ASHGATE PUBLISHING LTD

Reflecting on Darwin

von Sydow, Momme (2014). ‘Survival of the Fittest’ in Darwinian Metaphysics - Tautology or Testable Theory? (pp. 199-222) In E. Voigts, B. Schaff &M. Pietrzak-Franger (Eds.). *Reflecting on Darwin*. Farnham, London: Ashgate.

Chapter 11

‘Survival of the Fittest’ in Darwinian Metaphysics: Tautology or Testable Theory?

Momme von Sydow (Heidelberg)

Charles Darwin in Historiography – Scientist, Philosopher or Both?

Charles Darwin is often presented not only as a most eminent naturalist, but also as a prototypical empirical scientist, inductively deriving his theory of evolution based on empirical evidence rather than on theoretical, or even metaphysical or religious grounds. During his voyage on the HMS Beagle Darwin assembled a huge collection of animal-specimens, which contributed to the theory that changed our view of life. Such observations as those of the Galapagos finches were crucial for the paradigm shift linked to his specific theory of evolution, replacing belief in Genesis as well as pre-Darwinian theories of evolution. This positivist success story, dominating biology textbooks, sometimes presents Darwin’s theory of natural selection as simply a great victory of modern empirical science over earlier religious or philosophical prejudices. Although indeed this interpretation is broadly in line with today’s intellectual frontline between radicalized gene-Darwinism (Dawkins, 1976, 2007; Dennett, 1995, 2006) and radicalized religious literalism, the view involves gross simplifications. A detailed historical analysis shows that the relationship between Darwinism on the one side and religion, philosophy, and metaphysics on the other side has been much more volatile and intricate than this simple success story suggests (Desmond and Moore 1991; Depew and Weber 1995; Gould 2002; Knight 2004; Brooke and Cantor 2000; von Sydow 2005, 2012). Paradoxically, Darwin’s belief in natural theology, even after the voyage with the Beagle, had a strong impact on his theory of natural selection, even though this theory later subverted his religious tenets, ultimately rendering him agnostic at least (e.g., Ospovat 1995; Gould, 2002; von Sydow 2005). As a young man at the University of Cambridge, Darwin had studied theology. Although he was more interested in the ‘book of nature’ than in the Bible, the naturalist community in

Cambridge brought him into further contact with England's natural theology. Darwin read Paley's *Natural Theology* (1802) voluntarily and with delight, learning it almost by heart. Even later he wrote that he has 'hardly ever admired a book more than Paley's Natural Theology' (Darwin, 1985., vol. 7, letter to J. Lubbock, 22 November 1859: 388). Darwin's theory of natural selection appears in fact to have absorbed ideas from Paley's natural theology – among them his early beliefs in pan-adaptationism and in an unchangeable and universal law of natural selection (von Sydow 2005): The so-called 'Panglossian' perfectionism (cf. Gould 2002: 264) is found in Paley's *Natural Theology* (1802) and is linked to his argument that organisms provide evidence for an omniscient designer. Even after adopting the general idea of a transformation of species in 1837 (an idea discussed by romantic and Lamarckian biologists before) and sketching a first version of his theory of natural selection in 1838, Darwin still retained a Paleyan belief in the ubiquity of adaptations, which he retained perhaps until 1844 (Ospovat 1981/1995: xv, 60–86). For Paley (1802), universal and unalterable natural laws, quite similar to adaptations, suggest the existence of a designer. Influenced by Paley – as well as by the general predominant Newtonian approach of the time – Darwin fashioned his theory of natural selection as one based on a simple, unchanging, uniform and universal mechanism (von Sydow 2005) that seems to exclude, for instance, an evolution of evolutionary mechanisms (von Sydow 2012).

It is generally accepted today that Darwin did not adopt his theory of natural selection on the Galapagos Islands or while traveling on the HMS Beagle (1831–1836), but rather when ordering his observations in the light of then available theories. His influences included not only Paley, but also Erasmus Darwin, Robert Edmond Grant, Charles Lyell, Adam Smith and Thomas Malthus: Erasmus Darwin (Charles's grandfather) and Robert Edmund Grant were among those who introduced him to romantic and Lamarckian traditions of transmutation of species. Furthermore, Darwin's gradualism was influenced by Charles Lyell's geological uniformitarianism; and his early optimism regarding individual competition followed the British tradition of Adam Smith and Milne-Edwards. Finally, Darwin explicitly recounts in his autobiography that he got the basic idea for his theory of natural selection when re-reading Malthus's *Essay on the Principle of Population*. The theologian and economist Malthus had argued against socialist utopianism in the context of a theodicy resembling that of Paley's natural theology. The permanent, ceaseless struggle for existence is interpreted as a side effect of God's operating by general law to prevent human vice from obstructing the high purpose of creation. Reverend Malthus claimed that,

when one considered superfecundity, destruction and misery, the idea of a benevolent God could only be vindicated if he acted by general laws that lead to the improvement of the moral qualities of man. When Darwin in 1838 actually formulated his hypothesis of natural selection, he provided a missing link, not only for a theory of evolution, but – paradoxically – for Paley's and Malthus's theodicy as well (see Paley 1802, Chap. XXVI). However, partly based on his new theory Darwin in fact abandoned natural theology, eventually embracing agnosticism. Interestingly, in the fifth and sixth editions of the *Origin of Species* and in the *Descent of Man* Darwin states: 'I was not able to annul the influence of my former belief, then widely prevalent, that each species had been purposely created; and this led to my tacitly assuming that every detail of structure, excepting rudiments, was of some special, though unrecognized, service' (Darwin 1871: 153). He further concludes that these influences led him to 'extend the action of natural selection ... too far' (Darwin 1871; cf. von Sydow 2005).

