology* 10: 51–97.

Cosmides, L. & Tooby, J. (1992): 'The Psychological Foundations of Culture'. In J. Barkow, L. Cosmides & J. Tooby (Eds): *The Adapted Mind*. Oxford: Oxford University Press.

Cosmides, L. & Tooby, J. (1992b): 'Cognitive Adaptations for Social Exchange'. In L. Cosmides, J. Tooby, & J. Barkow (Eds): *The Adapted Mind*. Oxford: Oxford University Press, pp. 163–228.

Cosmides, L. & Tooby, J. (2005): 'Neurocognitive Adaptations for Social Exchange'. In D. Buss (Ed.): *Evolutionary Psychology Handbook*. Hoboken: John Wiley, pp. 584–627.

Cosmides, L. & Tooby, J. (2008): 'When Falsification Strikes: A Reply to Fodor'. In W. Sinnott-Armstrong (Ed.): *Moral psychology*, Vol. 1. Cambridge, MA: MIT Press.

Csibra, G. & Gergely, G. (2006): 'Social Learning and Social Cognition: The Case for Pedagogy'. In T. Munakata & M. H. Johnson (Eds): *Processes of Change in Brain and Cognitive Development: Attention and Performance, XXI*. Oxford: Oxford University Press, pp. 249–274.

Davies, P. (2002): 'Does Past Selective Efficacy Matter to Psychology?'. *Behavioral and Brain Sciences* 25: 513–514.

Dupré, J. (2001): *Human Nature and the Limits of Science*. Oxford: Oxford University Press.

Gangestad, S. W. & Thornhill, R. (1997): 'The Evolutionary Psychology of Extrapair Sex: The role of Fluctuating Asymmetry'. *Evolution and Human Behavior* 18: 69–88.

Gigerenzer, G. & Selten, R. (2001) (Eds): *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press, pp. 1–12.

[17] Interestingly, this rather restrictive conclusion also has some implications for the plausibility of defending heuristic applications of scientific theories more generally. Given the considerations presented in this paper, it becomes clear that it is far from obvious that scientific theories, in general, are *often* used in a heuristic way: this requires special conditions to be satisfied, and should not necessarily be assumed to be very widespread. However, defending this more general claim in detail calls for a paper of its own.

Gergely, G. & Csibra, G. (2009): 'Natural Pedagogy'. *Trends in Cognitive Science* 13: 148–153.
Gergely, G. Egyed, K. & Kiraly, I. (2007): 'On Pedagogy'. *Developmental Science* 10: 139–146.
Johnson-Laird, P. (1982): 'Thinking as a Skill'. *Quarterly Journal of Experimental Psychology* 34A: 1–29.
Kitcher, P. (1985): *Vaulting Ambition*. Cambridge, MA: MIT Press.
Machery, E. (forthcoming): 'Discovery and Confirmation in Evolutionary Psychology'. In J. Prinz (Ed): *The Oxford Handbook of the Philosophy of Psychology*. Oxford: Oxford University Press.
Meltzoff, A. (1988): 'Infant Imitation after a 1-Week Delay: Long-Term Memory for Novel Acts and Multiple Stimuli'. *Developmental Psychology*, 24: 470–476.
Pinker, S. (1997): *How the Mind Works*. New York: Penguin Press.
Pinker, S. & Bloom, P. (1990): 'Natural Language and Natural Selection'. *Behavioral and Brain Sciences* 13: 707–784.
Premack, D. & Premack, A. J. (2003): *Original Intelligence. Unlocking the Mystery of Who We Are*. New York: McGraw-Hill.
Richardson, R. (2007): *Evolutionary Psychology as Maladapted Psychology*. Cambridge, MA: MIT Press.
Samuels, R., Stich, S. & Faucher, L. (2004): 'Reason and Rationality'. In I. Niiniluoto, M. Sintonen & J. Wolenski (Eds): *Handbook of Epistemology*. Dordrecht: Kluwer, pp. 1–50.
Schulz, A. W. (2010): 'It Takes Two: Sexual Strategies and Game Theory'. *Studies in History and Philosophy of Biological and Biomedical Sciences* 41: 41–49.
Shapiro, L. & Epstein, W. (1998): 'Evolutionary Theory Meets Cognitive Psychology: A More Selective Perspective'. *Mind and Language* 13: 171–194.
Simon, H. (1957): *Models of Bounded Rationality*. Cambridge, MA: MIT Press.
Sober, E. (2008): *Evidence and Evolution*. Cambridge: Cambridge University Press.
Sterelny, K. (2003): *Thought in a Hostile World*. Oxford: Blackwell.
Tomasello, M (1999): *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.
Wason, P. (1966): 'Reasoning'. In B.M. Foss (Ed): *New Horizons in Psychology*. Harmondsworth: Penguin.
Wason, P. (1983): 'Realism and Rationality in the Selection Task'. In J. B.T. Evans (Ed): *Thinking and Reasoning: Psychological Approaches*. London: Routledge and Kegan Paul.

Chapter 11
Evolutionary Psychology and the Problem of Neural Plasticity

Chuck Ward

11.1 Introduction

Much of the recent work in Evolutionary Psychology is organized around a shared set of principles regarding the nature and design of the human mind.[1] Within this research framework the mind comprises a set of domain-specific, computational modules, each able to perform a specific cognitive task, and each of these modules is a genetically specified adaptation to the environment of our Pleistocene ancestors (Pinker 1997; Tooby and Cosmides, 2005). These principles have been criticized on developmental grounds (Donald 2000; Karmiloff-Smith, 2000; Buller and Hardcastle, 2000; Dupré, 2001; Buller, 2005). The basis of such criticisms is the phenomena of neural plasticity. Evidence is accumulating that neural structures and patterns of neural connections are constantly changing over the life of organisms in response to environmental conditions, individual experience, and the behavior of the organism itself. This fact is taken, by the critics of Evolutionary Psychology, as a challenge to some of the above tenets. In this paper I will examine the basis of the developmental criticisms of Evolutionary Psychology along with some responses that have been offered. I will then develop further a criticism sketched by John Dupré (2001) that specifically appeals to developmental plasticity related to culturally-mediated behavior. To do this I will review recent research that demonstrates significant neurological and cognitive effects of two culturally-mediated

[1] The phrase 'evolutionary psychology' is used ambiguously. On the one hand it could be used, quite descriptively, to refer to any approach that aims to give an evolutionary explanation of psychological or cognitive characteristics of organisms. More narrowly, and more commonly, it is used to refer to the work of a specific group of contemporary researchers that share a common set of guiding methodological and theoretical principles. I am usually using the term in the latter sense. Following Buller (2005) I will capitalize the phrase Evolutionary Psychology when using it in this way.

C. Ward (✉)
Department of Philosophy, Millersville University, Millersville, PA 17551, USA
e-mail: chuck.ward@millersville.edu

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*,
Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_11,

behaviors that have arisen since the Pleistocene: reading/writing and musical training. I will argue that phenomena such as these pose problems for the kind of evidence typically used to support the claim that cognitive modules are genetically pre-specified adaptations to an ancestral environment. More specifically I will argue that they counter the Evolutionary Psychological claim that natural selection acting on genetic variation is the only known cause of complex functional design. I will then examine some implications of these issues for understanding the evolution of human behavior in terms of a broader, less gene-centered view of heredity.

11.2 Evolutionary Psychology's Spin on Cognitive Modularity

Evolutionary Psychology is committed to a number of specific theses about the nature mind and its evolution (see Pinker 1997, 21; Tooby and Cosmides 2005, 16-18).

(1) The Computational Thesis (CT): Brains are computational systems. This means that cognitive processes (and mental processes generally) are to be understood as information processing, the conversion of input to output. To quote Tooby and Cosmides,

> The brain's evolved function is to extract information from the environment and use that information to generate behavior and regulate physiology. Hence, the brain is not just like a computer, it is a computer. (2005, 16).

Psychology/cognitive science undertakes to explain behavior by discovering the programs that facilitate these computational processes. We are not necessarily conscious of the operation of these programs, though their operation may be accompanied by conscious states. For example, the brains of children process information about language in very specific ways to produce competent language use. The children are conscious of their use of language, but they are not conscious of the way their brains processed the input that led to that ability. The computational thesis per se is not specific to Evolutionary Psychology. In fact it is the predominant view within cognitive science generally – though there is considerable debate about the nature of computational processes.[2]

(2) The Modularity Thesis (MT): Neural structures instantiate a modular set of cognitive processes. The mind is not one information processor, but rather a cluster of functionally distinct processors that are, somehow, integrated. Each cognitive task is carried out by a specialized "module" implementing a set of computational algorithms sufficient for its task. A module is a discreet computational unit, presumably realized in some discreet neural circuit.

[2]The principal theoretical approaches to computational cognitive science are the Representational Theory of Mind/symbolocist approach, the connectionist approach, the dynamic systems approach, and the distributed cognition approach. For a basic account of the former two approaches see Rapaport 2000. On connectionism and dynamic systems approaches see Elman 1998. On distributed cognition approaches see Clark 1998.

The concept of computational modules has been a central feature of cognitive science at least since Fodor championed it (Fodor 1983). However, there has been considerable debate over the nature of modules since that time. Fodor outlined several characteristics of modules. He didn't insist that all modules have all these characteristics, but argued that a system is modular to the degree that it exhibits more or less of these properties. Among the properties Fodor identified are that modules occupy specific areas of the brain; they process specific types of information; they are informationally encapsulated, meaning that they do not have access to all the information contained in other modular units in the mind; they are automatic (when the input comes in, it will be processed and generate the appropriate output); and their output is non-conceptual ("shallow output"). On the Evolutionary Psychologists account, modules need not conform to Fodor's analysis. Instead, modules are functionally specialized computational systems (Barrett and Kurzban 2006, 29). On this view modules are functional units defined by how they process the input they receive. So any neural circuit that has a specific computational function is a module whether or not it has a significant number of Fodor's properties. Several proponents of this view have added that the functional specificity of modules is the result of their being designed by natural selection to respond to input in certain ways. This has lead commentators to refer to these (non-Fodorian) modules as Darwinian Modules (Samuels 1998; 578; Machery 2007, p. 826).[3]

(3) The Domain Specificity Thesis (DST): These cognitive modules are domain-specific, which is to say that they are designed to carry out some task in a specific domain (viz. face recognition, mate selection, or cheater detection). This idea is distinguishable from the general MT (i.e. item (2) above), but for Evolutionary Psychologists, and indeed most proponents of modularity in cognitive science, they are tightly linked.[4] The various cognitive tasks are carried out by discrete modules rather than one central processor (this is the MT). Moreover, modules do not employ one set of general computational or logical principles. The very point of modularity is that the various modules can be designed to carry out specific tasks without having to worry whether the processes or principles used would work well in the performance of other tasks. So the programs do not need to adhere to general logical principles, as long as the rules they do employ work in the domain in question (this is the DST). So the MT claims that each module in fact carries out a specific task. The DST claims that the data/inputs and principles/programs used to do so are applicable only to the kind of task (or problem domain) at hand, and are not (necessarily) generalizable to other related tasks or problems.

Like the CT, the general MT, conjoined to the DST, is a widespread view within cognitive science. Considerably more controversial is the "Massive Modularity

[3] When modules are conceived at the outset as Darwinian adaptations, then what I am calling the MT and the Adaptationist Thesis (AT) are not really distinct theses. I discuss the AT further below.

[4] Fodor (1983) specifies nine properties that a cognitive system/process must have (or have most of) in order to be considered modular. Domain specificity is among them. While defenders of the Evolutionary Psychologists' sense of modularity have argued that modules need not conform to Fodor's criteria, domain specificity is generally accepted as a central feature of Darwinian modules.

Thesis" (MMT). The MMT claims that all or most of our cognitive processes are carried out by their own domain-specific modular systems (Samuels 1998, 581; Sperber 2001, 48; Fodor 2000, 55; Machery 2007, 827). Despite his being an early formulator and defender of modularity, Fodor is quite skeptical about the MMT. Specifically he believes that it is peripheral "input" processes that are good candidates for modularity but that "central" cognitive processes probably are not (1983, 2000). The "central cognitive processes" are those responsible for belief fixation and decision-making that yield behavioral output. "Input" modules function to structure information for use by central processes.

Evolutionary Psychology is committed to massive modularity. Tooby and Cosmides write that "natural selection will ensure that the brain is composed of many different programs, many (or all) of which will be specialized for solving their own corresponding adaptive problem" (2005, 17). They conceive of the mind as composed of many such modules functioning to produce behavioral responses to a wide range of adaptive problems. The Evolutionary Psychologists' take on the MMT is that the mind is a collection of integrated Darwinian modules, and that these modules underlay a range of our cognitive capacities, including peripheral processes and at least some more central cognitive processes (such as face recognition). It should be noted, however, that most Evolutionary Psychologists recognize that some cognitive processes may not be the result of Darwinian modules (Machery 2007, 827).

(4) The Genetic Specification Thesis (GST): Our cognitive architecture is the product of complex, evolved, genetic developmental programs. Selection acts to design the computational programs of the modules. But these programs are realized in neural circuits. The way that genes produce cognitive modules is by regulating the development of the brain and its neural wiring. Tooby and Cosmides write that "every time one gene is selected over another, one design for a developmental program is selected as well" (2005, 35). Natural selection has crafted developmental processes that can reliably produce (in environments such as ours) physical properties in the brain that implement the domain-specific programs described above. These developmental processes take place over an extended period of time and are often designed to be sensitive to environmental inputs which can activate (or deactivate) some gene thereby initiating its regulatory effect. The 'design' of the developmental program is stored in the genes that regulate development and in the stable features of the environment that interact with those genes.

Evolutionary Psychologists clearly consider our basic cognitive architecture to be the product of natural selection acting on genes that code for those features. This has led to charges of genetic determinism, and Evolutionary Psychologists are quick to reject that charge. They are keen to point out that genes do not regulate brain development in such a way as to produce rigid behavior. They claim to be well aware of the complexities of development in general and brain development in particular. Genes do not cause behavior directly, but rather do so by regulating brain development (Tooby & Cosmides 2005; Pinker 1997). And development involves both genes and the environment in which they act. Tooby and Cosmides write: "These elements [viz. genes] are transmitted from parent to offspring and together

with stable features of the environment, cause the organism to develop some design features and not others" (2005, 21). This recognition of the role of environment constitutes, in their minds, a transcendence of the "old" nature vs. nurture debates or concepts. Nevertheless it is clear that Evolutionary Psychologists consider our cognitive architecture to be specified in the genetic code. For example, while Steven Pinker recognizes that "[t]he genetic assembly instructions for a mental organ do not specify every connection in the brain as if they were a wiring schematic for a Heathkit radio" (1997, 35) he also asserts that "[t]he modules' basic logic is specified by our genetic program" (1997, 21). The assumption is that the evolution of the genetic programs takes advantage of the reliably stable features of the developmental environment.

(5) The Adaptationist Thesis (AT): These domain-specific cognitive modules are adaptations to the environment of human populations of the Pleistocene. This is often referred to as the environment of evolutionary adaptatedness (EEA). "Our modern skulls house Stone Age minds" (Cosmides and Tooby, 1997). They are designed by natural selection to generate specific types of behavior in certain environmental contexts, namely the context of the EEA. The results can at times be less than optimal in our present environment insofar as it differs from the EEA.

Research in Evolutionary Psychology seeks to discover these evolved, domain-specific cognitive modules, these "human universals" that make up human nature. These modules are genetically determined and species typical; though there will be some evolved modules that are not universal but are, rather, present in some portion of the population due their having a frequency dependent selective value. Evolutionary Psychologists are quick to point out that this view does not imply rigid, genetically determined behavioral patterns. They are talking about the way information is processed by the brain. Given different inputs we should expect different outcomes. So individual and cultural variation can arise due to variation in social environments and individual experience. But the universal aspects of human nature will be found in all "normal" individuals and serves as the basis for any variation that might arise. So the picture emerging from Evolutionary Psychology is one in which the underlying computational design of the human mind is genetically transmitted while cultural variation results from differential experiential inputs being processed through this common architecture.

11.3 Identifying Adaptive Modules

Accounting for the features of organisms by claiming that those features are adaptations is a pretty standard mode of explanation in biology. Such hypotheses are historical in nature: they claim that the feature in question was selected, relative to competing types, in ancestral populations, *because* it conferred a reproductive advantage with respect to some specific environmental challenge faced by members of those populations. But how are such adaptive hypotheses confirmed? The most

direct way to confirm an adaptive explanation would be to identify the specific agent of selection – i.e. the feature or features of the environment that exert selective pressure – and then to demonstrate that the proposed adaptive trait actually lead to reproductive advantage in the relevant populations and environment.[5] Getting such evidence of selective pressures and reproductive advantage of variant types in cases of past evolution is very difficult, so the case for most successful adaptive explanations is less than ideal. In the case of psychological and behavioral traits, it may be near impossible to obtain significant evidence of this sort. This makes things very difficult for Evolutionary Psychologists in terms of confirming their hypotheses that certain cognitive traits were adaptive in the Pleistocene. So, instead of confirming the presence of the phenotype in the ancestral population and the reproductive advantage it had over competitors, Evolutionary Psychologists employ an indirect argument of the character of an inference to the best explanation. The trait's being an adaptation to the ancestral environment is proposed as the best explanation for its universal presence in the modern population. And this claim is supported by the trait's being an example of a complex adaptive design.

The adaptive hypotheses advanced by Evolutionary Psychologists refer to computational modules. Generally speaking, Evolutionary Psychologists work at the level of the program or software. They seek to describe the computational principles employed by the adaptive modules in processing input and generating behavior. It is assumed that a module is realized in some neural network, but Evolutionary Psychologists don't generally aim to identify specific brain structures involved in cognitive processing. Another methodological point is that Evolutionary Psychologists describe their work as a kind of reverse engineering. Once an evolved adaptive module is identified, then the principles employed in the solution can be worked out. The first step in this sequence is the identification of an adaptive module. The evidence invoked in support of a claim that some behavioral pattern is caused by an evolved, domain-specific adaptive module is twofold. First the behavioral pattern must be fixed in the human population (or, alternatively, present in a definite proportion of the population in the case of frequency-dependent adaptations). Second the behavior must constitute a case of complex functional design. The first step is not sufficient alone since there are numerous cases of species typical behavior that did not evolve as an adaptive solution some problem in the EEA. Tooby and Cosmides cite the use of written language. Learning a spoken language, on their account (following Pinker and Bloom 1992) is the result of an adaptive neural program designed for that function. On the other hand "[t]he ability to read and write are by-products of adaptations for spoken language, enabled by their causal structure" (Tooby and Cosmides 2005, 26). Several Evolutionary Psychologists have argued that we can have "design evidence," i.e. evidence that some feature is an instance complex functional design, prior to identifying the selective forces that shaped the feature (Tooby

[5] See Brandon (1990) for an account of five specific categories of evidence needed, ideally, to support adaptive hypotheses. See Richardson 2007 for a discussion of these issues with respect to Evolutionary Psychology.

and Cosmides 2005, 27-28; Pinker and Bloom 1992, 454-455). The schematized argument that a module is an adaptation to the EEA runs as follows:

1. Natural Selection acting on genetic variation is the only available explanation for complex functional design.
2. The brain's cognitive architecture exhibits complex functional design (this claim can be applied globally or to a particular cognitive faculty, e.g. language use, see Pinker and Bloom, 1992)
3. So the elements of the brain's cognitive architecture are products of natural selection acting on genetic variation, i.e adaptations (in the neo-darwinian sense).
4. Evolution by natural selection takes a very long time.
5. So the adaptations in our brain/cognitive architecture are adaptations to the EEA for human beings, i.e. the environment of our Pleistocene ancestors.

The first three steps in this argument (in which 3 is inferred from premises 1 and 2) constitute an inference to the best explanation. Indeed it might be characterized as an 'only game in town' inference. Premise 2 is a claim about the presence of complex mechanisms that serve a current function in modern populations. Premise 1 invokes natural selection (for that function) as the only plausible explanation for such a state of affairs. The rest of the argument aims to place that history of selection in the EEA.

Once the conclusion is reached, the job of reverse engineering can begin, i.e., the task of identifying the design features of the module and its adaptive significance. This is initially an abductive process: a computational model (or set of alternative models) must be developed as hypothetical solutions to some adaptive problem faced by our Pleistocene ancestors (e.g. kin detection/inbreeding avoidance). Development of these models should be informed by the kind of information available to our ancestors in the EEA. Then a variety of techniques might be used to determine whether or not modern humans process environmental information of that sort in the way the model proposes (Tooby and Cosmides 2005, 28). The basic logic of the argument seems to be something like this: this computational process would be adaptive in the EEA; modern humans seem to employ this process; therefore a computational module that carries out this process probably evolved during the Pleistocene.[6]

But there is a problem: neural plasticity. Evidence is mounting from research in developmental cognitive neuroscience that the specific structures in the adult human brain are the result of the brain's own response to environmental inputs. In some instances these responses themselves may be shaped by cultural practices that potentially yield adaptive results. This introduces the possibility of changes in species typical neural structures that are not the result of modification in the genome and that have been introduced since the time when our ancestors occupied the EEA.

[6] See Buller (2005, Chapter 3) for a similar analysis of this kind of inference to adaptive hypotheses in Evolutionary Psychology.

11.4 Neural Plasticity and the Ever-Developing Brain

Despite their recognition of the importance of development in the production of neural structures, Evolutionary Psychologists have received criticism that charges their view with ignoring some significant features of brain development. In particular, it has been suggested that the phenomena of neural plasticity raises doubts about some of the theses that make up the Evolutionary Psychology framework (Donald 2000; Karmiloff-Smith, 2000; Buller and Hardcastle, 2000; Dupré, 2001; Buller, 2005). Neural plasticity refers to changes in the functional organization of the brain in response to sensory inputs and the brain's own activity patterns. Recent work has suggested that specific experience plays an important role in the determination of our neural wiring. This calls into question just how much of the computational architecture of our brains is specified in the genetic program.

For most of the twentieth century, the standard view about brain development held that humans were born with all the neurons they would ever have and that postnatal changes to the brain primarily involve cell death (Mohamed, et. al. 2002; Rosenzweig, 2003). Over the past thirty years or so this view has been replaced with a view of the brain as a dynamic system that undergoes a host of structural changes over the course of our lives. Many of these changes are responses to the experience of the individual.

The phenomenon of neural plasticity has been part of neurological theories for over a hundred years. William James was already discussing such things in his *Principles of Psychology* of 1890. There he posited habit as a central principle of mental phenomena and discussed patterns or currents of nervous activity as the physical aspect of habit.

> If habits are due to the plasticity of materials to outward agents, we can immediately see to what outward influences, if to any, the brain-matter is plastic. . . . The only impressions that can be made upon them are through the blood, on the one hand, and through the sensory nerve-roots, on the other; and it is to the infinitely attenuated currents that pour in through these latter channels that the hemispherical cortex shows itself to be so peculiarly susceptible. The currents, once in, must find a way out. In getting out they leave their traces in the paths which they take. The only thing they *can* do, in short, is to deepen old paths or to make new ones; and the whole plasticity of the brain sums itself up in two words when we call it an organ in which currents pouring in from the sense-organs make with extreme facility paths which do not easily disappear (James 1890, 107).

Those pathways that are traveled frequently become deeper and more apt to be traveled again. James' contemporary, neuroanatomist Santiago Ramón y Cajal, hypothesized the increased branching of neurons and the increasing the number of connections between them, as a result of training and experience (Rosenzweig 1996). This line of thought was translated into somewhat more contemporary terms involving neuron firing by Donald O. Hebb in 1949. Hebb coined the term *use-dependent plasticity* and hypothesized that the strength of synaptic connections is increased as two neurons (one pre-synaptic neuron and one post-synaptic neuron) fire together. This "strengthening" increases the efficiency of transmission of activity in the future. Starting in the early 1960s, experimental work began confirming this

hypothesis (Rosenzweig 1996, 2003; Mohammed et al. 2002; Elbert et al. 2001). Even more recently, beginning in the 1990s, evidence has begun to accumulate that even the adult human brain can add new neurons to itself (Schwartz and Begley 2002; Rosenzweig 2003). Work in this area is reversing what had been orthodoxy in mid-twentieth century neuroscience, namely the view that the number of neurons and the basic organization of the brain is fixed shortly after birth (Schwartz and Begley 2002, 167).

Perhaps the most researched form of use-dependent plasticity involves the modification of specific parts of the somatosensory cortex that map body activity in response to changing behavioral patterns (Elbert et. al. 2001). Early work in this area showed changes in the size of cortical zones mapping specific digits in monkeys due to decreased or increased sensory input from the hand (Merzenich et. al. 1987). Related to this is the role of plasticity in phantom limb sensations. V.S. Ramachandran and his colleagues have demonstrated that the cortical region that originally received and processed input from a now amputated limb can be recruited to receive sensory input from the face following amputation. Without input from the limb, that cortical region strengthens its connection to a neighboring region, which happens to map sensory input from the face. (Ramachandran and Ramachandran 2000). This process involves the strengthening of connections that had existed before but were relatively weak or inoperative. The process has come to be called 'unmasking'. Subsequent to the unmasking, stimulus to the face can trigger conscious phantom limb sensations. On the whole the brain is proving to be a very dynamic system in which structural and functional organization is the outcome of its own activity.

11.5 The Developmental Challenge to Evolutionary Psychology

Neural plasticity has been at the center of some recent criticisms of Evolutionary Psychology (Donald 2000; Buller and Hardcastle 2000; Karmiloff-Smith 2000; Dupré 2001; Buller 2005). Some of these criticisms challenge the idea of genetically prespecified cognitive modules. David Buller and Valerie Gray Hardcastle have argued that "we do not have lots of 'genetically specified,' domain-specific, informationally encapsulated, cognitive processing streams" (2000, 308). They describe aspects of neural development that show that the eventual location and functional role played by neurons is a function of cell competition and cell death resulting from that competition. So the specific organizational properties of the brain are not genetically prespecified. "Instead, during development we find a diffuse proliferation of connectivity, which later brain activity, guided by interaction with the environment, sculpts into its final form" (316). They cite some cases of a brain structure genetically "designed" to process one kind of information that, through a kind of unmasking process, is capable of performing its function by processing a different kind of information. This shows, they argue, that modules can be *domain-dominant* rather than *domain-specific*. This does not mean that human brains all end up with different computational properties. Indeed we do have some

set of species-typical neurological structures and corresponding cognitive faculties that serve adaptive functions. But this is not the result of genetic programs specifying the computational properties of those structures. Rather it is the result of the very plasticity of the brain itself, its ability to respond to environmental conditions in ways that shape its own structure and computational design. "[T]he degree to which the outcomes of human brain development have been regular through some of our evolutionary history is due to the fact that generally plastic brains have encountered recurrent environmental demands throughout that history" (317). The adaptive character of the brain is not some set of specific computational programs, but the brain's plasticity itself. (321).

Annette Karmiloff-Smith (2001) invokes neural plasticity to make a similar point. She addresses an important kind of evidence often employed to support the existence of domain specific modules, the phenomena of double-dissociation. This term applies to cases where two related capacities can be shown to operate independently by certain patterns of dysfunction. For example, children with Williams Syndrome (WS) are severely impaired in their ability to recognize objects, though they seem to have normal abilities in terms of face recognition. Other disorders might display the opposite effect: impaired face recognition capacity but seemingly normal object recognition capacity. This has been used to argue that there are two distinct modules in the brain that perform these two distinct functions. Karmiloff-Smith refers to research of her own and others to show that this argument breaks down under scrutiny. While WS children seem to have normal face recognition capacities, it has been shown that they use very different cognitive strategies to accomplish such tasks as compared with non-WS children. Similar considerations apply to the apparently normal aspects of WS children's language ability. Karmiloff-Smith argues that even if the adult brain is a set of domain-specific modules, this situation is the result of developmental processes, not genetic pre-specification. Through developmental processes the brain organizes itself into units that operate on specific inputs, and this organization is the result of the brain's ability to respond adaptively to environmental demands. Karmiloff-Smith suggests that the brain comes with a number of *domain-relevant* learning mechanisms which give rise to more localized *domain-specific* modules over time.

The modularity view of the Evolutionary Psychologists has been defended by John Sarnecki against the arguments summarized above. He argues that the distinction between domain-dominance (Buller and Hardcastle) and domain-relevance (Karmiloff-Smith) on the one hand and domain-specificity (Evolutionary Psychology) on the other hand, collapses in environments that remain relatively stable over time or show consistent variability (2007, 538). In such environments natural selection can act to select individuals with specific neural structures and this yields the propagation of the genes that serve to bring about the development of such structures in such environments. So even if the existence of these neural structures depends on certain environmental inputs, they can still be adaptations, in the neo-darwinian sense, to typical environmental conditions.

Sarnecki's argument is in keeping with the views of Evolutionary Psychologists concerning the interaction of genetic and environmental factors during development.

Recall Tooby and Cosmides claim that genes "are transmitted from parent to offspring and together with stable features of the environment, cause the organism to develop some design features and not others" (2005, 21). They recognize that genes can only be selected for the production of cognitive modules if the environmental factors relevant to their causing those modules to develop remain stable. Pinker spins this a bit differently, describing genes as taking advantage of the way the neurons in the developing brain process and respond to environmental input to bring about adaptive computational modules (1997, 35). This characterization corresponds to what Susan Oyama has called "standard (or traditional or conventional) interactionism" which she criticizes as still harboring an unacceptable degree of genocentrism (Oyama 2001).

11.6 What about Culturally-Mediated Developmental Environments?

While the two developmental critiques outlined in the last section appeal to neural plasticity, it is neural plasticity of a fairly generic sort. This is what opens the door to Sarnecki's rebuttal. If we can assume some constancy in the developmental environment (including the kinds of individual behaviors and experiences involved), then we can maintain a gene-centered approach to the development and evolution of neural structures. John Dupré has offered a developmental critique that specifically introduces the likelihood that environmental factors relevant to brain development have changed dramatically through cultural evolution. "[S]ince conditions under which contemporary brains develop are very different from the conditions under which human brains developed in the Stone Age, there is no reason to suppose that the outcome of that development was even approximately the same then as now" (Dupré 2001, 31). In this section I want to fill out this line of argument by appealing to some specific aspects of neural plasticity and examining some implications with regard to cognitive evolution.

But first it should be acknowledged that Evolutionary Psychologists do recognize that the social environment of human beings has changed since the Pleistocene. In fact they rely on this fact to explain the apparently maladaptive nature of some aspects of our evolved human nature.

> Although the behavior our evolved programs generate would, on average, have been adaptive (reproduction promoting) in ancestral environments, there is no guarantee that it will be so now. Modern environments differ importantly from ancestral ones, particularly when it comes to social behavior. We no longer live in small, face-to-face societies, in seminomadic bands of 20 to 100 people, many of whom are close relatives. Yet our cognitive programs were designed for that social world. (Tooby and Cosmides 2005, 17)

So environmental conditions come into their story in three ways. First the ancestral environment presents the design problems that natural section solves through designing adaptations. Second the environment constitutes the conditions under which development takes place. Here we must be looking at stable features of the

environment, features that are reliably present so that the genes can (co-) produce the designed programs. But the developmentally relevant features must (at least often) be different from those features that relate to the adaptedness of the programs. The third role of the environment in their story has to do with environmental change. There are a number of things that have changed in our environment over time, and some of those have made our design features less adaptive or even maladaptive (e.g. our craving for fats and sugars). These changes alter the adaptive nature of (some of) our programs, but those programs are still there. So the environmental conditions necessary to developmentally produce them must still be present. This is where the genocentric aspect or genetic-determinism aspect of the story becomes apparent. The assumption is that any adaptive modifications to the cognitive programs must be due to genetic mutation. The environmental changes that are recognized by Evolutionary Psychologists seem to have no role in determining the nature of the programs (recall Pinker's claim that "[t]he modules' basic logic is specified by our genetic program" (1997, 21)). If, however, culture propagates behavioral patterns and developmental environments that have significant effects on the kind of cognitive structures humans develop, and if some of these culturally mediated effects are more recent than the Pleistocene, then (following Dupré) we would have reason to believe that our brains and our minds are significantly different than our stone age ancestors.

Recall that the Evolutionary Psychologists' argument for the Adaptationist Thesis runs as follows:

1. Natural Selection acting on genetic variation is the only available explanation for complex functional design.
2. The brain's cognitive architecture exhibits complex functional design.
3. So the elements of the brain's cognitive architecture are products of natural selection acting on genetic variation, i.e adaptations (in the neo-darwinian sense).
4. Evolution by natural selection takes a very long time.
5. So the adaptations in our brain/cognitive architecture are adaptations to the EEA.

But suppose alterations in our cognitive architecture can result from changes to the developmental environment that have been introduced more recently than the EEA. Suppose further that those changes to the environment can be reliably replicated over generations through cultural transmission. And, finally, suppose that the resulting alterations prove to be adaptive in the general sense. These suppositions suggest that natural selection acting on genetic variation is *not* the only available explanation for the complex functional design of our cognitive architecture. The "only game in town" character of the Evolutionary Psychologists' argument evaporates.

But can the suppositions be substantiated? Recent work in developmental cognitive neuroscience supports them. There are two widely discussed examples of culturally-mediated behaviors that have potentially adaptive effects on neural development: (1) learning to read and write (use of a visual symbolic system), and (2) learning to play music. There is some evidence that modifications to neural structures that result from these culturally-mediated behaviors affect a wide range

of cognitive faculties. Some of these effects can be described as adaptive in the general sense (rather than the evolutionary sense) of allowing individuals to better meet some challenges of their environment.

Oral language is old enough to fall within the scope of the Evolutionary Psychologists' Adaptationist Thesis. But written language is a much more recent development, originating roughly five thousand years ago. Once the development of writing systems began, this cultural practice had epigenetic effects on our neural and cognitive systems. Merlin Donald has emphasized the importance of what he terms "the literacy brain":

> Some cultural changes can actually remodel the operational structure of the cognitive system. The clearest example of this is the extended and widespread effect of literacy on cognition. In this case, we know that the brain's architecture has not been affected, at least not in its basic anatomy or wiring diagram. But its functional architecture has changed, under the influence of culture. (Donald, 2000: 19)

In this passage Donald refers to both the physical and functional "architecture." His argument is that learning to use a complex symbolic system does not modify the general physical structure of the brain. Modern humans have the same basic brain structures as our pre-literate ancestors. But there is a functional reorganization, meaning that some of those structures have been "captured and redeployed" for different uses. The process of learning to operate in a symbolic system modifies the interconnections between certain areas of the brain, recruiting neural structures that evolved, biologically, for a different function, and teaching them to interact in new ways with each other. This cognitive reorganization is "mediated by basic neural-developmental processes such as synaptogenesis, displacement, and Hebbian learning (the strengthening of specific synapses by experience)" (Donald 2000, 23). While Donald's theorizing about cognitive evolution focuses primarily on the cognitive architecture (positing the existence of networks of cognitive modules to carry out specific functions), he clearly appeals to processes of neural plasticity to account for the non-genetically based evolution of our cognitive architecture through the influence of culture.

Maryanne Wolf (2007) makes very similar proposals regarding the effect of reading on the developing brain. Specifically, she reviews brain-imaging evidence that older neural structures that originally evolved for other functions are recruited in the process of learning to read. In making these suggestions she follows the work of Dehaene and his "neuronal recycling hypothesis" (Dehaene et al. 2005; Dehaene and Cohen 2007). It should be noted that Dehaene takes this hypothesis to be an alternative to one that posits large-scale effects of culturally-mediated behavior on an enormously plastic, domain-general neocortex. Instead, Dehaene considers the process to be fairly conservative. Brain structures that evolved for one function (e.g. object recognition) are recruited, through cultural and pedagogical practices, to very different but functionally similar tasks (e.g. letter and word recognition). He does recognize the role of plasticity in this process (Dehaene and Cohen 2007, 384). But he also concludes that the culturally imposed cognitive function will be highly constrained by the nature of the neural structures and pathways involved in the older, evolutionarily produced function. This conservatism accounts for the fact that

the same brain structures and regions are active in reading activity across all cultures, despite their very different writing systems. This cross-cultural invariance has been established through the use of fMRI imaging techniques (as well at meta-analyses of many individual imaging studies) to identify patterns of brain activity during reading. Bolger et al., have termed this the "universal reading network" (Bolger et al. 2005). There appears to be a common set of neural structures that are recruited to process script and decode it in relation to speech and meaning, despite the fact that not all writing systems are the same with regard to the relationship of script to the phonetic and semantic elements of language. Finer grained analysis does show some differences in the activity patterns within the universal reading network between a logographic script (such as Chinese) and alphabetic systems (Bolger et al. 2005; Tan et al. 2005). Tan and co-authors suggest that the processing of logographic scripts differs from processing of alphabetic scripts in terms of the way that the neural structures engaged in visual, phonological and semantic processing (as well as motor function) interact with one another: "Language form, cognitive process, and learning strategy seem to drive the development of functional neuro-anatomy" (Tan et al. 2005, 89).

These sorts of neuro-imaging data lead Wolf to suggest that the process of becoming literate produces a modified neural circuitry and a correlated modification to our cognitive architecture. She also invokes the work of cognitive developmental neuroscience to support the further claim that "the new circuits and pathways that the brain fashions in order to read become the foundation for being able to think in different, innovative ways" (Wolf 2007, 217). The new neural circuits support the cognitive processes that yield increased linguistic awareness which in turn allows for processes of classification and analysis. In other words, Wolf hypothesizes that a number of higher cognitive functions depend on the neural circuits developed through the culturally mediated practice of reading.

Music, musical performance and musical training are certainly examples of cultural practices. There have been claims that musical training has a positive effect on linguistic ability, spatial reasoning, mathematics and other cognitive faculties that fall outside of the domain of playing music (Asbury and Rich 2008; Schlaug et. al. 2005). There is considerable evidence that musical training and regular practice/performance leads to physical changes in the brain. Some results show an expansion of the areas of the sensorimotor cortex that represent and control the movement of the fingers and hand (see, for example, Pascual-Leone 2001). Brain imaging studies have indicated growth in the corpus callosum that sends signals between hemispheres (Schlaug 2001) as well as in areas that function in the control of motor function and auditory function among others (Krista, et. al. 2009). One might expect the changes in areas associated with motor control and auditory processing, since playing musical instruments requires refined skills in these processes.

The real question for present purposes, however, is whether the neurological changes brought about by musical training have cognitive effects outside the specific domain of musical performance. Evidence is accumulating that they do. Ho, Cheung and Chan (2003) report improved verbal memory effects of musical training in young children (while there were no measurable effects on visual

memory). More recently, musical training in children has been correlated with improved performance on standard cognitive tests for vocabulary and nonverbal reasoning (as well as auditory discrimination and fine motor skills) (Forgeard, et. al. 2008). So we do indeed have empirical findings indicating that this culturally mediated behavior has physical effects on the brain (via the mechanisms of neural plasticity) with associated effects on a range of cognitive faculties.

These are two key examples of recent findings that support the idea that there can be changes to our cognitive architecture brought about and propagated across generations by cultural practices. To be sure, neither literacy nor musicianship is a universal cognitive/behavioral trait in human beings. And I am *not* claiming that they are themselves adaptive in the evolutionary sense of increasing reproductive success. I do claim that these examples demonstrate the existence of processes that can serve to introduce and reliably propagate modifications in our cognitive architecture without genetic change. The central points brought out by these examples are, first, that certain cultural practices (including pedagogical practices) can reproducibly bring about physical changes in the brain, changing the way information is processed. Second, such changes can affect our cognitive performance in domains other than those that produced the changes. Third, some of these changes can be described as complex functional adaptations in the general sense that they promote the success of individuals in coping with social and/or environmental demands.

This creates a gap in the Evolutionary Psychologists' case for the adaptive significance of neural architecture being tied to the EEA. It does so by suggesting an alternative mechanism for producing and propagating adaptive cognitive processes that exhibit a complex functional design. A central premise in the Evolutionary Psychologists' argument that a behavior or cognitive module is an adaptation to the EAA is that only natural selection, acting on genetic variation, can produce such complex functional design. But we actually have reason to believe that this may not be the case. The Evolutionary Psychologists' explanation is *not* the only game in town.

It must be noted that we do not have, at present, an example of culturally mediated inheritance of neural or cognitive features that are species typical in the sense that they are universal in the human population. Evolutionary Psychologists have emphasized that their program aims to provide an evolutionary explanation of the species typical cognitive architecture (though they recognize that some of the traits will show variation of some sort and that selection may maintain some kind of stable polymorphism). Nevertheless, the cases that we have examined are sufficient for our present argumentative purpose, which is to show that we have some reason to believe that complex, adaptive, cognitive traits *could be* the result of mechanisms other than natural selection acting on genetic variation. One might describe the proposal as speculative, and in a sense it is. But it is not so purely speculative as to warrant the charge that, in Isaac Newton's terms (from his Fourth Rule of Reasoning in the *Principia*), we are attempting to evade an induction by hypothesis. First, we have evidence that mechanisms of use-dependent neural plasticity can in fact produce an adaptive neural and cognitive organization. So there is reason to think the proposal might be true independent of its capacity to serve as an explanation of

adaptive cognitive traits. And second, without the "only game in town" premise, the Evolutionary Psychologists' argument is abductive in nature rather than inductive.[7] It comes down to the claim that their hypothesis, that the behavior is caused by a cognitive module that is an evolutionary adaptation to the EEA, if true, can explain the phenomena. Moreover, it isn't too much of a stretch to think that the neurological and cognitive effects of literacy may be getting closer to being species typical. Literacy is by no means a universal human characteristic. But it has been increasing in the human population and continues to do so. And, as Donald has pointed out, our contemporary technological culture, with its forms of symbolic systems, may have even wider effects than traditional forms of literacy. I am not suggesting that this spread of symbolic culture is produced by natural selection (or that it is not so produced). Either way it is becoming more and more ubiquitous. If it has the neurological and cognitive effects indicated, then those may get close to being species typical characteristics in the future.

A likely response to these suggestions on the part of Evolutionary Psychology can be summarized with two statements from Pinker: "[N]atural selection is the only evolutionary force that acts like an engineer, 'designing' organs that accomplish improbable but adaptive outcomes" (1997, 36) and "[t]he evolution of information processing has to be accomplished at the nuts-and-bolts level by selection of genes that affect the brain-assembly process" (1997 176). Together these claims suggest the response that while cultural changes might bring about changes in our neural structures, these changes are not likely to be adaptive. If we find highly adaptive cognitive modules, these must be the product of natural selection acting on the genetic programs that regulate neural development. Tooby and Cosmides express a similar view. Recall their argument (summarized above) that both the genome and the stable features of the environment serve to propagate information used in the development of adaptive structures. Because of their recognition of the role of the environment, they argue that their view is not genocentric at all. But it is "very much natural selection-centered because it is natural selection that chooses some genes over others and, in so doing, orchestrates the interaction between the two inheritances [genes and environment] so that high degrees of recurrent functional order can emerge and persist, such as eyes or maternal love" (2005, 36). Here too we see the view that adaptive changes in cognitive programs must be the result of natural selection acting on the genome.

But there are other ways to think about evolution. Darwin himself defined evolution by natural selection *not* as the differential propagation of alleles, but as the differential propagation of heritable variations, i.e., phenotypes. The passage of a phenotypic trait from one generation to the next requires the transmission of genes, to be sure. But heritable variation need not be based on genetic variation, at least in the case of the neurological characteristics we are considering. Responsiveness of neural structures to culturally mediated behavior opens up the possibility of two

[7] See Ward and Gimbel (2010) for more discussion of the abductive nature of the Evolutionary Psychologists' adaptive hypotheses.

populations, or even subpopulations, that do not differ genetically but do differ in terms of neurological phenotypes. These two populations could have different sets of neurological structures, each perfectly typical within a given population. These different neurological phenotypes could realize different computational processes and so different cognitive capacities. If one variant is more adaptive, then we can imagine differential reproduction between the groups. If the behavioral factors that contributed to the occurrence of the phenotypes are reliably transmitted within the groups, we can imagine cumulative evolutionary change. Natural selection can act in the absence of genetic variation. It is not acting on genes, but it is, rather, acting on individuals in virtue of their heritable traits.

This alternative view, which allows for the possibility of heritable neurological (and hence cognitive) variation without genetic variation, fits more comfortably with the developmental systems view of evolution (Griffiths and Gray 2001; Dupré 2001) than with the Evolutionary Psychologists' view. On this developmental systems view, natural selection acts on individuals, not genes. To have long-term evolutionary effects those individuals must exhibit heritable variation, but heritability is construed more broadly than just the transmission of genes. It is, rather, the transmission of any stable developmental resource. This can include environmental factors that are causally relevant to developmental outcomes. Some species can have a role in maintaining such a developmentally relevant environmental factor across generations. In such cases this environmental factor becomes a heritable factor no less than physical genes. Culturally mediated practices seem a particularly good example of such a phenomena. Merlin Donald has put it as follows:

> If culture is essential in establishing the basic structure of the adult mind, it thereby becomes part of the mechanism of evolutionary replication and natural selection. Replicative mechanisms are central to evolutionary theory because natural selection acts on the entire process of replication, including non-genetic components (Donald 2000, 22)

Recent work on neural plasticity suggests that cultural practices are highly relevant in neural development and, indeed, relevant to computational properties exhibited by our brains and their evolution.