Although Darwin was careful not to taint science with crude ideology, his theory, as the above sketch indicates, provides not only an empirical synthesis but a theoretical one as well. It is suggested that Darwin's pan-adaptationism, as well as his advocacy of the universality and unchangeability of the law of natural selection, may historically be based on metaphysical assumptions that paradoxically seem to play a partly religious role in Darwin's earlier advocacy of natural theology. Darwin clearly was not only an eminent scientist, but an eminent theoretician or even philosopher as well. The sailors on the Beagle did not know how right they were when they nicknamed Darwin 'the philosopher'.

This historical sketch is a good preparation for considering next the possibility that Darwinism might partly have a metaphysical basis. The section that follows, however, more directly provides a brief systematic introduction to Darwinian metaphysics (cf. similarly: von Sydow, 2012, for a brief review). This, in turn, is followed by a main section, turning critically to the potentially tautological formulation of the concept of 'the survival of the fittest' as one important basis for a Darwinian metaphysics, if not for Darwinism in general.

Universal Darwinism and Darwinian Metaphysics

Darwinian metaphysics in a broad sense may be linked to the influence of Darwinism on philosophy. Although books on the history of philosophy still often only mention Darwinism in a footnote, it may well have played a key role in the history of philosophy of the late nineteenth and

twentieth centuries: for instance, in modern materialism, monism, pragmatism, and, with in part disastrous consequences, in various brands of social Darwinism. In the history of philosophy there has also been explicit but similarly heterogeneous criticism of Darwinism as in ontology or metaphysics (exemplified by such different authors as H. Drietsch, E. von Hartmann, H. Jonas, G.E. Moore, and A.N. Whitehead). Darwinism's main influence on the history of philosophy, however, may perhaps have been indirect, in providing the background for the change of focus of main philosophical schools of the second half of the twentieth century, who – with different arguments – all abandoned the field of philosophy of nature as their central constituent (this holds as well for logical positivism as for neo-Kantianism, phenomenology, existentialism and postmodernism). In this historical perspective the role of Darwinism seems important, but remains heterogeneous and difficult to assess.

Recent decades, however, have witnessed a renaissance of general naturalism in philosophy, often combined with the adoption of views from radicalized forms of Darwinism. The approaches of gene-Darwinism and process-Darwinism can be understood as two forms of a Darwinian metaphysics (von Sydow, 2012, 2013) because of their universal applicability across disciplines, their simple basic principles, their substantial changes of common-sense assumptions about what things exist and partly *a priori* (sometimes tautological) justifications.

First, what I call gene-Darwinism (cf. von Sydow 2012) is a biological approach most prototypically exemplified in the work of Richard Dawkins (1976, 1983; cf. e.g., Williams 1966) that has not only inspired considerable work in biology, but that is also often seen to have massive implications for the social, ethical and religious domains (e.g., Dawkins 2006). With regard to the unit-of-selection debate, gene-Darwinism radicalizes and purifies the existing reductionist tendencies of Darwin's theory by advocating a biological entity reductionism to the level of a single gene (cf. Gould 2002). Dawkins's Darwinism does not stop at the level of the individual organism, but rather at that of single selfish genes. Gene-Darwinism takes a nominalistic position on gene combinations, genomes, gene pools and groups, by which all are taken as ephemeral epiphenomena. Only genes, understood as short chunks of DNA, survived in the meiotic shuffle (gene-atomism). Moreover, in a radicalized interpretation of the central dogma of molecular biology, stating that information cannot explicitly be transferred from proteins back to the DNA, phenotypes (in contrast to genotypes) are regarded as mere 'puppets' or vehicles of the genes (germ-line reductionism).

Second, gene-Darwinism radicalizes the neo-Darwinian stress on natural selection by consistently advocating process reductionism, reducing all processes of information gain to processes of natural selection. Although Darwin clearly regarded natural selection as the core of his theory, he still allowed for a substantial causal pluralism, involving roles, for example, for the correlation of growth and use-inheritance. In addition, the modern evolutionary synthesis, particularly in its second phase with Dobzhansky and Mayr as main proponents, clearly put more stress than gene-Darwinism on the role of populations, phenotypes and causal pluralism (von Sydow 2012). In contrast, gene-Darwinism claims that all relevant evolutionary processes are essentially reducible to processes of *natural selection* (Darwinian process reductionism).