11.7 Conclusion

Evolutionary Psychologists consider the species-typical aspects of our cognitive architecture to be a set of genetically pre-specified, domain-specific, computational modules that are adaptations to the environment of our Pleistocene ancestors. The argument that Evolutionary Psychologists present for the claim that a particular cognitive module is in fact such an adaptation presupposes that the current environment is quite similar to the ancestral environment with respect to developmentally relevant factors. Some critics such as Dupré and Donald, along with those critics sympathetic to the developmental systems theory approach, have challenged this assumption. Recent work in developmental and cognitive neuroscience on behaviorally dependent neural plasticity strongly suggests that culturally mediated

behaviors can have profound effects on the structure of our brains throughout our lifetimes. This makes it quite likely that cultural evolution since the Pleistocene has introduced developmentally relevant changes to our social environment. This, in turn reduces the likelihood that our developmental environment is sufficiently similar to that of our Stone Age ancestors. Finally, it calls into question the assumptions and generic abductive arguments made by Evolutionary Psychologists in support of the Adaptationist Thesis. To support either a natural selection based adaptive explanation or a plasticity based cultural/developmental explanation would require the painstaking research that would yield evidence of the particular mechanism that produced our cognitive characteristics.

Acknowledgements The author would like to thank the editors of his volume and the anonymous reviewers for their very helpful criticisms and suggestions. He would also like to thank Steven Gimbel for comments on an early draft of this essay.

References

Asbury, C. and Rich, B. (Eds.) (2008): *Learning, Arts, and the Brain. The Dana Consortium Report on Arts and Cognition*. New York: Dana Press.

Barrett, H.C. and Kurzban, R. (2006): Modularity in Cognition: Framing the Debate. *Psychological Review* 113: 628–647.

Bolger, D.J., Perfetti, C.A., and Schneider, W. (2005): Cross-Cultural Effect on the Brain Revisited: Universal Structures Plus Writing System Variation. *Human Brain Mapping* 25: 92–104.

Brandon, R.N. (1990): *Adaptation and Environment*. Princeton: Princeton University Press.

Buller, D.J. (2005): *Adapting Minds. Evolutionary Psychology and the Persistent Quest for Human Nature*. Cambridge, Mass.: MIT Press.

Buller, D.J., and Hardcastle, V.G. (2000): Evolutionary Psychology, Meet Developmental Neurobiology: Against Promiscuous Modularity. *Brain and Mind* 1: 307–325.

Clark, A. (1998): Embodied, situated, and distributed cognition. In: Bechtel, W. and Graham, G. (Eds.): *A Companion to Cognitive Science*. Malden, Mass.: Blackwell, pp. 506–517.

Cosmides, L., and Tooby, J. (1997): *Evolutionary Psychology: A Primer*. Center for Evolutionary Psychology. http://www.psych.ucsb.edu/research/cep/primer.html (retrieved March 2008).

Dehaene, S. (2005): Evolution of human cortical circuits for reading and arithmetic; The 'neuronal recycling' hypothesis. In: Dehaene, S, Duhamel, J., Hauser, M., and Rizzolatti, G. (Eds.): *From Monkey Brain to Human Brain*. Cambridge, Mass.: MIT Press, pp. 133–157.

Dehaene, S. and Cohen, L. (2007): Cultural Recycling of Cortical Maps. *Neuron* 56: 384–398.

Donald, M. (1991): *Origins of the Modern Mind*. Cambridge, Mass.: Harvard University Press.

Donald, M. (2000): The Central Role of Culture in Cognitive Evolution: A Reflection on the Myth of the 'Isolated Mind'. In: Nucci L., Saxe, G., and Turiel, E. (Eds.): *Culture, Thought & Development*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 19–38.

Dupré, J. (2001): *Human Nature and the Limits of Science*. New York: Oxford University Press.

Elman, J.L. (1998): Connectionism, Artificial Life, and Dynamical Systems. In: Bechtel, W. and Graham, G. (Eds.): *A Companion to Cognitive Science*. Malden, Mass.: Blackwell Publishers, pp. 488–505.

Elbert, T., Heim, S., and Rockstroh, B. (2001): Neural Plasticity and Development. In: Nelson, C.A. and Luciana, M. (Eds.): *Handbook of Developmental Cognitive Neuroscience*. Cambridge, Mass.: MIT Press, pp. 191–202.

Fodor, J.A. (1983): *The Modularity of Mind*. Cambridge, Mass.: MIT Press.

Fodor, J.A. (2000): *The Mind Doesn't Work That Way*. Cambridge, Mass.: MIT Press.

Forgeard, M., Winner, E., Norton, A., and Schlaug, G. (2008): Practicing a Musical Instrument in Childhood is Associated with Enhanced Verbal Ability and Nonverbal Reasoning. *PLoS ONE* 3 (10): 1–8.

Griffiths, P.E. and Gray, R.D. (2001): Darwinism and Developmental Systems. In: Oyama, S., Griffiths, P.E., and Gray, R.D. (Eds.): *Cycles of Contingency. Developmental Systems and Evolution*. Cambridge, Mass.: MIT Press, pp. 195–218.

Ho, Y., Cheung, M., and Chan, A. (2003): Music Training Improves Verbal but not Visual Memory: Cross-Sectional and Longitudinal Explorations in Children. *Neuropsychology* 17: 439–450.

James, W. (1898): *Principles of Psychology*. Cambridge, Mass.: Harvard University Press.

Karmiloff-Smith, A. (2000): Why Babies' Brains are not Swiss Army Knives. In: Rose, H. and Rose, S. (Eds.): *Alas Poor Darwin: Arguments Against Evolutionary Psychology*. New York: Harmony Books, pp. 173–188.

Machery, E. (2007): Massive Modularity and Brain Evolution. *Philosophy of Science* 74: 825–838.

Mohammed, A.H., Zhu, S. W., Darmopil, S., Hjerling-Leffler, J., Ernfors, P., Winblad, B., Diamond, M.C., Eriksson, P.S., and Bogdanovic, N. (2002): Environmental Enrichment and the Brain. *Progress in Brain Research*: 138: 109–133.

Oyama, S. (2001): Terms in Tension: What Do You Do When All the Good Words Are Taken? In: Oyama, S., Griffiths P.E., and Gray, R.D. (Eds.): *Cycles of Contingency. Developmental Systems and Evolution*. Cambridge, Mass.: MIT Press, 195–218.

Pascual-Leone, A. (2001): The Brain That Plays Music and Is Changed by It. *Annals of the New York Academy of Sciences* 930: 315–329.

Pinker, S. (1997): *How the Mind Works*. New York: W.W. Norton & Company.

Pinker, S, and Bloom, P. (2001): Natural Language and Natural Selection. In: Barkow, J.H., Cosmides, L, and Tooby, J. (Eds.): *The Adapted Mind. Evolutionary Psychology and the Generation of Culture*. New York: Oxford University Press, pp. 451–494.

Ramachandran, V. S. and Rogers-Ramachandran, D. (2000): Phantom Limbs and Neural Plasticity. *Archives of Neurology* 57: 317–320.

Rapaport, W.J. (2003): Cognitive Science. In: Ralston, A., Reilly, E.D., and Hemmendinger, D. (Eds.): *Encyclopedia of Computer Science (4th ed.)*. Grove's Dictionaries, pp. 227–233.

Richardson, R.C. (2007): *Evolutionary Psychology as Maladaptive Psychology*. Cambridge, Mass.: MIT Press.

Rosenzweig, M.R. (1996): Aspects of the Search for the Neural Mechanisms of Memory. *Annual Review of Psychology* 47: 1–32.

Rosenzweig, M.R. (2003): Effects of Differential Experience on the Brain and Behavior. *Developmental Neuropsychology* 24: 523–540.

Samuels, R. (1998): Evolutionary Psychology and the Massive Modularity Hypothesis. *The British Journal for Philosophy of Science* 49: 575–602.

Sarnecki, J. (2007): Developmental Objections to Evolutionary Modularity. *Biology and Philosophy* 22: 529–546.

Schlaug G. (2001): The Brain of Musicians: A Model for Functional and Structural Plasticity. *Annals of the New York Academy of Sciences* 930: 281–299.

Schlaug G, Norton A, Overy K, Winner E. (2005): Effects of Music Training on the Child's Brain and Cognitive Development. *Annals of the New York Academy of Sciences* 1060: 219–230.

Sperber, D. (2001): In Defense of Massive Modularity. In Dupoux, E. (Ed.): *Language, Brain, and Cognitive Development. Essays in Honor of Jacques Mehler*. Cambridge, Mass.: MIT Press.

Tan, L.H., Laird, A.R., Li, K., and Fox, P.T. (2005): Neuroanatomical Correlates of Phonological Processing of Chinese Characters and Alphabetic Words: A Meta-Analysis. *Human Brain Mapping* 25: 83–91.

Tooby, J. and Cosmides, L. (2001): The Psychological Foundations of Culture. In: Barkow, J. H., Cosmides, L., and Tooby, J. (Eds.): *The Adapted Mind. Evolutionary Psychology and the Generation of Culture*. New York: Oxford University Press, pp. 451–494.

Tooby, J. and Cosmides, L. (2005): Conceptual Foundations of Evolutionary Psychology. In: Buss, D.M. (Ed.): *The Handbook of Evolutionary Psychology*. Hoboken, NJ: Wiley & Sons, pp. 5–67.

Ward, C. and Gimbel S. (2010): Retroductive Analogy: How to and How Not to Make Claims of Good Reasons to Believe in Evolutionary and Anti-Evolutionary Hypotheses. *Argumentation* 24: 71–84.

Wolf, M. (2007): *Proust and the Squid. The Story and Science of the Reading Brain*. New York: Harper.

Chapter 12
Free Will, Compatibilism, and the Human Nature Wars: Should We Be Worried?

Brian Garvey

12.1 Introduction

The 'human nature wars' are the controversies over sociobiology and its successor schools of thought – Evolutionary Psychology, biopsychology and the like. A recurring theme in these wars is the fear that characteristic claims of these schools of thought imply that we have no free will, or at least less free will than we might otherwise think we had. It is clear that, in many people's minds, sociobiology and its related schools have negative implications for free will, although it is not always clear whether this means that those implications negate free will entirely or merely mean that we have less than we might have thought. However, even if only the latter is the case, it is still *prima facie* a worry, since – assuming that it is a good thing to have free will – any news that we have less than we might have thought is bad news.

Standardly, the claim that sociobiology and related schools have negative implications for free will is based on the claim that those schools of thought have a commitment to genetic determinism. Rebuttals of this claim frequently take the form of denying that they have any such commitment, or of arguing that this commitment has no negative implications for free will. Richard Dawkins (1982, Chapter 2) employs both forms of rebuttal. He argues that genetic determinism is incompatible with a proper understanding of genetics, such as is perfectly well understood and endorsed by sociobiologists and their friends. But he also argues that, in any event, genetic determinism would be no more detrimental to free will than the opposing view that traits are shaped by the environment. Janet Radcliffe-Richards (2000, Chapter 6) employs the second form of rebuttal, using compatibilist arguments such as are familiar from previous philosophers. In the present paper I will argue that

B. Garvey (✉)
Department of Politics, Philosophy and Religion County South,
Lancaster University, Lancaster LA1 4YL, UK
e-mail: b.garvey@lancaster.ac.uk

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*,
Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_12,

neither of these forms of rebuttal are adequate when it comes to what is perhaps the most important (and certainly the most highly publicised) successor school of thought to sociobiology. This is evolutionary psychology, or to adopt David Buller's (2005a, b) usage: Evolutionary Psychology.

In Buller's usage, 'evolutionary psychology' is a field of inquiry, but 'Evolutionary Psychology' is a paradigm. A field of inquiry is a subject-matter that is studied; a paradigm is a set of agreed-upon theoretical assumptions that are used when we study it. How we explain observed facts is limited by the paradigm we are working in. For example, astronomy is a field of inquiry, but *Copernican* astronomy is a paradigm. Similarly, evolutionary psychology includes any programme of research into how evolution has shaped human psychology, but Evolutionary Psychology is the specific type of research programme that has among its notable advocates Leda Cosmides, John Tooby, Steven Pinker, David Buss and others. (For introductory accounts, see Cosmides and Tooby 1997, Pinker 1997, Buss 2008.) A central theoretical assumption that is distinctive of Evolutionary Psychology is the *massive modularity thesis* about the mind. This will be explained in more detail later but, in brief, it is a view of the mind as fundamentally a "Swiss Army Knife" – that is, as consisting of many special-purpose mechanisms that were "designed" by evolution to solve specific problems. As I will show, the massive modularity thesis plays a central role in Evolutionary Psychologists' accounts of human motivation and action. I will argue that it is because of this, rather than because of genetic determinism, that worries about Evolutionary Psychology having negative consequences for free will are justified. This means that other evolutionary schools of thought about human nature may not have these negative consequences. I will not here take up the issue of whether they do or not; instead I will concentrate on Evolutionary Psychology only. In order to determine whether my argument applies to other schools of thought, it would need to be worked out what – if any – account of human motivation those schools subscribe to.

12.2 Variants of the Worry

Before Evolutionary Psychology came on the scene, the worry about free will was articulated in responses to sociobiology, but (as will be seen) it was addressed in a very broad way to any theory that claims that human behaviour is underpinned by evolved mechanisms. In any event, much the same worry has been expressed specifically about Evolutionary Psychology, in very much the same terms and very often by the same people. The arguments of the present paper concern Evolutionary Psychology. However, both the accusations of denying free will and the defences against those accusations have been inherited from the debates around sociobiology. It should be noted that the worry is that *if* sociobiology and its related schools' claims are true *then* we have less free will than we might have thought. This would be an undesirable situation whether their claims are true are not. Many of the same critics who point out the (alleged) negative implications for free will also believe

that the relevant claims of sociobiology and its related schools are false. But it would of course also be possible to believe that they are true, and believe that the negative implications are true in consequence. So we can think of this worry as existing in two variants:

Variant 1: 'We're doomed!'
I.e. The relevant claims, the ones that have negative implications for free will, are true, and this is bad news.

Variant 2: 'It's irresponsible to say that!'
I.e. the relevant claims are false, or are at least unsubstantiated, so those who make them are spreading their tidings unjustifiably, which is liable to have bad consequences.

Why, though, should we think that if sociobiology or its related schools have these negative implications for free will, then that is something to be unhappy about? It can, I think, quite easily be shown to be plausible that *either* actually having less free will than you might have thought *or* falsely believing that that's the case, are bad. Admittedly there have been some – from the ancient Stoics, to Lao-Tzu, to Susan Blackmore (1999) – who have thought that believing in free will leads to unhappiness. However, I will not address these points of view in this paper. I will look briefly at some reasons for thinking that having less free will, or falsely believing one has less free will, are bad, with a view to determining whether those who are accused of denying free will deny it in the relevant way.

I will begin with reasons for thinking that if we *actually* have less free will than we might have thought, that is bad. These reasons, if they have relevance for the human nature wars, will *prima facie* have relevance via the 'we're doomed!' version of the free will worry.

12.2.1 Responsibility

Historically, the issue of whether we have free will has most frequently been linked to the issue of whether people can be held responsible for their actions. Classic examples of this include Hume's and Kant's discussions of free will, as well as those of J.J.C. Smart (1961) and Harry Frankfurt (1969). The basic thought here is that if people do not have free will then it makes no sense to hold them morally responsible for any actions, whether good or bad. Indeed, it has often been thought that the very idea of *morally* good or bad actions would make no sense if we did not have free will. Smart denied free will and argued that we ought to abandon the practice of morally praising or blaming people or their actions, although he thought that we could still praise or 'dispraise' them in non-moral ways, akin to rating an apple highly for its good flavour. This would imply that many of people's normal moral attitudes make no sense. Kant argued that our moral judgements necessarily presuppose free will, and consequently that we must presume that we have free will, at least when we are wearing our moral hat. But if we really do not have free will, then there is a fundamental

cognitive dissonance between our normal moral attitudes and a correct scientific view of the world. We would, then, have either to abandon our normal moral attitudes or to live in a state of cognitive dissonance. It might perhaps be thought that living with cognitive dissonance is an acceptable price to pay if it's necessary for preserving our normal moral attitudes. But if we live in a state of cognitive dissonance then we believe some things that are false, and, *prima facie*, that is less satisfactory than preserving our normal moral attitudes without cognitive dissonance. Moreover, and pertinently to the matter at hand, Evolutionary Psychologists often cite 'conceptual integration' as an advantage of their view, as against for example "the bold claims of autonomy made by the for the social sciences, accompanied by the institutionalised neglect of neighboring disciplines" (Cosmides, Tooby and Barkow 1992, p. 13, note 1). Moreover still, it is often claimed that evolutionary insights into human nature have important implications for ethics. This is one of the main messages of Edward O. Wilson's book *Consilience* (1998), for example. But this cannot be the case if conflicts between scientific findings and moral attitudes can just be ignored.

The same points apply *mutatis mutandis* if we in fact have less free will than we might have thought, rather than no free will. That would mean that *sometimes* people are not morally responsible, or deserving of moral praise or blame, when our normal moral attitudes would dictate that they are. It would mean that *sometimes* there is a cognitive dissonance between our normal moral attitudes and correct science, and so on.

12.2.2 Fatalism

Although the term 'fatalism' often has a more technical use in philosophy[1], I will here use it in the more everyday sense of the inexorability of fate. Some philosophers, for example Hume, have argued that determinism (or what Hume calls 'necessity') is not only compatible with moral responsibility but required by it. For Hume, this is because actions that are not determined are uncaused, and hence cannot be said to be any reflection of a person's character. Thus, he concludes, a person could not be morally responsible for actions that were undetermined. However, even if Hume's argument is successful, there may be other reasons why it would be bad not to have free will. To see why this is at least *prima facie* the case, imagine that you became persuaded by scientific or philosophical reasoning that people had no free will. Imagine, that is, being presented with reasons for denying free will that are

[1] Some philosophical arguments for fatalism appeal not to determinism but to the logical point that if it's true that I will do X on day Y then it always has been true that I will do X on day Y (e.g. Taylor 1962). It may be possible to bypass such arguments if one is willing to embrace the view that there are no truths about future events – i.e. that statements about a future event are neither true nor false, and only become true or false when the relevant time arrives. William James seems to have embraced this view. Be that as it may, I will leave this issue aside in the present paper.

so persuasive that you see no other option than to accept that we do not have free will. What would your feelings be? Certainly, you might be unhappy at discovering that your normal moral attitudes of praising and blaming, whether yourself or others, made no sense. But further to this, you might feel that you have no control over your own life or choices. The worry here is that, even though you might not want to X, you are fated to do so. This may go via the route of the thought: 'even though I don't want to X now, I may be fated to change, to want to X in the future'. For example, young people sometimes worry that, as they grow older, they will come to embrace their parents' values even though they reject those values now. Related to this is the worry that certain bad behaviours are inevitable. It may be believed, for example, that it is inevitable that criminals will re-offend. (This may, of course, also be thought to have implications for whether people are responsible for their actions – at least by anyone who does not find Hume's claim about moral responsibility requiring determinism convincing.)

Also related to this is the thought that certain programmes of social reform are futile, an issue that often comes up in the debates around sociobiology and related schools. It is often believed that proponents of these schools are saying that certain behaviours are 'programmed' into people's genes. Moreover, it is often believed that they are claiming this of certain *undesirable* behaviours – e.g. male chauvinist behaviour, going to war or forming social hierarchies. (See for example the first quotation from Rose in the next subsection.) This is in turn commonly taken to imply that attempts to eradicate these undesirable behaviours by education or other social-engineering means are doomed to fail. For example, in a critique of Wilson's *Sociobiology* by Richard Lewontin, Stephen Jay Gould and others, Wilson is accused of presenting "yet another defense of the status quo as an inevitable consequence of 'human nature'" (Allen et al. 1975). Interesting though this last issue is, it is not strictly speaking about free will. Rather, it is about whether certain programmes of social reform have any chance of success. So I will leave it aside here.

Once again, I take it to be clear that, if any of these accusations was true, that would be bad. Moreover, if – as the people who raise these worries usually think – they are false, but sociobiologists and related schools claim them nevertheless, that would also be bad. Most obviously, it would unnecessarily distress people to falsely tell them that their lives are fated to go certain ways whether they want them to or not. Further, there is the worry that to falsely tell people this would encourage people to falsely believe that they're not responsible for certain bad behaviours, and give them false excuses. And finally, telling people falsely that certain programmes of social reform are futile, would discourage people from following such programmes.

12.2.3 Examples

The worries I have just outlined can be illustrated by the following quotations. The first three are deserving of our attention in part because they appear in the relevant chapters by both Dawkins and Radcliffe-Richards, and in general these quotations can

be taken as representative of critiques of sociobiology and Evolutionary Psychology. In his review of Edward O. Wilson's *On Human Nature*, Stephen Rose writes:

> … for [Edward O.] Wilson human males have a genetic tendency towards polygyny, females towards constancy (don't blame your mates for sleeping around, ladies, it's not their fault they are genetically programmed). (Rose 1978, quoted in Dawkins 1982, p. 10)

Rose is here taking up the issue of responsibility. He is claiming that, according to Wilson, men are determined, because of their genes, to sleep around, and that this implies that they should not be blamed for doing so. Since Rose believes that this claim that Wilson (allegedly) makes is false, his worry is of the 'it's irresponsible to say that' variety.

In his discussion of the issue, Dawkins offers the following anecdote:

> A young woman asked the lecturer, a prominent 'sociobiologist', whether there was any evidence for genetic sex differences in human psychology. I hardly heard the answer, so astonished was I by the emotion with which the question was put. The woman seemed to set great store by the answer and was almost in tears. After a moment of genuine and innocent bafflement the explanation hit me. Something or somebody, certainly not the eminent sociobiologist himself, had misled her into thinking that genetic determination is for keeps; she seriously believed that a 'yes' answer to her question would, if correct, condemn her as a female individual to a life of feminine pursuits, chained to the nursery and the kitchen sink. (Dawkins 1982, p. 11)

Assuming that Dawkins' interpretation of the woman's tone of voice is correct, she apparently believes that if there are genetic sex differences in psychology, then she is fated to be a meek and subservient housewife, even though she doesn't want to be. Whether her worry is of the 'we're doomed' or the 'irresponsible' variety depends on whether she thinks the genetic determinism that her worry rests on is true.

In a similar vein is this quote from Stephen Jay Gould:

> If we are programmed to be what we are, then these traits are ineluctable. We may, at best, channel them, but we cannot change them either by will, education, or culture. (Gould 1978, p. 238)

Gould's claim here is very sweeping. He does not actually believe that we are 'programmed' to be what we are, so his worries are of the 'it's irresponsible to say that' variety. They also seem to be concerned with the issue of fatalism, since he refers to the impossibility of changing traits. Since he says that they can't be changed "by will", he seems to be saying that (sociobiologists' claims imply that) we as individuals are fated. But he also says that they can't be changed by education or culture, so he seems to be also saying that (those claims imply that) certain programmes of social reform are futile.

Although the quotations from Rose and Gould above relate to sociobiology, worries about free will have been raised by the same people in relation to Evolutionary Psychology. For example, in a more recent book devoted to criticisms of Evolutionary Psychology, Steven Rose asks: "Where does this strange free will come from in a genetically and evolutionarily determined universe?" (Rose in Rose and Rose 2001, p. 262). Indeed, many of the authors in that book, though it is explicitly described in the sub-title as "Arguments Against Evolutionary Psychology", devote much time

to criticising Edward O. Wilson, so indicating that they think that any criticisms that they have made of sociobiology in the past apply equally to Evolutionary Psychology.

The worry that Evolutionary Psychologists are giving excuses to bad people is greatly in evidence in some of the responses to Randy Thornhill and Craig Palmer's book *A Natural History of Rape* (2000). Because Thornhill and Palmer give an evolutionary explanation for rape, some critics conclude that they are giving excuses to rapists. For example:

> I can imagine that Thornhill's phone has been ringing off the hook with attorneys defending men accused of rape, asking him to be an expert witness for the defense. (Kimmel 2003, p. 232)

But both Dawkins and Radcliffe-Richards tell us of these worries in order to suggest that they are unfounded. Next, I will show how they set out to do this.

12.3 Dawkins' and Radcliffe-Richards' Rebuttals

Rather than saying: 'The news is bad but don't shoot the messenger', the defenders of sociobiology and related schools have usually defended themselves by arguing that their claims do not lead to the conclusion that we are any less free. In his most extended discussion of this, Dawkins (1982, Chapter 2), pursues two lines of argument: (1) He argues that genetic determinism is a straw man – i.e., that neither he nor anybody else thinks that environment plays no role in determining how an organism turns out; (2) he says *tu quoque* to his opponents – i.e. he argues that a trait that is a product of culture, upbringing, etc. is no less determined than one that is a product of genes, and consequently that his opponents' view is no less determinist than his own.

Dawkins argues for the first point by showing that we need to distinguish between genetic *selectionism* and genetic *determinism*. The former is the claim that, insofar as any traits of an organism are products of natural selection, they will be such as to promote the replication of the organism's genes. Thus, for example, genetic selectionism involves the rejection of group selection, and the endorsement of the claim that sexually reproducing organisms, insofar as their behaviour is a product of natural selection, are more likely to make sacrifices for kin than for non-kin, in the pattern predicted by Hamilton's rule (Hamilton 1964). These claims leave it completely open just which behaviours, or any other traits, are products of natural selection, and how important other factors, such as constraint and drift, are in trait-formation (a point carefully emphasised by Sterelny and Kitcher in their 1988 defence of genetic selectionism). But genetic determinism, by contrast, seems to be the view that, given that an organism possesses such-and-such a gene, it is inevitable that it will develop such-and-such a trait. It is a little difficult to precisely characterise this view, because no way of stating it comes remotely close to any view that anybody has ever held. Everybody from Genetics 101 upwards knows that the expression of

a gene depends on environmental factors, and it is difficult to see how anybody could have thought that anybody thought otherwise.[2]

Of the *tu quoque* argument, it can hardly be said that it defuses the worry about giving excuses to ne'er-do-wells, or that it alleviates any distress one might feel on being told that one is not free. At best, it spreads the blame for giving people the excuse, and for causing the distress, around a bit. Moreover, one might want to believe that at least some of one's actions are determined *neither* by one's genes nor by one's environment, so that any scientific claim that encroaches on this from either the biological or the sociological direction is bad news.

The strategy that is likely to occur to any philosopher is to appeal to a compatibilist argument regarding free will. In brief, it is to show that whether an action is determined by prior causes or not has no bearing on whether or not it is free. This strategy has been pursued by Janet Radcliffe-Richards in *Human Nature after Darwin* (2000). The aim of Radcliffe-Richards' book overall is to defuse many of the worries people commonly have about the claims of sociobiology and related schools, including, as she makes explicit, Evolutionary Psychology – worries about politically reactionary or quietist implications, for example. Knowing that one of these worries is that we are being claimed to be 'blameless puppets', she argues that this worry arises because of misunderstandings of what free will actually is. In arguing this, she uses standard arguments for compatibilism, such as are familiar from classic compatibilist accounts (e.g. Ryle 1949, Chapter III; Ayer 1954; Frankfurt 1969).Very briefly, compatibilists argue that free will is possible even if our actions are determined. They say that, when we are unfree, it is because of some specific circumstance, which might be, for example, being in prison, being subject to some psychological compulsion, etc., and they argue that anyone who thinks that determinism entails that free will is impossible is treating being determined by cause as if it were the same as one of those specific circumstances.

Radcliffe-Richards' type of response is specifically directed against the claim that a significant genetic component in determining behaviour means that we have no, or less, free will. Her argument consists of two strands: (1) she sets out to show that what the classical free will theorist wants – acts that are not determined, and are free – is incoherent, and hence cannot be had in any case; (2) she then sets out to show that an act can be determined and yet be free. She employs two time-honoured strategies for showing the first: (i) Hume's 'other fork': the argument that an event that is not determined is random, but a random event is not a free act; (ii) the argument that nothing can be the cause of itself: the classical free will theorist wants human actions to be *neither* determined by prior causes *nor* random, but this,

[2] Admittedly, Evolutionary Psychologists often claim that evolved cognitive mechanisms can be relied on to develop in a wide variety of different environments by virtue of being guarded against environmental vicissitudes that might disrupt development. The mechanisms by which they are so guarded are never specified beyond vague expressions such as 'feedback-driven compensation' (Tooby and Cosmides 1992, p. 81). I will leave this issue aside in the present paper. For a sceptical view on this claim of Evolutionary Psychologists', see Garvey 2005.

compatibilists hold, makes no sense. Since something cannot be the cause of itself, human actions, like any other event, must be either caused by something else or not caused at all, i.e. random. The strength of these two arguments is their extreme generality. They do not depend on particular scientific claims – not even very general claims such as that the world consists of matter in motion or that everything has a cause. But all that they show is that the classical free-will theorist is making demands that cannot be met; by themselves they do nothing to reassure us that we have free will. It is the second strand of the compatibilist argument that interests me here. (Strictly speaking, the first strand could just as easily be part of an argument for denying free will as for compatibilism, so only the second strand should be called 'the compatibilist argument' – which is what I will do from here on.)

The basic thrust of the compatibilist argument is to show that the hard determinist, in saying that we're not free because we're determined, is misunderstanding what it is to be free. This is sometimes cast as a misunderstanding about how we use the *word* 'free', but this is not the place to debate the merits of 'ordinary language philosophy' versus 'robust metaphysics'. For present purposes I am assuming that the compatibilist argument works, and that it proves something about the real nature of freedom, not just about the way we use words. Compatibilists often present their arguments as *reassurances* that any type of freedom 'worth wanting' is perfectly possible even if all our actions are determined (e.g. Dennett 1984). The hard determinist, it is alleged, conflates two very different circumstances in which one might say: 'I'm not free because …' On the one hand, the hard determinist would have us say 'I'm not free because my actions are determined by prior causes'. On the other hand, there are specific circumstances in which we might say 'I'm not free to do *x* because …' But, the compatibilist urges, if the *only* reason I can be said to be not free is because my actions are determined, then there is no reasonable cause for concern: I am not unfree in any sense that I should be worried about.

12.4 Why This Does Not Get Rid of the Problem

It is not the aim of the present paper to argue either for or against compatibilism. Rather, I want to determine whether, *if* compatabilism is true, the worries about Evolutionary Psychology's implications for free will are misguided. I will argue that compatibilism does not successfully defuse these worries. This is because, it will be argued, compatibilist arguments, even if successful, only show that it is possible to have free will in deterministic scenarios, not that we have free will in every deterministic scenario. If compatibilism is true, then the mere fact of being determined by prior causes does not make an action unfree. However, it does not follow from this that all actions are free. It may be that there are specific circumstances in which actions are not free, and all compatiblist accounts allow for this. Moreover, it could, consistently with compatibilism being true, be that a great many or even all of our actions fail to be free because of some general fact other than the fact of being determined. It could then, again consistently with compatibilism being true, be that

Evolutionary Psychology makes some general claims about human beings which implies that we are not free, or at least significantly less free than we might otherwise have thought.

12.4.1 Circumstances in Which We're Not Free, Even on a Compatibilist Account

There are various versions of compatibilism. All have in common that actions can be free even if determined, but all also allow that some actions are not free. Radcliffe-Richards' argument only shows that we can be free even if determined, and it doesn't follow from that that we are in general free. Still less does it follow that even if what Evolutionary Psychologists say is true we are are still as free as we thought. This is because all compatibilist theories allow that there are some circumstances where our actions are not free, or at least are less free than we would like.

The simplest version of compatibilism holds that we are free as long as we are moved by our own desires and not by anything else. Hume gave the classic formulation of this:

> By liberty, then, we can only mean *a power of acting or not acting, according to the determinations of the will*; this is, if we choose to remain at rest, we may; if we choose to move, we also may. (*Enquiry Concerning Human Understanding*, Section VIII, part I)

Thus, on this account, a person can be free even if their actions are caused by their desires, and those desires have prior causes, and even if the chain of causes goes all the way back to the beginning of time. In other words, even if their actions are fully determined by prior causes, a person can still be free. Most compatibilist accounts would require that one's actions be *caused* by one's desires in order for one to be free. That is, it would not be considered sufficient that they be merely *in accordance with* one's desires. This rules out cases where one was forced to do something that happened to be what one wanted to do. However, they allow that, even if those desires are in turn caused by something else, and the chain goes back to the beginning of time, one can still be free.

However, many people, whether compatibilists or not, think that this is insufficient for free will. The problem is that there are many cases where it looks as though one's actions *are* caused by one's desires, but it also looks as though one is not free. Among such cases are those where one is driven to act by *addictions* or *psychological compulsions*. For example, a person who is addicted to smoking may feel that they are *compelled* to smoke, and that their freedom of choice is reduced by this addiction. Similarly, a person who has OCD may feel compelled to count the paving-stones. The very name of the disorder – obsessive *compulsive* disorder – suggests that its sufferers are compelled by it to do things, and the testimonies of sufferers from OCD themselves indicate that they experience it in that way. Similarly again, a person may experience a *phobia* as a reduction of their freedom: an agoraphobic may be *unable* to leave the house, or at least find it very difficult to do so.

However, in these cases, it is not entirely clear that the person is being prevented from doing what they want. We might say that the smoker and the paving-stone counter *are* doing what they want to do, and that the agoraphobic is avoiding doing what she doesn't want to do. Yet the intuition that addictions, compulsions and phobias reduce one's freedom seems to be a strong one. There are two possible ways to accommodate this intuition. (1) One approach is to look more closely at what is happening in such cases, and argue that despite appearances the person is not actually acting on their own desires. Thus, the simple compatibilist definition of freedom is preserved, and the cases are interpreted to show that they do not fit it. (2) The other approach is to modify the definition itself, so that not all situations where one is acting on one's desires count as situations where one is free. But both approaches share the intuition that addictions, compulsions and phobias reduce one's free will; they merely differ on *why* this is so. I will look at a number of different answers to the question of why these conditions reduce one's freedom. What I aim to show is that, whichever of these answers one accepts, the mandatorily-arising desires which are said by Evolutionary Psychologists to be part of our legacy from evolution, reduce people's freedom for exactly the same reason, at least *prima facie*.

(1) One approach to explaining why addictions, etc., reduce one's freedom is to argue that, when one succumbs to an addiction, one is not doing what one wants, despite appearances. For example, (1a) one might describe cases such as a person addicted to smoking, or compelled to count the paving-stones, as situations where it is impossible for that person to have all the things she wants. For example, she wants to smoke, but doesn't want to incur the health risks. There may be some problems with this approach, however, for it is possible for a person to be completely indifferent to all the drawbacks of smoking and yet still be addicted.

(1b) An alternative possibility might be that, although the person wants to smoke, at the same time she wants to not want to smoke. Her desire to not want to smoke is a *higher order* desire, and it is this that she is unable, or finds it hard to, fulfil, because of her addiction. Anybody who has tried to give up smoking or any other addiction will be familiar enough with this. However, although it may be a correct description of some cases, it suffers from essentially the same problem as (1a). A person who has no such higher order desire – who is perfectly happy with wanting to smoke – might yet still be addicted. I will say more about higher-order desires a little later.

(1c) We might accommodate cases where the addicted person has no conflicting desires by arguing that the addicted person is doing what she wants, but that her wanting to do it is not what's causing her to do it. For example, a person may enjoy drinking, like the taste, enjoy the social accompaniments and even enjoy the sensation of being drunk, and any or all of these may be the reason that the person drinks. On the other hand, the reason that the person drinks may be that she is an alcoholic. It may be that, for example, she might change her mind about the pleasantness of the taste of drink, about the desirability of pub company, and so on, and not find any reasons for wanting to drink left, but still drink. If this were the case, it would be reasonable to conclude that, before the person changed her mind, her drinking was not caused by those reasons, and conclude from this that she is an alcoholic. So her drinking, we might as well say, coincided with what she wanted to do, but was no more a free act than if someone

held a gun to my head and made me sign an agreement to do something I wanted to do anyway. Although this may again be the correct description of some cases, it does not cover all of them, because an addiction may itself *produce* a desire. So, even if a person has no other desire either to drink or not to drink, she may, just because she is an alcoholic, want to drink and be caused to drink by that wanting.

(2) On the other hand, rather than trying to see addicted persons as not really driven by desires, one might instead say that their desires are what's causing them to do the thing, but that there is something amiss with their desires. This would require modifying the simple compatibilist definition of freedom, so that more is required, to be free, than just being caused to act by one's desires. The desires themselves have to be of the right kind.

We need to say a bit more about higher-order desires. If having difficulty satisfying a higher-order desire counts as a restriction on freedom, then the term 'higher-order desire' may be plausibly extended to cover desires that are *hypothetical*. That is, it may cover what, all things considered according to my own judgement, I *would* want. The part about '*my own* judgement' is important, because we don't want to confine the term 'free' to only actions that arise out of desires that are right according to some impersonal objective rational standard of which I'm unaware or which I would positively reject. But the part about '*would* want', rather than positively do want, is important too. For many actions may arise out of desires that have no actual higher-order desires attached to them at all. An obvious example is eating because I'm hungry. The sum total of my attitudes towards eating may be: I'm hungry, so I want to eat. But, presumably, eating is also what I would want to do, taking all my desires into consideration (e.g. I don't want to die). On the other hand, the sum total of an alcoholic's attitudes towards drinking may be: I want more drink. Moreover, the alcoholic may even want to drink for other reasons as well – as mentioned above. But even if all the person's attitudes about drinking are 'pro-attitudes', and one of those attitudes – the desire to drink itself – is what's causing the person to drink, the person can still be an alcoholic. We need some way to mark the difference between this and eating because one is hungry, or any number of other acts done out of unreflected-upon desires that are perfectly harmless. I suggest that the relevant difference is that some unreflected-on desires would be what we would still decide was best, or at least not harmful, were we to reflect on them.

But there is, I think, a deeper reason behind this – which is, that we would like to think that, were we to reflect and change our minds about the desirability of doing something, we would be able to act, or refrain from acting, as we saw best without being faced with obstacles from our own desires. Because of this, the fact that one is restricted stems from the fact that something would make it more difficult to do as one wanted, even if it isn't actually preventing one from doing anything that one wants to do now. Even in the case of the happy alcoholic who wants to drink because of the taste, the pub company, etc., her freedom to refrain from drinking is restricted because she *would* find it hard to refrain from drinking if she were to change her mind about the taste, the desirability of pub company and so forth. But, similarly, a person who is locked in is restricted in her freedom to leave the house, even if she doesn't want to, because she *would* find it hard to leave were she to change her

mind. So any freedom worth wanting has to involve being free from obstacles to doing something that I want, even if this wanting is only hypothetical, and whether those obstacles are external or internal. For reasons given above, the relevant hypothetical desires should not be thought of merely as what an abstract person of perfect judgement would want to do; in the final analysis, they should be thought of as what the individual person is liable to find herself wanting to do.

The upshot of this, then, is that one can be restricted in one's freedom to do something even if one doesn't want to do that thing. And conversely, one can be restricted in one's freedom to refrain from doing something, even if one wants to do that thing and has no actual desires that conflict with it. The key point is that, on any reasonable account, a person's freedom to X seems to be reduced to the degree that (1) that person would find it difficult to X if she wanted to, and (2) that person is liable to find herself wanting to X. And, *mutatis mutandis*, a person's freedom to refrain from X-ing is reduced to the degree that that person would find it difficult to, etc.

It may not be immediately obvious why the second condition is needed. One might think that, if it is difficult for me to X, then that is sufficient for my freedom to X to count as being reduced. However, the second condition needs to be added to avoid counting as restrictions on freedom things that pretty clearly aren't. This can be seen if we once again consider hunger. Clearly, most of us would find it difficult to refrain from eating even if we wanted to, because we would get hungry. But we don't usually consider this a restriction on our freedom. The same goes for the desire to sleep, the desire to urinate, and so forth. One might be tempted to write such desires off as 'normal', and hence not possible to count as addictions, and hence not as restrictions on freedom. However, it is clearly a *non sequitur* to go from 'this is not an addiction' to 'this is not a restriction on freedom'. Moreover, we at least owe the hard determinist the courtesy of allowing it to be *possible* that even perfectly normal circumstances can count as restrictions on freedom. That is, we shouldn't claim it as an *a priori* truth that what's normal can't be a restriction on freedom. In any event, the term 'normal' is notoriously slippery, carrying with it a danger of slipping between 'statistically average' and something like 'normative' or 'healthy'. Desires that are not statistically average do not just for that reason count as restrictions on freedom – otherwise we would have to count homosexual desires or very specialist tastes in music as restrictions on freedom. Further still, there might be *specific situations* where even desires that are 'normal' – in the senses of *both* statistically average and healthy – count as restrictions on freedom. A person might, for reasons that are very central to her world-view and ideals, decide to go on hunger strike, in which case hunger might be best thought of as a restriction on her freedom. Such things have been known to happen. Still, for most of us the desire to eat is not a restriction on our freedom, and I suggest that this is because it is unlikely that it is going to conflict with another desire. To repeat what I said above, even if we don't often consciously think about it, most of us want to stay alive, so the desire to eat is a desire to do something that we would be perfectly happy to do if we thought about it. So it is not sufficient for something to be a restriction on freedom, that it would make it hard to do something if we wanted to: the *degree to which we are liable* to actually want to do that thing is also a factor.

Note that I say 'to the degree that' and not '*only* to the degree that'. I do not wish to rule out other ways in which freedom may be considered to be reduced. Nonetheless, I believe this captures the reason that we have a strong intuition that addictions, compulsions and phobias reduce people's freedom. It need not be impossible for a person to avoid acting on a psychological compulsion, but it is difficult for them, and that difficulty is to the degree that the compulsion is strong. Neither need a psychological compulsion be in conflict with a person's desires, but it is liable to be so. And it is to the degree that it is liable to be in conflict with other desires that it constitutes a reduction of freedom. The desire to eat does not usually reduce freedom, but others, such as the addict's desire to smoke, more often do. This is *not* because the former is normal and 'natural' while the latter isn't, but because the desire to eat doesn't usually make it hard to fulfil other desires, whereas the desire to smoke often does.[3]

12.4.2 Evolutionary Psychology's Account of Motivation

In this section I will argue that, because of Evolutionary Psychologists' commitment to the massive modularity thesis, there is strong *prima facie* reason to think that, on their account, many perfectly normal human impulses to act are similar to addictions, phobias and so forth. Specifically, their view implies that those normal human impulses possess the very features of addiction and phobias that make them, on any reasonable compatibilist account, count as restrictions on free will. This is not affected by the fact that Evolutionary Psychologists do not subscribe to genetic determinism. It is their commitment to the massive modularity thesis, and not any genetic determinism, that leads to their views having *prima facie* negative implications for free will.

The massive modularity thesis is an absolutely central distinctive feature of Evolutionary Psychology. The latter's major proponents – Leda Cosmides, John Tooby, David Buss, Steven Pinker, Donald Symons, and others – all explicitly endorse the massive modularity thesis, and employ it with great frequency in their psychological theories. As I will argue, this fact by itself has *prima facie* implications for how Evolutionary Psychologists will view human motivation. Moreover, some Evolutionary Psychologists explicitly tell a story about human motivation along these lines. It is this view of motivation, I will argue, that justifies the worries that critics of Evolutionary Psychology have about it having negative implications for free will.

The massive modularity thesis is the thesis that the mind consists wholly or largely of special-purpose, dedicated, cognitive mechanisms; no even approximate

[3] A possible objection to this pair of conditions (which was actually raised to me by both Alex Neill and one of the referees for this volume) is the following: it might occur that, for some reason, I want to grow wings and fly, and on my account the fact that I can't would then count as a restriction on my freedom. In response to this I say: (1) it is not news to anyone that we are unable to grow wings, whereas the point at issue here is whether Evolutionary Psychology, if true, gives us grounds for thinking that we have less free will *than we would otherwise think we had*; (2) it is in any event unclear whether we can be said to *want* to grow wings, rather than that we *wish* we could.

number is specified, but we are to take it that there are a great many of them. For Evolutionary Psychologists, this thesis is explicitly grounded in adaptationist arguments, to the effect that adaptations are solutions to specific problems that arise at specific places and times (in the case of our cognitive modules, the relevant environment of evolutionary adaptedness is the Stone Age), and to the effect that decoupling of function is advantageous. So, strictly speaking, these arguments only have force insofar as our cognitive architecture consists of adaptations, allowing for much cognitive architecture that neither consists of adaptations nor is modular. But Evolutionary Psychologists typically believe that most of our architecture does consist of adaptations, and hence that it is modular.

One of the key features of cognitive modules is that their operation is *mandatory*. Jerry Fodor explains this with simple examples:

> You can't help hearing the utterance of a sentence (in a language you know) as an utterance of a sentence, and you can't help seeing a visual array as consisting of objects distributed in three-dimensional space. Similarly, mutatis mutandis, for the other perceptual modes: you can't, for instance, help feeling what you run your fingers over as the surface of an object. (Fodor 1983, pp. 52-53)

A consequence of this is that cognitive processes that are modular take place even in spite of other information that the mind might have. This can be illustrated with optical illusions. The Müller-Lyer lines, despite being the same length, *appear* to be different lengths, as a result of the arrows on the ends pointing in different directions Even when one has measured the lines and seen that they are the same, the optical illusion doesn't go away. This suggests that whatever part of the mind processes visual input does not receive all the information that is available to other parts of the mind. The knowledge that the lines are the same length does not seem to get through to the visual-processing mechanism – it still 'thinks' that they are different lengths. Evolutionary Psychologists would add to this story that it is because evolution hasn't prepared us for this trick that the Müller-Lyer lines appear to be different lengths in the first place. There presumably weren't any Müller-Lyer lines around in the Stone Age.

It is relatively uncontroversial that sense-perception and language comprehension are underpinned by cognitive modules. But Evolutionary Psychologists claim that a whole host of other things are as well. They claim that evolution has bequeathed us a host of automatic responses to situations, which are to be understood as responses that would have been fitness-enhancing for Stone Age humans.

> [O]ur minds consist of a large number of circuits that are *functionally specialized*. For example, we have some neural circuits whose design is specialized for vision. All they do is help you see. The design of other neural circuits is specialized for hearing. All they do is detect changes in air pressure, and extract information from it. They do not participate in vision, vomiting, vanity, vengeance, or anything else. Still other neural circuits are specialized for sexual attraction – i.e., they govern what you find sexually arousing, what you regard as beautiful, who you'd like to date, and so on. (Cosmides and Tooby 1997, p. 7. ; emphasis in original)

> [T]he reasoning circuits and learning circuits discussed above have the following five properties: (1) they are complexly structured for solving a specific type of adaptive problem,

> (2) they reliably develop in all normal human beings, (3) they develop without any conscious effort and in the absence of any formal instruction, (4) they are applied without any conscious awareness of their underlying logic, and (5) they are distinct from more general abilities to process information or behave intelligently. (Ibid., p. 9)

Evolutionary Psychologists often explicitly say that their commitment to the massive modularity thesis distinguishes their school of psychology from other evolution-based ones. For example, Donald Symons (1992) distinguishes Evolutionary Psychology from something which he calls "Darwinian Social Science". On Symons' account, the latter is committed to the idea that evolution has bequeathed us a general desire to survive or to reproduce. According to "Darwinian Social Science", this general desire causes us to have more specific desires, such as the desire to eat or to have sex. However, Symons and other Evolutionary Psychologists argue that evolution could not possibly have produced such a general desire. Rather, they argue, natural selection would favour special-purpose cognitive mechanisms that produce specific responses to specific conditions – for example, finding a certain food tasty or finding a certain person attractive. The fact that we possess these responses is explained by their contribution to our ancestors' survival and reproduction – *not* by any desire that we have to survive or reproduce.

Based on this, Evolutionary Psychologists have a standard template for explaining human psychological traits. It goes like this: (1) There is some problem that Stone Age humans had to solve in order to maximise their chances of surviving and reproducing – e.g. the problem of deciding who to mate with; (2) they did not solve this problem by conscious reasoning; instead, natural selection produced cognitive mechanisms that were dedicated to solving it – e.g. the cognitive mechanisms that cause us to find certain individuals sexually attractive; (3) those cognitive mechanisms operate non-consciously and mandatorily – e.g. we do not consciously calculate how beneficial to our genes it would be to mate with a certain person; rather, we simply find certain people sexually attractive, and we have no control over that fact; (4) those cognitive mechanisms are adapted to conditions in the Stone Age; they need not be fitness-enhancing in present-day conditions.

As an example of this template in action, consider Symons' account of humans' desire to eat sweet foods. Eating as much sugar-containing food as one could get would be a good strategy in an environment where there wasn't very much of it around, but it would be a very poor strategy today. Even Evolutionary Psychologists hold that we don't have to act on these automatic responses: the responses are desires, not actions. But they emphasise that the responses themselves are things we have no control over:

> Human behavior is flexible, of course, but this flexibility is of means, not ends, and the basic experiential goals that motivate human behavior are both inflexible and specific. For example, assume that we, along with many other primates, possess a specialized gustatory mechanism underpinning the sensation of sweetness. This mechanism was shaped by natural selection in ancestral populations because a sugar-producing fruit is most nutritious when its sugar content is highest, hence individuals who detected and liked sugar produced, on average, more progeny than did individuals who could not detect sugar or who actually preferred the taste of green or overripe or rotten fruit. Since human behavior is so flexible, we have been able to develop virtually an infinite number of ways of obtaining sugar; but the goal of eating sugar remains the same – to experience the sensation of sweetness.

> In modern industrial societies, where refined sugar is abundantly available, the human sweet tooth may be dysfunctional, but sugar still tastes sweet, and the goal of experiencing sweetness still motivates behavior. That's how we're made. We can decide to avoid refined sugar, but we can't decide to experience a sensation other than sweetness when sugar is on our tongues. …
>
> In summary, although human behavior is uniquely flexible, the goal of this behavior is the achievement of specific experiences – such as sweetness, being warm, and having high status. (Symons 1992, pp. 138-139)

So the claim is that, not just the sensation of sweetness, but the desire to eat sweet things arises mandatorily, just as the perception of a visual array as a three-dimensional object does. Note that Symons says that, even though human behaviour is flexible, the sweet taste of sugar and the goal of obtaining it remain unchanged. To go back to the particular example mentioned by Rose, the analogous claim is not that men can't help philandering, but that the temptation to philander arises mandatorily. Still, we might say, nobody is claiming that we are *compelled* to act on such temptations, so they do not count as restrictions of our freedom. Things are not quite that simple, however.

The worry that they might be restrictions on freedom arises from the Evolutionary Psychologists' claim that our cognitive modules are an inheritance from the Stone Age, and hence are likely to be adaptations to life in the Stone Age. But what was adaptive in the Stone Age need not be adaptive now, and nor need it coincide with what we want now. The desires to eat, drink and sleep would not normally count as restrictions on freedom, however mandatory they might be, because eating, drinking and sleeping are all still things that, all things considered according to our own judgement, we *would* want to do. But the same may not be true of all the things that it was good for our Stone Age ancestors to do. Eating as much sweet food as possible was something that promoted the well-being of Stone Age humans, and so was something that, all things considered according to their own judgement, they would want to do. Even if they didn't *know* that ripe fruit was more nutritious, they didn't know of any reason why eating it would be a bad thing (and, *ex hypothesi*, there usually wasn't any such reason). But nowadays a person, faced with a far greater amount of sugary foods, is reducing her fitness by pursuing the same strategy, acting on the same mandatorily-arising desires. That would not in itself make it a restriction on freedom, but there is the further fact that we now *know* that too much sweet food is bad for you. Hence, we are *liable* to not want to eat so much of it. Evolutionary Psychology implies that we have automatic responses to situations that are fundamentally inappropriate to those situations, even if we have information that would enable us to respond more appropriately, and even if we *want* to respond more appropriately. Since the automatic responses are said to be desires, not actions, we are not prevented from making the appropriate response – that is the straw man of genetic determinism. But if the evolved responses really are as mandatory as the Evolutionary Psychologists claim, then they are going to make it difficult to do things that we are liable to want to, and to refrain from doing things that we want to refrain from doing. Hence, it looks like they are restrictions on our freedom.

As a consequence of this, it also looks as though both the worries about responsibility and those about fatalism have some justification when it comes to Evolutionary Psychology. Addictions are generally taken to reduce responsibility. One may be

held responsible for *becoming* addicted, and hence for what one does when addicted, but that wouldn't make one responsible for inborn addiction-like tendencies, which is what the mandatorily-arising desires hypothesised by Evolutionary Psychologists appear to be. So what about Rose's worry – that an excuse is being given to philandering men? Since Evolutionary Psychologists see sexual desires as products of natural selection, and hence as based on modular cognitive architecture that kicks in automatically, it looks as though they are just as mandatory as the experience of sweet. Indeed, although Evolutionary Psychology aspires to being a complete theory of the underlying architecture of the human mind, the differences between the mating strategies of men and women are their number one favourite subject. They allege that it made sense for men in the Stone Age to be promiscuous, and for women to be highly selective, and that these strategies are embedded in preferences that are hardwired into the human mind. So, *prima facie*, it looks as though they are saying that men are perpetually tempted to philander. This suggests, in turn, that it is hard for men to refrain from philandering, even if they want to refrain. So it looks as though Evolutionary Psychologists are claiming that men's freedom is reduced in this regard. And it also looks as though what they are saying is that these desires are out of our control. Thus, they seem to be suggesting that no matter how much a man might not want to philander, the desire to philander will arise. This desire will in turn reduce men's ability to avoid philandering, the extent to which it reduces it depending on how strong the desire is. Hence we are all, on Evolutionary Psychology's picture, in a position analogous to that of the young person who is fated to embrace her parents' values even if she does not want to.

12.5 Conclusion: Evolutionary Psychology's Get-out Clause

Alert readers will have noticed that I have said 'looks like' and '*prima facie*' quite a lot. I have only been arguing that there is at least a plausible case that can be made that central claims of Evolutionary Psychology have negative implications for free will. Evolutionary Psychologists sometimes show an awareness of this problem, and gesture towards a solution by claiming that we have the ability to override the motivations generated by our evolved cognitive mechanisms. Some are fond of pointing out that *of course* they don't believe that it is inevitable that we will behave in the ways that our evolved cognitive architecture is designed to make us behave. Steven Pinker cheerfully points out that, although he is a healthy, high status male, he has yet to produce any offspring: "I am happy to be that way, and if my genes don't like it, they can go jump in the lake" he declares (Pinker 1997, p. 52). Similarly, Radcliffe-Richards says:

> Occasionally, of course, an emotion is so overpowering that a person is no longer capable of control, but that is a situation we count as mental disorder or illness, or, when temporary, a state of diminished responsibility. If evolutionary psychologists claimed that genetically ingrained emotions were typically of this kind – a kind that constituted mental illness – that would, of course, be enough to prove that evolutionary psychology was nonsense. But … evolutionary psychology makes no such claim. The claims of evolutionary psychology are about the *origins* of human dispositions, not about how strong they are. (Radcliffe-Richards 2000, p.115; emphasis in original)

She is thus denying the similarity that I am claiming exists between addictions, etc., and the mandatory desires that Evolutionary Psychology postulates. As a brief aside, it should be pointed out that Evolutionary Psychology does *not* just make claims about the origins of dispositions, but also about what dispositions we are likely to have and what form they take (cognitive modules). But the big problem is that it is not clear *how* they think we come to be able to override our evolved desires. Hence, it is not clear how, on their account, those desires are relevantly different from those of an addict or a person with OCD. It is a central pillar of their view that that all or nearly all of the underlying architecture of the human mind is modular. (For a sceptical view on this, see Fodor 2000.) But then, it is not clear just how the automatic responses get overridden. I am not here expressing scepticism about the claim that they *are* overridden, but pointing out that it is not clear what the mechanisms *by means of which* they are overridden are supposed to be. Unless we know this, we don't know just how easy or difficult the automatic desires are to override. If they are *very* easy to override, they don't count as reductions of freedom at all, and so we have nothing to worry about. But since the Evolutionary Psychologists have given us nothing that would supply an answer to the question of how difficult it is, we don't know to what extent – if any – they are giving excuses to ne'er-do-wells. Nor do we know just how hard it is for ourselves to escape biological destinies that might not appeal to us. Consequently, we don't know just how worried we should be.

This is one reason why we need a clearer account of what exactly Evolutionary Psychologists are claiming about motivation than has so far been given in the literature. They often say that we can override those desires that we have inherited from our Stone Age ancestors, but they give no account of how we are able to do this. It *is* possible, and indeed *prima facie* not all that difficult, for men to resist the temptation to philander. But Evolutionary Psychologists claim that their theories provide a causal account that explains why behaviours of this kind exist. More generally, they see their theories as providing insight into the causes of large swathes of human behaviour. If that's the case, then they owe us an account of why people behave in ways that are different from the ways their cognitive modules are 'designed' to make them behave. Evolutionary Psychologists might claim that we are self-deceived about how easy it is to override them; for example, they might claim that a lot more philandering goes on than we think, and that that is because of the difficulty of resisting the promptings of our evolved modules. If that is the case, then Evolutionary Psychologists can indeed claim to have uncovered significant causal factors in human behaviour. But equally, if that is the case, and to the extent that that is the case, they bring us bad news about free will. They cannot have things both ways. As things stand, *either* they have an account of motivation that has negative consequences for free will, *or* they have a seriously incomplete account of motivation, and hence of the mind generally.[4,5]

[4] It will be noticed that I have said nothing about whether any other scientific account of psychology is likely to lead to similar problems for free will. This is because I am concerned here with the issue of whether, specifically, Evolutionary Psychology raises *distinctive* problems for it. As I hope to have shown, Evolutionary Psychology's distinctive combination of modularity with the claim that the mind is fundamentally adapted to Stone Age conditions, means that it does.

[5] I am grateful to Kristian Ekeli and Alex Neill, as well as the three (necessarily anonymous) referees for this volume, for extremely valuable comments and criticisms.

References

Allen, E. *et al* (1975): 'Against "Sociobiology"', *New York Review of Books*, August 17.

Ayer, A.J. (1954): 'Freedom and Necessity.' In Ayer, A.J.: *Philosophical Essays*. London: Macmillan.

Barkow, J.H., L. Cosmides and J. Tooby (Ed.) (1992): *The Adapted Mind*. Oxford: Oxford University Press.

Blackmore, S. (1999): *The Meme Machine*. Oxford: Oxford University Press.

Buller, D. (2005a): 'Evolutionary Psychology: The Emperor's New Paradigm', *Trends in Cognitive Sciences* 9: 277–283.

Buller, D. (2005b): *Adapting Minds: Evolutionary Psychology and the Persistent Quest for Human Nature*. Cambridge, MA: MIT Press.

Buss, D. (2008): *Evolutionary Psychology: The New Science of the Mind*. Third edition, Boston: Allyn and Bacon.

Cosmides, L. and J. Tooby (1997): 'Evolutionary Psychology: A Primer.' http://www.psych.ucsb.edu/research/cep/primer.html. Accessed June 25, 2009.

Cosmides, L., J. Tooby and J.H. Barkow (1992): 'Introduction: Evolutionary Psychology and Conceptual Integration.' In: Barkow, J.H., L. Cosmides and J. Tooby (Ed.) (1992): *The Adapted Mind*. Oxford: Oxford University Press.

Dawkins, R. (1982): *The Extended Phenotype*. Oxford: Oxford University Press.

Dennett, D. (1984): *Elbow Room: The Varieties of Free Will Worth Wanting*. Oxford: Clarendon Press.

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

Frankfurt, H. (1969): 'Alternate Possibilities and Moral Responsibility', *Journal of Philosophy* 66: 829–839.

Garvey, B. (2005): 'Nature, Nurture and Why the Pendulum Still Swings', *Canadian Journal of Philosophy* 35: 309–330.

Gould, S.J. (1978): *Ever Since Darwin: Reflections in Natural History*. Harmonsworth: Penguin.

Hamilton, W. (1964): 'The Genetical Evolution of Social Behaviour' (I and II). *Journal of Theoretical Biology* 7: 1-16 & 17–52.

Kimmel, M. (2003): 'An Unnatural History of Rape'. In: Cheryl Brown Travis (Ed.) *Evolution, Gender and Rape*, Cambridge, MA: MIT Press.

Pinker, S. (1997): *How the Mind Works*. Harmondsworth: Penguin.

Radcliffe-Richards, J. (2000): *Human Nature After Darwin*. London: Routledge.

Rose, H. and S. Rose ed. (2000): *Alas, Poor Darwin: Arguments Against Evolutionary Psychology*. London: Jonathan Cape.

Rose, S. (1978): 'Pre-Copernican Sociobiology?', *New Scientist* 80: 45–46.

Ryle, G. (1949): *The Concept of Mind*. London: Hutchinson.

Smart, J.J.C. (1961): 'Free Will, Praise and Blame', *Mind* 70: 291–306.

Sterelny, K. and P. Kitcher (1988): 'The Return of the Gene', *The Journal of Philosophy* 85: 339–361.

Symons, D. (1992): 'On the Use and Misuse of Darwinism in the Study of Human Behavior.' In: Barkow, J.H., L. Cosmides and J. Tooby (Ed.) (1992): *The Adapted Mind*. Oxford: Oxford University Press.

Taylor, R. (1962): 'Fatalism.' *The Philosophical Review* 71: 56–66.

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

Tooby, J. and L. Cosmides (1992): 'The Psychological Foundations of Culture', In: Barkow, J.H., L. Cosmides and J. Tooby (Ed.) (1992): *The Adapted Mind*. Oxford: Oxford University Press.

Wilson, E.O. (1998): *Consilience: The Unity of Knowledge*. London: Little, Brown and Company.

Chapter 13
Altruistic Emotional Motivation: An Argument in Favour of Psychological Altruism

Christine Clavien

13.1 The Controversy

Human beings are capable of helping other beings in need at their own expense and, apparently, without thinking of their own short or long term interests. Philosophers and psychologists label this sort of action 'altruistic', or 'apparently altruistic'. There is a long standing debate over whether prima facie altruistic actions are truly other-directed or prove self-directed under closer analysis. In the first part of the paper, I will sketch out the main lines of this controversy before showing that it somehow cancels in a battle of a priori statements. The second part of the paper is a proposal to reframe the whole debate in order to overcome this deadlock and give way to an evolutionary argument in favour of the existence of altruistic actions.

Let us begin with some conceptual clarifications. The notion of 'psychological' altruism used in this paper should not be confused with 'biological' or 'behavioural altruism', as it is understood in biology or economics. Scholars in these research fields mostly define altruism in terms of *outcomes* on individual fitness (see Hamilton 1964; Sober and Wilson 1998) or well-being (Fehr and Fischbacher 2003), whereas psychological altruism is about the internal *motives* responsible for helping actions (Batson 1991).[1] While the notion of motive is not well defined in the literature, most authors would agree that the set of motives is a broad category that includes different things, such as desires, intentions or judgments. Motives underlie the whole procedure that eventually leads to action. Moreover, it is usually implicitly assumed that motives have an articulated conceptual content – some would prefer to say that they

[1] Sober and Wilson (1998) have popularized the distinction between 'psychological' and 'evolutionary' altruism. In (Clavien 2010b) I argue that the latter should be divided into two distinct categories: 'biological' and 'behavioural' altruism.

C. Clavien (✉)
Department of Ecology and Evolution, University of Lausanne,
UNIL Sorge, Biophore, 1015 Lausanne, Switzerland
e-mail: christine.clavien@unil.ch

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*,
Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_13,

are bound to beliefs – as well as an affective component, a bodily set of sensations that is felt as an urge to do something.[2]

The traditional debate over the possibility of psychological altruism divides participants into two categories: those who defend the possibility of genuine altruistic actions and those who think that all 'apparently altruistic' actions are, in fact, egoistically motivated. To make sense of these formulae, some definitions are needed. A key concept in the debate is the notion of the *primary*[3] – as opposed to the *instrumental* – motive of a helping action (see Sober and Wilson 1998: chap. 6-7; Kavka 1986: 42-44). A primary motive is the *first* motive of a causal chain that leads towards action; it is also the driving force that lasts until the action has come about. If the action is set off by more than one cause – or causal chain – a primary motive must be at least a necessary condition for the action to come about.[4] Here is an example:

> *Raymond seeks pleasure [primary motive] → Raymond thinks that if he does x, he will obtain pleasure [instrumental practical reasoning] → Raymond desires to do x [instrumental motive in order to achieve pleasure] → Raymond does x*

Motives are distinguishable in terms of their objects. If a primary motive is directed towards the needs and well-being of other individuals, it earns the label 'altruistic'. If a primary motive aims at some personal benefit for oneself, it is considered 'self-interested'.

'Psychological altruism' (PA) is the view according to which *at least some* actions are motivated by altruistic primary motives (Butler 1991 [1726]; Hutcheson 2004 [1725]; Nagel 1970; Smith 2002 [1759]; Batson 1991; Sober and Wilson 1998). On the contrary, 'psychological egoism' (PE) denies the possibility of primary altruistic motives (Hobbes 2005 [1651]; Mandeville 1997 [1714-1728]; Cabanac et al. 2002; Cialdini et al. 1987; Ghiselin 1974; Andreoni 1990; Rand 1964). According to this latter view, all human actions are motivated by the expectation of some personal benefit, usually conceived of in terms of pleasure and avoidance of pain – the hedonistic version – or such things as power, resources, or reputation.

It is worth noting that PE does not deny that actions motivated by self-interested motives can have positive effects for others. It is possible to seek one's own happiness without endangering others' well-being. PE does not deny the reality of non self-interested motives either, provided these motives are mediate objects of a primary self-directed motive. PE allows for sincere desires to help a person in need, but these

---

[2] However, it must be noted that, in this debate, authors are rarely clear about the particular features of motives. They simply take them as causal factors that lead toward action. Most of the time, motives are conceived as causally efficacious desires.

[3] In the philosophical literature, primary motives are usually called 'ultimate motives'. However, in order to avoid confusion with the notion of "ultimate cause" as described in biology (see footnote 29), I will avoid this formula here.

[4] A complementary way to capture the distinction between *primary* and *instrumental* motive is to think of their ends, of what they aim at. A primary motive is directed towards an *end in itself* whereas an instrumental motive – which is situated in the centre of a motivational causal chain – is directed towards an *intermediate end*, an end that is supposed to help in reaching the ultimate end of the primary motive.

desires can only be instrumental; they must be considered the best way to achieve a personal good – for example a fine reputation. In other words, others' well-being can be a mediate but not a primary object of one's motives.

Here, we can see that PE is a universal thesis about human motivation. It denies the reality of non self-interested primary motives; everything must be explained in terms of self-interest – for example, desire for applause, honour, pleasure, avoidance of pain. This universal aspect of PE makes it a very demanding claim because it is incompatible with *any* occurrence of a primary motive aiming at something other than one's own well-being; it must rule out any primary desire to help others, or to act in accordance with moral duties, even desires for self-destruction.

13.2 In Favour of Psychological Altruism

As such, PE is not a normative thesis; it does not take a stand on moral issues. However, when it is combined with the fairly widely-held thesis that an action is morally good only if it is caused by non self-interested motives, one cannot escape the conclusion that there is no morally good action. This is why PA is usually favoured over PE. It should be noted however, that, despite providing a good reason to dislike PE, these moral considerations offer no knockdown argument against PE. Firstly, one could question the very idea of describing moral action in terms of other-directed motives. Secondly, even if we accept this definition of moral action, PE might force us to admit that morality is only a matter of illusion. More is needed in order to convincingly reject PE. In this section, I will briefly present a representative panel of arguments against this view. However, I will not elaborate on these arguments, my purpose being to give a glimpse of the sort of objections that can be made against PE, before showing in the next section that all these objections can be rejected by a single argumentative line.

We have seen that PE is a variety of motivational monism: it claims that all of one's motives are of the same sort, namely self-interested. This demanding aspect of the theory has led most advocates of PA to search for counterexamples, particular actions or types of action that cannot be convincingly explained in egoistic terms. Indeed, PE could be proven false by showing that *at least* one action has been performed that was motivated by a non self-interested primary motive. In a thought experiment, Francis Hutcheson (2004 [1725]: treatise II, section II) aimed to provide one particular example of a helping action that could not be explained in egoistic terms. His story is as follows: imagine God told you that you were going to be annihilated in a few seconds, but that you had a last choice to make; you could choose to make your families, friends and humanity in general either happy or miserable in the future. However, you would not be able to feel any pleasure or pain as a consequence of your choice. Under these circumstances, he argues, many of us would choose the first option, that is, to make others happy. Such a choice cannot be explained by self-interested motives. Therefore, he concludes, PE is false – a critical response to the arguments summarized in this section follows in the next section.

Hutcheson's thought experiment was intended to provide *one particular example* of an action caused by an altruistic motive. In the literature, one can also find more general arguments, such as the attempt to show that some *types* of behaviours cannot be explained in egoistic terms. Take the 'argument from moral approbation' which is also to be found in Hutcheson's writings (2004 [1725]: treatise II, sections II & IV). According to him, when a person's actions have good effects, but she is merely motivated by self-interested desires, we do not morally approve of these actions. In contrast, we morally approve of some actions precisely because they do not seem self-interested. Since Hutcheson is convinced that we cannot always be completely wrong in our moral judgements,[5] he concludes that there must be some actions that are not motivated by self-interested desires.

Finally, more general arguments have been proposed, which do not focus on particular actions or types of behaviour. For example, Joseph Butler famously defended the following line of argument (Butler 1991 [1726]: § 415).[6] According to him, a preliminary condition for experiencing pleasure is to have a desire oriented towards an *external* object. For example, before obtaining pleasure from eating a piece of cake, one must first conceive a desire for a piece of cake – which is an external thing. Since pleasure can *only* emerge as an epiphenomenon of actions caused by desires for external things, it does not depend upon a desire for an internal state, such as the self-directed desire to experience pleasure. Indeed, pleasure from eating a piece of cake does not come from the desire to experience pleasure, but from the conjunction of a desire for a piece of cake *and* the satisfaction of this desire. Butler concludes that PE is not a valuable thesis and should be rejected.

Besides the thought experiments and formal arguments typically used by philosophers, psychologists have tried to prove the existence of actual cases where agents have no interest at all in helping others, but still choose to do so. For example, Daniel Batson and his colleagues have conducted a series of empirical studies designed to show that high levels of empathy – understood as feeling sympathetic, compassionate, warm or soft-hearted towards others – cause people to help others, even when – they argue – one cannot conceive of an egoistic interpretation for this sort of helping behaviour. For example, it was shown that many subjects are ready to endure electric shocks in place of another person towards whom they feel an empathic emotion, although they are given an easy opportunity to escape (Batson 1981). Other studies control for alternative egoistic interpretations such as reputation (Fultz *et al.* 1986) or the warm glow of the caretaker (Batson 1991).[7]

Plenty of other arguments and studies are to be found in the literature. It is not my purpose to discuss them all. I only intended to propose a brief review of the sort

[5] Hutcheson strongly believes in the existence of a moral sense.

[6] "That all particular appetites and passions are towards *external things themselves*, distinct from the *pleasure arising from them*, is manifested from hence; that there could not be this pleasure, were it not for that prior suitableness between the object and the passion." (Butler 1991 [1726]: 365)

[7] For more on theses studies see Batson (1991), Sober & Wilson (1998), Stich *et al*. (2010)

of objections that can be made against PE in order to capture the efficacy of the unique type of counterargument to PA to which we now come.

13.3 In Favour of Psychological Egoism

Sober and Wilson (1998) have famously argued that philosophical arguments and empirical data stemming from social psychology cannot prove PA because an "internal reward" explanation can always be invented to explain human action. By this, they mean that we cannot rely on introspection to identify our primary motives. Implicitly, they state a fairly well-known argument, according to which it is always possible to be mistaken about our true motives; any apparently altruistic motive could be caused by an unconscious selfish one, such as the avoidance of painful memories or the attainment of a warm feeling of self-satisfaction.

Let us briefly return to each argument presented in the previous section and see how they can be rejected with help of the notion of the unconscious.

Hutcheson's thought experiment does not allow for the fact that subjects may not be completely persuaded of the impossibility of being rewarded for their action. They might expect to collect 'good marks' for their afterlife. The point here is that introspection can be deceptive; we can be mistaken about our own motives.

The argument from moral approbation can be rejected on similar grounds. It is possible that humans who morally approve of moral actions are systematically mistaken about the true motives that have led to these actions, as well as what makes them approve of these actions. Any apparently altruistic action can – consciously or unconsciously – be caused by a self-directed motive; similarly, we can also be mistaken about what grounds our moral approbations.[8]

The unconscious motivation argument even proves powerful against Butler's attack on PE;[9] nothing precludes the possibility that the motives that lead us to seek consciously for *x* – where *x* is not pleasure, but an external thing that can elicit pleasure once obtained – are unconsciously self-directed.

In brief, a defender of psychological egoism could accept the fact that many actions do not seem self-directed, yet maintain his position on the grounds that conscious motives might be deceptive. In the case of apparently altruistic actions, the following causal chains would hold:

> *Primary self-directed motive (conscious or not) → Instrumental practical reasoning (conscious or not) → Instrumental motive directed towards other's well-being → Action [→ If the action obtains, pleasure]*

[8] To learn more on how much we fail to understand about ourselves, see Timothy Wilson's evocative book (2002)

[9] Note that Butler's assertion that pleasure can *only* emerge as an epiphenomenon of actions caused by desires for external things is controversial in itself. As Sober and Wilson argue, "satisfying the desire for an external thing is one way, among others, in which people obtain pleasure." (1998: 279)

More precisely, depending on the circumstances, the primary self-directed motive could involve two possible scenarios: the subject finds himself in an uncomfortable state, for example he feels bad at the sight of somebody suffering, and this situation motivates him – consciously or not – to rid himself of this state; alternatively, the subject anticipates – consciously or not – the fact that a particular action might be good for him, for example, helping a needy person will give him a pleasant feeling of self-satisfaction, and so he finds himself motivated to perform this action.

At the empirical level, even if it can be shown that empathy causes helping behaviour, it might not be as easy as Batson and colleagues think to avoid all possible egoistic interpretations of their experimental results. In fact, most of their data can be interpreted in the following egoistic terms: empathising with a needy person might create either a kind of sadness or a fear of feeling guilty afterwards that subjects know can only be successfully assuaged by helping that person (Hoffman 1991; Hornstein 1991). Moreover, there is always the possibility of questioning the design of particular experiments or demanding further corroboratory results stemming from independent research groups.[10]

Many philosophers question the plausibility of these 'internal reward scenarios' on the grounds that many are contrived and counterintuitive. Since, they argue, PE is intuitively less convincing than PA, the burden of proof lies with the proponents of PE. Sticking firmly to the logical possibility of an internal reward explanation reveals that PE is a dogma rather than an explanatory theory.

PA certainly has a slight advantage here, but it does not seem sufficient to settle the debate. Such a line of argumentation will only convince readers who already accept PA. Despite the efforts of PA's advocates, one should not overlook the fact that PE keeps cropping up, especially among psychologists (Cabanac *et al.* 2002; Cialdini *et al.* 1987) and economists (Andreoni 1990; Rand 1964).[11] Ghiselin's often quoted "scratch an 'altruist' and watch a 'hypocrite' bleed" (1974: 247) remains an evocative formula. Economic reasoning in terms of individual preferences ranked in a utility function renders the following reflection appealing to many readers:

> Whenever a man systematically (i.e., as a general rule) continues to sacrifice primary reward *x* to other people, he does so only because he usually obtains thereby some primary reward *y and* because *y* ranks higher than *x* on the person's preference scorecard, as determined in situations where no considerations of other people's interests and thus of sacrifice to other people's interests were involved. (Slote 1964: 533)

It is true that advocates of PE mostly stick to case-by-case argumentation and to the impossibility of *proving* that their interpretation is wrong. However, they can also claim that their hypothesis awaits proof, much in the same way that a theory of the illusion of colour perception awaits proof.

Moreover, there might be more in favour of egoism than is usually thought. As we shall now see, empirical results that, at first glance, seem to provide evidence

[10] For a more extensive discussion on this topic, see Sober and Wilson (1998: chap. 8; 2000) and Stich *et al*. (2010)

[11] On this topic, see Macpherson (1962).

in support of PA, add in fact more credence to PE when properly interpreted. A wide range of studies in experimental economics have tested people's pro-social versus self-regarding propensities to act. Human subjects were asked to play social dilemma games with each other via anonymous computer platforms; games such as the dictator game,[12] the trust game,[13] or the public goods game[14] were extensively used. These studies show that people are often ready to invest their money for the sake of the common good (Marwell and Ames 1981; Fischbacher et al. 2001; Ostrom 1990; Fehr and Rockenbach 2003; Henrich 2004) or in pro-social moves, even when they know that it is at their own expense and that they cannot gain anything in return (Fehr and Gächter 2002; Fehr and Fischbacher 2004a, b; Charness and Gneezy 2008; Hoffman et al. 1996).[15] At first glance, one might think that this empirical evidence could be used in favour of PA. However, it is important to distinguish between people's behaviour and their motives. Worries concerning subjects real motives are nicely illustrated in the 'nobody's watching' experiment conducted by Haley and Fessler (2005). The experimenters showed that very subtle cues can have a drastic impact on cooperative and pro-social behaviour. For example, when simple stylised eyespots are placed on the computer's desktop background while subjects are playing a dictator game, they cause a dramatic increase in pro-social behaviour: dictators were much more generous while 'being watched'. These eyespots can best be interpreted as cues relating to the presence of observers, thereby as elicitors of psychological mechanisms linked to reputation. This experiment suggests that people think of their own benefit even under the usual condition of anonymity.[16] In the light of these results, one cannot help thinking that many economics experiments might not have been sufficiently carefully designed to avoid similar cues.

[12] The dictator game is a two-person game in which the first player receives an amount of money and is asked to divide it between himself and the second player. The other player cannot reject the split proposed.

[13] In a trust game, two players receive the same amount of money. The first player is asked to decide how much of his money to pass on to the second player – the trustee. All money passed is increased by a multiplication factor of two to four – depending on the game. The trustee then decides how much of this to return to the first player. She is allowed to keep most or all the money for herself, in which case she would demonstrate free-riding behaviour.

[14] In a public goods game, all participants are free to contribute to a group project and once the group project is realized, every member of the group can benefit from it, even those who did not contribute.

[15] For more information about these experiments, see Klein (this volume) and Clavien and Klein (2010).

[16] Other studies speak in favour of this hypothesis. For example, it has been shown that experimental settings that more closely resemble everyday life – anonymity condition weakened or absent; more information provided about the other players – positively affect the dictator's generosity (Hoffman et al. 1996; Charness and Gneezy 2008). Therefore, there is reason to suspect the interference of other internal self-directed motives, such as guilt aversion, the unpleasant feeling resulting from acting too selfishly (Charness and Dufwenberg 2006), or aversion to disappointing the second player (Dufwenberg and Gneezy 2000; Koch and Normann 2008; Dana et al. 2006), or an expectation of feeling a warm glow, the pleasant feeling associated with the thought of oneself as a caretaker (Andreoni 1990; Eckel *et al.* 2005).

Other empirical findings that seem to capture some elements linked to the selfish unconscious are to be found in studies using brain-imaging as a research tool. In a study by Rilling and colleagues (2002), subjects' brains were scanned while they played an iterated prisoner's dilemma game, which is about choosing whether to cooperate or to defect.[17] It was shown that the choice to cooperate activated brain areas in the player that are linked with reward processing – including the 'caudate nucleus', well known to be associated with the *anticipation* of reward. According to the experimenters, activation of these parts of the brain positively reinforces reciprocity and helps to resist the temptation to defect (similar results have been obtained by King-Casas *et al.* 2005).[18]

Even if these studies do not directly address the question whether altruistic motivation exists, they indicate that behaviours that seemed to be good candidates for altruistic explanation are, in fact, best explained in terms of self-interest.

13.4 The Deadlock

We have seen that in order to respond to powerful arguments in favour of PA, PE needs to resort to the unconscious. This move is interesting for a defender of the latter view precisely because the unconscious cannot easily be sounded out. Therefore PE is not easily refutable. One can always appeal to an unconscious desire for internal rewards as an explanation for apparently altruistic actions. However, the unconscious has its drawbacks. It is a double-edged sword for the supporter of PE because, if the unconscious cannot be sounded out, there is no reason for favouring egoism over altruism! In the end, it seems that this line of reasoning is of no use to either a defender of PE or an advocate of PA, precisely because it destroys any means of settling the dispute between them.

There is some hope of overcoming this deadlock with experimental data, more specifically, with the help of the new brain imaging technology that is used extensively in young research fields, such as neuropsychology and neuroeconomics. In this respect, the aforementioned experiments seem to be of particular interest. Unfortunately, there are serious doubts about the real contribution of these studies to the particular philosophical debate over altruism. To begin with, the fact that people are highly sensitive to reputation cues (Haley and Fessler 2005) does not preclude the possibility that there is altruism in the absence of these cues. As for

[17] This two person game is considered a 'dilemma' because its payoff matrix is set up in such a manner that whatever the other player chooses to do – either cooperate or defect – it is always better to defect. However, the outcome obtained for each player is worse if both players defect than if both cooperate.

[18] In the same line, see also Moll et al. (2006). They found activation of the same reward-related brain area (the ventral striatum) both when subjects received and gave money. It is not clear however to what extent these phenomena are causally efficacious or merely side effects. I will come back to this difficulty in the next section.

Rilling and colleagues' prisoner dilemma study (2002), the experimental design is not fine-grained enough to discriminate between two concurrent interpretations: i) On an interpretation favourable to PE, the activation of the brain areas linked with reward processing represents both the anticipation of future reward and the direct cause of cooperation; ii) On an alternative interpretation favourable to PA, the activation of these brain areas is mainly a side effect of cooperation; even if there is some anticipation of reward, it is likely to be a minor motivating factor among other altruistic and more decisive factors. The feeling of reward experienced by the subject is hence mainly a side-effect of altruistic actions.[19]

It is an open question whether brain-imaging studies could bring novel and crucial arguments to the altruism versus egoism debate. In principle, it should be possible, proper technology and well-designed experiments permitting. However, the current state of knowledge about the neural systems involved in motivation does not allow for this level of subtlety. The relationship between an observed behaviour and specific brain activation is difficult to spell out; correlated events might not be directly causally related.[20] As with classical psychological experiments, we are faced with the difficulty of interpreting the results, and the challenge of modelling situations in which we can determine with near certainty whether the subjects think – unconsciously or not – of a possible advantage for themselves or whether they are truly interested in others' well-being.

Overall, at this stage of research, one has the impression that the debate over altruism cancels itself out in a battle of a priori statements. In what follows, I will try to show that there are replies to PE, but that in order to give them real force, we need to reframe the debate: instead of focusing on primary *motives*, I suggest concentrating on the more fundamental notion of *motivation*. As we shall see, such a reframing will make refutation of PE easier to obtain.

13.5 Two Ways of Conceiving the Motivational Causal Chain

Besides the deadlock just mentioned, there is another puzzling fact about the altruism versus egoism debate. Until this point, the causal chain underlying our choices of action has been explained in terms of primary and instrumental motives. Figure 13.1 depicts this view. The arrows describe the possible causal paths that lead a subject from the perception of a situation to the action.

However, the classical way of defining the altruism versus egoism debate may prove too superficial, as soon as one tries to grasp the starting point of the

[19] This hypothesis finds some support in recent neuroeconomic results: Harbaugh *et al.* (2007) observed activation of reward-related areas (the head of the caudate and the nucleus accumbens) when subjects merely observed a charity receiving money.

[20] On the difficulties linked to interpreting brain imaging results see Poldrack (2006), Henson (2006), Vul *et al.* (2009).

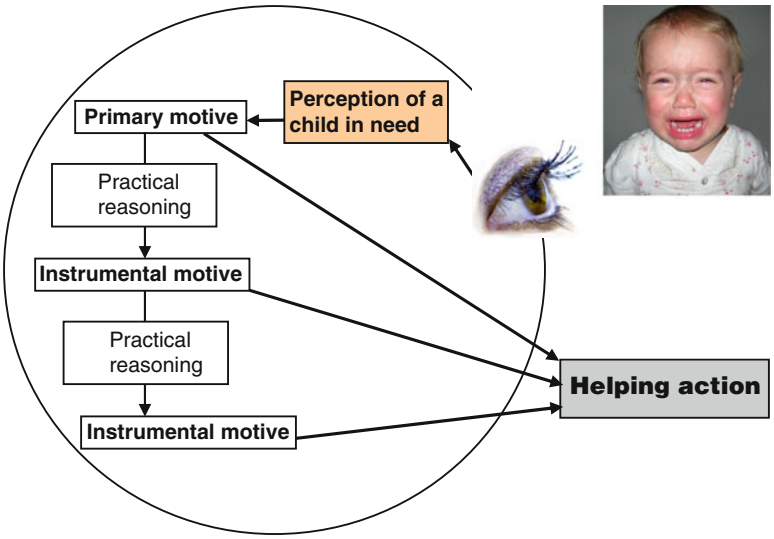

Fig. 13.1 The arrows describe the usual way of conceiving the causal routes of internal events starting from the perception of a person in need and ending with a helping action. In a simple scenario, the ultimate motive directly elicits the action, whereas in more complex scenarios, the causal path can include one or more instrumental motives

motivational process. Indeed, a motive is supposed to be the source of motivation. However, although usually assumed, it is not clear whether motivation always starts with a motive, that is, with a desire, an intention or a judgement. An analysis of the notion of motivation is needed here. Despite being widely used, the term 'motivation' is hardly ever defined in the philosophical or scientific literature. This prompted Ronald de Sousa to write in an article on emotion that the motivational aspect of emotion "is infected by the obscurity of the notion of motivation" (de Sousa 2005: 65).

Motivation might be conceived of as a *relational property*: 'D is motivated by *x* to do *y*'; this relation holds between *x* and an action and takes a direction from *x* to *y*. In the context of the altruism debate, *x* is usually understood as a motive such as a desire, an intention or a judgment.[21] However, *x* might also be an emotion. For example, being afraid often leads to avoidance and fleeing acts: Charles can be motivated by his fear of the neighbour's dog to take another path to get back home after his morning jogging. Hence, it seems that the relational property of 'being motivated' can stretch its arms beyond motives.

Alternatively, there is a more substantial understanding of motivation. One can point to the experiential aspect of motivation, which consists in the experience of being moved to do something. The dynamics of this experience implies that motivation does not

[21] A motivating judgement is typically considered to be the result of deliberation that is linked to an internalised norm or principle.

simply refer to an abstract relational property but to 'something' that makes one move. The most sensible way to make sense of this 'something' is to consider it an *affect*, a bodily set of sensations that incites the subject to act.[22] However, to count as motivating, this affect needs to be embedded in an intentional – in the sense of 'directed towards an object' – psychological state, such as an emotion, a desire, or possibly an affectively-laden judgment.[23] This dynamic account of motivation helps to put flesh on the bones of what is often referred to as the "motivational aspect of" emotions or desires.

The dynamic account is particularly interesting because it reveals that motivation to act might not – at least, not always – come from decisions based in the will, as is often assumed. Consider a situation in which Denise is deeply touched by seeing a starving child. It seems fair to say that it is the affective part of Denise's compassion that incites her to consider various possible helping actions, such as taking the child to her home, or giving money to his parents. The affective arousal – thus motivation – will cool down once Denise has realised one of these helping actions. Of course Denise's situation might be re-described in terms of desires or judgements. For example, one could say that she wants to help the child because she thinks it deserves a better fate. However, this description would not catch the fundamental source of motivation which seems to be the affective reaction itself; the desire to help the child is a direct product of Denise's compassion rather than the *primary* cause of her helping action. Motivation is present during the whole process: it starts with an emotional reaction and is carried over from this basic state of mind to more complex states of mind such as the conscious desire to help the child.

The distinction I propose to draw between motive and motivation is useful in two respects. Firstly, it provides an alternative explanation for the causal relationship between the first input – a person in need – and the final output – the helping behaviour. According to this interpretation, motivation is not necessarily accompanied by full-fledged desires or intentions. For example, when motivation is embedded in a basic emotional reaction such as an empathic feeling, there might be no articulated conceptual content of the sort necessary for a desire or an intention. Emotions can be primitive fast and frugal reactions towards our environment.[24] If we add some articulated conceptual contents to motivation, more complex states of minds such as desires, intentions or judgements can occur. Moreover, there is no need to draw a clear line between conscious and unconscious motivation. People often become gradually aware of their urges once they start to build cognitively around their affective reactions. Figure 13.2 depicts the possible causal paths that lead a subject from the perception of a situation to the action via an affective reaction. This picture can integrate classical

[22] I argue in favour of this idea in my (Clavien 2010a).

[23] Motivation should not be confused with a purely causal mechanism such as a reflex. Moreover, it is important to understand that there is no dichotomy between motive and motivation; motivation can be embedded in motives – such as desires – even though not all motivating states of mind are motives – i.e. emotions.

[24] For a detailed account of what an emotion is, see Robinson (2005).