Finally, gene-Darwinism has often been linked to biologism. E.O. Wilson, representing at least early on partly gene-Darwinian ideas, claimed that ethics should become 'biologized' and be based on the 'morality of the gene' (1975: 3–6). This appears coherent with gene-reductionism as well as Darwinian process-reductionism, although some gene-Darwinians have actually argued that this conclusion need not follow.[1] Whatever the conclusion, gene-Darwinism claims that we are like

> Chicago gangsters, our genes have survived, in some cases for millions of years, in a highly competitive world. This entitles us to expect certain qualities in our genes. I shall argue that a predominant quality to be expected in a successful gene is ruthless selfishness. This gene selfishness will usually give rise to selfishness in individual behaviour. (Dawkins 1989: 2)

All aspects of gene-Darwinism have been challenged. For instance, a modern multi-level approach gained influence challenging gene-reductionism (Hull, 1981; Sober and Wilson 1998; Gould 2002; Wilson, 2005; Okasha, 2006; Wilson and Wilson 2007; Nowak, Tarnita & Wilson, 2010; von Sydow 2012). Nevertheless, gene-Darwinism, with its simplicity of very few first principles and universality of postulated application, has remained influential within but also outside of biology (cf. Dawkins 2006; Dennett 2006).

Process-Darwinism signifies another class of Darwinian metaphysics, that can be defined by strict Darwinian process reductionism, advocating that Darwinian processes essentially provide an exclusive

[1] Richard Dawkins has resisted to advocate a morality of the gene and claimed ideas or 'meme's may have some autonomy (Dawkins, 1982: 110–12). However, it is at least questionable, whether within his highly reductionist framework a truly emergentist position could consistently be advocated.

and exhaustive explanation of knowledge acquisition in biology or other subject areas, while claiming the existence of such processes on at least one or several levels outside of biology (Campbell 1960; Dawkins 1983; Dennett 1995; Hull 1981; Hull, Langman and Glenn 2001; Popper 1972; Plotkin 1994; Skinner 1981; cf. von Sydow 2012). Although older roots of process Darwinism go back, for instance, to August Weismann, William James and Charles Sanders Peirce, it was Donald T. Campbell who first argued in a classical article that all inductive achievements and genuine increases in knowledge were basically 'blind-variation-and-selective-retention processes' (1960; cf. 1990). The term *process-Darwinism* may be used in the field of biology as well (for multi-level accounts that advocate strict process-reductionism). However, the term here signifies approaches of universal Darwinism postulating actual Darwinian processes in other disciplines as well. Table 1 below presents some subject areas in which Darwinian accounts have been influential, showing the units of selection, the alleged Darwinian process involved, and some main protagonists.

Table. 11.1 Process-Darwinism in Various Disciplines (cf. von Sydow, in press)

Subject Area	**Units of Selection**	**Darwinian Process**	**Authors**
Biology	Genes	Blind mutation and natural selection	G.C. Williams, R. Dawkins
Psychology	Acts, operants, associations	Trial and error	E.L. Thorndike, B.F. Skinner, D.T. Campbell
Philosophy of Science;	Theories	Conjecture and refutation	K.R. Popper, S. Toulmin

Epistemology			
History of Ideas	Ideas (or 'memes')	Blind variation and external retention	D.T. Campbell, R. Dawkins, D. Dennett
Economy	Firms, products, routines	Innovation and market selection	M. Friedman

Campbell (1960) posited in detail that the psychological processes of pattern recognition, creativity and operant conditioning (trial-and-error learning) are essentially Darwinian, analogous to the biological process of blind mutation and natural selection. Skinner, the father of operant conditioning, independently from Campbell concluded that the learning of operant behaviour corresponds to 'a second kind of selection' based on 'the first kind of selection' (natural selection) (Skinner 1981: 501; more recently, see, for instance, Hull et al. 2001). In economics the analogy between survival of the fittest (or natural selection) and market-selection is an old one, dating from early forms of social-Darwinism (see Hofstadter 1955; Greene 1981). It needs to be noted, however, that in the past as well as in the present not all accounts of evolutionary economics have strictly reflected Darwinian approaches (see Hodgson 1993; Knudsen 2002; Nelson 2007). Nonetheless, the idea of natural selection played a crucial role, for instance, in M. Friedman's influential advocacy of an unconstrained free capitalism and laissez-faire policy (1953: 22, cf. von Sydow, 2012). Moreover, there seem to be similarities between neo-Darwinism and neo-classical economics (for a critical view, see Khalil 1983). Darwinian economics advocates that innovations, routines, or businesses on the whole are selected by given consumer-preferences or the invisible hand of the market (Knudsen 2002; Hodgson 2002).

The advocacy of Darwinian processes in different domains has often been implicitly or explicitly accompanied by a commitment to the general research-program of process-Darwinism, with the claim that not only the described processes, but all forms of knowledge-acquisition essentially are Darwinian 'blind-variation-and-selective-retention' (Campbell 1960; Popper 1972; Dawkins 1983; Dennett 1995; cf. Hull et al. 2001). This approach may be called metaphysical, not only because it is a universal approach

(a ‘universal acid’, Dennett 1995) based on a second-level, extremely principled and purified ontological inventory (with one kind of process only), but because, as main proponents of process-Darwinism have argued, Darwinian approaches are not just true empirically, but also in principle (Popper 1972; Dawkins 1983). It is argued that conjectures must be blind and that, in principle, instructive learning is impossible. The theoretical, *a priori*, argumentations may perhaps be linked to two central philosophical or metaphysical issues (von Sydow, 2012, 2013): the fundamental problem of induction (going back at least to Hume) and the possible tautological interpretation of the concept ‘survival of the fittest.’ The remainder of this article will focus on the latter aspect only, now with regard to both biology and psychology.

The Problem of Tautological Formulations of Natural Selection

The British evolutionist Herbert Spencer, in his *Social Statistics* (1851), coined the phrase:

‘The survival of the fittest’ (1)

The formulation was later adopted by Darwin in 1869 in the fifth edition of the *Origin of Species* as synonymous to his central term, ‘natural selection’, yet without personifying nature (thus avoiding any religious imbroglio). This simple but resonant phrase provides the starting point for my argument here.[2]

Natural selection has often been defined in a much richer way, linked to other theoretical terms, such as common descent, blind mutation, gradualism, adaptation, and struggle for life. Nevertheless, survival of the fittest is an important explication of the central explanatory term of natural selection itself, which ever since Darwin has represented presumably the most central idea of Darwinism. Natural selection seems to have testable meaning of its own: Proposition 1 relates the explanandum, the question ‘what will survive?’ to an answer or explanans, ‘the fittest’. Proposition 1, to most laymen and presumably to most biologists, appears to be a clearly testable empirical hypothesis. Nevertheless, the

[2] Alternatively, *survival of the fittest* may be formulated in a more detailed and more precise way. Moreover, one may want to explicate a *relative* interpretation of fitness by comparing the fitness of two genotypes at a single locus. One may, for instance, reformulate Proposition 1: ‘For all organisms *x*, if and only if organisms *x* with genotype *A* are fitter than all organisms *x* with non-*A* genotype, then organisms *x* with genotype *A* tend to survive more frequently than organism *x* with genotype Non-*A*’. Such an elaborate formulation may well be helpful, but the main ideas of the text do not require such detailed—and less accessible—specifications.

concept of natural selection has long evoked criticism due to the potential for tautological interpretation (Scriven 1959; Popper 1972; Gould and Lewontin 1979; Rosenberg 1983; Lipton and Thompson 1988; cf. also Williams 1966).

The most pressing problem seems to be the definition of the 'respectable' scientific term 'fitness', used in several formulations and closely related to the notion of adaptation. Whereas adaptation is normally used retrospectively, fitness is used prospectively. Whether an entity is biologically more fit than another appears to be an empirical question. But how is one to decide this question? What is the ultimate measure of fitness?

Fitness, one may argue, should ultimately measure the ability to survive, which can be investigated by empirically assessing actual survival. Under such a definition, however, Proposition 1 as a whole becomes an untestable tautology. One could only predict:

The survival of the survivor(s) (those who will actually survive) (2)

As a prediction, that is not very bold: Proposition 2 may indeed have some connotations – an interesting issue we cannot discuss here – but it has no directly testable empirical content or predictive force. Whatever the world is like, this proposition holds true. Despite the apparently suitable definition of fitness, Proposition 2 immunizes natural selection. Such a formulation may also be used to justify an interpretation of a particular given feature of a survivor or a surviving population as an adaptation, in a *post hoc ergo propter hoc* explanation or a 'just so story' (see the classical paper on this related but different issue: Gould and Lewontin 1979; cf. Scriven 1959; Fodor et al. 2010). In any case, Proposition 2 interprets natural selection not as empirical theory at all, but at best as merely a metaphysical framework. In what follows several alternative formulations of fitness will be discussed that may prevent a circular or tautological interpretation of natural selection.

As a first objection, one may point out that 'fitness' actually has several meanings in ordinary language (for instance, 'physical fitness', referring to power, speed and other well defined terms). Using the everyday meanings recalls perhaps the proposal to use 'fitness' as a theoretical primitive (cf. an interesting discussion by Rosenberg, 1983: 464). If these meanings are adopted, Proposition 1 clearly ceases to be tautological. Based on Proposition 1, and employing common-sense features of fitness, that is, one may derive specific predictive propositions, such as:

The survival of the strongest (3)

The survival of the most vivid (4)

People may in fact imagine one of these testable interpretations when hearing Proposition 1. Charles Darwin, despite usually using Proposition 1, in *Variation under Domestication*, wrote: 'The strongest ultimately prevail, the weakest fail…' (1875: 5). But is Proposition 3 generally true? No. It is uncontroversial that the 'weak' may at least sometimes be more fit in evolution. For example, whereas dinosaurs became extinct, the weak predecessors of man (the size of a mouse) must have been quite successful (similar examples at the level of individuals may be provided). Interpreting Proposition 1, Proposition 3 is therefore either false or requires a theoretical system that allows specification of which proposition applies, contingent on some premise where power, speed, reaction-time, co-ordination, agility or any of their combinations is the determining factor for fitness. Briefly, Proposition 3 is thus either plainly false or underspecified or requires *post hoc* adjustments.