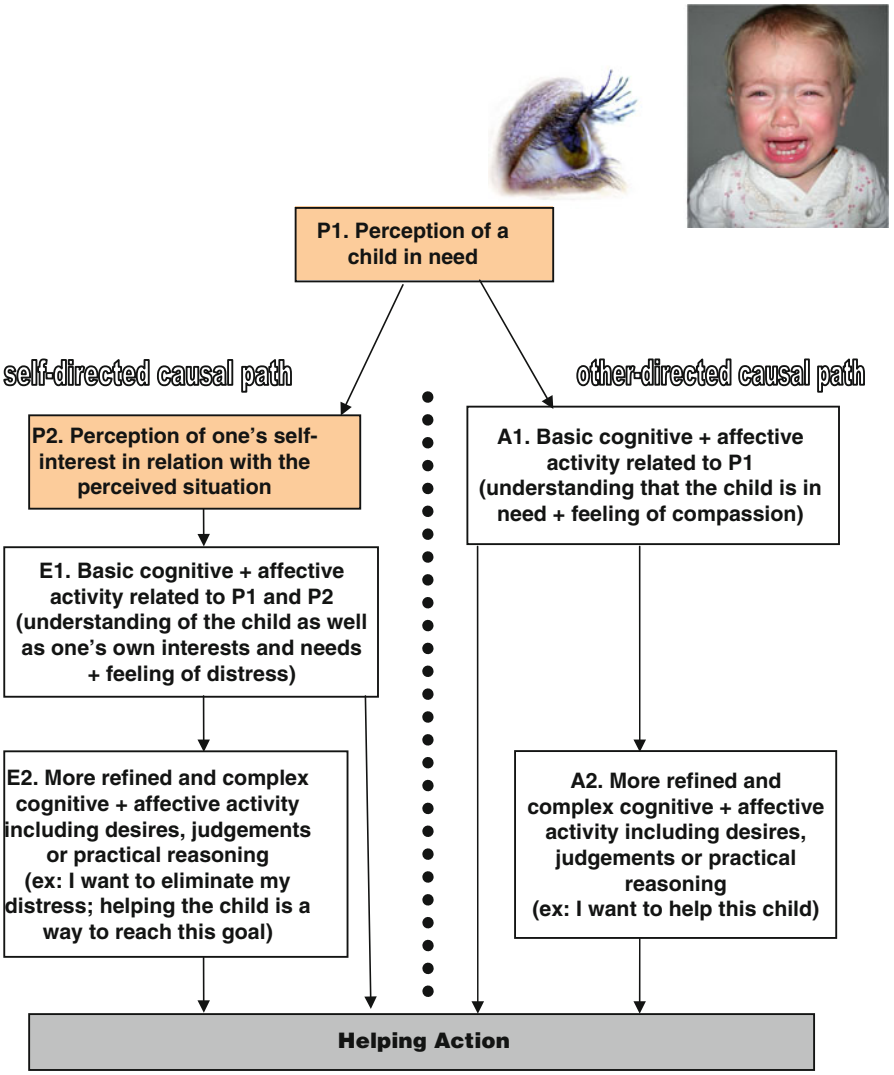

Fig. 13.2 The arrows describe two possible causal routes of internal events starting from the perception of a person in need and ending with a helping action. On the right side of the dotted line, the causal pathway is altruistic because the subject does not take his own interests and well-being into consideration; on the left side, the causal chain is self-directed because the perception of a person in need is associated with thoughts directed towards the subject's own interests and well-being

notions such as desires, judgements,[25] intentions or practical reasoning – boxes E2 and A2 – without being too explicit about the ways they are involved. These elements can appear further down a causal path, or maybe not, depending on the situation described. More explanations of this schema will be given in the next section.

Secondly, the distinction between motive and motivation is of importance because, as we shall see in the next section, if one shifts one's attention from motives to motivation, the altruism versus egoism debate can helpfully be reframed.

13.6 A Proposal to Reframe the Debate

As we have seen, the classical debate over altruism focuses on motives – usually conceived in terms of intentions and desires – and reaches a deadlock once the unconscious argument comes into play. To resolve this deadlock, my proposal is to make use of the aforementioned distinction between motive and motivation. The strategy I propose is a shift of focus from motive to motivation. I take it to be legitimate to reframe the debate in this way for two main reasons. Firstly, we have seen that motives are not necessarily the primary cause of our actions; the causal chain goes back to the source of motivation, which seems to be an affective reaction. If motives are not – at least, not always – the original motivating source, it seems more interesting to focus the debate over altruism on the possibility of altruistic *motivation*, rather than altruistic *motives*. Secondly, the motivational component seems to have causal priority over the motive. The controversy is over whether any human action can be called altruistic. To resolve this controversy, one considers how actions are brought about. What is to be found at the beginning of the causal chain of action is often – if not always – a basic affective state, rather than a motive.

The reframed debate over altruism versus egoism would then focus on the question of whether truly altruistic motivation can exist. By this, I mean the question whether there exist motivational causal pathways triggered by the awareness of others' needs and well-being, which do not include considerations of one's own self-interest. Such a causal route would start with a basic affective state and follow the "other-directed path" – right side of the dotted line in Fig. 13.2.

Now, one very interesting aspect of the proposed shift of perspective is that it allows emotions to enter into the analysis. Most – if not all – primary affective motivating elements are embedded in emotions.[26] Therefore, an easy strategy for an

[25] Some readers might want to consider emotions to be a form of judgement – in the sense of appraisal. This could easily be integrated into my picture.

[26] Although I will not argue for it here, it seems that most of our actions – possibly all – originate from emotional motivation, which amounts, more or less, to Hume's famous position. According to him, "it appears evident that the ultimate ends of human actions can never, in any case, be accounted for by reason, but recommend themselves entirely to the sentiments and affections of mankind" (Hume 1751: Appendix I). This is not to deny that emotional motivation can be monitored and affected by conscious deliberation.

advocate of altruism would be to demonstrate the existence of 'altruistic emotions', capable of leading someone to act *without the intervention of any further motivating factor*. This sort of emotion would need to be directly elicited by the perception of another's needs and well-being and would diminish once the other's needs and well-being had been satisfied. In other words, the debate over altruism can be thought of in terms of the two following questions: Are there altruistic emotions? Is the affective component of these emotions sufficiently strong to bring about action?

Before responding to these questions, let us elaborate a bit more on the two ways emotional reactions can motivate one to act altruistically. Consider the example of parental care. Human beings are naturally inclined to feel caring emotions towards their children. Usually, when a parent sees his child in need, a caring emotional reaction is elicited. The occurrence of this emotional reaction provides the first general instructions regarding the direction of the action that has to be taken. These general instructions can be followed in two ways.

In particular circumstances, the emotional reaction leads directly to a helping action. This direct motivating path is depicted in Fig. 13.2 with the arrow from box A1 to the action. Here, no particular desire or practical reasoning is needed in addition to the emotion in order to move the subject to act. In this case, one can speak of 'actions out of emotions' (see Döring 2003). For example, if a mother suddenly sees that her child is in great danger – say, being attacked by a wild lion – she might act spontaneously out of a caring emotion without forming any particular desire.

In most cases, however, the mental activity prior to action is more complicated. Emotional reactions can lead the subject to form complementary motives before acting – causal path A1-A2-Action. Recall Denise's example. Under emotional impulse, she builds cognitively both on her emotion and on her understanding of the child's critical condition. This mental activity leads her to form a proper motive such as a desire, an intention or a judgement which contains a more articulated conceptual content. There are also situations in which mental activity becomes even more highly complex; the agent might take time to employ practical reasoning before deciding to act.

13.7 The Existence of Altruistic Emotions

Let us now come back to the question of whether there are altruistic emotions. At first glance, it seems that the question can quickly be settled. Who would deny the existence of emotions such as love, sympathy or compassion? It would be ridiculous to deny, for example, that human beings are naturally inclined to feel caring emotions towards their children. These emotions are clearly caused by the perception of others' needs and well-being.

Nevertheless, an advocate of PE might raise two doubts about the altruistic character of these emotions, which would amount to denying the causal links represented on the right side in Fig. 13.2.

Firstly, the supporter of PE could contend that, when examined more closely, the apparently altruistic emotional systems prove to be self-directed – which

amounts to saying that box A1 should be placed on the left side of the dotted line. Compassion, for example, could be described as a feeling of uneasiness that motivates the agent to engage in actions that will eradicate this feeling. According to such an interpretation, motivation comes from the 'uneasiness' generated in compassion; the helping action is only performed because it enables the subject to rid himself of this uneasiness. On this account, it makes no sense to consider compassion an altruistic emotion.

There is a conceptual problem with this argument. It distorts the very notion of 'self-directed emotion' by focusing on the phenomenological aspect of the emotion, rather than on its eliciting cause. However, the fact that compassion has a phenomenology of 'uneasiness', which vanishes once the input changes, does not make this emotion self-directed. One should not overlook the important fact that, by definition, what makes a motivational system 'altruistic' is the way it has been elicited and is maintained, whatever the physical processes and endocrine systems involved in the course of the motivational process.[27] Indeed, the only sensible way to speak of altruistic emotion is to say that it is an emotion triggered by an understanding of others' needs or well-being. Compassion clearly meets this criterion. Another way to put it would be to say that what makes a motivational system 'altruistic' is the type of object towards which it is directed. For example, an emotion exclusively directed towards somebody else's needs can be considered altruistic.[28]

This leads us to the second objection, which questions the importance of the other-directed component of apparently altruistic emotions. Here the picture becomes a little more complicated. According to such a view, *apparently* altruistic emotions are in fact triggered by a combination of other-regarding perceptions reliably associated with self-regarding perceptions. Furthermore, only the latter are necessary ingredients for emotions to occur, therefore, only the latter ground motivation. Denise, for example, would only start to feel compassion once she had understood that the child was in need for help – box P1 – *and* that this situation was not advantageous for her – box P2. Both perceptions are needed for compassion – box E1 – to arise.

It is worth noting that these perceptions need neither be conscious nor conceptually well articulated: they can be simple apprehensions of relevant aspects of the situation observed. There is no need for inferential reasoning here; a simple mechanistic association of thoughts can trigger an emotion. Of course, if self-directed perceptions are needed for emotions to occur, there cannot be altruistic emotions – at least not of the sort wanted by an advocate of PA.

Moreover, another interesting aspect of such an account is that it is not always necessary to postulate the existence of self-directed *motives* – box E2 – in order to

[27] Similarly, it would be too trivial to reduce PE to the claim that agents are moved by their own motivation – and not that of others. As Kavka notes, such a position would amount to an uninteresting truism (Kavka 1986: 35).

[28] One might argue that, from an evolutionary point of view, uneasiness is the more reasonable ultimate cause of performing helping actions. However, this line of reasoning confuses ultimate with proximate causes (see footnote 29). It is important to keep in mind that the debate around psychological altruism is about proximate causes.

explain an action in terms of self-interest. Consider the case of the mother who sees her child endangered by a lion. To make her act out of a caring emotion, it is sufficient for her to have two preliminary perceptions: 'my child is in danger' and 'my child being in danger is not good for me' – causal path P1-P2-E1-Action.

To recapitulate, according to PE, any emotion – including caring emotions – can only be elicited once the subject has taken her personal interests into consideration. More particularly, the necessary eliciting ingredients for *apparently* altruistic emotions are: a real situation in which an individual is in need, a corresponding perception about that individual and an additional self-directed perception of the sort 'this situation is against my interests'. Without the additional perception, a caring emotion simply cannot be experienced, which makes altruistic emotions impossible.

To respond to this 'egoistic' view, one can refer to Sober and Wilson. According to these authors, the only way to ground psychological altruism is to use an evolutionary line of argument. Their strategy is to focus on the evolutionary *proximate mechanisms* that cause apparently altruistic behaviours.[29]

They argue that there are good evolutionary reasons to think that highly social actions, such as human parental care, are set off by other-regarding proximate mechanisms instead of self-directed mechanisms. This assertion is based on what can be named the 'reliability argument' (Sober and Wilson 1998: chap. 10).

A preliminary remark is needed here. Sober and Wilson's original argument was formulated in the context of considering the possibility of primary altruistic *desires*. So, one might think that it is not relevant in a reframed context that focuses on emotions instead of motives. However, when one looks more closely at the details of their argument, it appears that the notion of desire does not play a significant role after all. This will allow me to reformulate the reliability argument in a discussion about emotions.

In fact, I am convinced that Wilson and Sober's argument has even greater relevance in the reframed context proposed here. Their evolutionary argumentative strategy based on desires was not received with much enthusiasm and has encountered numerous objections (Brunero 2002; Stich 2007; Jamieson 2002b; Rottschaefer 2002). The scepticism of these readers stems partly from the fact that, in focusing on articulated desires, Sober and Wilson overlook other possible proximate mechanisms responsible for caring behaviour, such as simple encapsulated input-output systems. As Dale Jamieson points out, even non-psychological mechanisms could do the job: "Parental care behaviour is widely dispersed across species, and it is likely that it occurs in many organisms that are not minded at all" (2002a: 703). This said, the sort of mechanisms producing other-directed behaviour that have evolved in a social species, capable of feelings and complex mental activity, are very likely

[29] Proximate mechanisms are the direct causal mechanisms underlying a behaviour. They are to be distinguished from ultimate causes which refer to the fitness consequences of a behaviour in the evolutionary past; the latter are the ones with which biologists are usually concerned. Both causes are complementary; they are two causal aspects that help understanding the occurrence of a behaviour (for this distinction see Mayr 1961).

to be *psychological mechanisms*. The real question is what sort of psychological mechanisms they are[30] and what level of complexity they can reach.

We have seen that at least some 'apparently' altruistic actions seem to be mediated by emotional systems. Let us take for granted that such systems exist and are the results of evolution; basic parental love is a typical example of an adaptive emotional system (Lazarus and Lazarus 1994). It is fairly easy to understand that the biological function of a caring behaviour is to enhance the number of fit offspring who survive to adulthood. In an environment where competition is intense and resources are scarce or difficult to reach, parents need to develop capacities to respond quickly to the necessities of their offspring. In a species capable of feelings and minimal cognition, a quick motivational proximate mechanism such as parental love – or a set of caring emotions – is an excellent response. We now need to ask whether it makes sense to expect that this system has evolved in a self-directed form rather than an other-directed form. Here, we have two competing emotional mechanisms, an altruistic and a self-directed one, and the question remains which of them is responsible for the occurrence of caring behaviour whenever a child is in need.

In principle, it is possible that both mechanisms have evolved; evolution does not always exclude redundancy. However, if two motivational mechanisms are both capable of generating the same type of behaviour, one of them might be more likely than the other to be selected. There are good reasons to think that this is precisely what has happened in the present case. As Sober and Wilson point out, one important selection criterion is the *reliability* of a system: among various mechanisms, the most reliable – that is the one that realises its function with the greatest probability – is much more likely to evolve (Sober and Wilson 1998: 221-223).

Let us compare our two competing emotional mechanisms for reliability. Consider first the self-directed mechanism. Recall that, according to PE, in order for a subject to feel a caring emotion towards his children, three ingredients are needed: the children must be in need of help; the subject must have a corresponding perception of the sort "my children are in need of help"; the subject must have an

[30] Sober and Wilson think that they are primary desires "produced by natural selection" (1998: 303) and the evolutionary explanation they provide is based on simple replicator-based models. However, this idea of altruistic motives as the pure results of evolution is controversial. Of course, from an evolutionary perspective, a type of motive reveals a proximate mechanism, but not all proximate mechanisms can be given convincing evolutionary explanations – at least not with the present state of scientific knowledge. It might be possible to provide a more differentiated evolutionary explanation of motives by resorting to complex models of cultural evolution and the Baldwin effect (Ananth 2005). This is surely the most promising way of trying to explain the evolution of fine-grained proximate mechanisms underlying particular types of desires. However, the task is not easy – if not impossible – because of the many intricate parameters that have to be considered. Explanatory complexity often comes at the expense of clarity (see Sterelny and Kitcher 1988). To enter into this complex debate would take us too far afield, but it is worth keeping in mind how difficult it is to reliably account for cultural products such as types of desires or intentions with the mere use of evolutionary tools. It is less controversial to provide evolutionary explanations for basic and easily observable psychological mechanisms, such as simple emotions. To say that these evolved psychological mechanisms have a causal influence on people's motives is uncontroversial. This is the less speculative line of reasoning that I propose to take in the revised 'reliability argument'.

additional self-interested perception of the sort "this situation is not good for my interests". Unless these three conditions are met, the motivational mechanisms will not be put into motion and the subject will not engage in parental care at all – which is not desirable from the evolutionary point of view. In contrast, the altruistic motivational mechanism is much simpler. In order for a subject to be motivated to care for his children, only two ingredients are needed: the children must be in need of help and the subject must have a corresponding perception of the sort "my children are in need of help".[31]

There is evidence that the simplest or more direct of two competing strategies is likely to do a more reliable job than the more complex, indirect one. This remark is especially relevant in the present case because it is not clear at all why the thought association postulated by PE would have occurred in the first place. Moreover, it seems that the process underlying self-interested parental care is quite vulnerable to disruption. If an individual fails to have the self-directed associated perception, the link towards action is broken and he will fail to care for his children. If these cases of imperfect correlation happen regularly – a very plausible hypothesis – natural selection will be likely to opt for the alternative altruistic mechanism.[32]

In brief, the altruistic emotional mechanism seems much more reliable than the self-directed one and is hence more likely to have evolved; in the light of evolutionary considerations, it does not make sense to expect only self-directed emotions to result from natural selection processes.

13.8 The Motivational Power of Altruistic Emotions

We have seen that there are good evolutionary reasons to think that highly social actions, such as human parental care, are set off by purely other-regarding emotional proximate mechanisms, rather than with the help of self-directed emotional mechanisms. Evolution has influenced motivational mechanisms in such a way that parents typically react altruistically towards their children via caring emotions. Since a single counterexample is sufficient to reject PE, PA seems to be the adequate way to explain altruistic action.

However, even if caring emotions are genuinely altruistic, the question remains whether they are strong enough to set off action. Indeed, one might still contend that altruistic emotional motivation always co-occurs with other self-directed motivations and that the latter are stronger.

---

[31] Note that this way of approching the problem saves from thinking in terms of a precise causal steps process that leads towards a helping action. It is sufficient to track the types of ingredients needed for a helping action to come about – however the exact causal chain is realised.

[32] Note that such an imperfect correlation might not prove selectively disadvantageous in all cases. We find in the animal world species whose adults sometimes eat their own offspring – trout, for example. However, such a behaviour would not be selectively adaptive in the case of humans; this is due to the considerable prenatal investment needed for the production of each child.

Again, a simple evolutionary argument enables a response to this objection. Emotions are proximate behavioural mechanisms. Without doubt, some altruistic emotions – such as compassion towards one's children – exist and are adaptive. The evolution of psychological mechanisms, such as basic emotions, is best explained in terms of the behavioural impact of these mechanisms; they have been selected *because* the behavioural propensities they induce are usually beneficial in terms of fitness to the subjects who possess them. Therefore, one can be sure that at least some altruistic emotions are causally efficacious. Simple altruistic emotional mechanisms would not have been selected if they did not have a behavioural impact. Besides, many empirical researchers in behavioural psychology have demonstrated the effect of empathic emotions on behaviour (see for example Batson 1991; Eisenberg 1986; Eisenberg and Fabes 1998). This is enough to provide convincing evidences in favour of psychological altruism.

13.9 Conclusion

I have argued that the altruism versus egoism controversy reaches a deadlock as soon as one makes use of the unconscious argument. In order to loosen this deadlock, I have proposed a shift away from an over-intellectualisation of the proximate motivational mechanisms responsible for altruistic action. Instead, I suggest a move towards an emotional account of altruistic decision-making. In the context of the controversy over altruism, this move proves fruitful because it allows the debate to focus on self-directed versus altruistic emotions. This focus provides firm ground for a defence of PA; evolutionary arguments in favour of the existence of motivating altruistic emotions are sufficient to convincingly argue against PE. This conclusion depends on the acceptance of a shift of perspective from motive to motivation, which leads to a revised motivational causal chain beginning with simple affective reactions such as emotions, rather than motives. Incidentally, I also hope to have shown that Sober and Wilson's reliability argument assumes full relevance in the proposed reframed context.

Acknowledgements are due to two anonymous referees and to the editors for their helpful suggestions and recommendations. I also would like to thank Philip Kitcher who helped me to improve my proposal to reframe the debate, Christian Maurer who drew my attention to Hutcheson's arguments in favour of altruism, and Rebekka Klein with whom I discussed the distinction between motive and motivation. Many thanks as well to Chloë FitzGerald for correction, advice, and comments on previous versions of this paper.

References

Ananth, M. (2005): Psychological altruism vs biological altruism: narrowing the gap with the baldwin effect. *Acta Biotheoretica* 53: 217–239.

Andreoni, J. (1990): Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal* 100: 464–477.

Batson, C.D. (1981): Is empathic emotion a source of altruistic motivation? *Journal of Personality & Social Psychology* 40: 290–302.
Batson, C.D. (1991): The altruism question: Toward a social psychological answer. Hillsdale: Lawrence Erlbaum.
Brunero, J.S. (2002): Evolution, altruism and "internal reward" explanations. *The Philosophical Forum* 33: 413–424.
Butler, J. (1991 [1726]): Fifteen sermons. In *British moralists, 1650-1800: selected and edited with comparative notes and analytical index*, ed. D. D. Raphael, 325-377. vol. Vol. 1. Oxford: Clarendon Press.
Cabanac, M., Guillaume, J., Balasko, M., and Fleury, A. (2002): Pleasure in decision-making situations. *BMC Psychiatry* 2: 7.
Charness, G., and Dufwenberg, M. (2006): Promises and partnership. *Econometrica* 74: 1579–1601.
Charness, G., and Gneezy, U. (2008): What's in a name? Anonymity and social distance in dictator and ultimatum games. *Journal of Economic Behavior & Organization* 68: 29–35.
Cialdini, R.B., Schaller, M., Houlihan, D., Arps, K., Fultz, J., and Beaman, A.L. (1987): Empathy-based helping: Is it selflessly or selfishly motivated? *Journal of Personality and Social Psychology* 52: 749–758.
Clavien, C. (2010a): An affective approach to moral motivation. *Journal of Cognitive Science* 11: 129–160.
Clavien, C. (2010b): Je t'aide moi non plus: biologique, comportemental ou psychologique, l'altruisme dans tous ses états. Paris: Vuibert.
Clavien, C., and Klein, R. (2010): Eager for fairness or for revenge? Psychological altruism in economics. *Economics and Philosophy* 26: 267–290.
Dana, J., Cain, D., and Dawes, R. (2006): What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes* 100: 193–201.
de Sousa, R. (2004): Emotions: What I know, what I'd like to think I know, and what I'd like to think. In *Thinking about feeling: contemporary philosophers on emotions*, ed. R.C. Solomon, 61–75. Oxford; New York: Oxford University Press.
Döring, S.A. (2003): Explaining action by emotion. *The Philosophical Quarterly* 53: 214–230.
Dufwenberg, M., and Gneezy, U. (2000): Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior* 30: 163–182.
Eckel, C.C., Grossman, P.J., and Johnston, R.M. (2005): An experimental test of the crowding out hypothesis. *Journal of Public Economics* 89: 1543–1560.
Eisenberg, N. (1986): *Altruistic emotion, cognition, and behavior*. Hillsdale, New York: Erlbaum.
Eisenberg, N., and Fabes, R.A. (1998): Prosocial development. In *Handbook of Child Psychology: Social, Emotional, and Personality Development*, eds. William Damon, and Nancy Eisenberg, 701–778. New York: Wiley & Sons.
Fehr, E., and Fischbacher, U. (2003): The nature of human altruism. *Nature* 425: 785–791.
Fehr, E., and Fischbacher, U. (2004a): Social norms and human cooperation. *Trends in Cognitive Sciences* 8: 185–190.
Fehr, E., and Fischbacher, U. (2004b): Third-party punishment and social norms. *Evolution and Human Behavior* 25: 63–87.
Fehr, E., and Gächter, S. (2002): Altruistic punishment in humans. *Nature* 415: 137–140.
Fehr, E., and Rockenbach, B. (2003): Detrimental effects of sanctions on human altruism. *Nature* 422: 137–140.
Fischbacher, U., Gächter, S., and Fehr, E. (2001): Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters* 71: 397–404.
Fultz, J., Batson, C.D., Fortenbach, V., McCarthy, P., and Varney, L. (1986): Social evaluation and the empathy-altruism hypothesis. *Journal of Personality & Social Psychology* 50: 761–769.
Ghiselin, M.T. (1974): *The economy of nature and the evolution of sex*. Berkeley: University of California Press.
Haley, K.J., and Fessler, D.M.T. (2005): Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior* 26: 245–256.

Hamilton, W.D. (1964): The genetical evolution of social behaviour. *Journal of Theoretical Biology* 7: 1–52.
Harbaugh, W.T., Mayr, U., and Burghart, D.R. (2007): Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science* 316: 1622–1625.
Henrich, J.P. (2004): *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*. Oxford: Oxford University Press.
Henson, R. (2006): Forward inference using functional neuroimaging: dissociations versus associations. *Trends in Cognitive Sciences* 10: 64–69.
Hobbes, T. (2005 [1651]): *Leviathan*. Peterborough, Ont.: Broadview Press.
Hoffman, E., McCabe, K., and Smith, V. (1996): Social distance and other-regarding behavior in dictator games. *The American Economic Review* 86: 653–660.
Hoffman, M. (1991): Is empathy altruistic? *Psychological Inquiry* 2: 131–133.
Hornstein, H. (1991): Empathic distress and altruism: Still inseparable. *Psychological Inquiry* 2: 133–135.
Hume, D. (1751): *An enquiry concerning the principles of morals*. London: Printed for A. Millar.
Hutcheson, F. (2004 [1725]): *An inquiry into the original of our ideas of beauty and virtue*. Indianapolis, Ind.: Liberty Fund.
Jamieson, D. (2002a): Morality's progress: essays on humans, other animals, and the rest of nature. Oxford, New York: Oxford University Press.
Jamieson, D. (2002b): Sober and Wilson on psychological altruism. *Philosophy and Phenomenological Research* 65: 702–710.
Kavka, G.S. (1986): *Hobbesian moral and political theory*. Studies in moral, political, and legal philosophy. Princeton, N.J.: Princeton University Press.
King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R., and Montague, P.R. (2005): Getting to know you: Reputation and trust in a two-person economic exchange. *Science* 308: 78–83.
Koch, A.K., and Normann, H.-T. (2008): Giving in dictator games: Regard for others or regard by others? *Southern Economic Journal* 75: 223–231.
Lazarus, R.S., and Lazarus, B.N. (1994): *Passion and reason: Making sense of our emotions*. New York: Oxford University Press.
Macpherson, C.B. (1962): The political theory of possessive individualism: Hobbes to Locke. Oxford: Clarendon Press.
Mandeville, B. (1997 [1714-1728]): *The fable of the bees: and other writings*. Indianapolis: Hackett Publishing compagny.
Marwell, G., and Ames, R.E. (1981): Economists free ride, does anyone else? Experiments on the provision of public goods. *Journal of Public Economics* 15: 295–310.
Mayr, E. (1961): Cause and effect in biology. *Science* 134: 1501–1506.
Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souzat, R., and Grafman, J. (2006): Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences of the United States of America* 103: 15623–15628.
Nagel, T. (1970): *The possibility of altruism*. Oxford: Clarendon Press.
Ostrom, E. (1990): Governing the commons: The evolution of institutions for collective action. Cambridge: Cambridge University Press.
Poldrack, R. (2006): Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences* 10: 59–63.
Rand, A. (1964): *The virtue of selfishness: a new concept of egoism*. New York: New American Library.
Rilling, J.K., Gutman, D.A., Zeh, T.R., Pagnoni, G., Berns, G.S., and Kilts, C.D. (2002): A neural basis for social cooperation. *Neuron* 35: 395–405.
Robinson, J. (2005): Deeper than reason: Emotion and its role in literature, music, and art. Oxford: Oxford University Press.
Rottschaefer, W.A. (2002): It's been a pleasure, but that's not why I did it. In *Evolutionary origins of morality: cross-disciplinary perspectives*, ed. L.D. Katz, 239-243. Thorverton: Imprint Academic.

Slote, M.A. (1964): An empirical basis for psychological egoism. *The Journal of Philosophy* 61: 530–537.

Smith, A. (2002 [1759]): *The theory of moral sentiments*. Cambridge texts in the history of philosophy. Cambridge ; New York: Cambridge University Press.

Sober, E., and Wilson, D.S. (1998): *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, Mass.: Harvard University Press.

Sober, E., and Wilson, D.S. (2000): Morality and 'Unto others'. Response to commentary discussion. *Journal of Consciousness Studies* 7: 257–268.

Sterelny, K., and Kitcher, P. (1988): The return of the gene. *Journal of philosophy* 85: 339–360.

Stich, S.P. (2007): Evolution, altruism and cognitive architecture: a critique of Sober and Wilson's argument for psychological altruism. *Biology and Philosophy* 22: 267–281.

Stich, S.P., Doris, J.M. and Roedder, E. (2010): Altruism. In *The moral psychology handbook*, ed. John M. Doris, 147-205. Oxford: Oxford University Press.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009): Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science* 4: 274–290.

Wilson, T.D. (2002): *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Chapter 14
The Neurobiology of Altruistic Punishment: A Moral Assessment of its Social Utility

Rebekka A. Klein

14.1 Introduction

The article deals with the experimental model of altruistic punishment and social norm enforcement which has recently been designed in the fields of experimental economics and neuroeconomics. By using this model, neurobiologists and economists investigate the close relationship between neurobiological mechanisms in the brain and the enforcement of cooperation norms in human social behavior. They have shown experimentally that the implementation of a costly punishment tool in social dilemma experiments provides strong evidence for the impact of altruistic and prosocial behaviors at the level of group interaction and cooperation. The biological and behavioral interpretation of this evidence will be critically questioned in this article from the point of view of moral philosophy. The following argument will be presented: an exclusive concern for biological motivational mechanisms and behavioral outcomes of punishment fails to discriminate between good and bad punishment in a moral and legal sense, because it does not provide us with an appropriate criterion by which to evaluate the social utility of punishment. Hence, the moral aspects of this behavior have to enter the picture in order to allow us to arrive at a full judgment on its social utility.

R.A. Klein (✉)
Institute for Systematic Theology, University of Halle-Wittenberg,
06099 Halle/S., Germany
e-mail: kleinrebekka@hotmail.com

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*,
Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_14,

14.2 The Study of Altruism in Experimental Economics and Neuroeconomics

The understanding of altruism in experimental economics is based on a consideration with regard to the economy of human behavior. It says: if a human being is altruistic he will incur personal costs in order to increase the benefit of other individuals.[1] Hence, the economic notion of altruism differs from the biological concept by remaining basically on the individual level[2], whereas biologists account for altruistic behavior in terms of Darwinian fitness, group selection and the number of offspring. In addition, the economic notion of altruism can be distinguished from the notion of altruism in psychology and philosophy because it does not deal with the motives, e.g. the beliefs, desires, and reasons behind actual behavior which are crucial for calling a behavior altruistic in psychology and philosophy.[3] Instead, the economic notion of altruism focuses on the outcomes of behavior and measures them in terms of costs and benefits to the individual.

Recently, economists have applied this concept of altruism to the study of human social behavior in experiments that have been conducted according to behavioral game theory[4]. Their observation of social interactions and transactions in social dilemma games was guided by an interest in modeling the social preferences of individuals. In economic theory, preferences are used to measure people's choices and their valuations of certain goods such as food, money, prestige, etc.[5] To determine the actual preferences of people, economists observe people's choices in an experimental environment in which real money is at stake. They do so because they particularly focus on the monetary outcomes of behavior. With regard to social preferences, these outcomes have to have a relation to other individuals' outcomes, and thus are referred to as 'social.'

However, the growing interest on the part of experimental economists in the study of altruism and social preferences is a rather provocative enterprise within their own discipline, because the standard approach in economics, the neoclassical theory of human behavior, usually does not take into account non-selfish, altruistic or even social preferences. Instead, they assume that human rational behavior is exclusively exhibited in the form of egoistic rational choices (the homo economicus model of human agency). This rather narrow understanding of human behavior, which reduces it to the economic principle of self-interested profit-maximization, is due to the habit of neoclassical theory to found its explanation of human behavior

[1] See Fehr & Fischbacher (2003), 785.

[2] See a more detailed analysis of the three different notions of altruism in biology, psychology, and economics in Clavien & Klein (2010). The authors investigate the contribution of experimental economics and neuroeconomics to the debate on psychological altruism, and point out that so far there is neither evidence for nor against psychological altruism in economic experiments.

[3] See the difference between biological and psychological altruism in Sober & Wilson (1998).

[4] An introduction in behavioral game theory can be found in Camerer (2003).

[5] See Camerer & Fehr (2004), 55.

on simple axiomatic approaches.[6] This has led to a model of human agency which is not at all convincing as regards human psychology. To show that the neoclassical approach does not provide proper tools to account for actual human behavior, some experimental economists have begun to systematically test these axiomatic assumptions about human agency in the field and in laboratory experiments using a game-theoretical framework. Their work, which has been done in conjunction with anthropologists and ethnologists, has shown that the assumption of a purely selfish rational agent is not appropriate in most human societies across the globe.[7]

In addition to the behavioral experiments, some experimental economists have also tried to reinforce their view of human social behavior by naturalizing human agency, and investigating the biological roots of people's choices. To this end, they integrated new methodologies and research strategies from the natural sciences into their experimental framework, and participated in the foundation of the trans-disciplinary research approach of 'neuroeconomics.'[8] This approach allows the combination of the methodological tools of neuroscience and those of experimental economics in a shared experimental environment.[9] It can help to uncover the psychological motivational mechanisms behind people's choices, and is very useful in terms of integrating psychological parameters into the economic model of human behavior. Nonetheless, neuroeconomics as a behavioral and brain science does not claim that neoclassical economics is wrong as a whole, but that all its theoretical assumptions and predictions of human behavior can be verified or falsified by empirical research. One of the main objectives of neuroeconomics thus is to provide "…an alternative theoretical approach for predicting behavior and a methodology for testing those theories."[10]

Beyond the engagement with its own discipline, neuroeconomics participates in the major endeavor of explaining the nature of human altruism and the evolution of cooperation across human species. From the point of view of evolutionary anthropology, human cooperation not only differs from non-human mammalian species with respect to intensity and frequency, but rather is of a different kind: it shows a great variability in scale and domain and was probably developed in a non-genetic evolutionary process which cannot be observed in other species.[11] As a consequence,

[6] Glimcher et al. (2009) give a short introduction into the history and development of neurobiological studies in economics and refer to the axiomatic approach of neoclassical economics as one of the main causes of this development.

[7] An overview of the field experiments on social preferences can be found in Henrich et al. (2004). This book documents a global study on the validity of cooperation and fairness norms in social exchange practices. It shows that the economic assumption that individuals exhibit purely selfish preferences in their behavior is violated in all of the fifteen small-scale societies that have been investigated.

[8] See Glimcher et al. (2009) for how wide-spread the approach of neuroeconomics is and the different research questions it can be applied to.

[9] See Gintis (2007).

[10] Glimcher et al. (2009), 6.

[11] See Henrich & Henrich (2006), 223-224.

humans live in large-scale societies which are built on anonymous encounters between genetically non-related individuals.[12] Human cooperation flourishes in these societies in spite of anonymity and non-relatedness, because group interaction is based on social norms. Stability and coordination in social interaction among humans is, therefore, established through the enforcement of norms.

In modern societies, this enforcement is done in two ways. In the case of legal norms, these norms are maintained because their violation is formally sanctioned by the law and penalty system of society. In the case of social norms, which back up the enforcement of legal norms by providing an informal basis for them, enforcement takes place in an autonomous and self-organized process of monitoring and control in local communities, as has been shown by Elinor Ostrom's field studies in the 1990s.[13] Ostrom studied independent systems of social monitoring and control in several long-standing common property regimes, including Swiss grazing pastures, Japanese forests, and irrigation systems in Spain and the Philippines. She could show that the establishment of cooperative institutions in these regimes is organized by the resource users themselves. Hence, the maintenance of social norms and their adaptation as rules of behavior is not secured through formal sanctions by state policy, but through self-governance such as, for instance, social monitoring and interpersonal sanctioning in local communities and (ethnic) groups.

Starting from this insight, the sciences of experimental economics and neuroeconomics have developed a wide range of experimental tools to study the relevant behavioral patterns of social norm enforcement. For obvious reasons, they account for norms in terms of social preferences and individual choices.[14] Thus, they investigate the maintenance of norms as a "second-order public good"[15] in social interaction. By definition, goods are referred to as 'public' in experimental economics if every individual participating in the interaction has a benefit from them "…including those who did not pay any costs of providing the good."[16] Thus, public goods such as natural resources or social infrastructure in human societies are prone to be exploited by free-riders and have to be protected by social norms which govern their use. But norms cannot be chosen by people in the same way as material goods are. Rather, they have to be established and monitored as stable behavioral patterns through the initiative of individuals. Thus, they are not given in advance, but are constituted in social interaction ('second-order public goods'). The behavior of altruistic punishment, which will be focused upon in this article, has been proven to be one of the key patterns for the maintenance of social norms in human interaction.

[12] See Fehr & Fischbacher (2004), 185.

[13] See Ostrom (1990) and Ostrom et al. (1992).

[14] For a philosophical concept of social norms which is in accordance with game theory, see Bicchieri (2006). Bicchieri also integrates various psychological dispositions in her model of norms as preferences of the individual. Thus, her account might also be very valuable for the study of norms in neuroeconomics.

[15] Fehr & Gächter (2002), 137.

[16] Ibid.

14.3 The Correlation of Norm Enforcement and Altruistic Punishment

Several experimental studies on cooperation and prosociality in economics have shown that altruistic punishment plays a key role in understanding the evolution of norm enforcement in human societies.[17] Altruistic punishment does not directly benefit the welfare of an individual person, but society as a whole. Therefore, it is referred to as a 'prosocial' behavior. The term 'prosociality' is used in experimental economics to indicate a behavior that does not directly benefit others (as does cooperation), but the well-being of group interaction as a whole.[18] The behavioral pattern of altruistic punishment has been clearly shown to be of great significance for the study of prosociality in a series of behavioral experiments in economics and neuroeconomics.[19] These have been conducted in different behavioral laboratories since the first study on altruistic punishment was published by Ernst Fehr and Simon Gächter in 2002.[20]

In this study, altruistic punishment is defined as a non-selfish act of punishment which "[provides] …a material benefit for the future interaction partners of the punished subject but not for the punisher."[21] In an experimental setup with 240 participants[22] at the University of Zurich, Fehr and Gächter tested their subject's individual willingness to punish altruistically in a 'public goods' experiment. In this type of experiment, several people have the option of investing a certain amount of money in a group project. Afterwards, the sum of all contributions is to be shared among the group members equally. The experiment in Zurich was conducted in twelve sessions and the group composition was changed after each session. The latter guaranteed that none of the subjects could again meet the same subjects during the experiment. This ensured that the subjects' decisions and behaviors were not based on a preference for reputation-building among group members. The opportunity to punish group members who did not invest in the group project, but benefited from its gain, was offered at the end of each session. In order to test whether the subjects' willingness to punish did include the willingness to suffer personal cost, the punishment was not only costly for the free-rider, but also for the punishing subject himself, because he had to pay for it from his own gain.

---

[17] The claim that social reciprocity (prosocial norm enforcement) provides the best explanation for the evolution of punishing behaviors has been defended in Carpenter et al. (2004).

[18] A definition of the distinction between prosociality and cooperation can be found in Henrich & Henrich (2006). For a model explaining the cultural evolution of prosociality and cooperation see Gintis (2003).

[19] Fehr & Gächter (2002); Fehr & Fischbacher (2003); Fehr & Rockenbach (2003); De Quervain et al. (2004). An assessment of the evolutionary origin of altruistic punishment can be found in Boyd et al. (2003).

[20] Fehr & Gächter (2002).

[21] Ibid., 139.

[22] All of the participants in the experiment were undergraduate students from the University of Zurich.

The results of the experiment were as follows: over twelve sessions, the opportunity to punish social free-riding behavior was taken by 84.3% of the subjects at least once, and even 34.3% of the subjects punished more than five times.[23] A minority of 9.3% of the subjects punished more than ten times. Thus, the experimental results provide strong evidence that altruistic punishment is a stable behavioral pattern among humans. Additionally, a significant effect of altruistic punishment was shown in the later sessions of the experiment. After having been punished, the punished subjects invested a higher amount of money in the group project and changed from non-cooperative to cooperative behaviors in the following sessions. Thus, altruistic punishment caused a substantial increase in terms of the average cooperation level of the group over time. This was highly correlated with the subject's investment strategies and can, therefore, be considered among the facilitating conditions of the evolution of human cooperation. Hence, the remarkable result of the study by Fehr & Gächter (2002) was that the opportunity to punish free-riders altruistically has a significant impact on the maintenance of the norm of cooperation and equity, even in anonymous encounters.

With regard to the interpretation of this evidence, the experimenters suggested that the evolution of social norms has to be explained further in terms of the level of the individual's preferences. Thus, the experimenters asked how the willingness to punish might be triggered on a psychological level. As a suggestion, they hypothesized that the subjects' negative emotions concerning the free-riding behavior of others might be the source of their decision to punish. Emotions such as anger and outrage could provide a proximate mechanism of altruistic punishment.[24]

To elicit the correlation between punishment and the individual's emotions, the experimenters prepared a questionnaire which was given to the subjects after the experiment, and asked them to indicate their intensity of anger concerning the free-riding behavior on a seven-point scale. As a result, 47% of the subjects indicated the highest intensity of anger. Hence, the experimenters concluded that these emotions might be a psychological trigger for punishment. This led them to seek a research tool to further investigate this correlation, which in turn led them to engage in a new research field investigating the neurobiology of prosocial and cooperative behaviors in humans.

14.4 The Neurobiological Explanation of Altruistic Punishment

In a follow-up study[25] to the first experiment on altruistic punishment in 2004, economists Ernst Fehr and Urs Fischbacher started to work together with neuropsychologists for the first time. They added a neuroimaging tool to the experimental

[23] See Fehr & Gächter (2002), 137.

[24] A definition of proximate causes of evolution can be found in Mayr (1961), 1503.

[25] De Quervain et al. (2004).

setup of their study on social norm enforcement, and observed the neurological foundations of people's choices. The idea of combining experiments on norm enforcement with the neurological investigation of the human mind had already come up in a study in 2003 when neuroscientists Alan Sanfey, James Rilling and colleagues adapted an experimental design from economics, and started to investigate the neural substrates of the cognitive and emotional processes involved in decision-making concerning altruistic punishment. After they brain-scanned the subjects with functional magnetic resonance imaging (fMRI), they found an increased activity in the 'anterior insula'—a brain area associated with negative emotional feelings. Hence, they concluded that emotions might be the psychological and neurological driving force behind this behavior, a view which was still consistent with the 2002 findings of Fehr and Gächter.

However, the follow-up study by Fehr, Fischbacher and de Quervain in 2004 led to a rather different neurological finding. The procedure of this experiment was as follows: the subjects were brain-scanned during their decision to punish free-riding behavior by using positron emission tomography (PET). They were placed in a scanner immediately after the interaction with another player was over. The scanning started when subjects learned about the free-riding behavior of the other participant and it finished when they had determined the punishment. In the observation of the neural circuits of the subjects' brains, it could be shown that not the 'interior insula,' but a brain area linked to the anticipation of reward—the 'caudate nucleus'—played a prominent role when people decided to punish. Subjects who exhibited stronger activation of the 'caudate nucleus' were ready to incur more personal costs to punish a free-rider in comparison with subjects who exhibited low caudate activation. Hence, the experimenters interpreted the finding as evidence of the anticipation of "hedonic rewards"[26] being the benefit that altruistic punishers weigh against the costs of punishing. The punishing subjects seemed to feel relief when the violated social norm was established again through an act of retributive justice.

Thus, experimenters concluded that, according to the underlying neurological processes, the subjects' decision-making was driven by hedonic motivation. Hedonic motivation is one of the key features in an evolutionary explanation of behavior, because there is natural selection for avoiding pain and unpleasantness. Therefore, the correlation between hedonic motivation and altruistic punishment might function as a proximate mechanism of the evolution of human cooperation. But this has to be explored further in future research, and cannot be concluded from a single study.

In my view, a much more pressing question with regard to the interpretation of the result of the neuroeconomics study concerns the assignment of psychological motivational states to the neurological findings, and their validity for determining the social utility of this behavior. My question is whether it is really justifiable to conclude from the consequentialist and neurobiological explanation of punishment in neuroeconomics that punishment is a prosocial and thus beneficial act in terms of the welfare of human society. In the following sections of the paper, I will try to cast

[26] Ibid., 1257.

some doubt on this conclusion. I will show that the behavioral and neurobiological explanation of punishment might lead to a shortened (reduced) judgment when it comes to determining the social utility of this behavior. Thus, external reasoning about its motivation and consequences has to be integrated into the picture in order to form a judgment about the purely positive evaluation of its prosocial outcomes.

As we have seen, the behavioral pattern of altruistic punishment as investigated in economics is different from that of reciprocal (direct) and reputation-based (indirect) altruism as investigated in evolutionary biology. Its manifestation in human behavior is dependent on the revealed preference of an individual to incurring personal costs which are never likely to be recovered, in order to sanction another for his norm violation or social free-riding behavior. Thus, the punisher is referred to as an altruistic person in a consequentialist sense which means that his personal motivation for the decision to punish does not enter the picture. The study by Fehr and Gächter (2002) has shown that this kind of altruistic behavior has a remarkable effect on human interaction: it increases the average cooperation level of group interaction in the long run. From the perspective of neuroeconomics, altruistic punishment is among the proximate (individual) causes of the evolution of human cooperation and is due to a neural mechanism which explains why the human species maintains such a high degree of cooperation among non-relative individuals, which is different in kind from that of all other species.[27]

But the investigation of the neural mechanism underlying altruistic punishment has also shown that there is not only cost but also benefit to the punisher: he anticipates a strong feeling of satisfaction when expecting the free-rider to be punished and the norm of cooperation and equity being re-established. Thus, a behavior which is altruistic in the consequentialist sense seems to be motivated by hedonic reward anticipation on the psychological level. Thus, the study is ambivalent in its result: the individual's motivation for altruistic punishment is obviously self-concerned in the first place. During decision-making, the punishing subject anticipates his own state of mind which will occur after the punishment is carried out.[28] Hence, as several interpretations of the neuroeconomic study of de Quervain and colleagues (2004) have shown, it is not absolutely clear from the neurological findings whether the punishing subject's feeling of satisfaction is primarily related to the (indirect) establishment of the cooperation norm, or whether it is primarily related to a desire for revenge—longing for a compensation of the cost he has suffered as a result of the initial social free-riding. In other words: is the motivation for altruistic punishment grounded in a desire for social norm enforcement or a desire for revenge?