A second objection to the tautological interpretation of Proposition 1 seems that fitness today is defined, not by survival but by reproduction-rate (reproductive survival). One may calculate the absolute fitness of a genotype by the number of individuals possessing it after selection divided by the number before selection: $\omega_{abs} = N_{after} / N_{before}$; or one may calculate a relative measure: $\omega_{abs} = \omega_{abs} / average(\omega_{abs,\ all\ genotypes})$. The time span involved is usually one generation (until the filial generation reaches reproductive age[3]). If this interpretation of fitness is applied to Proposition 1, the understanding of survival (the explanandum) needs to shift in the same way for reproductive survival, which results once again in a tautological claim (cf. Rosenberg and Bouchard 2008):

Those organisms leave most (or more) offspring who leave most (or more) offspring (5)

Instead of claiming that 'survivors survive', the claim now is that 'reproducers reproduce' ('those who reproduce better, reproduce better'). Interestingly, the meaning of the first part of the sentence (the explanandum) is assimilated exactly to the second part (the explanans). By linking reproductive fitness to reproductive survival, the resulting interpretation of Proposition 1 becomes again a tautology. Alternatively, one may of course stick with the simple survival-interpretation, at least on one of the two sides; for instance: 'Those organisms that tend to survive longer (organismic survival), reproduce better

[3] There may be additional problems concerning an adequate choice of timespan. See Dawkins (1982/1989: 184) for an interesting discussion of the matter.

(reproductive survival)'. Long lifespan of animals (such as elephants), however, neither implies high reproduction rate nor high reproductive fitness. Thus this brings one back to a position between the Scylla of tautological formulation and the Charybdis of rendering natural selection plainly false.

Nevertheless, one may argue that we do not want to predict an organism's direct reproductive success (the number of offspring) by its personal fitness, but the occurrence of an organism's genes in the next generation, also influenced by an organism's effects on non-descendent kin (like the support of siblings); thereby potentially affecting copies of the organism's genes in other bearers. On the explanandum side, one is now interested in a general probability of the survival of an organism's genes, whether this is effected by the organism itself or by non-descendent kin (*inclusive reproductive survival*). Yet this modification on the explanans side requires analogous changes to the explanandum side as well. Personal fitness becomes changed into Hamilton's inclusive fitness. Deviations from predictions based on inclusive fitness will usually be attributed to auxiliary hypotheses not to the concept of inclusive fitness itself. To avoid rendering Proposition 1 plainly false, one needs to produce a tautological formulation:

Organisms with higher inclusive reproductive survival have higher inclusive reproductive survival rate

(6)

A third objection is that in evolutionary biology there are clearly abundant *specific theories and hypotheses* that are testable. Used fitness values need not be defined based on survival or reproduction, but rather on some additional, specific theory of design. As an example, assume a chart, depicting an observed frequency distribution of the gradation of beak-sizes for a species of Darwin finches. One may further assume an existing specific biological theory allowing the expectation that the fitness-level of large beaks is greater than that of small beaks. Based on this theory and on the refined mathematics of population-genetics, a precise prediction can be derived for further generations of Darwin finches. The predictions are not tautological (although the equations of population genetics, like all mathematics, may be interpreted to be tautological); yet if the prediction were falsified, what would usually be abandoned would not be the idea of natural selection (Proposition 1), but the specific biological theory instead (thousands such specific theories have been falsified in the past). Specific theories about evolution or fitness are by normal standards testable; this says little, however, about whether the principle of natural selection is testable (von Sydow 2012). Rosenberg (1994) stressed earlier that the problem of biological theories using specific optimal design arguments was that they 'easily lead to misidentifying the more fit

as the less, and vice versa' (461); also that accounts of optimal design are too heterogeneous to count as a unified theory, at least testing something other than 'natural selection' itself. It appears that 'survival of the fittest' per se may remain a metaphysical framework as long as conflicting evidence is normally taken to refute auxiliary hypotheses rather than the metaphysical core. Interestingly, however, during the testing of specific theories, the actual meaning of 'fitness' changes both due to substantial modifications (cf. the inclusive fitness approach or the propensity approach discussed next) and due to reference to more specific associated theories. 'Survival of the fittest', that is, would not have a fixed meaning specified by the principles of the theory, but rather changed its meaning over time, in an *ad hoc* way, to account for empirical findings and theoretical changes – thus again rendering it irrefutable.

Finally, a fourth objection to the tautological interpretation of Proposition 1 is that fitness must be defined not by survival or reproductive survival, but by a probabilistic propensity or disposition to survive or leave offspring (Mills & Beatty, 1979; cf. also Rosenberg 2008). Accordingly, absolute individual fitness values reflect an organism's expected number of offspring. Technically, the expected value reduces a given probability distribution over propensities to a single value. The relative fitness of organism *x* being larger than the fitness of organism *y*—$\omega(x) > \omega(y)$—is stipulated as equivalent to the probabilistic survival relation that '*x* is expected to leave more offspring than *y*'. Using probabilities allows one to state that a more fit individual may (by chance) not survive, while yet allowing the fitness claim to be sustained, by using information on the type level. Although this reasonable move rules out strict falsification, it does not rule out statistical tests. This testability, however, again seems to concern specific claims about fitness; and a refutation usually will lead to another specific and again testable hypothesis about fitness values without allowing for dismissing the principle of survival of the fittest. In the finch example, let us assume that a (fit) finch with a large beak does not survive. It may be argued that it is nonetheless a fit animal. Yet, a low survival-rate of similar finches 'tested' over the long run would lead to a modified fitness-value rather than to a falsification of the principle of natural selection. A propensity-formulation, may well have its advantages, but used in this way, does not fundamentally change the dependency of the explanans on the explanandum, and essentially leads to the following interpretation of survival of the fittest that does not – at least not in a relevant way – resolve the problem of tautology:

Organisms that survive with a high probability have a high probability to survive (7)

As has been shown up to this point, a tautological interpretation of ‘survival of the fittest’ can be defended quite well against a number of objections. Tautological interpretations, therefore, may have played a considerable role in immunizing Darwinism in general and process-Darwinism in particular.

Such tautological interpretations of ‘survival of the fittest’, however, would even accommodate for evolutionary processes that have been in stark opposition to any strictly Darwinian understanding of evolution, such as drift, divine creation, directed variation, strict instructivism, saltation, group selection, internal constraints, synthetic (rather than variational) processes, and orthogenetic, self-organizational or self-determining tendencies (for a historic account, see Gould 2002; von Sydow 2012). Even if the tautology problem up to this point has not been resolved for ‘survival of the fittest’, the meaning of natural selection may be supplemented by neo-Darwinian tenets to make it testable. Even formulating the claim that ‘survival of the fittest’ does not rule out processes regarded to be non-Darwinian, actually implies that there are meanings of Darwinism (or natural selection) that can be delineated from alternative processes.

One may, for instance, test the concept of drift (random change of gene or feature frequency) against natural selection (actually together with drift) *contrastively* by theoretically constructing two likelihood functions for a given phenotypic feature (e.g., beak size) (Sober 2008: 189–263). If the empirically found difference went in a direction other than that predicted by natural selection, and if the changes remain too close to the original state, it follows that drift has to be favoured over natural selection. Tests may indeed need to be contrastive; and given the likelihood functions, the above example provides a possible empirical test of natural selection. First, however, ‘survival of the fittest’ would remain a tautology if one continues to define ‘fitness’ by reproductive outcome only (cf. Rosenberg 1983). If one were to adjust the likelihood function based on the data, one would never − even in contrastive tests − be able to refute natural selection. The proposal actually *assumes* that there is a solution to our central question in the first place (how fitness could be defined independently from survival) by specifying a given likelihood function. Nonetheless, in practical terms, contrastive testing allows for cases where immunization will presumably not occur. Experimenters may actually refrain from accommodating their selection hypothesis in the light of the outcome if from the outset they explicitly aim to test against drift. Again, this needs to be based on a specific biological theory (here treated as

auxiliary hypothesis). Changing the auxiliary hypothesis, however, is neither logically excluded nor necessarily always wrong. Thus even for such contrastive testing, problems are not fully solved.

Second, the distinction between natural selection and drift is *conceptually* not the most central issue. Historically, Darwinism has mainly been associated with a relatively undirected chance process; and contrasting natural selection only against a process that is even more based on chance may be possible but seems to miss the point. Even in Darwin's time, John W.F. Herschel, astronomer and highly influential philosopher of science, disdainfully called natural selection the 'law of higgledy-piggledy'. The contrast to drift would not preclude to falsely identify 'survival of the fittest' with, for example, directed mutation, saltation or an evolutionary direction based on internal constraints. Thus, contrasting natural selection with drift does not provide a sufficient specification of the actual neo-Darwinian meaning of the term 'natural selection', nor makes it testable in a way that excludes its main historical alternatives.

To account for central aspects of neo-Darwinian theory, one may refer to natural selection as a full Darwinian process of blind variation and external selection (von Sydow 2012), not just as its second step (natural selection in the narrow sense). A Darwinian process is a search algorism in a design-space that appears to have testable aspects. We examine the two involved processes separately.

Blind variation, the first step of such a Darwinian process – albeit not part of natural selection in the narrower sense – has always been essential to neo-Darwinism, delineating it from more directed accounts of evolution. The claim of blindness of variation appears testable, whether one thinks of refutations of naïve Lamarckism or of recent suggestions that perhaps rehabilitate some role for use-inheritance linked, for example, to epigenetics. Nonetheless, testing strict blindness of variation in general is not a trivial issue. For the most part, blindness has been tested against 'omniscience' only, using the radical alternative hypothesis that organisms produce variations almost perfectly directed towards adaptation. Nevertheless, one may argue that strict blindness (of mutations or other trials in process-Darwinism) is conceptually only one extreme of a continuum of myopia ranging from complete blindness to omniscient production of variation, trials or conjectures (cf. von Sydow 2012). Yet dismissing this dichotomous alternative, strict Darwinism could not be proven correct by refuting only an extreme antithesis of perfect use inheritance, even if done in several biological cases. It is much easier to show that variation is *not* omniscient than to test for strict blindness. A proper evaluation of the theory that all