Unfortunately, no further neuroeconomic research has been done to answer this question concerning the psychological motive underlying altruistic punishment.[29]

---

[27] See Fehr & Fischbacher (2003); Fehr & Fischbacher (2005).

[28] Knutson (2004) has already pointed towards this ambivalence of the study's results. The claim that there is no evidence explaining the causal chain of motivation behind the behavior is developed further in Clavien & Klein (2010).

[29] For a distinction between motive and motivation see the article on "Altruistic Emotional Motivation" by Christine Clavien in this volume.

And from the point of view of neuroeconomics, the question might also be irrelevant because the outcome of both of these motives, the enforcement of a social norm, is the same. On purely consequentialist grounds, it doesn't matter that the enforcement of a social norm is merely a secondary (instrumental) motive or even an unintended outcome of people's choices. The only thing that matters is whether the causal chain that leads to this outcome works reliably. Whether it is grounded in self-concern, or even selfish motivational states or motives, does not influence the evaluation of the prosocial outcomes of punishment behavior. But the disregard of the issue of motivation makes the use of the term 'altruism' with respect to punishment behaviors in economics highly questionable.

In contrast to the view that the neurobiological investigation of motivational states is sufficient to judge on the social utility of punishment behavior, I will point out now that the neuroeconomic approach is too short-sighted. In the following section I will show that—in contradiction to the neuroeconomic interpretation—the proof that punishers act out of hedonic motivation is the crucial point when it comes to the moral assessment of the consequences of punishment behavior. My thesis will be based on the argument that the motivation for a punitive act, and the motive behind that act, are not negligible in an assessment of its consequences. Hence, I have to clarify in what sense the questions of motive and motivation are crucial concerning the distinction between good and bad punishment in a moral and legal sense.

14.5 Moral Philosophical Assessment of Altruistic Punishment

In this section, I will introduce the moral perspective of judgment on social behaviors as a supplementary approach to the behavioral and brain sciences. The moral philosophical approach adds certain crucial aspects to the experimental study of behavior, whose understanding and explanation will improve the evaluation of its social utility and will help to avoid misjudgments concerning its overall prosocial consequences. The moral assessment of human behaviors not only deals with the question of whether certain behaviors have a prosocial or antisocial outcome concerning the common good or society's welfare, but also concerning the welfare of a single individual. Hence, it judges the social utility of human behaviors, not only in terms of 'general others,' which are represented by the anonymous social structures and institutions of society, but also in terms of 'concrete others,' who are affected in their individual well-being by the actions of others.

In this regard, the moral motive behind a punitive act shapes the social character and outcome of this behavior in a twofold sense: (a) it marks the boundary of the punitive act as regards its consequences for the well-being of concrete others, and (b) it prevents punishment from becoming an act of sheer violence which goes awry in the sense that it is extended beyond the scope of the moral and legal measures of social interaction. Hence, the motive or intention behind punishment is crucial for determining how it is acted out with respect to others as regards their individual

right to well-being and intactness (a), and their individual right to the adequacy of punishment of their offence (b). Thus, the empirical question of who is harmed, the extent of such harm and whether this can count as a prosocial or antisocial act, cannot be answered from a moral perspective without taking into account the intention and motivation accompanying the punitive act on the part of the punisher.

The moral question concerning punishment becomes even more pressing when we recognize that in every act of punishment—whether it is legally justified or not—there is individual leeway with regard to how the one who imposes the sanction can strengthen or weaken its consequences for others. Sometimes this leeway is acted out by the individual in the form of a very subtle psychological mistreatment of the other, and sometimes it is done in a very offensive and exposing way, involving dehumanization. Nonetheless, both modes can count among the varieties of human cruelty, insofar as they violate the individual's well-being and intactness with lasting effect.

To consider an example for the question I have in mind, we can see how the behavioral pattern of prosocial and altruistic punishment is demarcated from the sadistic behavior that was exhibited in Abu Ghraib Prison in Iraq in 2004. In Abu Ghraib, the societal institution of penalty became an excuse and a means for an excess of sheer violence.[30] The imprisoned criminals were held in a kind of lawless state. They were physically tortured and sexually abused by their prison guards. Although this treatment violated the norms of prisoner treatment outlined in the 'Geneva Convention' (1949), it was well known and accepted among the military police authorities and in the U.S. government.[31] The guards could, therefore, rely on official tolerance or, rather, official neglect of their behavior.

Abu Ghraib represents punishment which is certainly not beneficial to society because the legal institution of punishment is turned into its opposite—a violation of legal norms. Although the imprisoned criminals of war might have legally deserved punishment in terms of imprisonment, they received a much harder (physical) punishment than the one that would have been legally imposed on them—including acts of debasement and dehumanization. What is interesting about the case in the context of my argument is the following: the unlegislated punishment became possible because the prison guards established a social norm among their group members, considering it acceptable to punish the prisoners in order to nourish their own sadistic appetites. Maybe their behavior was rationalized afterwards by arguing that the prisoners deserved this kind of punishment because they are criminals. Hence, the prison guards considered it as a collective goal to maximize their pleasure at the cost of others who do not share their religious, national and ethnic background and who have failed to be respected as human beings in terms of their human dignity.

The incidents in Abu Ghraib show how important it is to safeguard the notion that the purpose of punishment in society is to enforce social norms which do not

[30] See Taguba (2004). The *Taguba Report* on the torture scandal in Abu Ghraib judges the behavior of the prison guards from the point of view of the *Geneva Convention Relative to the Treatment of Prisoners of War* (1949).

[31] See the discussion in Denner (2004).

violate the moral or legal norms of egalitarian cooperation. This is because the latter are also established precisely in order to protect individual's rights. In the case of Abu Ghraib, norms were not officially established. Hence, a form of self-governance took place among the group members. In this regard, the situation in Abu Ghraib is similar but not identical to the paradigmatic case study of social norm enforcement in the 'public goods' experiments. The crucial difference between the 'public goods' situation and the situation in Abu Ghraib in terms of societal welfare is that it was not a prosocial but an antisocial norm[32] that evolved out of the lawless state people were placed in. However, one could argue that this was due to the circumstance that prison guards and prisoners were placed in an environment where the one side had executive power over the other, whereas in 'public goods' situations, all individuals belong to the same group and therefore start from a level of egalitarian interaction. I think this argument points in the right direction, and can be substantiated by experimental evidence.

In a large-scale field experiment with fifteen native societies around the world, it has already been shown that differences of culture matter a lot when it comes to social norm enforcement.[33] Furthermore, there is evidence from a field experiment in Papua New Guinea (Bernhard et al. 2006) that the enforcement of norms across the boundaries of culture, nation and group membership seems to work less effectively than in 'public goods' situations, where people belonging to the same group establish sanctioning behaviors which protect the commons that their collective life is dependent on. Hence, especially in intercultural and inter-group interactions on the local and global level, it is important to safeguard the idea that punishment, as a means of norm enforcement, should not have antisocial or inhuman side effects for individuals. Accordingly, one could argue that the solution to this problem might be that social norm enforcement has to be prevented from violating the superordinate norms of justice and equity which are universally held in all human societies. This means that the evolution of 'particularistic norms' among social groups has to be governed by the maintenance of universal 'societal norms' such as equity, fairness and reciprocity.

Focusing on the distinction between particularistic and universal norms would lead us in the end to a solution of the conflict between prosocial and antisocial consequences of punishment on the level of transnational political and legal institutions. The latter was certainly not under consideration when behavioral economists designed their experimental research tools to study the self-governance of social norm enforcement among individuals. In contrast to a non-individual, state-governed or cosmopolitan solution, they have proposed that it is not the responsibility of centralized institutions alone to govern the evolution of social norms. Rather, individuals and groups can develop a system of monitoring and controlling the maintenance of

[32] The norm is antisocial only with respect to the wider group of people that includes the guards as well as the prisoners. With respect to the population of the guards alone, the norm is actually prosocial, because it increases their status. Hence, the fact that a particular action is prosocial with respect to a limited peer group does not say that it is morally unproblematic in general.

[33] See Henrich et al. (2004).

norms for themselves. Hence, to take the experimental economist's research work seriously, and to account for the distinction of socially beneficial and non-beneficial punishment on the individual level, we again have to look more closely at the individual rationale and psychological motivation of the punishing subjects.

Beyond the institutional level it is the responsibility of individuals to prevent social norm enforcement through punishment developing into acts with antisocial side effects. Coming back to the case of Abu Ghraib we can ask: did the prison guards' desire to satisfy their own sadistic appetites simply override their rational faculties with regard to weighing the costs and benefits of punishment against each other? Or did they establish an individual rationale for their behavior which made it reasonable to establish an antisocial norm, promoting collective fulfillment of their sadistic appetites at the cost of others? Undoubtedly, there were personal costs to the prison guards in Abu Ghraib: they risked prosecution and eventually lost their jobs and were accused of breaking international law. But undoubtedly, there were some benefits for them as well: the punishers expected reputational gain from their fellows when they abused prisoners while also fulfilling their own sadistic appetites.

Although it represents a different kind of punishment than the one that was in mind in the economic experiments on altruistic punishment, the torture scandal in Abu Ghraib is a good example of the dangers of highlighting the prosocial consequences of punishment in a purely consequentialist sense.[34] The paradigmatic case shows how the establishment of a social norm among group members can turn into a norm violation itself: the violation of human rights. Of course, such a situation as in Abu Ghraib was not modeled in the neuroeconomic experiment presented earlier (De Quervain et al. 2004). The experimental setting only allowed for financial punishment which means that the degree of harm which could be imposed on a non-cooperative subject was limited and controlled externally, ensuring that the punishing subject could not overrun the given conditions of punishment. This means that the punishment in the experiment was not shaped by the punishing subject's individual preferences determining the mode of punishment (psychological, physical, financial, etc.) and its heaviness was not independent of the experimenter's setting. Thus, an immoral excess of punishment, i.e. a punitive act which overrides the boundaries which the moral sense of the other imposes on punishment, could not even be modeled.

Furthermore, the argument could be raised that the experimental study by De Quervain et al. (2004) neglected to pose the morally crucial question of how the motivational states of punishing subjects might shape and influence the different modes of punishment, as well as the difference between excessive and limited punishment. The study simply did not take into account the fact that the motivational states of the subjects might make a difference with regard to the act of punishment itself. However,

[34] See the experiments related to punishment in prison in Milgram (1963). As far as I can see, the experimental economic study of punishment has not been related to this social psychology study of the excess of physical punishment.

the neurobiological investigation of these motivational states has positively shown that the punishing subjects are looking for some personal benefit besides the prosocial effect of their punishment behavior. They weigh the material costs of punishment (personal loss) against its being the cause of a feeling of reward. But it remains rather unclear in the neuroeconomic study whether the anticipated reward is an appraisal of the social utility of the punishment (norm enforcement) or a cause for personal satisfaction (revenge). Nonetheless, experimental economics and neuroeconomics claim that the objective utility of this behavior which can be observed in its outcome (increasing the average cooperation level over time) is sufficient to appraise punishment as a social utility tool in human societies.

Contrary to this position, I have claimed that a critical moral evaluation of the social utility of punishment has to start from the negative observation that the punishment of people who deserve it in terms of their preceding antisocial behavior is not in itself a socially valuable act. The moral costs of punishment may outweigh the social benefit, because punishment always involves someone being harmed—either physically, psychologically or financially. This raises the moral question of a possible violation of an individual's rights associated with, or even inherent to, punishment—a problem which is more or less concealed in the euphemistic term 'altruistic punishment.' Since it is the major distinction between a liberal and a dictatorial concept of society that cooperation and 'public goods' are not maintained to the disadvantage of individual's rights, we have to ask for a justification of any cost the punisher imposes upon others. In order to consider punishment as 'prosocial,' it is not enough—as the neuroeconomic concept of altruism claims—that we ensure that the punisher obtains no reputational or financial benefit from the punishment, especially if there is evidence of a hedonic reward mechanism governing his decision-making. The fact that there is a material cost to the punisher does not safeguard that his behavior will not have intolerable antisocial side effects in terms of the outcome, for example when punishment is acted out in order to satisfy a sadistic appetite.

Hence, the moral assessment of punishment requires external reasoning about the motivation and intentions of the punishing subject. This reasoning can, of course, be substantiated or falsified by neurological findings, but it is, in principle, indispensable when it comes to an evaluation of punishment behaviors from a moral perspective. The moral question concerning the motivation of punishment is whether the punishing subject is still concerned with the individual welfare of the other. This moral concern for the other should occur even when the punished subject deserves punishment, because it prevents the punitive act from going awry, i.e. turning out to have ambiguous, prosocial and antisocial consequences at the same time.

14.6 Punishment and the Welfare of a Just Society

In the preceding section, I have shown why the question of motive or intention behind punishment is not insignificant. I have argued that the distinction between justified punishments and acts of unjustified violence shall be upheld by external

reasoning about the motivation for, and motive behind, a punitive act. By considering neuroeconomic findings about the motivational causes of a punitive act, I have pointed out that to harm someone for a good reason can include the motive of revenge as well as the motive of preventing further norm violations. The first one is a selfish motive leading (instrumentally) to prosocial consequences, whereas the second is a purely prosocial motive. In this section, I will argue that the only way to ensure that norm enforcement by punishment has purely prosocial outcomes on both the individual and the societal level, is not to show that it is altruistic, but to prove that it is primarily driven by an egalitarian motive. Hence, the moral assessment of punishment has to distinguish between (a) punishment as a means of establishing egalitarian cooperation and (b) punishment as an excess of sheer violence (retaliation, revenge, sadistic appetite and the like).

Thus, the following steps associated with the assessment of punitive acts have to guide the evaluation of its social utility: (a) determine the (neurobiological) motivational causes of the punitive act, (b) look for a moral concern included in these motivational causes, (c) clarify the intention or motive behind the punishment behavior, (d) consider the conformity of the motive to moral and legal norms of a just society's welfare, (e) evaluate the prosocial or antisocial consequences of a punitive act. It should be obvious from these assessment steps that the investigation of the neurobiological motivational causes of behavior alone has not clarified the prosociality of the intention or motive behind punishment behavior. A moral assessment of these motivational causes is needed in order to provide a valid judgment of its social utility. But nor is this enough. One more step has to follow: the consequences of punishment behavior as well as its motivational causes have to be assessed from a moral perspective. Let me point out briefly how this can be done by bringing together the experimental economist's and the moral philosopher's approaches.

Instead of using the term 'prosociality,' moral and social theory account for the welfare of a just society in terms of a high level of egalitarian cooperation, because they conceive of justice as the equal distribution of the liberties and rights of individuals in societal cooperation. Thus, egalitarian cooperation is cooperation which aims at producing social and economic equity in society without violating the rights of the individual. Hence, moral and social theory judges the social utility of punishment behavior in terms of its contribution to the maintenance and increasing of egalitarian cooperation. In order to determine what this contribution is, moral and social theory requires the justification of an agent's decision to punish, by investigating his underlying 'egalitarian motives.'[35] In the behavioral study of prosocial decision-making, egalitarian motives can be represented by modeling the consequences of behavior in terms of an increase in equality and a decrease in inequality in the aggregate level of distribution.[36] Hence, egalitarian motives are correlated to

[35] See the behavioral experiment on egalitarian motives in Dawes et al. (2007). For future research, it would be necessary to investigate the neurobiological underpinnings of this behavioral model of egalitarian motives.

[36] See Masclet & Villeval (2008).

the financial outcome of social decision-making. They can be inferred from the equal distribution of an exchanged good. In the experimental study by de Quervain et al. (2004), which investigated the biological motivational causes of punishment, no egalitarian motives were in play. The punishing subjects did not align the other subject's outcome to their own by punishment. Rather, they decreased both outcomes equally in order to harm the social free-rider. Hence, their punishment was not grounded in egalitarian motives which, at very least, casts a severe doubt on its social utility from the perspective of moral philosophy.

In this article, I have demonstrated how punishment and norm enforcement go awry when a moral concern is missing in the motivational causes of this behavior. I have shown how judgments on the moral nature of motivational causes change the interpretation of altruistic punishment as a socially beneficial behavior. Furthermore, I have argued that even the consequences of punishment cannot be judged to be socially valuable from the perspective of moral philosophy unless they are evaluated in the light of their underlying egalitarian motive. From these arguments, it can be concluded that the legitimacy of social norms is not bound to their effective enforcement, but requires some external reasoning about the motivation of, and motive behind, enforcement on the individual level. Otherwise, their legitimacy could not be demarcated from their misuse in a dictatorial system of social disciplinary power, which would extinguish the norms' crucial function of providing the 'breeding ground' of egalitarian cooperation within modern and democratic societies.

In a society's penal system, the political power of law is enforced by sanctions which are bound by the law. Thus, legal punishment is distinguished from illegal punishment by its conformity to the law, and not to some other subjective rationale. In contrast, punishment is not simply bound by the law as part of the society's system of informal social norm enforcement. Rather, in this context, bad punishment is demarcated from good punishment by its impact on the welfare of a just society. Hence, a society's system of informal norm enforcement is referred to as 'just' if it establishes welfare in terms of its 'public goods' as well as in terms of its individual's rights. Thus, there is a positive and a negative condition for the prosociality of punishment: in order to be 'prosocial' the punitive act has to (a) increase the average cooperation level of social interaction and it has (b) to do so by not violating the rights of individuals. Both conditions refer to the moral aspects of this behavior on the individual and at the societal level.

14.7 Summary of the Argument

In the experimental study of the motivational causes of punishment behavior, an exclusive concern for biological motivational mechanisms and behavioral outcomes of behavior fails to discriminate between good and bad punishment in a moral sense. The article has substantiated this claim concerning a recent study of the neurobiological underpinnings of altruistic punishment (De Quervain et al. 2004). This study has revealed the biological motivational causes of punitive acts with a background

in social norm violation. It has shown that the punishing subjects are looking for some personal benefit besides the prosocial effect of their behavior. Although this provides valuable new insights into the psychological motivation of altruistic subjects who anticipate satisfaction from the punishment of norm violation, it challenges the claim for the purely altruistic and socially beneficial nature of this behavior. The subject's anticipation of reward can refer to a selfish (e.g. retaliation, revenge) as well as a prosocial motive for punishment (norm enforcement). To handle this ambiguity, the article has claimed that external reasoning about the moral concern of punishment is required to judge its social utility. The use of the terms 'altruism' and 'prosociality' with respect to punishment in economics does merely conceal this ambiguity. Hence, the consideration of moral motives and intentions should enter the neurobiological and behavioral explanation of punishment behavior. The article concludes with the argument that the assessment of the intention or motive behind punishment behavior is not insignificant for the question as to whether it has a positive or a negative impact on the welfare of a just society.

References

Bernhard, H., Fischbacher, U., and Fehr, E. (2006): Parochial Altruism in Humans. *Nature* 442: 912–915.

Bicchieri, C. (2006): *The Grammar of Society. The Nature and Dynamics of Social Norms*. New York: Cambridge University Press.

Boyd, F.R., Gintis, H., Bowles, S., and Richerson, P.J. (2003): The Evolution of Altruistic Punishment. *Proceedings of the National Academy of Sciences of the United States of America* 100: 3531–3535.

Camerer, C.F. (2003): *Behavioral Game Theory. Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

Camerer, C.F., and Fehr, E. (2004): Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists. In *Foundations of Human Sociality*, ed. J. Henrich et al. Oxford: Oxford University Press, 55–95.

Carpenter, J.P., Matthews, P.H., and Ong'ong'a, O. (2004): Why Punish? Social Reciprocity and the Enforcement of Prosocial Norms. *Journal of Evolutionary Economics* 14: 407–429.

Clavien, C., and Klein, R.A. (2010): Eager for Fairness or for Revenge? Psychological Altruism in Economics. *Economics & Philosophy* 26: 267–290.

Dawes, C.T., Fowler, J.H., Johnson, T., Mc Elreath, R., and Smirnov, O. (2007): Egalitarian Motives in Humans. *Nature* 446: 794–796.

De Quervain, D.J.F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., and Fehr, E. (2004): The Neural Basis of Altruistic Punishment. *Science* 305: 1254–1258.

Denner, M. (2004): *Torture and Truth: America, Abu Ghraib, and the War on Terror*. New York: New York Review Books.

Fehr, E., and Fischbacher, U. (2003): The Nature of Human Altruism. *Nature* 425 (2003): 785–791.

Fehr, E., and Fischbacher, U. (2004): Social Norms and Human Cooperation. *Trends in Cognitive Sciences* 8: 185–190.

Fehr, E., and Fischbacher, U. (2005): Human Altruism – Proximate Patterns and Evolutionary Origins. *Analyse & Kritik* 27: 6–47.

Fehr, E., and Gächter, S. (2002): Altruistic Punishment in Humans. *Nature* 415: 137–140.

Fehr, E., and Rockenbach, B. (2003): Detrimental Effects of Sanctions on Human Altruism. *Nature* 422: 137–140.

Gintis, H. (2003): Solving the Puzzle of Prosociality. *Rationality and Society* 15: 155-187.

Gintis, H. (2007): A Framework For the Unification of the Behavioral Sciences. *Behavioral Brain Science* 30: 1–61.

Glimcher, P.W., Camerer, C.F., Fehr, E., and Poldrack, R.A. (2009): Introduction: A Brief History of Neuroeconomics, in *Neuroeconomics. Decision-Making and the Brain*, ed. P. Glimcher et al. Amsterdam: Elsevier/Academic Press, 1–12.

Henrich, J., and Henrich, N. (2006): Culture, Evolution and the Puzzle of Human Cooperation. *Cognitive Systems Research* 7: 220–245.

Henrich, J., Boyd, R., Bowles, S., Camerer, C.F., Fehr, E., and Gintis, H. (2004): *Foundations of Human Sociality. Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press.

Knutson, B. (2004): Sweet Revenge? *Science* 305, 1246–1247.

Masclet, D., and Villeval, M.-C. (2008): Punishment, Inequality, and Welfare: A Public Good Experiment. *Social Choice and Welfare* 31: 475–502

Mayr, E. (1961): Cause and Effect in Biology. *Science* 134: 1501–1506.

Milgram, S. (1963): Behavioral Study of Obedience. *Journal of Abnormal and Social Psychology* 67: 371–378.

Ostrom, E. (1990): *Governing the Commons. The Evolution of Institutions for Collective Action*. New York: Cambridge University Press.

Ostrom, E., Walker, J., and Gardner, R. (1992): Covenants With and Without A Sword: Self-Governance Is Possible. *American Political Science Review* 86, 404–417.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. (2003): The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science* 300: 1755–1758.

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

Taguba, A.M. (2004): The "Taguba Report" On Treatment Of *Abu Ghraib* Prisoners In Iraq. ARTICLE 15-6 Investigation of the 800th Military Police Brigade, May 2004, http://news.findlaw.com/hdocs/docs/iraq/tagubarpt.html (28.6.2008).

Part V
Neurobiological Explanations of Behavior

Chapter 15
Behavioral Traits, the Intentional Stance, and Biological Functions: What Neuroscience Explains

Marcel Weber

15.1 Introduction: The Intentional Stance and the Observation of Animal Behavior

According to some philosophers, we take an "intentional stance" towards animals, especially but not exclusively towards rational ones (Dennett 1989). This means that we bring some of their activities and movements under intentional concepts, i.e., concepts that involve an attribution of intentional states to the entity whose behavior we want to explain. Intentional states present some object as *being so*, for example, an incoming object as being a dangerous predator. Even though we may not have any direct evidence for the presence of such states, treating living beings *as if* they had such states allows us to predict them much more easily. Consider a gazelle that is being hunted by a leopard. Attributing to the gazelle intentional states to the effect that it *thinks* that the charging object is a dangerous cat, or at least something as dangerous as a snake but much, much faster, makes its ensuing behavior more predictable. The point is not that attributing intentional states to the gazelle is the best *explanation* of its behavior, it is merely that taking some other stance toward it, for instance, the stance that a physicist takes towards two moving masses, make its movements much more difficult to predict for us. Let's refer to this as the *predictive value of the intentional stance*.

Here is a second, related idea that is different enough to merit separate consideration. Looking at animals with our intentional goggles on unveils certain *general patterns* that would otherwise remain invisible. This thesis has been defended by

M. Weber (✉)
Department of Philosophy, University of Geneva, 2 rue de Candolle, CH-1211, Geneva, Switzerland
e-mail: marcel.weber@unige.ch

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*, Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_15,

Hilary Kornblith in his *Knowledge and its Place in Nature* (Kornblith 2002). Kornblith considers reports from ornithologists such as this one:

> George Schaller told me of watching raven pairs in Mongolia *cooperate* in snatching rats from feeding raptors. Similarly, in Yellowstone Park, Ray Paunovitch reported seeing a re-tailed hawk with a ground squirrel. Two ravens approached. One *distracted* the hawk from the front while the other handily snachtched the squirrel from behind. Carsten Hinnerichs saw the same maneuver repreated three times in a row in a field near Brücke, Germany, where a fox was catching field mice. Terry McEneaney, Yellowstone Park ornithologist, observed two ravens circling an osprey nest where the female osprey was incubating. One raven landed on the nest rim and took a fish, then while the osprey was *distracted* by this thief, the other ravens swooped down and stole an osprey egg. (Kornblith 2002: 31; emphasis mine)

Kornblith emphasizes in particular that the use of intentional notions such as "cooperate" and "distract" allows us to recognize what quite different behaviors in anatomically and behaviorally quite different species have in *common*. For example, he draws attention to the fact that in the description of the "distracting"-behavior, the ornithologists do not even mention any bodily motions:

> Heinrich does not speak of the ravens moving their beaks, or wings or bodies in certain patterns. Indeed, there is no reason to think that the manner in which the ravens distracted their various targets involved any commonality at all at the level of bodily motion. There are bits of animal motion that may be described in such terms, but this does not seem to be such a case. Instead, what is common to the various episodes described can only be appreciated by attributing certain intentional states to the animals involved. If we see the behavior as a case of one bird distracting another, we are able to make sense of it in a way that a description in terms of moving beaks, wings, and bodies fail to capture. (Kornblith 2002: 33)

The point is not that without the intentional stance, there are no patterns discernible at all. Kornblith allows descriptions such as "wing-flapping, squawking, pecking, and so on" as descriptions that are entirely free of intentional notions. The point is rather that, if viewed in non-intentional terms, these patterns are *heterogeneous*. Only an intentional category such as "distraction" allows us to view these patterns as instances of a single kind, a kind that also includes behaviors in animals that lack wings and beaks, and are utterly bad at squawking.

Thus, Kornblith's contention is not that animal behavior can only be *explained* in intentional terms. The point is rather that it is more often than not *described* in such a way, before anyone even ventures to propose an explanation.

Kornblith goes as far as to suggest that his point stands even for behaviors that seem to admit of fairly simple descriptions, because they can be described without any reference to complicated movements of the part of an animal relative to each other, but simply with reference to the spatio-temporal location of the entire animal. The examples he discusses here are mostly animals that are capable of homing.

Thus, Kornblith's point is that the intentional notions allow us to bring patterns that are quite heterogeneous at the physical level under *general kinds*. Let us refer to this property as the *classificatory value* of the intentional stance.

The goal of Kornblith's argument is to show that intentionality and cognate notions pick out perfectly fine natural kinds that are on a par with better-known natural kinds from science, such as gold or oxidation-reduction reactions.

Kornblith's use of the ethological literature is quite objectionable; because he does not check what role the reports he cites actually play in these investigations.

It is often a mistake to take scientific language too literally. Nonetheless, this kind of arguing does raise some interesting issues in philosophy of science, namely the problem of the theory-ladenness of observation and the question of what constitutes a good explanandum for (neuro-)biological explanations of behavior in the first place.

In this chapter, I want to address these issues in the context of a recent example from neuroscience, namely social behavior in the nematode *C. elegans*. In Section 15.2, I shall motivate the problem of trait individuation a little further and critically examine some arguments from the literature. This will make clear that trait individuation in biology is often based on the concept of function. In Section 15.3, I discuss a concept of function that I claim is suited for this task. In Section 15.4, I present my central neurobiological example and show that it illustrates functional trait individuation. Furthermore, I show that research on this model organism actually transformed what the neuroscientists were trying to explain.

15.2 The General Problem of Trait Individuation

Kornblith's point is an instance of what philosophers of science refer to as the theory-dependence or more metaphorically the "theory-ladenness" of observation. The point is that ornithologists' reports of animal behavior do not merely deploy pure observation terms – even if there should be such terms. Rather, these reports employ theoretical notions. The theory that supplies these notions, according to Kornblith, is *folk psychology* (this is my term, not Kornblith's). For notions such as "cooperation" and "distracting" are precisely those that we use in our everyday interactions with our fellow human beings, and sometimes also with animals. (It is a good idea to try to "distract" a bear that is approaching a small child. It is a very, very bad idea to try to push the bear away by brute force).

Now it is possible to accept the claim that descriptions of behavior in biology are theory-dependent while, at the same time, rejecting the idea that the theory that such descriptions depend on is folk psychology.[1] Of course, such a reply would have to explain why intentional vocabulary (while it was banned from both psychology and ethology for a long time, namely from the heydays of behaviorism until quite recently) is so widely used in the scientific literature, and why cognitive ethology is flourishing to the extent that is has been in recent years. However, it should be noted that cognitive ethology is precisely the field that problematizes accounts of animal behavior that involve intentionality, so it may not be representative.

Before I take on the task of developing an alternative view as to what theoretical notions (ought to) inform behavioral observations, I shall briefly lay out the general problem of trait individuation as it has been discussed in the philosophy of biology.

[1]That certain ways of classifying animal behavior are theory-laden has also been pointed out by Philip Kitcher in *Vaulting Ambition*, his powerful critique of sociobiology (Kitcher 1985). He argues that the way in which certain behavioral traits such as rape are singled out for evolutionary explanations is often informed by ideological preconceptions and false stereotypes, e.g., that rape is sexual behavior.

The problem we are dealing with is an instance of the more general problem of how the traits of organisms are individuated. By "traits" I do not mean only genetic traits, not even traits that vary in a population. The question is rather how certain parts of an organism such as organs are identified as some sort of a unified structure in the first place. While this may seem obvious in some cases like the leaves on the beech tree in front of my window, it is difficult to state some general or even medium-range principles as to what is a good way of cutting up an organism along its natural joints. Clearly, Socrates's advice to proceed by trying not to splinter any parts, "as a bad butcher might do" (Plato 1997, *Phaedrus*, 265e) is not helpful at all, for we have no theory-independent way of knowing when we have splintered something. One way of putting our problem is thus whether he or she who cuts behaviors with intentional knives is a good butcher or a bad one. The case is particularly difficult in behavioral biology, as behaviors can involve different parts of an organism, and may consist of non-contiguous sequences of events that spread out in time.

What are the options? What are the conceptual tools used by a good butcher of animal behavior? While there is hardly any literature on this specific question, some philosophers of biology have addressed the problem of trait individuation in general. They have come up with two kinds of answers: *Proper functions* and *homology*.

An adherent of the first answer is Alex Rosenberg (Rosenberg 2006). He argues that parts of organisms are individuated via their biological functions, which he understands in terms of proper functions. On this view, some item X has a function F in organism S exactly if X does F and the fact that some earlier tokens of X have done X is a cause of X's presence in S. The way in which earlier tokens can cause the presence of some item in later generations, of course, is natural selection. Thus, Rosenberg's view is that natural selection is not only needed to explain why some organism S came to have a part X, but to speak of X as some kind of unity in the first place. It is for this reason that Rosenberg thinks that the theory of natural selection is fundamental for the whole of biology. This, of course, includes behavioral biology. According to Rosenberg, the description of behavioral traits is laden and/or ought to be laden by theoretical hypotheses about selection history. A trait such as a wing is individuated by the fact that it was selected for flying, no matter what other capacities it may have (for instance, it's capacity of being flapped so as to distract some other animal). On this view, descriptions of an organism's trait are laden by the theory of natural selection and assumptions about the evolutionary past.

Paul Griffiths (2009) has argued that this view puts the cart before the horse. The parts of organisms and their causal capacities must be understandable independently of natural selection. Otherwise, the following regress threatens:

1. Selected effect functions are ascribed by causal analysis of the capacities of the parts of ancestral organisms and a determination of their fitness contribution.
2. Thus, we must already be able to individuate the parts. This cannot be done on the basis of the ancestors to the ancestral organisms, because it would generate a regress.
3. But if we are able to individuate parts for ancestral organisms independently of their selection history, then this is possible for living organisms. (This is modified from Griffiths, 2009)

So if natural selection is not fit for the individuation of organismic parts, what is? One possibility is another fundamental concept of biology, that of *homology* (Griffiths 2006; Love 2007; Brigandt 2007). Richard Owen defined homologues as "the same organ in different animals under every variety of form and function". For example, a bird's wings are said to be homologous to our arms, while the panda's famous "thumbs" are not homologous to our thumbs. Obviously, Owen's definition is purely formal; it does give us any criteria when two organs count as "the same". It merely says that homology is a sameness-of-kind relation that abstracts away from form *and* function. The standard view of homology sees it as a relation of *shared ancestry*. However, this view has some known difficulties.

The main difficulty consists in saying what it means for a *part* of an organism to have the same ancestry as some part of another organism. Parts of organisms do not reproduce, at least not independently of the individual of which they are a part. Thus, it seems that, not unlike in the case of selected effect functions, biologists must *first* apply some *other* concept to individuate characters and *then* trace that part through the phylogenetic tree via the ancestry of the organisms that have these parts. Some have suggested that this elusive other individuating property might be something like developmental units, however, this has turned out to be difficult to spell out. At any rate, this seems highly unsuitable for behavioral traits, which are normally not at all viewed in developmental terms.

The issue of homology is, of course, extremely complex. My considerations here are only meant to show that it is unlikely that behavioral traits normally often individuated by homology.

So now we have considered the following candidates of theories or theoretical notions that have been claimed to inform – for better or for worse – the classification of behavioral traits: (1) folk psychology, (2) ideology, (3) natural selection, (4) the theory of descent and modification, (5) theories about development. It is not my goal here to provide a critical assessment of these various possibilities. Rather, what I would like to do is to sketch an alternative.

This alternative consists of two movements. The first movement will be to grant biological functions an important role in the individuation of behavioral traits. However, the salient concept of function will not be that of proper functions, but a modified version of the concept of causal role functions (Section 3). The second movement will consist showing that the investigation of a behavioral trait can actually transform the explanandum. This will be shown on an actual example (Section 4).

15.3 Biological Functions

I have argued that a construal of functions in terms of selected effect functions is unfit for the task of trait individuation. As an alternative, I suggest a modified version of causal role functions (Weber 2005a, b). This account starts with Cummins's (1975) analysis according to which functions are such capacities that are capable of explaining a capacity of some containing system. The paradigm is the heart's

capacity to pump blood figuring in any adequate explanation of the circulatory system's capacity to transport nutrients, oxygen and blood cells through the body. According to Cummins, the pertinent capacity of the containing system is a matter of an interest-based choice to be made by the investigator. I have modified this account by suggesting that this system's capacity should be made dependent not on the investigator's interests, but on the role that the containing system itself plays in the self-reproduction of the *whole* organism (see Davies (2003) for a similar idea). I argue that this is what turns Cummins-functions into *biological* functions. Cummins-functions can be applied to any kind of system. But only biological systems are capable of self-reproduction. In order for self-reproduction to occur, an organism's functions must work together. The specific contribution that some organ's causal capacities make to self-reproduction makes will depend on what other organs do. For example, if there were subsystems of an organism that would use the heart's heat production towards something that itself makes a contribution of self-reproduction, then the heart would (also) have the function of producing heat. It is the place that such a causal capacity plays in a whole network that gives it its function (perhaps much in the way in which a linguistic expression's meaning is given by the inferential role that the expression plays in a network of other expressions, as claimed by inferentialists and semantic holists).

I have argued that introducing such a global constraint on a system of functions might make the interest-dependence vanish, provided that there is exactly one way of laying a network of cooperating functions over an organism. Of course, this is hard to prove; but I suggest that it might be possible by using a notion of maximal explanatory coherence (Weber 2005b).

I have argued in my (2005a) that it is such a concept of function that underlies much of experimental biology. I would now like to suggest that such a concept also underlies the individuation of behavioral traits, at least in some cases. To demonstrate this, I shall finally turn to a real example.

15.4 Example: Social Behavior in *C. elegans*

The example comes from research on the soil nematode *Caenorhabditis elegans*. This is a tiny worm, about 1mm long, which feeds on bacteria. Even though its nervous system is extremely simple, consisting of only 302 neurons, this roundworm exhibits a form of social behavior: If placed on a bacterial lawn (which is their favorite food), the worms clump together for feeding. The functional significance of this is not entirely clear until this day. It seems that this behavior is somehow regulated, as it does not occur when the bacteria are spread (as opposed to forming a lawn) or when there are no bacteria at all. Furthermore, the worms do not merely move about randomly and then get stuck to each other. It seems that they move towards each other and respond to the vicinity of clumps by slowing down. What is striking is that there exist strains of the worm that do not clump together at all. Worm scientists Mario de Bono and Cornelia Bargmann showed that this difference

can be causally attributed to a single gene, dubbed *npr-1*, for which there is natural variation (de Bono & Bargmann 1988). The gene encodes a receptor for a small peptide that showed a high sequence similarity to a protein also found in other animals including humans: the so-called neuropeptide Y. This peptide is a neurotransmitter, in other words, it is involved in the transmission of signals between neurons. Most interestingly, it seems to be involved in the regulation of food intake in mammals, including humans.

What lessons can we draw from this example? The most important one, in my view, is this:

> *The classification of this phenomenon as* behavior *already involves bringing it under functional concepts. If the worms would just move around randomly and get stuck to each other, it would not be classified as behavior but as some sort of surface adhesion phenomenon that interferes with the worms' mobility. To speak of* behavioral *phenomenon implies the existence of a specific regulatory mechanism that responds to certain environmental stimuli (here: bacterial lawns and conspecific worms) and produces a functional response.*[2]

By bringing the clumping of worms under the concept of behavior, the biologists are not yet committed to what the actual function of this regulatory mechanism is. In order to attribute a specific function, more knowledge is required, at least according to the account of functions that I have briefly summarized in the last section.[3] All that is involved in seeing a behavioral trait here is that there exists some hypothetical regulatory mechanism that has *some* biological function.

So far, it is not clear that we are already dealing with a well-individuated trait. But at the very least, it is distinguished from other kinds of phenomena that might look similar where animals clump together for purely accidental reasons. But it might not yet be conceptually distinguished from other forms of animal clumping.

Another question is what warrants the classification of this behavior as 'social'. It should be quite obvious that it is social in quite a weak sense, namely in the sense that the worms regulate their feeding behavior such as to functionally respond to the presence of conspecifics. More elaborate forms of social behavior will involve recognition and differential responses to different individuals (which enables certain solutions to game theoretic situations like the infamous 'tit-for-tat' solution to the prisoner's dilemma). Even more elaborate forms will involve a theory of mind, self-consciousness and, as some people think, collective intentionality. But my point here is that the classification of certain phenomena as *behavior* in the first place involves the concept of biological functions rather than an intentional stance or any of the other options that we have discussed. Thus, the description of behavioral phenomena is sometimes laden with functional notions.

---

[2] Some philosophers such as (Millikan 1984) would already see the operation of such a control mechanism as a simple form of intentionality (provided that it is an evolutionary adaptation), but that's controversial, to say the least.

[3] It is not necessary to understand all of an organism's functions on my systemic account. It is enough to know some of the next nodes in a function graph. Here, in all likelihood the most salient systems capacity will be foraging for food.

I do not want to make any strong claims to the effect that this is always the case, or that this is the One Right Way of individuating behavioral traits. My reasons for this are not so much a general pluralistic stance, but another fundamental insight into the classification and explanation of phenomena in science. This insight is found, along with the general idea of theory-ladenness, in Paul Feyerabend's classic article "Explanation, Reduction, and Empiricism" (Feyerabend 1962). It is time to put a spotlight on the salience of this insight for the practice of biology, especially behavioral biology.

The insight is that scientific explanations of a phenomenon do usually not leave the explanandum phenomenon the same. In other words, the following picture will be rejected: That biologists can first pick out some pattern or phenomenon for explanation and then provide an explanation for that phenomenon, for example, some mechanism or some adaptive evolutionary story. Rather, the phenomenon to be explained is *transformed* in the very attempt of explaining it.

In our example from worm biology, the inquiry into the underlying mechanisms of the "social feeding" behavior in *C. elegans* made the biologists realize that they only had the elephant's trunk (or perhaps its tail, or …) when they started to investigate the phenomenon. The following additional findings from the worm study mentioned should make this clear.

De Bono and Bargmann used the isolated *npr-1* gene to determine in which strains it was present and absent, respectively. They compared 17 different strains, 15 of which were natural isolates from different parts of the world. These strains fell into two groups: those that exhibit social feeding and those that do not. Remarkably, this variation could be fully accounted for by a single amino acid substitution in the *npr-1* gene. The distribution of these alleles provided reasons for thinking that this mutation arose only once in evolution. No matter which allele was ancestral, both alleles were not only able to be maintained, but to actually spread to different parts of the world and quite different habitats. In some locations, both kinds of strains (social and solitary) were found coexisting. This is strong evidence for the hypothesis that both variants are *adaptive* under at least some conditions. Thus, the so-called "solitary" are not somehow defective; they merely have a different reaction norm to some environmental cues.

It should also be mentioned that when bacteria do not form a lawn but are so scarce as to become growth-limiting, all strains exhibit the clumping, and in the total absence of bacteria all of them move around solitarily. There are also indications that different worms respond differently to different bacterial species, which they may recognize by the chemical profiles of their metabolites. De Bono and Bargmann conclude from these findings: "It is likely that the behaviors we observed represent a narrow view of a more complex behavioral choice."

The "complex behavioral choice" alluded to by the worm biologists may be an intricate regulatory mechanism that chooses between different behavioral repertoires depending on the availability of and quality of food and the density of worms. An important piece of evidence for this hypothesis is the fact that the neuropeptide receptor encoded by the *npr-1* gene seems to be involved in the regulation of food intake and metabolism in a wide variety of animal species from extremely remote taxonomic

groups. It is also thought that, in the worms, it activates the formation of so-called *dauer*-forms of the worm, which is a dormant state that allows the worms to survive starvation. There also seem to be connections to the pheromone response.

The case thus allows us to draw another lesson:

> *As a result of the investigation, the behavioral phenomenon that the scientists want to explain changed: It started as a functional response to bacterial lawns and other worms, namely clumping. Now we are dealing with a more complex phenomenon that includes not just clumping but also swarming, the onset of reproduction and the formation of the dormant state. What seemed to be different phenomena at first are now treated as one phenomenon. What unified them was a combination of causal knowledge, extrapolating gene-function relations from other organisms, knowledge about biogeography and the autecology of the worms, and the pursuit of other questions raised by the initial findings.*

None of the initial findings really reveals anything about the underlying mechanism of the worms' behavioral response to the availability of food and the presence of conspecifics. Perhaps even more remarkable is the fact that it is not yet entirely clear what the functions of the initially observed clumping behavior actually are. One family of hypotheses says that the function of clumping must have something to do with the fact that it creates a microenvironment with a reduced oxygen level, perhaps to avoid oxidative stress (Gray et al. 2004). Alternatively, it might be a side-effect of the animals seeking a low-oxygen environment because that's where the food normally is (anaerobic bacteria). That way, they might also end up in clumps. So it is not clear that the clumping itself has a function.

This is not in contradiction to my earlier claim that to classify the clumping phenomenon as behavior involves an application of the concept of function. The hypothesis that generates this classification is weaker, perhaps something like this: There exists a regulatory mechanism M with respect to a function F that explains the occurrence of the worm clumps under conditions C. Note that regulatory mechanisms are necessarily connected to some function, otherwise they would not be *regulatory* mechanisms. Regulation always needs a set of *goal states*, which can only be states that have some functional significance for the organism (on pains of applying the notion of regulation in its original, intentional sense; see (Canguilhem 1988)).

15.5 Conclusions: A Plurality of Stances?

If the modification of behavioral phenomena by the very attempt of explaining them occurs regularly, we might be able to give a more pluralistic answer to our question of what is a good way of individuating behavioral traits or, more generally, for carving out behavioral phenomena. There may not be One Right Way of doing this. If the initial classification will change anyway as the underlying causes are unveiled, perhaps it is perfectly all right to start the classificatory task with folk psychology, or with some other stance. But my suspicion is that, as they learn more about a phenomenon that was initially picked out under the intentional stance, biologists will quickly move towards using functional concepts for classifying behaviors, as I

demonstrated it for the worm case. Folk psychology may have a great predictive value, when it comes to naturally classifying the characteristics of organisms it seems wiser to use functions. For what could be more biological than that?[4]

Such an account of trait individuation is needed to complement existing accounts of explanation in neuroscience, such as Carl Craver's (2007) ground-breaking account. Craver gives an excellent analysis of what it means to explain some (neuro-) biological phenomenon. A phenomenon consists of the regular behavior of some biological entity, and explaining it consists in exhibiting a mechanism that produces this behavior. Craver brilliantly analyzes the conditions such a mechanism must fulfill. However, he doesn't say much about what constitutes a phenomenon that is worth explaining in the first place. He does show that mechanisms sometimes form inter-level hierarchies (see especially Chapter 5). These hierarchies often top-off at the level of behavior and bottom-out at the level of atoms. But what makes biologists pick some top-level behavior as their explanandum in the first place? To this question, I have tried to give an answer here. It should be noted that the significance that functionally individuated behavioral traits bestow on the mechanisms that explain them trickles down the inter-level hierarchies that are characteristic of mechanistic explanation in neuroscience and elsewhere in biology.

Acknowledgements I wish to thank Daniel Sirtes, Katie Plaisance and Thomas Reydon for helpful comments.

References

de Bono, M. & Bargmann, C. I. (1998): "Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in C. elegans." *Cell* 94: 679–89.

Brigandt, I. (2007): "Typology now: homology and developmental constraints explain evolvability." *Biology and Philosophy* 22: 709–725.

Canguilhem, G. (1988): *Ideology and Rationality in the History of the Life Sciences*. Cambridge (MA): MIT Press.

Craver, C. (2007): *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Davies, P. S. (2003): *Norms of Nature: Naturalism and the Nature of Functions*. Cambridge (MA): MIT Press.

Dennett, D. C. (1989): *The Intentional Stance*. Cambridge (MA): MIT Press.

Feyerabend, P. K. (1962): "Explanation, Reduction and Empiricism." In *Scientific Explanation, Space, and Time (Minnesota Studies in the Philosophy of Science, Vol. III)*, Minneapolis: University of Minnesota Press, pp. 28–97.

Gray, J. M. et al. (2004): "Oxygen sensation and social feeding mediated by a *C. elegans* guanylate cyclase homologue." *Nature* 430: 317–322.

[4] I don't mean to suggest that this is a general solution to the problem of trait individuation. As Ken Waters has convinced me, there is no general solution to this problem. Trait individuation depends on context. For example, sometimes biologists will pick traits precisely because they have no function and are thus neutral with respect to fitness and natural selection. I wish to thank David Sloan Wilson for bringing this kind of example to my attention.

Griffiths, P. E. (2006): "Function, Homology, and Character Individuation." *Philosophy of Science* 73: 1–25.
Griffiths, P. E. (2009): "In What Sense Does "Nothing Make Sense Except in the Light of Evolution"?" *Acta Biotheoretica* 57: 11–32.
Kitcher, P. (1985): *Vaulting Ambition: Sociobiology and the Quest for Human Nature*. Cambridge (MA): MIT Press.
Kornblith, H. (2002): *Knowledge and its Place in Nature*. Oxford: Clarendon Press.
Love, A. (2007): "Functional homology and homology of function: biological concepts and philosophical consequences." *Biology and Philosophy* 22: 691–708.
Millikan, R. G. (1984): *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge (MA): MIT Press.
Rosenberg, A. (2006): *Darwinian Reductionism*. Chicago: The University of Chicago Press.
Weber, M. (2005a): "Holism, Coherence, and the Dispositional Concept of Functions." *Annals in the History and Philosophy of Biology* 10: 189–201.
Weber, M. (2005b): *Philosophy of Experimental Biology*. Cambridge: Cambridge University Press.

Chapter 16
From Reactive to Endogenously Active Dynamical Conceptions of the Brain

Adele Abrahamsen and William Bechtel

16.1 Introduction

Observe a living organism—from a bacterium to a fellow human being—and you see an endogenously active system. Introspect and you will observe, as did William James, a continual flow of thoughts. If pressed, most neuroscientists and psychologists will acknowledge that neural systems are endogenously active, generating activity even in the absence of any stimulus. But for decades they have tended to disregard this key characteristic, pursuing programs of research in which they present discrete stimuli in structured tasks designed to focus on the neural and behavioral effects of experimental manipulations. In this paper we contrast this perspective (which we call *reactive*) with a dynamic perspective emphasizing endogenous activity. In neuroscience these have a long history of coexistence, but only recently has the endogenous perspective become less isolated as powerful new strategies for pursuing it have begun changing the overall research landscape. We provide a selective tour of this history from the vantage point of the new mechanistic philosophy of science, in which we highlight the interplay between basic and dynamic mechanistic explanation.

The reactive perspective has deep historical roots and is widely pursued in both neuroscience and psychology. Neuroscientists, following a tradition initiated by the British neurophysiologist Charles Scott Sherrington (1923), commonly treat the brain as a reactive mechanism in which sensory input initiates processing along a

A. Abrahamsen (✉)
Center for Research in Language, University of California,
San Diego, 9500 Gilman Drive, 0526 La Jolla, CA 92093-0526, USA
e-mail: adele@crl.ucsd.edu

W. Bechtel
Department of Philosophy, Center for Chronobiology, and Interdisciplinary
Program in Cognitive Science, University of California, San Diego,
9500 Gilman Drive, 0119 La Jolla, CA 92093-0119, USA
e-mail: bill@mechanism.ucsd.edu

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*,
Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_16,

neural pathway, terminating in a motor response. One of their core techniques is to present stimuli and record neural responses in the brain area of interest; another is to manipulate neural activity and record motor responses. Psychologists more often target the whole organism, presenting stimuli and recording behavioral responses without determining the intervening neural activity. Most North American psychologists treated the gap between stimulus and response as a black box during the behaviorist era and as information processing thereafter, but for a minority (yesterday's psychobiologists and today's cognitive neuroscientists) the gap is filled by neural activity that should be investigated.

There is no doubt that this reactive framework has been enormously productive for both neuroscience and psychology. It has served to identify many of the parts and operations within the mechanisms responsible for cognitive phenomena, as we will show in the case of vision in section 16.2. But there are also indications of its limitations. One is the considerable variability researchers commonly observe in both behavioral and neural responses. While this variability tends to be construed as noise to be eliminated from experimental data by averaging across time and subjects, if examined rather than concealed it can reveal compelling signatures of endogenous activity.

Laboratory research on endogenous activity, while relatively sparse, has historical roots nearly as deep as those of the reactive approach. Most notably, Thomas Graham Brown (1914) studied neural mechanisms for motor behavior in decerebrate cats alongside Sherrington in his laboratory at Liverpool from 1910 to 1913—but arrived at quite different conclusions. Sherrington was commited to a sequential reflex mechanism, by which peripheral input (e.g., to the cat's feet when placed on a moving treadmill) produced a sequence of neural signals (to the spine, within the spine, and out to flexor and then extensor muscles). Each cycle of stepping resulted in renewed input (sensory feedback) and hence ongoing, rhythmic stepping movements. Brown discovered that he could obtain similar rhythmic stepping even after isolating the spinal cord from afferent (peripheral) input by cutting the dorsal root nerves. This impressive demonstration of endogenous control led him to propose a neural mechanism that later would be recognized as the first description of a *central pattern generator—central* because the key components were in the spine (sensitive to but not dependent on peripheral input); *pattern* because it produced an ongoing oscillatory pattern (observed as rhythmic stepping, in which flexion alternates with extension); and *generator* because this mechanism could initiate production of the pattern. More specifically, Brown proposed what would now be described as two coupled networks of spinal neurons—one for flexion and one for extension—which oscillated in inhibiting each other's activity. However, Sherrington resisted distraction from his own pursuit of a sequential reflex account of motor behavior;[1] as his

[1] Early on Sherrington (1913, p. 207) acknowledged that Brown's view "demands careful attention" but demurred on grounds that his own line of explanation "would be led too far afield by its consideration now."

reactive approach became more entrenched, Brown was increasingly marginalized (for discussion, see Stuart & Hultborn, 2008).

It was a half-century before Brown's emphasis on endogenous activity was revived by researchers who converged on central pattern generators as the explanatory mechanism of choice for a variety of rhythmic motor behaviors. Wilson and Wyman's (1965) landmark account of flight in locusts was followed by identification of central pattern generators in the brain stem and spinal cord for such activities as walking, swimming, respiration, and circulation (Grillner, 2003). Almost another half-century passed before neuroscientists investigating sensory processing and central cognition turned their attention to endogenous activity in cerebral cortex and were rewarded with multiple streams of evidence from single cell recording, EEG, fMRI, and other techniques. We introduce some of this evidence in section 3, and emphasize that the resulting conception of the brain as endogenously active poses a profound challenge to the purely reactive perspective that has dominated much of psychology as well as neuroscience.

The slow spread of the endogenous perspective is unsurprising considering the history of other sciences. Max Planck (1949, pp. 33-34) famously suggested that "A new scientific truth does not triumph by convincing its opponents … but rather because its opponents eventually die." He exaggerated for effect, presumably, but it is not uncommon for scientists to bemoan delays in the uptake of new approaches. Less remarked upon is the delayed impact of changes in the sciences on *philosophy* of science. This is a young field (its first journal, *Philosophy of Science*, began publication in 1934), and it has been slow to move beyond its initial roots in twentieth-century physics to incorporate quite different influences from the biological and cognitive sciences. We suggest that this delay has been excessive and detrimental to its own development as a field of inquiry. Philosophers of science did not even recognize the dominant mode of explanation in these sciences—*mechanistic explanation*—until pioneering work by William Wimsatt, who pointed out that "At least in biology, most scientists see their work as explaining types of phenomena by discovering mechanisms …" (Wimsatt, 1976, p. 671). His influence on a cohort of students gave rise in the 1990s and especially after 2000 to the *new mechanists*, who have drawn on biology and cognitive science rather than physics in constructing a new mechanistic philosophy of science (Bechtel & Richardson, 1993/2010; Bechtel & Abrahamsen, 2005; Glennan, 1996, 2002; Machamer et al., 2000; Thagard, 2003; Wimsatt, 2007).

Recently we have argued that further developments in these sciences—especially computational modeling of the dynamics of cognitive and neural mechanisms—require extending the mechanistic framework to incorporate *dynamic mechanistic explanation* (Bechtel & Abrahamsen, 2010, 2011). Thus, in what follows we begin by distinguishing between basic mechanistic explanation, in which target systems are treated as reactive mechanisms, and dynamic mechanistic explanation, which has the resources to characterize endogenous as well as reactive activity and to do so with greater precision (section 16.1). We then discuss investigations of brain mechanisms in particular, contrasting those that exemplify the reactive perspective (section 16.2) with those targeting endogenous activity (section 16.3). We consider

certain implications of the endogenous perspective for how we understand cognitive activity (section 16.4). Finally, we return to the philosophical understanding of dynamic mechanistic explanation and how it can illuminate research that takes an endogenous perspective on the brain (section 16.5).

16.2 Two Conceptions of Mechanism

The new mechanists have primarily focused on basic mechanistic explanation, in which investigators *decompose* a system into a set of component parts, each of which performs one or more operations, and *recompose* it by figuring out the spatial organization of the parts and temporal/causal organization of the operations (Bechtel & Abrahamsen, 2005, 2009). The idea is that going down to a lower level provides the most useful explanation of how the system's activity generates a phenomenon of interest.

What makes these explanations "basic" is that the accounts of organization are mostly qualitative rather than quantitative. Thus, a typical *structural decomposition* into parts would be recomposed into a spatial ordering (e.g., the spine's lumbar vertebrae are designated as L1, L2, L3, L4, L5) or a schematic layout (e.g., a eukaryotic cell is depicted as a membrane enclosing one nucleus and numerous organelles in cytoplasm). A typical *functional decomposition* into operations would be recomposed most simply into a temporal ordering in which the product of one operation is operated upon by the next (e.g., the chain of biochemical reactions comprising intermediary metabolism). The act of constructing a basic mechanistic explanation of a phenomenon is complete[2] when the investigator can specify which parts perform which operations. This task of *localization* sometimes is integral to the discovery process, but may instead be deferred (pending development of necessary tools, for example, the electron microscope). Once achieved, a well-supported basic mechanistic account is an important research milestone.

One example, discussed at greater length in section 16.2, is a pathway through the visual system that is responsible for the phenomenon of object recognition. Parts of the pathway have been identified in the retina, lateral geniculate nucleus (LGN), occipital lobe (visual areas V1, V2, and V4) and temporal lobe. A very simplified version of the basic mechanistic account has each of these parts in turn performing one or more operations on the output of the preceding operation:[3] the retina represents

[2] "Complete" does not imply "final." An important role for such an account is to provide a framework for further research that elaborates and corrects it and eventually may replace it. Darden and Craver (2002) referred to incomplete accounts as *mechanism sketches* and traced how two different sketches for protein synthesis in the 1950s were gradually modified and brought together in a basic mechanistic account that was completed (but not final) in the 1960s.

[3] The outputs of operations arguably are (a special class of) parts. This is clearer in the case of biochemical reaction pathways, in which the outputs are molecules, than in the case of neural pathways, in which the most useful characterization of the output often is an abstract representation.

a stimulus object topographically, the LGN modulates or gates the representation, V1 extracts several types of features, V2 analyzes contours, V4 analyzes form and color, and inferior temporal cortex performs higher-level, integrative operations that yield a percept recognized as a particular type of object.

It is possible to find ordered components with no beginning or end: the beads in a bracelet, the bases in loops of mitochondrial DNA, people circle-dancing, and so forth. But something in us likes an ordering to be not only invariant but also bounded and unidirectional. Scientists are no exception, showing a preference for basic mechanistic accounts in which operations are ordered with a beginning and an end. We will reserve the term *sequence* for this type of organization in time or space.

Sequential organization is especially prominent in the definition of mechanism offered by Machamer, Darden, and Craver (2000):

> Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions.

Note that their terms "entities" and "activities" are equivalent to parts and operations respectively. (We use the term "activity" to refer to the overall behavior of a mechanism as distinguished from the component operations.)

What we wish to highlight here is their explicit stipulation of a beginning and end. In the case of protein synthesis, as discussed by Darden and Craver (2002), this seems appropriate. The start or set-up conditions are not itemized, but they would seem to include the availability of ribosomes and amino acids in the cytoplasm, the availability of several kinds of RNA where needed, and (crucial to initiation of the process) the appropriate RNA polymerase coming into proximity with the DNA segment that codes for the protein. Highlights of the "regular changes" (sequence of operations) enabled by those conditions include the RNA polymerase unzipping and transcribing the DNA into a complementary mRNA base sequence, the transport of the mRNA into the cytoplasm, each codon (sequence of three bases) on the mRNA forming a weak hydrogen bond with an appropriate tRNA, guiding its attached amino acid to form a peptide bond with the previous tRNA's amino acid. These last two operations are repeated for each codon in turn, hence synthesizing the protein one amino acid at a time. When the last peptide bond has been formed, the key termination condition of the protein synthesis mechanism has been satisfied and it stops.

This case and the definition itself exemplify the reactive perspective, insofar as a sequence of activity is initiated by satisfaction of set-up conditions and ends with satisfaction of termination conditions. Many cases in biology are less good exemplars. Machamer, Darden and Craver (2000, p. 11) acknowledged that set-up conditions "may be the result of prior processes" but justified requiring them on grounds that "scientists typically idealize them into static time slices taken as the beginning of the mechanism." They further noted that "the bulk of the features in the set-up … are not inputs into the mechanism but are parts of the mechanism." A focus on internal components is indeed a strength of any mechanistic account, as contrasted with purely functional accounts of input-output relations. Nonetheless, set-up and termination conditions misleadingly suggest that the system targeted for explanation is passively

awaiting initiation of activity that, once underway, reaches a stopping point. Since biological mechanisms typically function continually, what investigators have treated as start-up conditions are better viewed as perturbations to ongoing endogenous activity.

To build a mechanism capable of sustained, endogenous activity a minimal first step is to allow at least one operation posited as later in the sequence to feed back on operations posited as earlier. Adding even a single negative feedback loop to an otherwise feedforward mechanism can produce ongoing dynamic activity, most notably oscillations. In a mechanism with appropriately weighted feedback and openness to energy, these oscillations can be regular (exhibiting, for example, a stable frequency of 10 Hz: 10 cycles of rise and fall per second) and self-sustained (i.e., not dampen to a steady state over time; see Goodwin, 1965). Many actual biological systems are well-characterized by a mechanistic account in which positive as well as negative feedback loops are added to a sequential backbone of operations. Carbohydrate metabolism, for example, is achieved by a chain of reactions that begins with glycogen and ends with pyruvate. At least *in vitro*, sideloops regulate the system such that the amount of pyruvate produced oscillates with a frequency of about one cycle per minute. (Examples from metabolism are further discussed in Bechtel & Abrahamsen, 2011). It should be note that this glycolytic oscillator is harmonic—the amount of pyruvate changes at a constant rate. Neural oscillators, in contrast, are relaxation oscillators—also regular, but with pulselike activity (spikes) against a low-activity background.

The addition of feedback loops is not a trivial adjustment: it is a key means of moving beyond a purely sequential conception of mechanism to a more dynamic conception. Our own earlier characterization of mechanism gestured in the direction of dynamics in referring to "orchestrated functioning of the mechanism," but we recently augmented it as follows:

> A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism, **manifested in patterns of change over time in properties of its parts and operations**, is responsible for one or more phenomena. (Bechtel & Abrahamsen, 2010).

The phrase in boldface was added to explicitly cover a broader range of mechanistic accounts offered by scientists: not only sequential accounts but also those in which the parts and operations are organized so as to generate endogenous oscillations or other interesting dynamics. This led us directly to consideration of how the dynamics might be characterized. Those scientists who emphasize laboratory research tend to look first to their data for this, as in the example above of pyruvate concentrations oscillating at about one cycle per minute. They then attend to what operations and organization might be responsible for the dynamics observed. (In this example, they were able to show that the feedback loops involving one particular enzyme early in the reaction pathway were crucial.). Another approach is important as well. A computational biologist can use mathematical tools to construct a computational model that is explicitly grounded in a mechanistic account. The model offers a precise (and potentially falsifiable) characterization of the mechanism's

dynamics. The variables in the model are more or less directly aligned with properties of parts and operations in the mechanistic account.[4] A computational modeler can capture various oscillatory patterns produced by biological mechanisms and then determine whether there are realistic values of the model's parameters for which the oscillations are self-sustaining. This is of particular interest when endogenous oscillations are claimed. In brief, dynamic mechanistic explanation encompasses both laboratory-based and computational research. Ideally (but not usually) these are carried out collaboratively.

Endogenously active mechanisms typically can be affected by exogenous inputs, but how they respond to these depends upon their current endogenous state, which may vary systematically or irregularly over time. It is important to understand the underlying endogenous behavior of the mechanism in order to understand how it responds to perturbations. The situation becomes even more important when the endogenous behavior of one mechanism is affected by endogenous activity in other mechanisms with which it is dynamically linked (e.g., within an organism's body, or within an ecological network in which the organism is behaving). We return to the discussion of how understanding endogenous activity is relevant to understanding the responses of mechanisms to exogenous inputs in section 16.4.

16.3 Traditional Experimental Approach to the Brain

Although we will focus on shortcomings of the reactive conception of mechanism, research programs grounded in that conception have been enormously productive. Indeed, researchers inclined to dynamic mechanistic accounts typically are not in a position to advance serious proposals about the integrated, dynamical behavior of a mechanism until researchers pursuing the reactive approach have provided a rich understanding of the parts and operations within it. Further, a premature emphasis on the whole integrated system can be counterproductive. Brain research in the 18th-20th centuries was marked by ongoing tension between mechanists who sought to localize specific mental functions in specific areas of cerebral cortex and holists who argued that the activity required for particular functions was broadly distributed. For example, Ferrier's (1876) localization of a number of sensory and motor functions based on ablation and stimulation experiments in monkeys were countered in the 1880s by Goltz's claim that such functions were preserved in dogs with extensive ablations. Post-mortem examinations supported Ferrier, and some degree of localization of sensory and motor functions became widely accepted (see Finger, 1994, pp. 54-56). The debate continued with respect to localization of intellectional

[4]Most simply there is direct correspondence between variables and properties; for example, *c* may denote the concentration of the product of a reaction and *r* the rate of the reaction. Sometimes, though, it is a more complex expression in the model (e.g., a variable multiplied by a scaling parameter) that corresponds to a property in the mechanistic account (e.g., the rate of a reaction).

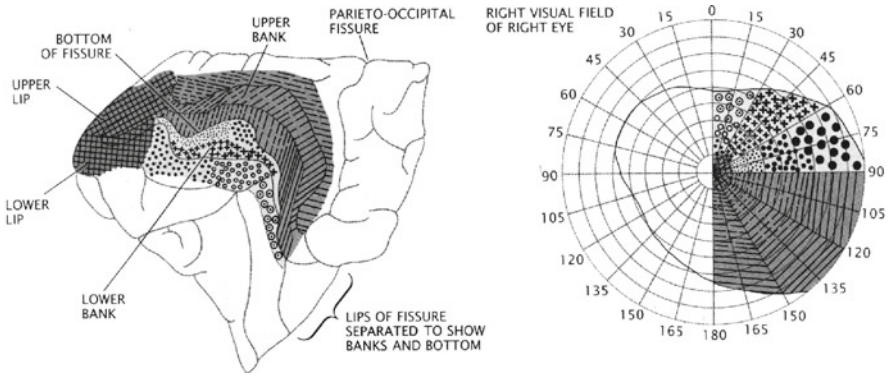

Fig. 16.1 Goldon Holmes' (1918) map indicating how different areas of the right visual field (the right side of the image on the right) project onto particular regions of the left occipital lobe (shown on the left)

functions, such as memory and reasoning, in remaining areas of cerebral cortex. The claim that what matters is the amount of tissue destroyed in ablations, not which tissue, received its best-known expression in Lashley's (1929) "law of mass action" and goes back at least to Flourens (1824). But whatever its merits, this holistic view that large parts of cortex act as a distributed, integrated system did not generate a positive program of inquiry. History has adjudicated that it was the researchers pursuing localization of functions in the brain who achieved results that could be built upon, leading to our current mechanistic accounts of how the brain performs cognitive tasks.

The reactive perspective on the brain is well exemplified in research on visual processing. In the late 19th century it was established that a key area for visual processing was a region of the occipital lobe distinguished by its pattern of striation (hence, *striate cortex*; now called *V1*). Neural pathways were traced from the eyes to this region, and lesions to it produced visual deficits in both humans and animals. Salomen Henschen (1893) determined that damage to particular regions within striate cortex resulted in blindness to specific parts of the visual field, leading him to propose that striate cortex was organized in terms of a topographic map of the visual field. While the idea of topographic maps has endured, the particular map Henschen proposed turned out to be inverted from the ones subsequently developed by Tatsuji Inouye (1909) and Gordon Holmes (1918) based on correlations between visual deficits and brain damage in soldiers (Fig. 16.1).

Later in the 20th century, electrophysiologists developed techniques for recording activity from individual neurons in response to carefully selected visual stimuli. Single-cell recording enabled researchers to determine not only *where* each neuron was responsive (yielding much finer-grained topographic maps) but also *how* it responded. It turned out that the topographic mapping strategy was relied upon in multiple regions—retina, LGN, striate cortex, and beyond—but that none produced a simple pixel-like representation. Each region had its own distinctive

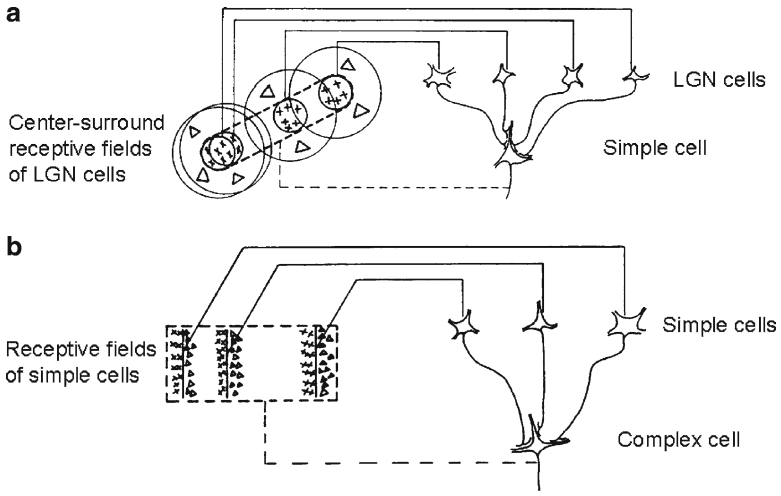

Fig. 16.2 Hubel and Wiesel's (1962) proposed simple and complex cells in striate cortex. (**a**) Center-surround cells in LGN that detect spots are connected to simple cells that detect location- and orientation-specific bars. (**b**) Simple cells are connected to complex cells that detect orientation-specific bars, especially those moving in a particular direction

computations awaiting discovery. In pioneering work, Kuffler (1953) found that retinal ganglion cells in cats respond best to light spots on dark backgrounds or dark spots on light backgrounds. He proposed that these *center-surround cells* processed stimulation at the center of their receptive fields and stimulation at the immediately surrounding area in an antagonistic manner (responding maximally if one was dark and the other light).

Hubel and Wiesel demonstrated that the center-surround design is replicated in cats' LGN (the lateral geniculate nucleus of the thalamus), but failed to find it in striate cortex. The edge of a misoriented slide sparked their realization that striate neurons respond not to spots but to linear stimuli (light or dark bars or edges). They proceeded to differentiate simple cells (those responsive to a bar at a specific locus and orientation) from complex cells (which respond to bars anywhere within a broader area of the visual field but especially to those moving in a preferred direction). As illustrated in Fig. 16.2, Hubel and Wiesel (1962) proposed that a bar in the visual field spanned the receptive fields of several center-surround cells in the LGN. These were connected to at least one simple cell that detected their joint activation, and multiple simple cells (ideally with closely adjacent receptive fields) in turn were connected to each complex cell. A complex cell was especially responsive if the simple cells connected to it were triggered in sequence by a bar moving in the appropriate direction. The simple and complex cells can be thought of as engaging in feature detection at two levels. In a subsequent study Hubel and Wiesel replicated these findings of simple and complex cells in monkeys but also reached the conclusion that their activity "represented a very elementary stage in the handling of complex forms" and must be followed by further processing "at later stages in the visual

path" (Hubel & Wiesel, 1968, p. 242). They began referring to striate cortex as V1 (visual area 1) and the areas involved in later stages of processing as V2, V3, V4 and MT (medial temporal area)—clearly embracing the conception of the visual system as a sequential processor, with its sequence of operations initiated by presentation of a visual stimulus and culminating in a percept.

Hubel and Wiesel's strategy found numerous applications in subsequent years as researchers inferred function from the classes of visual stimuli that drove responses in specific regions of occipital, temporal, and parietal cortex. For example, neurons in area V4 were found to achieve color constancy: in addition to responding to variations in the incoming wavelength due to changes in the color of an object (like V1) they compensate for variations due to changes in its illumination. Similarly, neurons in area MT were discovered to respond to the perceived direction of movement of complex stimuli, whereas those in V1 presented with the same stimuli respond only to the direction in which components of the stimuli move across the visual field. Subsequent research revealed regions in the temporal lobe that respond to specific classes of objects and regions in the parietal lobe that respond to their spatial location (each with distinct pathways from subareas of LGN, V1 and so forth). By the 1990s over thirty different brain areas in the macaque had been identified as engaged in processing visual stimuli, and for many of these areas research pursuing the approach just described succeeded in determining the specific features of stimuli that evoked a response (van Essen & Gallant, 1994). Each successive brain area was regarded as operating on the products generated in previous areas to extract new information about the visual stimulus.

By any measure, this research endeavor that treated the visual system as reacting to visual stimuli was extremely successful (for a detailed account of this century of research, see Bechtel, 2008). There are reasons to suspect, however, that the resulting explanatory accounts may be incomplete and require non-trivial revision. First, the approach assumes sequential processing of inputs by a succession of processing centers. But researchers have long known that in addition to forward axonal projections there are extensive backwards and collateral projections in this system (Lorente de Nó, 1938). Hubel and Wiesel had found that the neurons they recorded were organized into columns traversing the six layers of cortex and that neurons within a column responded to stimuli in the same part of the visual field. Forward, backwards, and collateral projections could be differentiated by the layers from which and to which they projected, which helped researchers uncover the complex pattern of connections between visual areas in the macaque. The findings were displayed visually in the well-known "subway map" diagram by Felleman and van Essen (1991), as shown in Fig. 16.3. The other indication that sequentially-based accounts require revision is that researchers have constantly confronted the problem that neural responses are highly variable. This variability is generally regarded as noise which needs to be removed so as to reveal the signal – but a very different research program emerges if instead it is taken to indicate that much more may be going on within the mechanism than is revealed by what is regarded as signal.

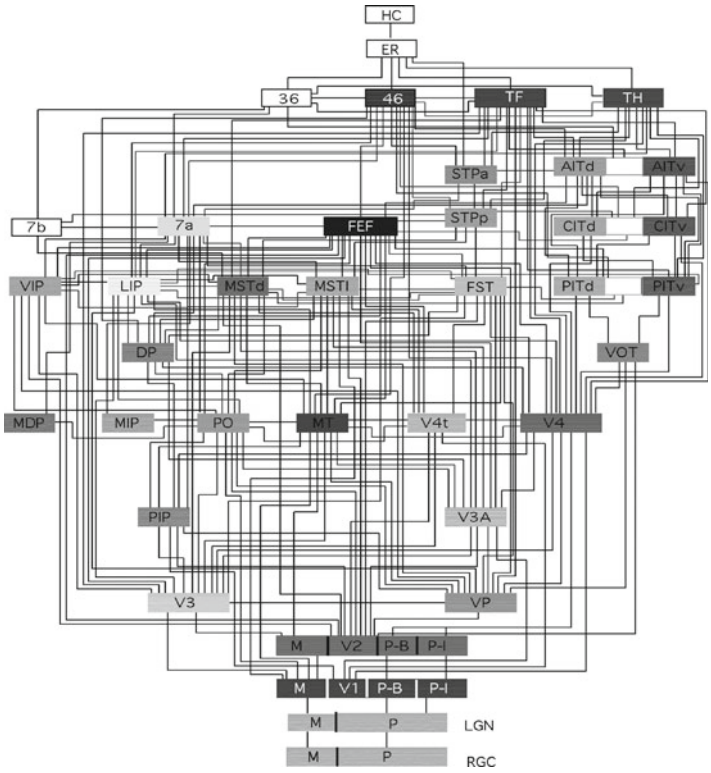

Fig. 16.3 Felleman and van Essen's (1991) representation of 32 cortical visual areas identified in the macaque and the known connections between them. In most cases the connections are bidirectional, with separate bundles of axons running between different layers for feedforward vs. feedback signals (not shown)

16.4 Reconceptualizing the Brain as Endogenously Active

The alternative conception of the brain as an endogenously active mechanism is being pursued by a growing vanguard of neuroscientists. They are rethinking brain dynamics, are redirecting the tools of their trade towards detection of endogenous activity, and are devising analyses that can describe that activity and tease out interactions with activity evoked by stimuli. We will discuss three key technologies in the order in which they began to be directed to uncovering endogenous activity in the brain. Since they differ in their temporal and spatial range and resolution, we conclude this section by asking how activities at multiple timescales might interrelate.

16.4.1 Electroencephalography (EEG)

The vision researchers discussed above implanted electrodes so as to record the activity of individual neurons, but there is an even longer tradition of inferring aggregate activity from electrodes placed on or into the scalp or (in animals or surgical patients) on the cortical surface. The difference in electrical potential between two electrodes fluctuates over time, providing a measure of electric currents in the brain with high temporal but low spatial resolution. In pioneering research with rabbits and monkeys, Richard Caton (1875) experimented with various placements of pairs of electrodes connected to a mirror galvanometer that represented the currents visually. Despite primitive tools, he made the first observations both of continuous spontaneous activity ("feeble currents") and of localized "negative variation" evoked by a stimulus.[5]

Caton's technique was reinvented more than once, but did not give rise to an ongoing program of research until psychiatrist Hans Berger adapted it to humans in the 1920s. Berger initially inserted needle electrodes into subcutaneous tissue, often one at the front and one at the back of the head, but found that he could obtain similar results with the less intrusive procedure of affixing lead foil electrodes to the scalp. With electrodes connected to a string or double-coil galvanometer that was attached to a recording apparatus, he could permanently capture oscillations in the current as lines on long strips of paper (with some delay since the recording involved a photographic process). In his first publication (Berger, 1929), he called this an *Elektrenkephalogramm* or, in English, an electroencephalogram (EEG), in recognition of existing electrocardiogram instrumentation which he had adapted. In patients and healthy individuals at rest with eyes closed, he repeatedly observed two distinct waveforms. In his next report (Berger, 1930), he coined the term *alpha waves* for the approximately 10 Hz oscillations that most intrigued him and the term *beta waves* for smaller, faster 20-30 Hz oscillations. Moreover, Berger discovered what was later called *alpha blocking*: when the eyes were opened alpha waves declined precipitously, leaving beta waves to predominate. Even with eyes closed, events in other sensory modalities or attention-demanding tasks such as mental arithmetic could produce this effect. For example, Fig. 16.4 shows the "striking change" from predominantly alpha to beta waves that Berger obtained by stroking his subject's hand with a glass rod.

Berger conducted extensive control studies to show that EEG oscillations were not due to artifacts but in fact provided a window on the brain's endogenous and

[5]One of Caton's objectives was to evaluate Ferrier's claims regarding localization of motor commands, and he reports (p. 278): "on the areas shown by Dr. Ferrier to be related to rotation of the head and to mastication, negative variation of the current was observed to occur whenever those two acts respectively were performed. Impressions through the senses were found to influence the currents of certain areas; e. g., the currents of that part of the rabbit's brain which Dr. Ferrier has shown to be related to movements of the eyelids, were found to be markedly influenced by stimulation of the opposite retina by light."

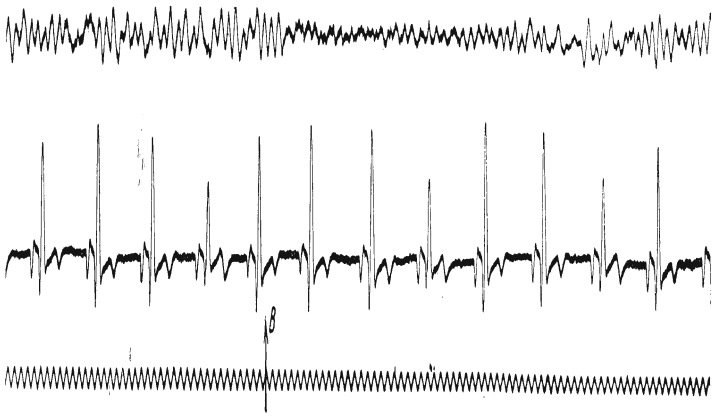

Fig. 16.4 In this eight-second extract of a recording made by Berger (1930), the upper line is a subcutaneous EEG. It shows three seconds of predominantly alpha waves that were blocked 0.27 seconds after he stroked the subject's hand with a glass rod (indicated by arrow "B" on the 10 Hz timing signal at the bottom). For at least the next two seconds the EEG shows lower-amplitude, higher-frequency beta waves, not alpha waves. The middle line is an electrocardiogram recorded simultaneously. Extracted from Fig. 16.5 in Gloor's translation (1969, p. 82) of Berger (1930)

evoked activity. (In Fig. 16.4, for example, the EEG was not obviously correlated with the electrocardiogram displayed beneath it.) Few were convinced until he attracted the attention of Edgar Adrian, a leading investigator in neurophysiology at Cambridge University. Adrian and Matthews (1934) were initially skeptical, based on their own recordings in rabbits and cats, but found compelling evidence in humans of the alpha rhythm and attributed it (p. 384) to "the spontaneous [synchronous] beat of an area in the occipital cortex which is normally occupied by activities connected with pattern vision."[6]

With Adrian's imprimatur, human EEG research attracted other pioneering investigators; within a decade, three additional rhythms had been investigated and named. The term *gamma rhythm* was proposed by Jasper and Andrews (1938) to designate frequencies above 30 or 35 Hz, but high-quality evidence for functionally

[6] Adrian and Matthews limited their focus to the alpha rhythm, which they called "the Berger rhythm," and characterized it as disappearing in task conditions. They further claimed that "a group of cortical cells in some part of the occipital lobe … tend to beat synchronously when they are undisturbed, but visual activity or widespread non-visual activity in the brain breaks up the rhythm by exposing the cells to a mosaic of excitations which makes synchronous action impossible. Berger, if we have interpreted him correctly, regards the waves as having a much wider and less specific origin, but the evidence as to localization is the only important point on which our results seem to differ from his" (p. 356). Later researchers confirmed a primary localization in the occipital lobe, but also found other origins and/or broader activity under some conditions. Others confirmed that, although alpha rhythms are not prominent in most animal species, Berger was correct in reporting that they were prominent in dogs as well as humans (see especially pp. 239 and 256 in the review by Shaw, 2003.)

distinguishing gamma from beta rhythms is much more recent and often involves evoked rather than endogenous activity. (Proposals regarding functions of evoked gamma activity have included object perception, cross-modal perception, feature binding, aspects of short-term memory, and even—focusing more on endogenous than evoked activity—consciousness and ongoing information processing.) The other two rhythms involved slower waveforms that Berger had already associated with brain lesions and sleep. W. Grey Walter (1936), in reporting EEG studies of awake humans with brain tumors, proposed the term *delta rhythm* for waves lower in frequency (and typically higher in amplitude) than the alpha rhythm. Later he designed an automated frequency analyzer and, deploying it on EEGs from a variety of patients, differentiated a primarily subcortical *theta rhythm* (4-7 Hz) from the slower (<4 Hz), primarily cortical delta rhythm (Walter & Dovey, 1944).

The subsequent years brought improvements in recording technologies (e.g., digital EEG in the 1960s) and in analysis of complex EEG waveforms. Methods generally proceed from the assumption that these waveforms can be decomposed into sinusoidal components of different frequencies. Even Berger had noted that irregularities in the alpha waves ("notches" on their descending limbs too small to see in Fig. 16.4) indicated that they were always mixed with the smaller beta waves. However, automated analysis as pioneered by Walter is far more revealing than visual inspection of EEG recordings. Since the 1960s computers have allowed efficient calculation of power density (a measure of amplitude) for each frequency or frequency band within a time window. Current variations on this method use a fast Fourier or wavelet transform of the EEG waveform. Herrmann, Grigutsch, & Busch (2005) provide an introduction to wavelet analysis and suggest the following as well-established frequency bands: delta (0-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), gamma (30-80 Hz). Additional, less standardized bands are sometimes specified by those taking advantage of recent advances in technology; examples include infraslow (0.01 - 0.1 Hz), very slow (0.25 - 0.5 Hz), and very fast (100-500 Hz).

One early application of EEG was in differentiating stages of sleep, with the first comprehensive proposal advanced by Loomis, Harvey, and Hobart (1937). Most recently (Silber et al., 2007), the American Academy of Sleep Medicine has identified five stages: wakefulness, rapid eye movement (REM) sleep, and three stages of lower-frequency, non-REM sleep (NREM 1, 2, and 3).

When people wind down at night, the beta and gamma waves of the cognitively active brain typically yield to more relaxed alpha waves, which in turn become mixed with yet slower theta waves as sleep approaches. By definition, an individual is awake (stage W) as long as alpha exceeds theta activity, and makes the transition to NREM 1 sleep (stage N1) when theta exceeds alpha activity. For those with little alpha activity, typical accompaniments such as slow rolling eye movements are counted instead. NREM 2 sleep (stage N2) is characterized by one or more half-second (or longer) episodes of high-amplitude patterned activity superimposed on a low-amplitude, mixed-frequency background. One pattern, the *sleep spindle*, is a train of rhythmic 11-16 Hz waves that increase and then decrease in amplitude, producing a spindle shape in the EGG. The other pattern, the *K-complex*, is a sharp negative wave immediately followed by a slower positive component. NREM 3

sleep (stage N3) begins when at least 20% of activity is *slow wave sleep*, restricted by definition to the lower end of the delta range (0.5-2 Hz, with peak-to-peak amplitude above 75 μV). Typically these slow waves rise to over 50% of EEG activity, treated as the threshold to a separate stage in older systems. Finally, REM sleep (stage R) has a complex definition emphasizing three characteristics that tend to co-occur: its namesake rapid eye movements (during which the most memorable dreaming can occur), low muscle tone, and low-amplitude, mixed-frequency EEG activity (usually predominantly theta as in stage N1, but alpha or sawtooth waves may be prominent). After its discovery in 1953, REM sleep was called *paradoxical sleep* because the sleeper could not move even though the brain and other systems were active.

The sequence of stages in a prototypical night begins with W, then N1, then four or five repetitions of N2-N3-N2-R, but variations are common. It should be kept in mind that the stages are rigorously defined in part to assure comparability across research laboratories; in an actual night's sleep the passage from one stage to another often is gradual or ambiguous. The dynamic character of sleep, and its relative isolation from environmental influences, make it a highly relevant context for investigating edogenous activity in the brain.

In summary, EEG research first showed its worth as a means of differentiating overall brain states. In the early decades it was most usefully applied to the discovery and characterization of stages of sleep in terms of endogenous alpha, theta, and delta rhythms. From Berger forward, researchers also recognized that a variety of cognitively active states were marked by faster rhythms in the beta and gamma ranges. They lacked tools for distinguishing between these states, however, as would be needed to move towards a brain-based account of cognitive operations and their orchestration in complex tasks.

This changed beginning in the 1960s (for sensory processing) and 1970s (for more complex cognitive processing). Instead of recording ongoing brain activity—primarily endogenous in origin—investigators presented carefully chosen exogenous stimuli and looked for systematic changes in the EEG pattern, especially within the first half-second or so following stimulus onset. The time-locked waveform is referred to as an *evoked potential (EP)*, *evoked response potential (ERP)*, or in cognitive investigations, *event-related potential* (also *ERP*).[7] This dramatically repositioned the EEG technology: now it could serve those who found a reactive conception of the brain most promising for rapid gains in knowledge. Since response to a stimulus was only one among many influences on the highly variable EEG waveform, however, it was essential to average the waveforms obtained over multiple trials in which similar stimuli were presented. Computer processing of digital EEG made this practical. In a classic experiment, Kutas and Hillyard (1980) presented subjects with 160 seven-word sentences in which the final word was either anomalous or semantically appropriate. Comparing the average ERPs, they discovered a robust negative deflection peaking approximately 400 milliseconds after onset of

[7]There was interest in such waveforms as far back as Berger (1920s) and Davis (1930s); what was new in the 1960s was powerful new tools for identifying and interpreting components.

an anomalous word but not after an appropriate word. They interpreted this N400 component as signaling reprocessing of anomalous semantic information and were able to distinguish it from positive deflections signaling disconfirmation of an expectation (the already well-known P300) or following a change in font size.

For our purposes, what is most noteworthy here is the considerable variability across trials in the endogenous components of the EEG waveform that makes it challenging to extract the response specifically evoked by an exogenous stimulus. Although this variability is viewed as noise from a reactive perspective, as mentioned above, it can instead be viewed as reflecting the varied endogenous origins of the brain's ongoing activity. The challenge is to detect and analyze patterns in this activity and uncover their origins and functions. For example, thalamocortical oscillations seem to play a pivotal role in regulating communication between cortical areas (Buzsáki, 2006). We discuss other proposals regarding the functional importance of endogenous activity in section 4.

16.4.2 Recording from Individual Neurons

EEG rhythms were assumed to reflect neural activity, but what kind of neural activity? Two opposing explanations were pursued in the 1940s and 1950s (see Kaada, 1953). One explanation credited individual cortical neurons with the capacity for endogenous generation of rhythmic firing. If such neurons synchronized their activity, this would be sufficient to produce the overall rhythms observed in an EEG. The other explanation relied on a reactive conception of the neuron now known as *integrate-and-fire.* It was assumed that each neuron continuously performed an essentially linear integration of inputs received from other neurons at synapses on its dendritic tree. When a threshold was exceeded it "fired"— that is, it sent an electric pulse (called an *action potential* or *spike*) along its axon towards yet other neurons. This model was an important part of the conceptual framework for work on the mammalian nervous system, which relied heavily on single-cell recording studies of motor neurons in the spinal cord. The rate of spiking indicated responsivity to a stimulus, and it was relatively straightforward to study propagation through circuits of such neurons. Cortical neurons had much more complex connectivity patterns, but the same overall framework was assumed to apply to them. To account for alpha rhythms, for example, it was suggested that a closed, self-re-exciting chain of integrate-and-fire neurons in cortex, driven by thalamic input, could keep neural impulses circulating rhythmically. In the version proposed by Eccles (1951, p. 462)[8]

[8]Eccles bemoaned the confusing accounts of cortical potentials and advocated the study of motor neurons in the spinal cord as affording a clearer perspective on neural activity: "It is the thesis of this paper that basically the responses of neurones are similar throughout the central nervous system, and that the more easily analysed responses of motoneurones provide the data for a satisfactory explanation of the electrical responses evoked in the cerebral cortex by all conditions of stimulation: by direct electrical stimulation; by afferent volleys; and by antidromic volleys" (Eccles, 1951, p. 449).

a 10 Hz firing pattern—the alpha rhythm—was assured by the fact that each neuron in the chain required approximately 100 msec. of recovery time before it could respond to above-threshold input with an action potential. On this account, endogenous rhythmicity was not required to explain observed rhythmicity.