variations, mutations, trials and conjectures are blind needs not only to access the former but the latter as well. This holds for the different fields of process-Darwinism as well as for biology itself. But what would an empirical test of the blindness of a specific type of variation look like? As an example, assume that (a) neutral DNA – the so-called 'junk DNA', – was actually shown to be usable, with only minor modifications, to construct important elements in gene-regulation (as found, e.g., by Eichenlaub and Ettwitter 2011); that (b) mutations often transform whole strands of silent DNA into active DNA; and that (c) strands of such DNA have a significantly higher (albeit still low) fitness than purely random point mutations affecting the same number of nucleotides. Would this count as a violation of strict blind variation? If blindness is intended to be an empirical claim, such or similar examples must presumably be interpreted as disconfirming. Otherwise, another kind of immunization is present. Consider a second hypothetical experiment: (a) a species in its evolution has repeatedly profited from variations on a particular dimension (e.g., size); (b) evidence confirms that this species has an increased probability to produce mutations within this dimension; and (c) evidence confirms that mutations of this dimension have a significantly higher (albeit still very low) fitness than purely random point mutations affecting the same number of random nucleotides. Again, an adapted dimension of variation that actually turns out to increase the probability of survival of the organism or evolutionary line in the future, may be interpreted as a disconfirmation of strict blindness of variation (cf. von Sydow, 2012). Alternatively, however, one may plausibly defend the blindness assumption by arguing (perhaps based with reference to the problem of induction) that one is here still concerned with blind variation alone, even if the study controlled for chance fluctuations on various possible dimensions of change. Yet would not such a move turn a supposedly empirical claim into a merely tautological one? Of course, using tautological claims in science is not necessarily deplorable (think of mathematical equations), but this would require a truly metaphysical justification, not a supposedly empirical one. It remains regrettable, however, that many common-sense advocates of the idea of strict blindness employ this concept with empirical meaning without providing means testing it (in this respect, Dawkins [1983] may be seen as a positive counter-example, in consistently advocating a metaphysical basis to Darwinian claims). If, however, one does not want to advocate Darwinism metaphysically, the operationalization of more subtle tests of a strictly defined understanding of blindness needs to be implemented. This is a difficult yet interesting task: for example, how might one control for chance fluctuations of the fitness of different dimensions of

variation? Although the burden of proof is commonly placed on those criticizing the blindness assumption, for empirically minded advocates of the latter it is equally important to show the claim to be adequately testable. In conclusion, although 'blindness of variation' on the one hand seems a bold, non-tautological claim, it may be fully or partly immunized, by either advocating the concept of blindness metaphysically or by simply not testing against more subtle deviations. Taking such problems seriously, a more clearly testable version of the claim of blindness of variation (mutations, trials, conjectures, etc.), however, may well not stand up to more a refined empirical scrutiny (von Sydow 2012; cf. Fodor et al. 2010).

Alternatively one may want to abandon this central tenet of neo-Darwinism, as is sometimes done to extend process-Darwinism to fields outside of evolutionary biology (cf. Hull et al. 2001). Clearly this does not increase the testability of process-Darwinism. Yet natural selection may perhaps only require variation in the sense of a 'filter account', where more conjectures are made than are accepted. Darwin's Malthusian stress on a general overproduction of organisms seems testable,[4] but the question remains as to how a filter account could be refuted in general? Assuming, counterfactually, that almost all variations (mutations in biology, trials in psychology and conjectures in science) would be highly directed, and that almost all would survive (which is clearly not the case), this (by any historical standard) very non-Darwinian picture would be fully coherent with a filter interpretation of Darwinism, as long as only one variation would not get into the next generation. It would not be favorable to neo-Darwinism to be based only on refuting a straw-man hypothesis of totally omniscient trials. This would make fully or partly Lamarckian, orthogenetic, or saltationist accounts coherent with a neo-Darwinian account, which as a concept is inacceptable.

Is there a reasonable quantification, how small the percentage of surviving mutations, trials, or conjectures has to be, to count evolution as variational? Despite such difficulties, one may argue that a filter account at least provides a general idea of a simple algorism that can lead to evolution. Even if this abstract idea seems underspecified, as outlined before, it is correct and a merit of Charles Darwin. Nonetheless, this does not help us solve our main problem. First, the criterion of variation as such now

[4] One may argue that this has actually been criticized historically in the work of Wynne-Edwards. But even if his approach would partly be true, it is not necessarily to be understood as a refutation of a general filter idea.

does not distinguish natural selection even from mere drift defined by a random survival of variants. Second, as long as we do not add a criterion for fitness (the explanans) that is independent of survival, 'survival of the fittest' makes *no prediction* about *which* organisms survive (the explanandum).

This leads one back to the second step of a Darwinian process: natural selection or survival of the fittest (Proposition 1) in its narrow sense and its potentially tautological interpretations. It has been seen above, for instance, that disconfirming evidence will often be attributed to specific theories rather than to the principle of natural selection. Despite taking these tautological tendencies seriously and emphasizing their role, one may define natural selection itself using some further general attributes associated with neo-Darwinism, now linked to the second step of a Darwinian process, such as individual or genic selection, or the role of external environment (Darwinian externalism; cf. von Sydow 2012; Gould 2002). By counting progeny alone such distinctions are ignored (Rosenberg 1988).