Support for the alternative conception of the neuron as endogenously active came eventually, from research on invertebrates that was rooted in the dominant reactive perspective but led in new directions. The initial goal was to achieve a finegrained understanding of the action potential itself as an event across time, and the giant axon of the squid provided easy access for intracellular recording under controlled conditions. A key set of researchers were less interested in how fast the electrical pulse traveled down the axon than in the timecourse and mechanism of voltage changes at any point along the axon as the pulse passed. The primary mechanism turned out to be ion movements across the axonal membrane, as captured by Hodgkin and Huxley (1952) in an elegant set of equations. The membrane's resting potential is approximately -65 millivolts, reflecting a normal predominance of negative ions inside and positive ions outside. As a pulse arrives the voltage becomes less negative (depolarizes), triggering an *influx* of sodium ions (Na^+) from the extracellular space that drives the membrane potential into the positive range (very quickly, due to positive feedback). At a short delay a less rapid *efflux* of potassium ions (K^+) brings voltage back into the negative range, first overshooting (hyperpolarizing) and then returning to the resting potential.

Thus, the shape of the action potential is derived as the net effect of incoming sodium and outgoing potassium currents. It was not until the 1970s to 1980s that researchers discovered the molecular mechanism behind these dynamics: proteins in the membrane act as specialized ion channels, opening (or closing) as a function of voltage in the immediate vicinity and thereby collectively offering high (or low) conductance to their particular type of ion (see Hille, 2001). Voltage changes at one location on the membrane trigger channel-opening nearby, resulting in propagation of the action potential along the axon.

Although the Hodgkin-Huxley equations were nonlinear, they fit into the dominant reactive framework of the era in that the dynamics they described were those of a neuron firing in response to a depolarizing input. Invertebrate researchers began moving towards an appreciation for endogenous activity in the 1960s, however, as new findings emerged from intracellular recording and related experiments. Notably, investigators found specialized *pacemaker* neurons that generated their own rhythmic action potentials (Alving, 1968), as well as an unexpected variety of voltage-gated and other currents producing complex dynamics not only in axons but also in neurons' dendrites and cell bodies (reviewed by Kandel, 1976).

Mammalian researchers initially doubted the generality of such findings, but took notice when Rodolfo Llinás and colleagues found a variety of functionally important ion currents in neurons of the inferior olive and cerebellum in mammals (and birds) in the 1970s and 1980s. Most were spatially distributed and gated by voltage differently than the sodium and potassium channels in the axon, equipping them for functions other than the direct generation of action potentials. Notably, the dentrites were endowed with channels providing high-threshold conductance to calcium (Ca^{2+}) ions, enabling dynamically complex dendritic excitation in contrast

to earlier assumptions of passive transmission of signals from synapses.[9] Moreover, the cell bodies of some neurons in the inferior olive had a different kind of calcium channel with a seemingly paradoxical low-threshold conductance that, in interaction with sodium and high-threshold calcium conductances, enabled these neurons to function as single-cell oscillators "capable of self-sustained rhythmic firing independent of synaptic input" (Llinás, 1988, p. 1659).[10] They sent these rhythmic action potentials to target neurons in the cerebellum that were able to respond at the same frequency, qualifying them as *resonators* in the dynamical lexicon championed by Llinás—reacting, but in ways shaped by their internal properties. Llinás also investigated spontaneous oscillations in electrical potentials elsewhere in the brain. Research on the thalamus and thalamocortical relay neurons (e.g., Jahnsen & Llinás, 1984) proved particularly useful for linking dynamic behavior of individual neurons to the large-scale dynamics seen in EEG.

Llinás' research offers a radically different picture of neural activity than that featured in the reactive framework. Neurons, on his account, are complex dynamic systems that are constantly changing their states and spontaneously generating action potentials. To generate oscillations at the different frequencies found in EEG requires the synchronization of many individual neurons, but before neurons can synchronize they must first oscillate. By showing how they do so, Llinás identified the needed foundation for the overall brain to exhibit complex dynamics. A number of researchers have subsequently built upon this foundation, some examining how multiple ion channels contribute to the intrinsic oscillatory activity of neurons and others determining how different neurotransmitters and receptors affect the temporal dynamics of synaptic activity (see Destexhe & Sejnowski, 2003, for a review and theoretical framework for modeling how these endogenous and reactive processes interact in producing synchronous thalamocortical oscillations). We return to this in section 5.

16.4.3 Functional Magnetic Resonance Imaging (fMRI)

Single cell recording has been the workhorse technique of neuroscientists using animal models to understand information processing in the brain, as seen in section 2, but with rare exceptions it is too invasive for studying the human brain itself. It was the advent of functional neuroimaging techniques that catalyzed study of the

[9]This linked nerve excitability with the Ca^{2+}-dependent second messenger system that is important for regulating general cellular functions. It also provided a mechanistic explanation of the suggestion (Bremer, 1958) that EEG primarily reflects synchronized post-synaptic potentials in dendrites—not, as originally thought, action potentials in axons.

[10]For further exposition, see Buzsáki (2006, pp. 181-183), who comments: "These findings … illustrate that nature went to a lot of trouble bringing together these channels at the right densities and location just to serve one purpose: oscillation." For evidence extending the findings to sensory neurons in various mammalian species, see Huguenard (1996).

involvement of different brain areas in a variety of human cognitive performances and gave rise to the distinct field of cognitive neuroscience. An existing technology, positron emission tomography (PET), was adapted to this end in the 1980s by Marcus Raichle and colleagues (as reviewed by Posner & Raichle, 1994), followed in the 1990s by functional magnetic resonance imaging (fMRI). Each of these techniques detects changes in blood flow in the brain that serve as a proxy for neuronal processing; most commonly used today is the BOLD (blood oxygen level dependent) signal from fMRI. Until recently the primary strategy in both PET and fMRI studies has been to identify areas exhibiting higher signal intensity (greater blood flow) when a subject is performing a target task than when performing a control task. Those areas are said to be *activated* by whatever type of processing distinguishes the two tasks.

If the stimulus is a word, for example, providing a semantically associated word versus merely reading the stimulus word aloud would call upon semantic processing and thereby activate brain areas responsible for the relevant semantic operations (Petersen et al., 1989). Often passive viewing of the same stimulus is included as a control; in the same experiment, reading a word aloud versus passively viewing it identified areas responsible for certain phonological operations. Subtracting brain images to find activated areas has been a powerful strategy for localizing cognitive operations in the brain, currently reaching a spatial resolution as small as 2 mm. It is a clear success story for the reactive paradigm.

There is a twist, though, that has brought the endogenous perspective back into the story. In the decades before neuroimaging was deployed on humans responding to stimuli in cognitive tasks, a few researchers measured blood flow and brain activity in what has been called the *resting state*. For Berger (and his successors even today) this was the most appropriate state for detecting endogenous activity, especially alpha waves. In contemporary experimental designs, the term references a control condition in which a subject is still with eyes closed and is presented with no stimuli or task requirements (in variations, the eyes are open with or without a fixation point). In a pioneering study using the xenon 133 inhalation technique to measure regional cerebral blood flow, Ingvar (1975) discovered that subjects at rest exhibited high levels of frontal activity. He surmised that this activity reflected "undirected, spontaneous, conscious mentation, the 'brain work,' which we carry out when left alone undisturbed" (quoted by Buckner et al.,2008, p. 2). In an even earlier study using nitrous oxide to measure cerebral metabolism, Sokoloff, Mangold, Wechsler, Kennedy and Kety (1955) found that performing mental arithmetic did not increase metabolism, indicating that the background activity was as energy demanding as any operations involved in performing a cognitive task. Recently, Raichle and Mintun (2006) drew upon these and other findings to argue that while the brain consumes 20% of the energy utilized in the body, it increases its consumption very little when performing tasks rather than resting. Humans exhibit, it is then inferred, a great deal of endogenous activity even in the resting state.

The early results indicating substantial endogenous activity received little further attention once PET and fMRI burst on the scene and researchers focused on identifying brain areas activated in experimenter-defined tasks. Often, though, they

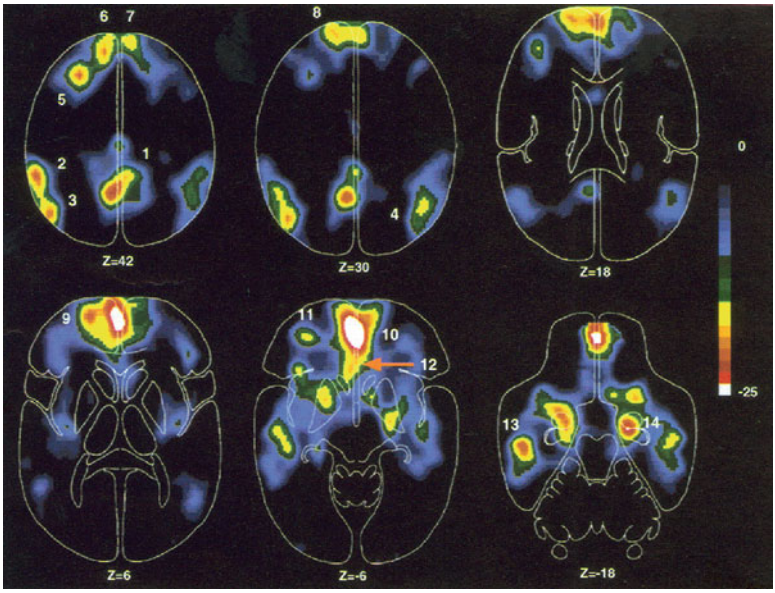

Fig. 16.5 Metaimage from Shulman et al. (1997) in which areas showing decreases in blood flow during task performance versus resting state are indicated in yellow and red. Components of this default network are labeled by numerals as follows: Junction of posterior cingulate and precuneus (1); Inferior parietal cortex (2, 3, 4); Left dorsolateral prefrontal cortex (5); Medial frontal strip that continues through the inferior anterior cingulate cortex (6, 7, 8, 9, 10, 12); Left inferior frontal cortex (11); Left inferior temporal gyrus (13); Right amygdala (14)

included the resting state as a control condition, presuming that it would elicit less activation than any of the tasks. A few of these researchers were intrigued to find that certain brain areas around the midline instead exhibited greater activation (stronger BOLD signal) in the resting state condition than in task conditions (e.g., Ghatan et al., 1995; Baker et al., 1996). They characterized these areas as *deactivated* during tasks, but it should be kept in mind that this referred to low relative activity (not negative activity). The choice of term reflects surprise that some of the subtractions (task condition minus rest condition) would yield negative numbers.

To determine whether a common set of brain areas manifested task-induced deactivation, Shulman et al. (1997; see also Mazoyer et al., 2001) conducted a meta-analysis of PET studies in which a task condition was compared to a non-task condition in which the same stimulus was present (which turned out to be similar in result to a resting condition with no stimulus). They found that the following areas were reliably *less* active in task situations, as shown in Fig. 16.5: the junction of precuneus and posterior cingulate cortex (PCC), inferior parietal cortex (IPC), left dorsolateral prefrontal cortex (left DLPFC), a medial frontal strip that continued through the inferior anterior cingulate cortex (inferior ACC), left inferior frontal cortex, left inferior frontal gyrus, and the amygdala.

Shifting the focus from the fact that these areas are less active during tasks to the fact that they are more active in the absence of task requirements, Raichle and his collaborators (Raichle et al., 2001; Gusnard et al., 2001) proposed that these areas constitute a *default network* – one which performs actual functions best carried out when there are no external task demands. There are clues to those functions in Ingvar's 1975 study, discussed above, and more directly in a neuroimaging study of autobiographical memory. Andreasen et al. (1995) found that the areas exhibiting heightened BOLD responses in a resting state condition were also relatively active in an episodic memory task. In contrast, a different set of areas exhibited heightened BOLD responses in a more typical semantic memory task. In an attempt to figure out what functions might elicit increased activity during rest, the researchers queried the subjects. Their reports pointed towards "a mixture of freely wandering past recollection, future plans, and other personal thoughts and experiences"—activities that plausibly draw upon episodic memory. Subsequent research has confirmed that thinking about one's own experiences is among the characteristic functions of the default network.

The studies discussed so far focused on relative amount of activity in the default network under various conditions, but not on the micro-temporal dynamics of this activity. Synchronized oscillations would be a salient criterion for network status, but finding them with fMRI initially seemed challenging due to the sluggish nature of the hemodynamic response. The feasibility of such a temporal analysis of fMRI data was demonstrated first for networks activated by tasks. Biswal, Yetkin, Haughton, and Hyde (1995) obtained BOLD signal values every 250 msec. for two minutes following a simple motor task (moving a hand). They identified spontaneous very low-frequency oscillations bilaterally in sensorimotor cortex, i.e., less than one cycle every 10 seconds (< 0.1 Hz). These oscillations were synchronized across the left and right hemispheres and also with oscillations in other motor areas. The researchers interpreted their results as indicting functional connectivity among the regions studied. Cordes et al. (2000) found similar oscillations in resting state BOLD signals in networks of areas previously identified by their synchronized activity in sensorimotor, visual, receptive language, or expressive language tasks. Moreover, their *functional connectivity MRI (fcMRI)* analysis – applying correlational statistics to resting state BOLD time series data to determine patterns of synchronization – yielded functional networks very similar to those identified from activity during tasks. That is, areas within the same network had correlated patterns of activity across time (rising and falling in synchrony) regardless of whether overall level of activity was relatively high (e.g., the sensorimotor network while moving a hand) or relatively low (e.g., the same network in a resting state condition).

To begin assessing whether the regions proposed to constitute a default network likewise met the criterion of synchronized oscillation, Greicius, Krasnow, Reiss, and Menon (2003) employed fcMRI with two seed areas, the PCC and inferior ACC. They regarded their results as providing "the most compelling evidence to date for the existence of a cohesive, tonically active, default mode network" (p. 256) and argued that the PCC was a critical node in this network. When it was used as the seed area for statistical analysis, its resting state oscillations were correlated with

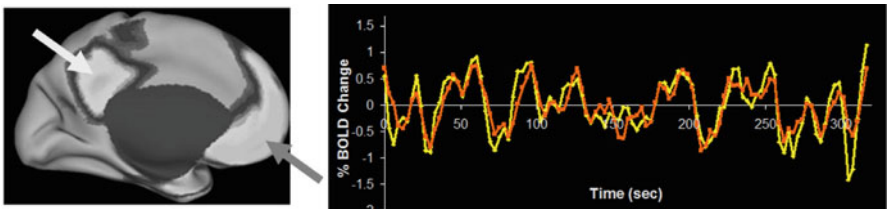

Fig. 16.6 On the left, two areas of the default network are indicated: Posterior cingulate cortex (white arrow) and ventral medial prefrontal cortex (grey arrow) On the right, it can be seen that patterns of activity for the two areas are highly correlated. From Raichle and Synder (2007)

those in much of medial prefrontal cortex (including inferior ACC and orbitofrontal cortex), left DLPFC, IPC bilaterally, left inferolateral temporal cortex, and left parahippocampal gyrus. (One of these synchronies is illustrated, using data from another study, in Fig. 16.6.) This is almost the same set of areas as those deactivated in task conditions according to Shulman's et al.'s meta-analysis. Greicius et al. argued that this default network performs higher cognitive functions, especially various forms of endogenously directed memory retrieval. Turning to their findings for ventral ACC as the seed area, the correlated areas included the PCC, medial prefrontal cortex/orbitofrontal cortex, the nucleus accumbens, and the hypothalamus/midbrain. They argued that these primarily paralimbic and subcortical areas comprised a network important for calibrating affective and autonomic operations, and further suggested that the strong connection between inferior ACC and PCC provided a crucial link between the two networks.

Finally, the investigators confirmed that both networks became deactivated during a working memory task but not in an eyes-closed or eyes-open resting state or even when passively viewing a blinking checkerboard. Besides indicating a rather high threshold of cognitive demand for deactivation of the default network, these findings clearly distinguished the default network from the neural system responsible for alpha rhythms in EEG (which diminish when subjects open their eyes). Subsequently Greicius and Menon (2004) found that the default network included the hippocampus and Vincent et al. (2006) determined that by seeding an analysis with a hippocampal region they could find correlated activity in the rest of the default network.[11] Buckner, Andrews-Hanna, and Schacter (2008, pp. 4-5) summed up the perspective provided by this research: "The default network is a brain system much like the motor system or the visual system. It contains a set of interacting brain areas that are tightly functionally connected and distinct from other systems within the brain."

Most of the known networks in the brain, in contrast to the default network, show more BOLD activation during tasks than at rest. But even at rest there is enough

[11]By combining time-series analysis of fMRI with diffusion tensor imaging, van den Heuvel, Mandl, Kahn, and Pol (2009) have recently shown that eight of nine areas in the default network are directly connected by fiber tracts (see also Greicius et al., 2009).

activity to assess whether the constituent areas of any such task-activated network fluctuate in synchrony with each other (but not with the default network). Fox et al. (2005) selected a network that was especially active during attention-demanding tasks (intraparietal sulcus, frontal eye field, middle temporal region, supplementary motor areas, and the insula). Examining those areas in the resting state, they found that fluctuations in their BOLD signals indeed were correlated. Moreover, fluctuations in that network and in the default network were anticorrelated. That is, the areas that were positively correlated within each network were negatively correlated with areas in the other network—an outcome more interesting than a zero correlation. Fox et al. emphasized that:

> anticorrelations may be as important as correlations in brain organization. Little has been said previously in the neuronal synchrony literature regarding the role of anticorrelations. While correlations may serve an integrative role in combining neuronal activity subserving similar goals or representations, anticorrelations may serve a differentiating role segregating neuronal processes subserving opposite goals or competing representations (Fox et al., 2005, p. 9677).

This pattern of results in resting state data is a strong indicator that within both the default network and the network involved in attention-demanding tasks, coordinated activity of some kind goes on in the absence of external stimulation—activity that is is different for the two networks.

Subsequently, researchers have used the strategy of finding correlations in resting-state fluctuations to identify yet other networks. For example, temporally correlated activity was found by Vincent et al. (2006) in the hippocampus and parietal memory systems and by Fox, Corbetta, Synder, Vincent, and Raichle (2006) in the dorsal and ventral attention systems. Deploying an alternative method, independent component analysis, on resting state fMRI data, Mantini, Perrucci, Del Gratta, Romani and Corbetta (2007) differentiated six different networks.[12] Fox and Raichle (2007, p. 701) concluded: "A consistent finding is that regions with similar functionality—that is, regions that are similarly modulated by various task paradigms—tend to be correlated in their spontaneous BOLD activity."[13] An important unanswered question is how the anticorrelation is achieved—via a separate control system or as a direct result of competition between networks.

[12] An interesting finding in their study was the differentiation of the default network from what they characterize as a "self-referential network" that contains areas often associated with the default network: medial-ventral prefrontalcortex, the pregenual anterior cingulate, the hypothalamus, and the cerebellum.

[13] Fox and Raichle also note that finding correlations in the resting state ("spontaneous BOLD activity") is a good way of identifying the full range of components of a network, since many of them may not be activated by any particular task. "Interestingly, most memory tasks implicate only a subset of regions, whereas the hippocampal formation resting-state correlation map reveals the full distribution of memory-related regions assessed across multiple experiments. Patterns of spontaneous activity could thus serve as a functional localizer, providing *a priori* hypotheses about the way in which the brain will respond across a wide variety of task conditions" (p. 702).

16.4.4 Relating Endogenous Oscillations at Different Timescales

Spontaneous oscillations in brain activity for awake humans are found across a wide range of timescales, depending on measurement technology and practices. Characteristic oscillations are slowest in fMRI data (< 0.1 Hz), midrange in EEG (1-80 Hz), and fastest in single-cell recording of action potentials (>100 Hz). This raises the question of whether neural activities as captured by different methods at different timescales are independent or related.

Intriguing results have been reported by researchers who asked whether the networks inferred from correlated BOLD signals in fMRI exhibit distinctive EEG signatures. Laufs et al. (2003) first found that activation of the default network is associated with strong activity in the mid-beta range, and activation of attention-related frontal and parietal areas is associated with weak activity in the alpha range, i.e., alpha blocking. Mantini, et al. (2007) expanded this kind of analysis to the six networks they had identified from resting state fMRI data, seeking associations across as a broader range of EEG frequency bands. They found that "Each brain network was associated with a specific combination of EEG rhythms, a neurophysiological signature that constitutes a baseline for evaluating changes in oscillatory signals during active behavior" (p. 13170). Their methods were sensitive enough to find alpha and beta rhythms together providing a broader positive signature for the default network and negative signature for the attention network. Beta recombined with delta and theta rhythms as the positive signature for the auditory network and was the sole positive signature for the somatomotor network. Gamma rhythms were the only positive signature for the self-referential network and the only frequency band excluded for the vision network. It was assumed that different sets of neurons within an area were responsible for different rhythms, but also that the resting state BOLD signal indirectly reflected the endogenous activity of those same neurons.

In another line of research, direct-current-coupled full-band electroencephalography has been used to detect infraslow oscillations (.01 -.1 Hz, meaning a single cycle lasts 10 to 100 seconds). The rising phase of a single ultraslow cycle is associated with (a) higher-amplitude activity in each of the traditional EEG frequency bands and (b) improved accuracy at detecting a somatosensory signal (Monto et al., 2008). Infraslow EEG oscillations are similar in timescale to BOLD oscillations in fMRI, and there are indications that these are sensitive to some of the same activity. For example, He, Snyder, Zempel, Smythe, and Raichle (2008) found a positive correlation between slow cortical potentials (infraslow and delta activity combined) and spontaneous fluctuations in the BOLD signal.

A final example of interrelatedness can be found in analyses of local field potentials, which are recorded in animals by means of electrodes implanted in the extracellular space between neurons. A variety of neural activities contribute to this signal. If dendritic and other lower-frequency activities are of primary interest, high frequency oscillations reflecting action potentials (>150 Hz) can be removed. The highest-amplitude oscillations contributing to the remaining signal in monkeys turn out to be in the same frequency range (<0.1 Hz) as the oscillations in the human BOLD signal (Leopold et al., 2003).

16.5 The Significance of Endogenous Oscillatory Brain Activity

In the previous section we reviewed evidence from EEG, single-cell recording, and fMRI studies pointing to endogenous oscillatory activity in the brain. An advocate of the reactive framework might acknowledge these findings but downplay their significance, maintaining that these oscillations are appropriately regarded as noise with respect to any functional analysis of neural performance. For example, one would expect basic metabolic activities to be maintained even in the absence of task demands, so perhaps the endogenous electrical activity is merely an epiphenomenon of ongoing metabolism. However, the evidence we reviewed indicated that the oscillations within individual brain areas are periodic, not random, and that those areas are organized into networks within which oscillations are correlated and between which they are anticorrelated. As well, these networks exhibit distinctive EEG signatures. This intricate organization is highly persuasive that the endogenous activity in the brain is functional, and invites the thought that this functional activity is so important that it cannot be ignored in seeking to understand how the brain performs its functions. Here we briefly explore four different ways in which endogenous activity may be crucial for understanding the brain mechanisms involved in cognition.

First, it seems obvious that for any mechanism that responds to stimuli with an increase (or decrease) in activity, but also exhibits ongoing endogenous oscillations, the magnitude of any particular response will depend in part on the phase of that oscillation at the moment the stimulus arrives. Fox, Snyder, Zacks, and Raichle (2006) devised an innovative strategy for demonstrating that this is true of neural activity in somatomotor cortex (SMC). Specifically, they showed that trial-to-trial variability in fMRI BOLD response to an exogenous stimulus could be attributed largely to spontaneous (endogenous) fluctuations. Data were available from subjects who had been instructed to press a button with the right hand each time an event was detected. This evoked a BOLD response in left SMC, peaking on average at eight seconds following the button press and then returning to baseline. However, there was considerable variability across trials. Typically such variability is treated as random noise, but Fox et al. suspected that it reflected endogenous oscillations on which the event-related responses were superimposed. That is, the endogenous BOLD signal might happen to be relatively high or low or inbetween at the moment a particular button press was required. The researchers surmised that simultaneous spontaneous fluctuations in right SMC should serve as a good proxy for the endogenous contribution to the activity in left SMC, since these were correlated in resting state and only left SMC was involved in production of a right-hand button press. Indeed, they found that the task-related increase in the left SMC BOLD signal could be analyzed as a linear addition to the amplitude of the spontaneous fluctuation. A subsequent study (Fox et al., 2007) was designed to permit calculations of trial-to-trial variability in the BOLD amplitude at each of eight timepoints following a button press. As illustrated in Fig. 16.7, variability in left SMC was significantly reduced by subtracting out the corresponding activity in right SMC. This indicated that much of the variability reflected ongoing spontaneous fluctuations – not

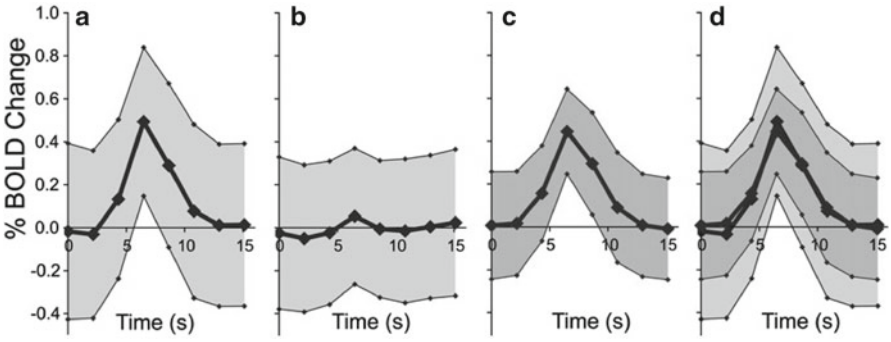

Fig. 16.7 Fox et al.'s (2007) demonstration that (**a**) the variability (average standard deviation) around the mean BOLD response in left SMC in a button press task is largely explained by (**b**) the spontaneous activity in the right SMC. Panel (**c**) shows the effect of subtracting (**b**) from (**a**). Panel D is an overlay of panels (**a**) and (**c**) showing in dark grey the variability attributable to spontanous activity and in light grey the remaining unexplained variability in BOLD activity in the left SMC

noise – and that event-related activity is superimposed on these fluctuations.[14] Moreover, they were able to extend previous findings of an association between behavioral and BOLD responses by showing that the force with which the button was pressed was correlated with the spontaneous fluctuation component in their own BOLD data. There were some complications in this BOLD-behavior effect that await further research for firm interpretation, but the main message is clear: spontaneous fluctuations in brain activity are synchronized within networks and contribute to the variability of event-related responses in both brain and behavior.

Second, at the end of the previous section we noted that oscillations recorded through EEG are positively correlated with the slower BOLD oscillations registered in fMRI. These correlations may reflect systemic coherence in brain functioning. It has been found that when a mammalian EEG waveform is decomposed into simpler component waveforms (e.g., by fast Fourier transform), the amplitude of each component is inversely proportional to the frequency ($1/f$). (See Freeman et al., 2000.) Even more interesting, there is evidence that the amplitude of higher-frequency oscillations is modulated by the phase of lower-frequency oscillations. Specifically, gamma waves (30-50 Hz) are strongest during the rising-positive portion of a single theta wave (4-10 Hz) within which they are embedded, and in turn the small number of theta waves concomitant with a single delta wave (1-4 Hz) are strongest during its rising-positive phase. (These data and the "oscillatory hierarchy hypothesis" proposed to explain them are from Lakatos et al., 2005; see also Canolty et al., 2006.)

[14]There had been similar findings earlier at the level of individual neurons. In a single cell recording study using cats, Tsodyks, Kenet, Grinvald, and Arieli (1999) showed that a given neuron's response to a visual stimulus was affected by the spontaneous state of the visual system at the time of stimulation, as indicated by local field potentials measured by means of optical imaging with voltage sensitive dye.

Such coupling can be particularly important when the brain is perturbed by a stimulus, since a modulation in low-frequency oscillations can, through phase-locking with higher-frequency oscillations, yield rapid changes at those frequencies. In addition, oscillations at lower frequencies tend to synchronize over more widely distributed areas of the brain than those at higher frequencies:

> [The] *1/f* power relationship implies that perturbations occurring at slow frequencies can cause a cascade of energy dissipation at higher frequencies and that widespread slow oscillations modulate faster local events. These properties of neuronal oscillators are the result of the physical architecture of neuronal networks and the limited speed of neuronal communication due to axon conduction and synaptic delays. Because most neuronal connections are local, the period of oscillation is constrained by the size of the neuronal pool engaged in a given cycle. Higher frequency oscillations are confined to a small neuronal space, whereas very large networks are recruited during slow oscillations" (Buzsáki & Draguhn, 2004, p. 1926)

In sum, as a result of endogenous activity maintaining synchronized oscillations at different frequencies, the brain is able to generate coordinated responses to perturbations.

Third, endogenous activity in the brain's default network is the most obvious candidate for the neural underpinnings of *mindwandering* (Antrobus et al., 1970). Ingvar's (1975) interpretation of his early blood flow studies noted above suggested such a connection. Further support was provided by Andreasen et al.'s (1995) subjects, whose informal reports after they had been imaged in the resting state emphasized recollection and planning. These activities involve episodic memory, and episodic memory tasks are among the few highly cognitive tasks for which the default network's BOLD signal remains as strong as in the resting state. Buckner and Carroll (2007) in fact found that remembering the past, envisioning future events, and considering the thoughts and perspectives of other people produced selective activation within the default network. They construed these results as supporting the view that activity recorded during the resting state reflects thinking—but thinking that is "engaged when individuals are left to think to themselves undisturbed" rather than elicited by specific stimuli or tasks (Buckner et al., 2008, p. 1). They then hypothesized "that the fundamental function of the default network is to facilitate flexible self-relevant mental explorations—simulations—that provide a means to anticipate and evaluate upcoming events before they happen" (p. 2).[15] In defending this view they cite not only Andreasen at al.'s results but also correlations found by Mason et al. (2007) between stimulus independent thoughts (as initially characterized by Antrobus et al., 1970) and activity in the default network. Intriguingly, Li, Yan, Bergquist, and Sinha (2007) correlated trials on which subjects failed to detect stop signals in behavioral tasks with increased activity in the

[15]They also presented, but did not pursue, an alternative view that activity in the default network generates low-level generalized awareness or watchfulness (Gilbert et al., 2007). This view gains support from Hahn, Ross, and Stein's (2007) findings of increased activity in the default network in a target-detecting task when the target could appear anywhere, but not when it was expected in a specific location.

default network, as one would expect if that network were involved in a person thinking distracting thoughts about past and future experiences.

One factor that renders problematic such a characterization of the activity of the default network is that its oscillatory behavior is well maintained in sleep (Fukunaga et al., 2006; Larson-Prior et al., 2009) and under anesthesia (Vincent et al., 2007)—circumstances in which spontaneous conscious thoughts (such as Andreasen et al.'s subjects report) presumably are not occurring. This suggests that the default network's spontaneous activity is more foundational than originally supposed. Fox and Raichle (2007) considered three possible interpretations of synchronization in such activity between different areas in this network, as evidenced in correlated BOLD signals; these interpretations have in common that they reflect cognitive processing but need not be characterized in terms of conscious thought:

> One possibility is that spontaneous activity serves as a record or memory of previous use, showing correlations between regions that have been modulated together in a task-dependent manner. Another possibility is that spontaneous activity serves to organize and coordinate neuronal activity and that this coordination is more prominent between regions that commonly work in concert. This is similar to the temporal binding, although spontaneous BOLD occurs at a much slower, broader and more permanent scale. Finally, spontaneous activity may represent a dynamic prediction about expected use, with correlations occurring between regions that are likely to be used together in the future (Fox and Raichle, 2007, p. 709).

Fourth, endogenous brain activity might be crucial for building and maintaining certain types of organization in the nervous system required for cognitive activity. There is growing evidence that the brain exhibits *small-world* organization (Watts & Strogratz, 1998) in which most connections are between neighboring neurons, creating clusters that can collaborate in processing specific information, but a few long range connections enable overall coordination. There also is evidence that while most brain areas have connections to only a few other areas, some have a large number of connections, thereby constituting hubs. Both neuroanotomical and neurophysiological studies provide compelling evidence of such an architecture at different scales in the brains of different organisms. Watts and Strogatz, for example, identified a small-world architecture in the neural network of the nematode worm *Caenorhabditis elegans*, whose structure had been identified by White, Southgate, Thomson, and Brenner (1986) using serial reconstruction of electron microscopy sections. Also, as shown in Fig. 16.8, Sporns and Zwi (2004) developed a connection matrix for the 30 cortical areas and 311 connections Felleman and van Essen (1991) identified in the macaque visual cortex and demonstrated that it exhibited the features of a small-world network: short characteristic path lengths and high clustering. They determined that some areas, such as V4, exhibit an atypically large number of connections to other areas, qualifying them as hubs.

Such an architecture provides a highly efficient organization for information processing. It is notable that the default network itself exhibits a small-world architecture, as indicated by neuroanatomical studies on macaque homologues to the areas composing the human default network. Some of these areas (PCC/Rsp, vMPFC, and IPL) are hubs that link the other areas into a network. An important question is how such organization might arise. Rubinov, Sporns, van Leeuwen,

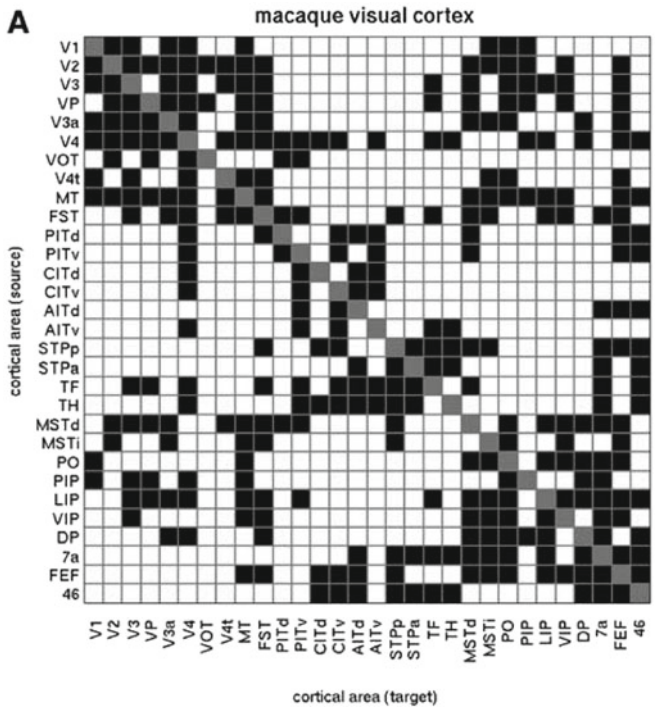

Fig. 16.8 Sporns and Zwi's (2004) connection matrix based on Felleman and van Essen's (1991) depiction of connectivity in the macaque visual cortex (shown in Fig. 16.3). Black squares indicate a known connection and white squares indicate that no connection has been identified. The connections are neither random nor purely local, but clustered so as to indicate a small-world network architecture

and Breakspear (2009) advanced the intriguing suggestion that oscillatory neurons, by developing connections when synchronized, might self-organize into a small world network with hubs. In support of this proposal they described a model by Gong and van Leeuwen (2004) that employed a logistic map activation function for individual units that can individually exhibit chaotic dynamics. Including a coupling factor enables the emergence of temporary patterns of synchronized oscillations. A Hebbian learning procedure establishes new connections between pairs of units whose activity is synchronized and prunes those between unsynchronized units. Even when these networks begin with random connectivity, they develop clusters linked to each other through hubs.

Once this structure appears, it in turn generates more coherent patterns of synchronized activity in subsequent processing, which leads via further Hebbian learning to a structural organization that is likely to exhibit yet more coherent synchronized patterns. At the larger scale of human brains, symbiotic interaction between the generation of functional synchrony and the building of structural connectivity could well result in a highly coordinated brain capable of maintaining multiple anticorrelated

networks. The initial state in real brains presumably is already biased towards a small-world pattern of connectivity, rather than random connections as in the simulations, and experience further shapes the emerging organization by Hebbian or other kinds of learning. Even so, the endogenous activity may be an important molding force so that the very architecture of the information processing system is in an important way a consequence of both endogenous and evoked activity.

In this section we have considered four suggestions as to how endogenous activity in the brain may contribute to its functioning as a cognitive system. Although it is too early to judge which construals will prove most fruitful, clearly the time for dismissing endogenous activity as mere noise has passed.

16.6 Conclusion: Endogenous Brain Activity and Dynamic Mechanistic Explanation

In the last two sections we provided an overview of the now substantial body of evidence that the brain is an endogenously active mechanism (or assemblage of mechanisms), one that is perturbed by stimuli or task impositions but changes its activity in ways that depend not just on these "inputs" but also on internal dynamics. This contrasts with the reactive framework of the vision research discussed in section 2, in which brain activity is treated as a response evoked by a stimulus. Llinás suggested viewing the interaction between stimuli and the brain's endogenous activity as a conversation:

> Although sensory nerve pathways deliver messages to the CNS that are quite invariant with respect to given sensory stimuli, the manner in which the CNS treats these messages depends on the functional state of each relay station. Thus, rather than a simple mirror of the external world, the CNS embodies a dialogue between the internal states generated by the intrinsic electrical activity of the nerve cells and their connectivity, which represents the internal context, and the information that reaches the brain from the senses (Llinás, 1988, p. 1633).

As with spoken dialogue, we are unlikely to understand how the brain subserves cognitive activity if we listen to only one side of the conversation: the message conveyed by a stimulus. It is equally important to identify endogenous activity and its contribution to what the system does next.

In this concluding section we return to the question of what conception of mechanism is best suited to understanding the brain's endogenous activity. Neuroscience has a long history of offering explanations in terms of basic mechanisms: parts and operations organized to produce a phenomenon, with that organization described qualitatively rather than quantitatively. The definition offered by Machamer, Darden, and Craver (2000) adds the stipulation that the mechanism's activity leads from set-up to termination conditions, which suggests sequential organization and situates basic mechanistic explanation within the reactive perspective that has long dominated research. This approach has been highly productive, but it offers insufficient

resources for understanding the dynamics of endogenous activity. Oscillations and other complex dynamics arise only in mechanisms with nonlinear component operations exhibiting some sort of nonsequential or cyclic organization (particularly feedback loops). Thus, a conception of mechanism adequate for endogenously active systems must reveal how nonsequentially organized, nonlinear operations interact quantitatively to generate the overall behavior of the mechanism. But when we as humans try to understand a particular system's behavior, we find it difficult to go beyond mental rehearsal of a sequence of operations; that is, our cognitive limits predispose us to basic mechanistic explanation. To understand endogenously active mechanisms we must turn to strategies that extend our capabilities.

The main scientific strategy for understanding how mechanisms generate complex behavior is computational modeling, in which researchers mathematically describe the dependency of changes in certain variables on changes in other variables, often by means of differential equations. As discussed in section 1, some computational models are explicitly grounded in mechanistic accounts; that is, particular variables or other terms in their equations correspond to particular properties of the mechanism's parts and operations. Given such an alignment of variables with properties, the modeler can simulate the functioning of the mechanism by assigning initial values to variables and performing numerical integration with an appropriate time step. The time evolution of each variable (the timecourse of its values) will depend in part on these initial values, but also on the constants chosen for the model's parameters. Sometimes biologically plausible values that have worked well in other simulations are known; if not, the modeler can determine best-fit values or, more interesting, determine how different values affect the model's behavior. Not infrequently, computational models can be best understood by employing the concepts and representational tools of complexity or dynamical systems theory, with possible implications for the mechanistic account. For example, when the time evolution of two of the model's variables meet criteria for a limit cycle in phase space, this indicates that the mechanism is capable of producing sustained oscillations.

Thus, the type of explanation required is one that integrates these strategies: dynamic mechanistic explanation. Mechanistic explanation identifies the parts and operations of a mechanism and how they are organized, and computational modeling and tools of dynamical systems theory reveal how such a mechanism will behave. This distinctive type of explanation is playing an increasingly important role in the sciences and warrants attention in the new mechanistic philosophy of science as well (Bechtel & Abrahamsen, 2010, 2011).

An instructive example of dynamic mechanistic explanation is provided by accounts of the endogenous oscillatory capabilities of thalamocortical (TC) relay neurons. These neurons produce sustained pacemaker oscillations – slow but regular firing at a frequency of 1-3 Hz – during human sleep. The mechanism has been uncovered by *in vitro* investigations using cat and rat TC neurons (Leresche et al., 1991). These oscillations rely upon the coordination of two voltage-gated currents: (1) I_h is an influx of sodium and potassium ions through channels opened when the membrane gets hyperpolarized following a spike. (2) I_T is an influx of calcium ions

brought about when the depolarization due to I_h opens low-threshold calcium channels. The ensuing calcium spike[16] further depolarizes the membrane, causing the various ion channels to close. This results in temporary hyperpolarization of the membrane, and the cycle repeats.

The basic mechanistic account provides a spatial layout of the relevant parts (membrane, ion channels, and ions) and specifies the sequence of their operations. The overall activity of the TC neuron—its firing pattern—will depend on quantitative properties of these parts and operations. Several different computational models aligned with the basic mechanistic account have succeeded in simulating the pacemaker firing pattern. Dextexhe, Bal, McCormick and Sejnowski (1996), for example, used Hodgkin-Huxley style equations that included variables for the conductance and activation of each type of channel, membrane voltage, and other properties of the TC neuron's component parts and operations.[17]

One of the parameters specified the maximum conductance of the I_h current (essentially, the membrane's capacity to move sodium and potassium ions if the maximum number of relevant channels were open). There were values of this parameter at which they succeeded in simulating the pacemaker oscillation pattern. But even more interesting was the fact that values in a higher range changed the dynamics of the conductance of the I_h current relative to that of the I_T current in such a way that the system now exhibited spindle-like oscillations. A series of spikes at about 1-3 Hz with a more or less spindle-shaped amplitude envelope would be observed for a few seconds, followed by a quiet period of five seconds or longer, followed by another series of spikes, then a quiet period, and so forth. This is the pattern that surprised Llinás in the 1980s and drew him and others to a more dynamic conception of the neuron. Finally, there was yet a higher range of values for the parameter that produced a quiet resting state.

Destexhe and Sejnowski (2003) provide a comprehensive guide to the many innovative models and research investigations of thalamocortical neurons including subsequent empirical evidence that their different patterns of activity are regulated in the manner proposed in their 1996 model. Buzsáki (2006) makes the case that complex dynamics in various brain regions—not only endogenous oscillations but also synchronized activity—are crucial to understanding how the brain works.

---

[16]This refers to a rapid depolarization of the membrane due to an influx of calcium, in contrast to the even more rapid depolarization that characterizes action potentials and involves channels specialized to sodium ions. In both cases the membrane then repolarizes less rapidly as the channels close. Hence, a plot of membrane voltage shows a rapid rise followed by a somewhat less rapid decline, typically overshooting—hyperpolarization—before returning to baseline.

[17]Neither this basic mechanistic account nor the computational model enumerates individual ions and ion channels—there are simply too many. Thus, the basic mechanistic account states that there is large number of low-threshold calcium channels that can open or close and can vary in permeability. The computational account includes a conductivity variable that is specific to the type of channel and is presumed to represent the collective effect of the number of channels and their properties. Accounts at the level of individual channels are also available, but they focus on how a single channel works.

One point to emphasize in closing is that adding a focus on dynamics is not intended to replace the importance of traditional mechanistic research directed at identifying the parts, operations, and overall organization of a mechanism. There are some who disagree, arguing for dynamical explanation as a self-contained, successful competitor to mechanistic explanation (e.g., Chemero & Silberstein, 2008). Dynamical accounts that are not grounded in accounts obtained by decomposing mechanisms into their parts and operations do describe the activity of possible systems, but those systems may not be be like those that are actually functioning in the world. For a dynamical account to offer explanation, it must characterize activity of the actual mechanism producing the phenomenon of interest. When the variables and terms in a computational model are grounded in properties of well-established parts and operations, we have a better basis for trusting the explanation. On the other hand, the model may reveal that the mechanism, as delineated so far, accounts for some but not all aspects of the actual behavior. Tinkering with the model to determine what sort of mechanism would account for these additional aspects can then serve as a discovery heuristic by predicting the occurrence of new parts, operations, or organizational relationships. Some of these predictions may be borne out, some not, generating a dialectical engagement between mechanistic research and computational model building. Such dialectical engagement may provide the only way neuroscience can account for the endogenous dynamics exhibited in brains and ultimately the cognitive behavior the brain supports.

References

Adrian, E. D., & Matthews, B. H. C. (1934): The Berger rhythm: Potential changes from the occipita lobes in man. *Brain* 57: 355–385.

Alving, B. O. (1968): Spontaneous activity in isolated somata of Aplysia pacemaker neurons. *Journal of General Physiology* 51: 29–45.

Andreasen, N. C., O'Leary, D. S., Cizadlo, T., Arndt, S., Rezai, K., Watkins, G. L., et al. (1995): Remembering the past: two facets of episodic memory explored with positron emission tomography. *American Journal of Psychiatry*, 152: 1576–1585.

Antrobus, J. S., Singer, J. L., Goldstein, S., & Fortgang, M. (1970): Mindwandering and cognitive structure. *Transactions of the New York Academy of Sciences*, 32: 242–252.

Baker, S. C., Rogers, R. D., Owen, A. M., Frith, C. D., Dolan, R. J., Frackowiak, R. S. J., et al. (1996): Neural systems engaged by planning: a PET study of the Tower of London task. *Neuropsychologia*, 34: 515–526.

Bechtel, W. (2008): *Mental mechanisms*. London: Routledge.

Bechtel, W., & Abrahamsen, A. (2005): Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36: 421–441.

Bechtel, W., & Abrahamsen, A. (2009): Decomposing, recomposing, and situating circadian mechanisms: Three tasks in developing mechanistic explanations. In H. Leitgeb & A. Hieke (eds.), *Reduction and elimination in philosophy of mind and philosophy of neuroscience* (pp. 173-186). Frankfurt: Ontos Verlag.

Bechtel, W., & Abrahamsen, A. (2010): Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science Part A*, 41: 321–333.

Bechtel, W., & Abrahamsen, A. (2011): Complex biological mechanisms: Cyclic, oscillatory, and autonomous. In C. A. Hooker (ed.), *Philosophy of complex systems. Handbook of the philosophy of science, Volume 10*. New York: Elsevier.

Bechtel, W., & Richardson, R. C. (1993/2010): *Discovering complexity: Decomposition and localization as strategies in scientific research*. Cambridge, MA: MIT Press. 1993 edition published by Princeton University Press.

Berger, H. (1929): Über das Elektroenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, 87.

Berger, H. (1930): Über das Elektroenkephalogramm des Menschen. Zweite Mitteilung. *Journal für Psychologie und Neurologie*, 40.

Biswal, B., Yetkin, F. Z., Haughton, V. M., & Hyde, J. S. (1995): Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, 34: 537–541.

Bremer, F. (1958): Cerebral and cerebellar potentials. *Physiological Reviews*, 38: 357–388.

Brown, T. G. (1914): On the nature of the fundamental activity of the nervous centres; together with an analysis of the conditioning of rhythmic activity in progression, and a theory of the evolution of function in the nervous system. *The Journal of Physiology*, 48: 18-46.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008): The Brain's default network. *Annals of the New York Academy of Sciences*, 1124: 1–38.

Buckner, R. L., & Carroll, D. C. (2007): Self-projection and the brain. *Trends in Cognitive Sciences*, 11: 49–57.

Buzsáki, G. (2006): *Rhythms of the brain*. Oxford: Oxford University Press.

Buzsáki, G., & Draguhn, A. (2004): Neuronal oscillations in cortical networks. *Science*, 304: 1926–1929.

Canolty, R. T., Edwards, E., Dalal, S. S., Soltani, M., Nagarajan, S. S., Kirsch, H. E., et al. (2006): High gamma power is phase-locked to theta oscillations in human neocortex. *Science*, 313: 1626–1628.

Caton, R. (1875): The electric currents of the brain. *British Medical Journal*, 2.

Chemero, A., & Silberstein, M. (2008): After the philosophy of mind: Replacing scholasticism with science. *Philosophy of Science*, 75: 1–27.

Cordes, D., Haughton, V. M., Arfanakis, K., Wendt, G. J., Turski, P. A., Moritz, C. H., et al. (2000): Mapping functionally related regions of brain with functional connectivity MR imaging. *American Journal of Neuroradiology*, 21: 1636–1644.

Darden, L., & Craver, C. (2002): Strategies in the interfield discovery of the mechanism of protein synthesis. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 33: 1-28.

Destexhe, A. D., Bal, T., McCormick, D. A., & Sejnowski, T. J. (1996): Ionic mechanisms underlying synchronized oscillations and propagating waves in a model of ferret thalamic slices. *Journal of Neurophysiology*, 76: 2049–2070.

Destexhe, A. D., & Sejnowski, T. J. (2003): Interactions between membrane conductances underlying thalamocortical slow-wave oscillations. *Physiological Reviews*, 83: 1401–1453.

Eccles, J. C. (1951): Interpretation of action potentials evoked in the cerebral cortex. *Electroencephalography and Clinical Neurophysiology*, 3: 449–464.

Felleman, D. J., & van Essen, D. C. (1991): Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1: 1–47.

Ferrier, D. (1876): *The functions of the brain*. London: Smith, Elder, and Company.

Finger, S. (1994): *Origins of neuroscience*. Oxford: Oxford University Press.

Flourens, J. P. M. (1824): *Researches Expérimentales sur les Propriétés et les Fonctions du Système Nerveux dans les Animaux Vertébris*. Paris: Crevot.

Fox, M. D., Corbetta, M., Snyder, A. Z., Vincent, J. L., & Raichle, M. E. (2006): Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *Proceedings of the National Academy of Sciences*, 103: 10046–10051.

Fox, M. D., & Raichle, M. E. (2007): Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience*, 8: 700–711.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005): The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102: 9673–9678.

Fox, M. D., Snyder, A. Z., Vincent, J. L., & Raichle, M. E. (2007): Intrinsic fluctuations within cortical systems account for intertrial variability in human behavior. *Neuron*, 56: 171–184.

Fox, M. D., Snyder, A. Z., Zacks, J. M., & Raichle, M. E. (2006): Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses. *Nature Neuroscience*, 9: 23–25.

Freeman, W. J., Rogers, L. J., Holmes, M. D., & Silbergeld, D. L. (2000): Spatial spectral analysis of human electrocorticograms including the alpha and gamma bands. *Journal of Neuroscience Methods*, 95: 111–121.

Fukunaga, M., Horovitz, S. G., van Gelderen, P., de Zwart, J. A., Jansma, J. M., Ikonomidou, V. N., et al. (2006): Large-amplitude, spatially correlated fluctuations in BOLD fMRI signals during extended rest and early sleep stages. *Magnetic Resonance Imaging*, 24: 979–992.

Ghatan, P. H., Hsieh, J. C., Wirsén-Meurling, A., Wredling, R., Eriksson, L., Stone-Elander, S., et al. (1995): Brain activation induced by the perceptual maze test: A PET study of cognitive performance. *Neuroimage*, 2: 112–124.

Gilbert, S. J., Dumontheil, I., Simons, J. S., Frith, C. D., & Burgess, P. W. (2007): Comment on "Wandering minds: The default network and stimulus-independent thought". *Science*, 317: 43b-.

Glennan, S. (1996): Mechanisms and the nature of causation. *Erkenntnis*, 44: 50–71.

Glennan, S. (2002): Rethinking mechanistic explanation. *Philosophy of Science*, 69: S342–S353.

Gloor, P. (1969): *Hans Berger on the electroencephalogram of man. Amsterdam:* Elsevier.

Gong, P., & van Leeuwen, C. (2004): Evolution to a small-world network with chaotic units. *Europhysics Letters*, 67: 328–333.

Goodwin, B. C. (1965): Oscillatory behavior in enzymatic control processes. *Advances in Enzyme Regulation*, 3: 425–428.

Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003): Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 100: 253–258.

Greicius, M. D., & Menon, V. (2004): Default-mode activity during a passive sensory task: Uncoupled from deactivation but impacting activation. *Journal of Cognitive Neuroscience*, 16: 1484–1492.

Greicius, M. D., Supekar, K., Menon, V., & Dougherty, R. F. (2009): Resting-state functional connectivity reflects structural connectivity in the default mode network. *Cerebral Cortex*, 19: 72–78.

Grillner, S. (2003): The motor infrastructure: from ion channels to neuronal networks. *Nature Reviews Neuroscience*, 4: 573–586.

Gusnard, D. A., Akbudak, E., Shulman, G. L., & Raichle, M. E. (2001): Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*, 98: 4259–4264.

Hahn, B., Ross, T. J., & Stein, E. A. (2007): Cingulate activation increases dynamically with response speed under stimulus unpredictability. *Cerebral Cortex*, 17: 1664–1671.

He, B. J., Snyder, A. Z., Zempel, J. M., Smyth, M. D., & Raichle, M. E. (2008): Electrophysiological correlates of the brain's intrinsic large-scale functional architecture. *Proceedings of the National Academy of Sciences*, 105: 16039–16044.

Henschen, S. E. (1893): On the visual path and centre. *Brain*, 16: 170–180.

Herrmann, C. S., Grigutsch, M., & Busch, N. A. (2005): EEG oscillations and wavelet analysis. In T. Handy (ed.), *Event-related potentials: a methods handbook* (pp. 229–259): Cambridge, MA: MIT.

Hille, B. (2001): *Ion channels of excitable membranes. Sunderland*, MA: Sinauer.

Hodgkin, A. L., & Huxley, A. F. (1952): A quantitative description of membrane current and its application to the conduction and excitation of nerve. *Journal of Physiology*, 117: 500–544.

Holmes, G. M. (1918): Disturbances of visual orientation. *The British Journal of Ophthalmology*, 2: 449–468.

Hubel, D. H., & Wiesel, T. N. (1962): Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160: 106–154.

Hubel, D. H., & Wiesel, T. N. (1968): Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195: 215–243.

Huguenard, J. R. (1996): Low-threshold calcium currents in central nervous system neurons. *Annual Review of Physiology*, 58: 329–348.

Ingvar, D. H. (1975): Patterns of brain activity revealed by measurements of regional cerebral blood flow. In D. H. Ingvar & N. A. Lassen (eds.), *Brain work: The coupling of function, metabolism, and blood flow in the brain: Proceedings of the Alfred Benzon Symposium VIII, Copenhagen, 26-30 May 1974, held at the premises of the Royal Danish Academy of Sciences and Letters, Copenhagen* (pp. 397-413): New York: Academic.

Inouye, T. (1909): *Die Sehstöhrungen bei Schussverletzungen der kortikalen Sehsphähre nach Beobachtungen an Verwundeten der letzten japanischen Kriege. Leipzig:* Engelmann.

Jahnsen, H., & Llinás, R. R. (1984): Electrophysiological properties of guinea-pig thalamic neurones: an in vitro study. *The Journal of Physiology*, 349: 205–226.

Jasper, H. H., & Andrews, H. L. (1938): Brain potentials and voluntary muslce activity in man. *Journal of Neurophysiology*, 1: 87–100.

Kaada, B. R. (1953): Electrical activity of the brain. *Annual Review of Physiology*, 15: 39–62.

Kandel, E. R. (1976): *Cellular basis of behavior: An introduction to behavioral neurobiology*. San Francisco: W. H. Freeman.

Kuffler, S. W. (1953): Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16: 37–68.

Kutas, M., & Hillyard, S. A. (1980): Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, 207: 203–205.

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., & Schroeder, C. E. (2005): An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology*, 94: 1904–1911.

Larson-Prior, L. J., Zempel, J. M., Nolan, T. S., Prior, F. W., Snyder, A. Z., & Raichle, M. E. (2009): Cortical network functional connectivity in the descent to sleep. *Proceedings of the National Academy of Sciences*, 106: 4489–4494.

Lashley, K. S. (1929): *Brain mechanisms and intelligence*. Chicago: University of Chicago Press.

Laufs, H., Krakow, K., Sterzer, P., Eger, E., Beyerle, A., Salek-Haddadi, A., et al. (2003): Electroencephalographic signatures of attentional and cognitive default modes in spontaneous brain activity fluctuations at rest. *Proceedings of the National Academy of Sciences of the United States of America*, 100: 11053–11058.

Leopold, D. A., Murayama, Y., & Logothetis, N. K. (2003): Very Slow Activity Fluctuations in Monkey Visual Cortex: Implications for Functional Brain Imaging. *Cereb. Cortex*, 13: 422–433.

Leresche, N., Lightowler, S., Soltesz, I., Jassik-Gerschenfeld, D., & Crunelli, V. (1991): Low-frequency oscillatory activities intrinsic to rat and cat thalamocortical cells. *The Journal of Physiology*, 441: 155–174.

Li, C.-S. R., Yan, P., Bergquist, K. L., & Sinha, R. (2007): Greater activation of the "default" brain regions predicts stop signal errors. *Neuroimage*, 38: 640–648.

Llinás, R. R. (1988): The intrinsic electrophysiological properties of mammalian neurons: Insights into central nervous system function. *Science*, 242: 1654–1664.

Loomis, A. L., Harvey, E. N., & Hobart, G. A. (1937): Cerebral states during sleep, as studied by human brain potentials. *Journal of Experimental Psychology*, 21: 127–144.

Lorente de Nó, R. (1938): Analysis of the activity of the chains of internuncial neurons. *Journal of Neurophysiology*, 1: 207–244.

Machamer, P., Darden, L., & Craver, C. F. (2000): Thinking about mechanisms. *Philosophy of Science*, 67: 1–25.

Mantini, D., Perrucci, M. G., Del Gratta, C., Romani, G. L., & Corbetta, M. (2007): Electrophysiological signatures of resting state networks in the human brain. *Proceedings of the National Academy of Sciences*, 104: 13170–13175.

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007): Wandering Minds: The Default Network and Stimulus-Independent Thought. *Science*, 315: 393–395.

Mazoyer, B., Zago, L., Mellet, E., Bricogne, S., Etard, O., Houdé, O., et al. (2001): Cortical networks for working memory and executive functions sustain the conscious resting state in man. *Brain Research Bulletin*, 54: 287–298.

Monto, S., Palva, S., Voipio, J., & Palva, J. M. (2008): Very Slow EEG Fluctuations Predict the Dynamics of Stimulus Detection and Oscillation Amplitudes in Humans. *J. Neurosci.*, 28: 8268–8272.

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., & Raichle, M. E. (1989): Positron emission tomographic studies of the processing single words. *Journal of Cognitive Neuroscience*, 1: 153–170.

Planck, M. (1949): *Scientific autobiography, and other papers* (F. Gaynor, Trans.). New York: Philosophical Library.

Posner, M. I., & Raichle, M. E. (1994): *Images of Mind*. San Francisco: Freeman.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001): A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*, 98: 676–682.

Raichle, M. E., & Mintun, M. A. (2006): Brain work and brain imaging. *Annual Review of Neuroscience*, 29: 449–476.

Raichle, M. E., & Snyder, A. Z. (2007): A default mode of brain function: A brief history of an evolving idea. *Neuroimage*, 37: 1083–1090.

Rubinov, M., Sporns, O., van Leeuwen, C., & Breakspear, M. (2009): Symbiotic relationship between brain structure and dynamics. *BMC Neuroscience*, 10: 55.

Shaw, J. C. (2003): *The brain's alpha rhythms and the mind: A review of classical and modern studies of the alpha rhythm component of the Electroencephalogram with commentaries on associated neuroscience and neuropsychology*. Amsterdam: Elsevier Publishers B.V.

Sherrington, C. S. (1913): Further observations on the production of reflex stepping by combination of reflex excitation with reflex inhibition. *The Journal of Physiology*, 47: 196–214.

Sherrington, C. S. (1923): *The integrative action of the nervous system*. New Haven: Yale University Press.

Shulman, G. L., Fiez, J. A., Corbetta, M., Buckner, R. L., Miezin, F. M., Raichle, M. E., et al. (1997): Common blood flow changes across visual tasks. II. Decreases in cerebral cortex. *Journal of Cognitive Neuroscience*, 9: 648–663.

Silber, M. H., Ancoli-Israel, S., Bonnet, M. H., Chokroverty, S., Grigg-Damberger, M. M., Hirshkowitz, M., et al. (2007): The visual scoring of sleep in adults. *Journal of Clinical Sleep Medicine*, 3: 121–131.

Sokoloff, L., Mangold, R., Wechsler, R. L., Kennedy, C., & Kety, S. S. (1955): The effect of mental arithmetic on cerebral circulation and metabolism. *Journal of Clinical Investigation*, 34: 1101–1108.

Sporns, O., & Zwi, J. D. (2004): The small world of the cerebral cortex. *Neuroinformatics*, 2: 145–162.

Stuart, D. G., & Hultborn, H. (2008): Thomas Graham Brown (1882-1965), Anders Lundberg (1920-), and the neural control of stepping. *Brain Research Reviews*, 59: 74–95.

Thagard, P. (2003): Pathways to biomedical discovery. *Philosophy of Science*, 70: 235–254.

Tsodyks, M., Kenet, T., Grinvald, A., & Arieli, A. (1999): Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science*, 286: 1943–1946.

van den Heuvel, M. P., Mandl, R. C. W., Kahn, R. S., & Pol, H. E. H. (2009): Functionally linked resting-state networks reflect the underlying structural connectivity architecture of the human brain. *Human Brain Mapping*, 30: 3127–3141.

van Essen, D. C., & Gallant, J. L. (1994): Neural mechanisms of form and motion processing in the primate visual system. *Neuron*, 13: 1–10.

Vincent, J. L., Patel, G. H., Fox, M. D., Snyder, A. Z., Baker, J. T., Van Essen, D. C., et al. (2007): Intrinsic functional architecture in the anaesthetized monkey brain. *Nature*, 447: 83–86.

Vincent, J. L., Snyder, A. Z., Fox, M. D., Shannon, B. J., Andrews, J. R., Raichle, M. E., et al. (2006): Coherent spontaneous activity identifies a hippocampal-parietal memory network. *Journal of Neurophysiology*, 96: 3517–3531.

Walter, W. G. (1936): The location of cerebral tumours by electro-encephalography. *The Lancet*, 228: 305–308.

Walter, W. G., & Dovey, V. J. (1944): Electro-encephalography in cases of sub-cortical tumour. *Journal of Neurology, Neurosurgery and Psychiatry*, 7: 57–65.

Watts, D., & Strogratz, S. (1998): Collective dynamics of small worlds. *Nature*, 393.

White, J. G., Southgate, E., Thomson, J. N., & Brenner, S. (1986): The structure of the nervous system of the nematode *Caenorhabditis elegans. Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 314*: 1–340.

Wilson, D. M., & Wyman, R. J. (1965): Motor output patterns during random and rhythmic stimulation of locust thoracic ganglia. *Biophysical Journal, 5*: 121–143.

Wimsatt, W. C. (1976): Reductive explanation: A functional account. In R. S. Cohen, C. A. Hooker, A. C. Michalos & J. van Evra (eds.), *PSA-1974* (pp. 671–710): Dordrecht: Reidel.

Wimsatt, W. C. (2007): *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, MA: Harvard University Press.

Index

A
Abduction, 241, 250, 252
Acceptance and Commitment Therapy (ACT), 159, 160, 162
Action potential, 344–346, 352, 360
Adaptation, 156, 167–184, 237, 239, 246, 247, 252
Adaptive
  explanation, 210, 239–240, 252
  sensory biases, 194, 198, 202–206
Adaptiveness, 208, 212
Addiction, 19, 264–268, 271–273
Aesthetics, 16, 167, 171–172, 175, 183, 189–191, 194, 199, 200, 202, 203, 207, 211
Affect, 5, 8, 21, 34, 35, 52, 98, 109, 111, 129, 130, 132–135, 137, 139, 143, 158, 170, 196, 202, 206–207, 246, 247, 249, 250, 268, 275–276, 281, 285, 287, 293, 305, 335, 346, 350, 354, 359
Affective activity, 286
Alberts, J.R., 112
Allen, C., 4, 6, 12, 85, 135
Altruism
  vs. egoism debate, 283, 287
  evolution of, 21, 173, 179
Altruistic emotions, 21, 275–293, 304
Altruistic motivation, 287, 292
Animal, 4ff., 44, 85, 125, 170, 292, 317, 336
  cognition, 116, 125
Anthropocentric stance, 13–14, 126, 127
Art
  adaptationist explanations of, 16, 169–171
  and brain lesions, 171
  byproduct explanations of, 16, 171–173
  evolution of, 15–17, 169, 171, 178, 189–213
  neuroimaging studies of, 170–171, 349
Artistic behavior, 4, 5, 16–17, 190, 191, 201, 202, 204, 207, 209, 211–213
Association studies, 46
Assumption, 7, 18–20, 29–36, 50, 52, 57, 69, 73, 79, 86, 93, 116, 163, 173, 200, 203, 208, 239, 246, 251, 252, 256, 299, 320, 342, 345–346
Ayer, A.J., 262

B
Batson, C.D., 275, 276, 278, 280, 292
Behavior
  foraging, 132, 138
  genetics, 5, 26, 30, 31, 38, 62
  social, 28, 87, 245, 297–299, 319, 322–325
Behavioral biology, 3–24, 320, 324
Biology, 3ff., 26, 58, 67, 85, 157, 190, 239, 255, 275, 297, 319, 331
Biophilia, 206–207, 209, 211
Blackmore, S., 257
Bloom, P., 168, 218, 240, 241
Bowerbirds, 199, 200
Boyd, R., 17, 175, 177, 191, 192, 198–203, 207, 211, 212
Brain, 5ff., 28, 50, 105, 125, 170, 194, 236, 282, 297, 329
Brain-imaging, 247, 248, 282, 283
Buller, D., 217, 218, 223–225, 227, 256
Buss, D., 17–18, 218, 226, 227, 256, 268
Butler, J., 276, 278, 279

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*, Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4,

C
Caenorhabditis elegans, 23, 322, 356
Care, 11, 13, 59–61, 75, 76, 87, 95, 97, 104, 109, 110, 113, 140, 142, 149, 153–154, 159, 160, 183, 196, 226, 227, 230, 232, 233, 261, 281, 288, 290, 292, 336, 343
Caring emotion, 288, 290–292
Causal
  explanation, 9, 65–81
  inference, 49, 53, 57, 74
  space, 8, 29–34, 36, 37
Cause, 7ff., 29, 45, 65, 91, 127, 151, 168, 200, 219, 236, 262, 276, 299, 320, 355
Central pattern generator, 330, 331
Cheater detection, 18, 222–227, 230, 237
Children, 14, 45, 47, 49, 50, 52, 53, 55, 56, 58, 70, 71, 89, 116, 129, 132, 157, 158, 160, 161, 168, 169, 178, 200, 230, 236, 244, 248, 249, 288, 291–293
Chimpanzees, 12–14, 87, 116, 126, 128–143
Choice, 11, 56, 66, 67, 69, 72–77, 90, 101, 128, 136, 159, 191–194, 197, 198, 203, 259, 264, 267, 282, 283, 298–300, 303, 305, 322, 324, 331, 348
Cichlids, 196, 197, 204–206
Coevolution, 194, 197, 198, 200, 202
Cognition, 62, 87, 88, 91, 97, 98, 102–105, 107, 116, 117, 125, 126, 129, 140, 155, 169, 170, 210, 231, 236, 247, 291, 331, 353
Cognitive, 3ff., 38, 50, 86, 125, 155, 170, 205, 223, 235, 258, 285, 303, 319, 330
  activity, 130, 139, 356, 358
  architecture, 19, 88, 231, 238, 239, 241, 246–249, 251, 269, 272
Comparative psychology, 12, 86, 87, 89, 103, 116, 128
Compatibilism, 19, 255–273
Complex functional design, 236, 240, 241, 246, 249
Computational modeling, 331, 359
Computational modules, 235, 237, 241, 245, 251
Conditional, 52, 79, 152, 222–225
Conscious, 126, 128, 137, 159, 205, 236, 243, 270, 279, 280, 285, 287, 289, 342, 347, 356
Context, 4, 9–12, 16, 17, 22, 23, 51, 55, 56, 61, 65, 67, 69, 71–78, 80, 81, 86, 89, 92, 94, 95, 101, 104, 110, 114, 115, 130, 131, 135, 136, 138, 139, 171, 180, 184, 191–198, 203, 206, 211, 219, 221, 224–226, 231, 232, 239, 284, 290, 293, 306, 311, 319, 343, 358
Contrasts, 11, 17, 21, 23, 34, 43, 66–76, 79–81, 89, 100, 102, 105, 127, 136, 141, 171, 172, 191, 204, 219, 222, 223, 227, 261, 278, 292, 305, 307, 311, 329, 331, 333, 334, 345, 349, 350, 358, 360
Correlation, 7, 9, 26, 35, 43, 47, 49–52, 57–59, 66, 70, 71, 74, 134, 212, 292, 301–303, 336, 349, 351, 352, 354–356
Cosmides, L., 17, 18, 169, 170, 217, 218, 221–227, 230, 231, 236, 238, 240, 241, 245, 250, 256, 258, 262, 268, 269
Csibra, G., 18, 219, 222, 227–232
Cultural
  runaway, 191, 200, 201, 207
  transmission, 192, 200, 202, 207–209, 246

D
Darwin, 44, 157, 159, 163, 190–192, 199–201, 211, 250, 262
Darwinian Social Science, 270
Darwin machine, 15, 150, 151, 154, 157
Dawkins, R., 135, 158, 176, 192–194, 199, 210, 255, 259–263
Decomposing mechanisms, 332, 361
Default network, 348–352, 355–356
Dennett, D., 138, 263, 317
Design, 10, 30, 44, 49, 52–54, 57–59, 89, 97, 102–104, 112, 150, 151, 160, 190–192, 202, 204, 205, 210, 211, 228, 229, 235–241, 243–246, 249, 250, 256, 269, 272, 273, 278, 280–283, 297, 303, 307, 329, 332, 337, 341–342, 347, 353
Desire, 20, 21, 70, 71, 93, 127, 137, 150, 159, 210, 230, 264–268, 270–273, 275–279, 282, 284, 285, 287, 288, 290, 291, 298, 304, 308
Determinism, genetic, 149, 150, 238, 246, 255, 260, 261, 268, 271
Development, 3ff., 26, 43, 65, 85, 135, 149, 169, 203, 227, 235, 262, 299, 321, 331
Developmental niche construction, 13, 88, 96–98
Developmental psychobiology, 12–13, 86, 87, 90, 94, 102, 107, 108, 117
Developmental systems theory (DST), 6–8, 13, 26, 33, 34, 37, 100, 114, 115, 237, 251
Dictator game, 281
Discovery, 49, 61, 92, 98, 117, 221, 224, 332, 336–337, 343, 361
Disorder, 8, 35–38, 47, 154, 155, 160, 161, 244, 264, 272

Disposition, 11, 12, 29–31, 33, 65–67, 76–81, 113, 130, 135, 192, 272, 273, 300
Dissanayake, E., 169, 170, 189, 190, 199, 208
Domain specificity, 198, 235, 237–240, 243, 244, 251
Down's syndrome, 11, 62, 69, 75
DST. *See* Developmental systems theory
Dual inheritance theory, 200, 202
Dutton, D., 168
Dynamic mechanistic explanation, 24, 329, 331, 332, 335, 358–361
Dynamic perspective, 329

E
Economic experiment, 298, 308
EEG. *See* Electroencephalography
Eexperimental economics, 21, 281, 297–301, 309
Egg spots, 196–198, 204, 205, 211
Egoism, 20, 276, 279–283, 287, 293
Electroencephalography (EEG), 340–344
Emotions, 16, 21, 54, 127, 162, 172, 192, 205, 211, 260, 272, 275–293, 302–304
Endogenously active, 5, 329–361
Environment, 6ff., 26, 45, 65, 87, 126, 149, 169, 205, 223, 235, 255, 285, 298, 323, 343
Environment of evolutionary adaptedness, 18–20, 180, 239, 245, 246, 252, 269–273. *See also* Stone Age
Environment Wide Association Studies (EWAS), 9, 50–53, 55–59, 61
Epigenesis and preformationism, 92–93
Epigenetics, 13, 43, 65, 69, 89, 91–101, 108–110, 247
Epistemic, 12–14, 37, 38, 98, 125–132, 135–143
Epistemology/epistemological, 13, 14, 25, 36, 37, 46, 129, 130, 140–143
Ethnic markers, 176, 178–180, 183
Ethology, cognitive, 12, 86, 319
Etiology, 7, 29, 35, 47, 55, 56, 61
Event related potential (ERP), 343
Evidence, 12, 31, 43, 71, 104, 126, 158, 175, 200, 217, 235, 260, 280, 297, 317, 331
Evolution, 3ff., 26, 58, 75, 85, 127, 148, 169, 189, 217, 235, 256, 275, 299, 319, 359
Evolutionary
  argument, 21, 230, 275, 290, 293
  psychology, 4, 5, 12, 15, 17–20, 86, 87, 154, 170, 217–232, 235–252, 255, 256, 260–264, 268–273
Evolution of art, 17, 189–213
Evolvability, 170
Experience, 8, 10, 12, 13, 18, 26–28, 36, 53, 70, 76, 85–117, 127, 128, 134–138, 151–152, 169, 192, 198, 201, 206, 208, 209, 211, 235, 239, 242, 245, 247, 258, 264, 270–272, 278, 283, 284, 290, 349, 356
Experiment, 10, 27, 44, 85, 126, 158, 167, 195, 220, 242, 277, 297, 322, 329
Explanandum, 23, 24, 33, 67, 79, 99, 319, 321, 324, 326
Explanans, 33, 67, 68, 73, 79
Explanation, 3ff., 25, 45, 65, 86, 127, 169, 196, 221, 235, 260, 279, 298, 317, 329
  developmental, 3, 5, 12–14, 94, 252
  evolutionary, 5, 6, 15–21, 169, 235, 249, 291, 303, 319
  genetic, 5, 9–12, 45, 66–68, 73, 76, 77
  neurobiological, 3, 5, 21–24, 302–305
  scientific, 4–6, 61, 93, 324
Explanatory, 8–11, 18, 37, 65–67, 69, 71–73, 77, 79–81, 94–97, 113, 184, 191, 219–222, 224, 226, 230, 280, 291, 322, 338

F
Face recognition, 172, 205, 206, 209–211, 237, 238, 244
Fatalism, 258–260, 271
Feelings, 259, 269, 278–281, 283, 285, 286, 289–291, 303, 304, 309
Fisher-Zahavi model, 192–193
Flexibility of human behaviour, 126, 142
fMRI. *See* Functional magnetic resonance imaging
Fodor, J., 116, 237, 238, 269, 273
Folk psychology, 319, 321, 325–326
Frankfurt, H., 257, 262
Free riding, 223, 281, 302–304
Free will, 4, 5, 15, 19, 20, 255–273
Functional magnetic resonance imaging (fMRI), 248, 303, 331, 346–354
Functions
  biological, 22, 23, 291, 317–326
  causal role, 22, 114, 115, 321
  proper, 22, 320, 321
Fusiform face area, 205

G
Gamble, C., 180, 183
GBG. *See* Good behavior game
Gene, 3ff., 26, 43, 65, 87, 125, 149, 169, 191, 229, 235, 255, 300, 320

Genetic
determinism, 148, 150, 238, 246, 255, 260, 261, 268, 271
dispositions, 11, 12, 65, 76–81
quality, 193, 196, 197, 199
traits, 9–12, 65–81, 320
Genome wide association studies (GWAS), 5, 7, 9, 10, 27, 43–62
Genotype, 7, 8, 10, 29–36, 43, 58, 70, 95, 98, 151, 177
Gergely, G., 18, 219, 222, 227–232
Gigerenzer, G., 226
Gilbert, S.F., 95, 96, 101, 102, 107, 114, 115
Good behavior game (GBG), 160
Gottlieb, G., 33, 37, 86, 87, 92, 93, 96, 101, 107
Gould, S.J., 156, 208, 259, 260
Grau, J.W., 103, 105
Green beards, 17, 176–178
Griffiths, P.E., 85, 93–95, 100, 101, 107, 115, 117, 251, 320, 321
Group selection
cultural, 17, 167–184
GWAS. *See* Genome wide association studies

H
Hamilton's rule, 261
Help, 8, 12, 15, 37, 38, 51, 57, 58, 89, 91, 94, 105, 107, 109, 113, 150, 151, 159, 161, 179, 180, 219, 221, 225, 230, 231, 269, 270, 275–293, 299, 305, 320, 338
Helping behaviour, 278, 280
Henrich, J., 173, 175, 200, 212, 281, 299, 301, 307
Heritability, 7, 9, 26, 30, 35, 44–49, 56, 60, 62, 66, 74, 75, 251
Heuristic, 18, 217–233, 361
Higher-order desires, hypothetical, 266
Homology, 22, 95, 320, 321, 356
Human, 4ff., 25, 43, 76, 87, 125, 149, 167, 190, 219, 235, 255, 275, 297, 319, 329
action, 19, 20, 262, 263, 276, 279, 287
evolution, 155, 200, 206, 208, 210, 211, 213
height, 57
nature, 239, 245, 255–273
Human Genome Project, 43, 46, 50
Hume, D., 257–259, 262, 264, 287
Hutcheson, F., 276–279

I
Ice Age, 180, 183, 184
Iconic representations, 192, 204–208
Iindirect benefit, 17, 191–196, 198, 202–204, 207, 211, 213
Imitation, 104, 105, 132, 191, 227, 228
Indirect bias, 200
Inference to the best explanation, 240, 241
Innate/innateness, 12–14, 77, 80, 86, 91, 92, 94, 95, 100, 108, 113, 116, 117, 125, 130, 192, 205, 206, 208, 228
Innate *vs.* acquired, 85, 87, 90, 91, 94, 95, 107, 111, 113, 115
Instrumental motive, 276, 279, 283, 284, 304
Intention, 4, 20, 22, 50, 89, 104, 116, 117, 126, 127, 149–163, 168–170, 177, 178, 275, 284, 285, 287, 288, 291, 305, 309, 311, 312, 317–326
Intentionality, 117, 318, 319, 323
Interdisciplinary, 3, 4
Internal reward, 279, 280, 282
Ion channels, 345, 346, 360

J
Jablonka, E, 97, 101, 110, 150
Judgement, 257, 266, 267, 271, 278, 284–288
Justice, 303, 307, 310

K
Kalahari !Kung, 179, 180
Kant, I., 257
Kimmel, M., 261
Kitcher, P., 66, 73, 217, 261, 291, 319
Knowledge, 5ff., 25, 66, 86, 125, 150, 208, 219, 269, 283, 318, 343

L
Language, 14, 86, 88, 93, 105, 112, 127, 128, 131, 175, 178, 179, 212, 236, 240, 241, 244, 247, 248, 263, 269, 319, 349
Lao-Tzu, 257
Late Pleistocene, 173, 175, 179
Learning, 5, 12–14, 19, 85–117, 132, 138, 154, 155, 159, 161, 162, 169, 191, 192, 200, 208, 209, 212, 227–231, 240, 244, 246–248, 267, 357
Learning as development, 12, 13, 91, 111, 116
Lehrman, D.S., 85–87, 90
Lewontin, R.C., 74, 97, 100, 170, 208
Linguistic, 14, 126, 128, 131, 171, 248, 322

Linkage studies, 35, 46
Local field potentials, 354
Localization of function, 336

M
Magdalenian, 173, 176, 180–184
Maladaptive, 191, 208, 210, 211, 245, 246
Male display, 190–198, 203, 204
Massive modularity thesis, 19, 256, 268, 270
Mate choice, 192, 193, 203
Mating preferences, 193–195, 197, 202
McGonigle, B., 88, 105
Meaney, M.J., 87, 94, 98, 99, 109, 110
Mechanisms, 5ff., 75, 86, 126, 153, 173, 196, 223, 244, 256, 281, 297, 323, 329
Mechanistic explanation, 24, 62, 93, 326, 329, 331, 332, 335, 346, 358–361
Mechanistic philosophy of science, 329, 331, 359
Mendelian inheritance, 44
Mental bias, 192, 205, 210, 211, 213
Midgley, B.D., 87, 107, 111
Miller, G.F., 110, 169, 170, 189, 191, 192, 195, 198–204, 207, 208, 211
Mimicry, 196–198, 205, 211
Mindreading, 230
Missing heritability problem, 47–49
Model bias, 200, 203
Modularity, 19, 99, 170, 236–239, 256, 268, 270
Module, 16, 19, 88, 172, 223, 225, 226, 235–241, 243–247, 249–251, 273
Modules, cognitive. *See also* Massive modularity thesis
  as adaptations, 239, 249, 250
  mandatoriness of, 269
  role in evolutionary psychological explanations, 243, 271
Molecular biology, 26
Monism, 8, 36, 37, 277
Moral assessment of behavior, 310
Morality, 277
Motivation, 4, 20, 21, 161, 172, 184, 256, 268–273, 275–293, 297, 299, 303–312
Motivational causal chain, 276, 283–287, 293
Motive, 20, 21, 275–291, 293, 298, 304, 305, 309–312
Mozart, 11, 12, 77–81

N
Nativism, 92, 116
Natural pedagogy, 18, 227–232
Natural selection, 21, 87, 95, 96, 155, 156, 158, 170, 176, 192, 236–239, 241, 244, 246, 249–252, 261, 270, 272, 291, 303, 320, 321
Nature *vs.* nurture, 37, 77, 86, 87, 93–95, 113, 115, 239
Neural plasticity, 18, 109, 113, 235–252
Neurobiological, 3, 5, 16, 21–24, 28, 32, 35, 297, 299, 302–305, 309, 310, 312, 319
Neuroeconomics, 282, 283, 297–300, 303–305, 308–310
Neuropeptide Y, 323
Neuroscience, 5, 36, 87, 115, 171, 241, 243, 246, 248, 251, 299, 317–326, 329–331, 347, 358, 361
Neurotransmitter, 7, 23, 27, 29–32, 34, 36, 110, 346
NHST. *See* Null hypothesis significance testing
Nonshared environment, 10, 52–56, 58
Norm enforcement, 297, 300–304, 307–312
Norm of reaction, 95, 151
Novel, 92, 93, 108, 113, 115, 194, 205, 229, 230, 233, 283
npr genes, 323, 324
Null hypothesis significance testing (NHST), 9, 48, 50, 51, 57

O
Objectivity, 71–73
Obsessive compulsive disorder (OCD), 264, 273
Occurrence, 49, 103, 182, 212, 221, 223, 224, 232, 251, 277, 288, 290, 291, 361
Optical illusions, 269
Organism, 4ff., 33, 47, 68, 86, 126, 149, 169, 194, 235, 261, 290, 319, 329
Oscillations and oscillators, 334, 335, 340, 346, 349, 351–357, 359, 360
Oyama, S., 33, 75, 93–95, 100, 101, 115, 245

P
Palaeolithic art
  beads, 168, 176, 178, 180, 182, 184
  dance, 168, 169, 171, 176
  music, 168, 171, 176, 248
  sculpture, 168
Palmer, C., 261
Parental care, 13, 76, 110, 288, 290, 292
PE. *See* Psychological egoism
Peacock, 169, 191, 200, 201, 203
Perceptual bias, 204, 206

Phenotype, 13, 56, 58, 66, 86, 91, 93, 95, 96, 108, 109, 111, 115, 151, 152, 171, 194, 199, 240, 250, 251
Phenotypic plasticity, 15, 86, 96, 97, 108, 115, 150–156
Phenylketonuria (PKU), 76
Philosophy
of behavioral biology, 3–24
of biology, 3, 4, 6, 22, 319
of science, 3, 10, 22, 25, 66, 319, 329, 331, 359
Phobias, 265, 268
Pinker, S., 171, 172, 189, 208, 217, 218, 226, 235, 236, 238–241, 245, 246, 256, 268, 272
PKU. *See* Phenylketonuria
Pluralism, 8, 26, 36–39
Population, 9–11, 16, 21, 26, 30–31, 43, 45, 46, 48–49, 51, 53–55, 57, 58, 61, 66, 69, 70, 74, 75, 95–97, 143, 151, 155, 156, 159, 161, 174, 176–179, 181–184, 193, 195, 200, 201, 204, 209, 210, 212, 213, 232, 239–241, 249–251, 270, 307, 320
genetics, 26, 57, 58, 193, 200
stratification, 49
Povinelli, D.J., 87, 88, 104
Pragmatism, 7, 8, 26, 37–39
Predict, 43, 50, 54, 178, 201, 207, 317
Price equation, 174, 175
Primary motive, 276–277, 279, 283
Primate, 12, 88, 104, 125–143, 155, 156, 172, 203, 270
Prisoner's dilemma, 283, 323
Proximate, 6, 25, 26, 38, 289–293, 302–304
Psychological altruism, 20, 21, 275–293, 298
Psychological egoism (PE), 20, 276–283, 288–293
Psychology, 4–5, 12, 13, 18, 20, 43, 50, 51, 53, 85–89, 91, 97, 102, 103, 107, 108, 116, 117, 128, 194, 255, 256, 260, 270, 273, 279, 293, 298, 299, 308, 319, 321, 325, 326, 329–331
Psychosensory bias, 203, 207
Public goods game, 281
Punishment, 21, 104, 162, 175, 297–312

R

Radcliffe-Richards, J., 255, 259, 261–263, 272
Reactive perspective, 23, 329, 331, 333, 336, 344, 345, 358
Recomposing mechanisms, 332
Regulation, 90, 93, 98–100, 108, 109, 111, 113, 115, 323–325
Reliabilism, 14, 126, 140, 142
Reliability argument, 290, 291, 293
Religion, 157, 158, 178, 192, 210–211
Reproductive success, 17, 157, 174, 175, 191–193, 201, 208, 212, 249
Reputation, 150, 276, 277, 281, 282, 301, 304, 308, 309
Responsibility, moral, 258, 259
Reward, 104, 154, 160–162, 172, 191, 196, 279, 280, 282, 283, 303, 304, 309, 312, 331
Rigidly flexible, 15, 150, 152–154
Rose, S., 259, 260
Runaway process, 191–193, 195, 196, 200–203, 211
Ryle, G., 262

S

Schneirla, T.C., 90
Selection, 15, 29, 44, 71, 86, 132, 150, 169, 190, 222, 236, 261, 291, 298, 320, 329
Selective breeding, 44–45
Self-directed, 275, 276, 278–281, 286, 288–293
Self-exploitation, 192, 206, 207
Sense of beauty, 190, 191
Sensory bias, 17, 191, 194–196, 198, 202–208, 211
Sensory exploitation, 17, 189–213
Sequential organization, 333, 358
Sexual selection, 17, 169, 170, 189–213
Sexual strategies, 226
Simple heuristics, 226
Single-cell recording, 331, 336, 344, 346, 352, 353
Single nucleotide polymorphism (SNP), 46, 48, 49, 57, 58, 60
Skinner, B.F., 104, 154, 155
Sleep stages, 342, 343
Smart, J.J.C., 257
SNP. *See* Single nucleotide polymorphism
Sober, E., 67, 74, 150, 193, 217, 226, 275, 276, 278–280, 290–291, 293, 298
Social Darwinism, 150, 157–159, 161
Social-environmental approaches, 7, 28, 32
Social exchange, 223, 225, 299
Social learning, 88, 97, 98, 115, 138, 191, 192, 200, 209
Social science, 9, 10, 43–63, 258, 270

Society, 15–17, 21, 117, 162, 168, 172, 173, 175, 184, 199–201, 210, 245, 271, 299–301, 303, 305–307, 309–312
Sociobiology, 19, 86, 255–257, 259–262, 319
Spandrel, 17, 207–209, 211, 212
Sterelny, K., 65–66, 73, 75, 88, 94, 170, 174, 218, 227, 261, 291
Stone Age, 18–20, 180, 239, 245, 246, 252, 269–273
Stotz, K., 4, 6, 7, 12–13, 85, 88, 98–100, 114, 117
Structural, 13, 88, 109, 111, 220, 242, 243, 332, 357–358
Sweatt, J.D., 109, 110
Sweet, experience of, 271, 272
Symons, D., 226, 268, 270, 271
Synchronized oscillations, 349, 355, 357

T
Talent, 11, 12, 70, 78–81
Taste, 138, 172, 229–231, 265–267, 270, 271
Teaching, 5, 100, 228–231, 247
Theory-ladenness, 22, 319, 324
Thornhill, R., 227, 261
Timberlake, W., 103
Tomasello, M., 88, 89, 104, 131, 132, 137, 139, 155, 227, 228
Tooby, J., 17, 18, 169, 170, 217, 218, 222–227, 231, 235, 236, 238–241, 245, 250, 256, 258, 262, 268, 269
Trait, 4ff., 26, 44, 65, 87, 149, 172, 190, 224, 255, 317
Trait, behavioral, 5, 7–10, 15, 16, 22, 23, 91, 93, 102, 240, 248, 317–326
True biological effects, 48, 57, 60–62
Trust game, 281
Twin studies, 26, 46, 53, 55, 56, 58

U
Ultimate, 6, 20, 26, 57, 93, 94, 98, 100, 103, 117, 174, 221, 276, 284, 287, 289, 290, 361
Unconscious, 20, 157, 279, 282, 283, 285, 287, 293
Use-dependent plasticity, 242, 243

V
Variation, 7–10, 26, 27, 30–33, 35, 44, 45, 53, 56, 57, 60, 61, 66, 74, 92, 95, 150–154, 157, 159, 160, 162, 174, 175, 193, 200, 236, 239, 241, 246, 249–251, 323, 324, 338, 340, 342, 343, 347
Venus figurines, 206
Violence, 27, 28, 50, 157, 305, 306, 309, 310
Vision, 117, 134, 269, 330, 340, 341, 352, 358

W
Warm glow, 278, 281
Wason selection task, 222–226
Well-being, 16, 271, 275–277, 279, 283, 286–289, 301, 305, 306
West, M.J., 87, 96–98, 102, 112
Whallon, R., 179, 180
White, R., 180
Wilson, D.S., 4, 15, 16, 149, 150, 154, 156, 158, 218, 226, 275, 276, 279, 280, 290, 304, 326
Wilson, E.O., 155, 173, 206, 210, 258, 259, 261
Within-family designs, 10, 53, 54, 58, 59
Woods, T., 81

Z
Zeki, S., 171