Darwinism may be defined as a theory that is opposed to group selection, since Darwin's focus on the selection of individual organisms favoured "the most reductionistic account available" at that time (Gould 2002: 14, 125f., but cf. Sober 2011). Survival of the fittest can be turned into a testable claim by using 'Darwinian fitness' instead of fitness in a multilevel framework by using only the individual's or gene's fitness and by excluding a potential additive component of group fitness. Survival of the individually fittest organism or gene is testable, at least if one starts with a multi-level account (Sober and Wilson 1998; Gould 2002; Wilson and Wilson 2007). Although tautological arguments may have played a role in the unit-of-selection debate as well (von Sydow 2012), many evolutionary biologists want to define group level selection in a way that can be investigated empirically. Group selection since Wallace has been regarded to be a testable claim, although it has indeed fallen into disrepute since gene-Darwinism became popular (Dawkins 1976; Williams 1966). The recent revival of multi-level accounts, however, integrated selection on the level of the gene or individual with selection on the level of sub-populations, and allowed for slightly more altruism within species (Sober and Wilson 1998; Gould 2002; Wilson and Wilson 2007; cf. also Fehr, Fischbacher & Gächter, 2002; Nowak and Sigmund 2005). Thus defining natural selection based on a strictly reductive position in the unit-of-selection debate would render 'survival of the fittest' not only a universal law that is clearly testable, but also, presumably, plainly false.

One final way to achieve a testable, general definition of 'survival of the fittest', usable in biology as well as in the other fields of process-Darwinism, would be to understand 'fitness' as defined by an externally given, actual environment (von Sydow 2012). Historically, Darwinism has always stressed that organisms are determined by their environment, as opposed to other historical evolutionary approaches (orthogenesis, etc.) that stress the internal structure of an organism or population (Gould 2002; Fodor et al. 2010; von Sydow 2012). Variations may 'fit', in different degrees, into previously defined ecological niches. The general idea of distinguishing external from internal causes seems to be highly intuitive as well as allowing for testable externalist definition of natural selection. Nevertheless, practical testing of the survival of those organisms that best fit an *externally given* environment in fact poses many problems. In all situations that are slightly more complex than marbles and a predefined external sieve (almost all situations), one may easily resort to defining external fitness by survival, which would beg the question once again. Even though an internal-external distinction seems plausible and historically significant, evolution always involves a kind of dialectic interaction between genes, organism and groups (as internal causes), on the one hand, and their environments (external causes) on the other. One might even argue that an environment or niche could not be understood without an organism involved, and vice versa. Obviously, Darwinian externalism (Gould 2002, von Sydow, 2012) can only be tested by distinguishing both sides. This issue is at present far from completely resolved. Although recent decades have witnessed an increasing interest in the also historically important distinction between adaptations to an external environment and effects of internal morphological constraints (laws of form, correlation of parts) (see Gould and Lewontin 1979; Gould 2002), what remains disputable is how to distinguish such adaptations from internal constraints and so-called exaptations (features with changed function, Gould and Vrba 1998). It has recently been argued by Fodor and Piattelli-Palmarini (2010) that at least in the biological domain no experimental procedure or counterfactual argument can distinguish between adaptations and coextensive non-adaptations. Although Fodor and Piattelli-Palmarini have explicitly detached this problem from the tautology debate sketched here (2010: 210), these issues seem closely related. Their interesting proposal and its criticism cannot be discussed in detail here, but it appears that Fodor and Piattelli-Palmarini are nonetheless elaborating a new, at least potentially immunizing aspect of Darwinism (for further immunizing aspects, cf. von Sydow 2012). On a more pragmatic level, however, there seem to be ways to distinguish adaptations from exaptations at least intuitively (but quite inter-

subjectively). For instance, the exoskeleton of arthropods is plausibly described to impose a rather *internal* constraint on the size of organisms, whereas the form of the whale (originally a land-based mammal) rather seems to be an adaptation to an externally given environment. Although such explanations remain post-hoc, they seem to improve the just-so stories where adaptation is the only option, by now allowing for alternative either adaptive or exaptive stories. Although one is still concerned with a kind of historical, singular explanation, we are now at least concerned with a kind of contrastive testing of hypotheses (cf. Sober 2008). The structure of such post-hoc arguments indeed requires further explication (presumably involving probabilistic reasoning, counterfactual thinking, comparisons between classes of species, etc.), but the arguments themselves appear to have *prima facie* plausibility, potentially ruling out not only a tautological understanding of adaptationism, but of natural selection as well.

Moreover, it seems possible to formulate competing *predictions*, by testing internal (e.g., developmental) constraints against natural selection (given that we continue to use an externalist interpretation of this term). Although a disconfirming empirical result for natural selection may again be attributed to auxiliary hypotheses alone, researchers may not necessarily draw such immunizing conclusions without additional argument, if their intention from the outset has been to investigate the contrast of the two mentioned main hypotheses.

For the externalist definition of natural selection, however, we may leave aside the intricate issue of internal morphological constraints (and the interesting problem of co-extensionality; Fodor et al. 2010), since evolutionary biologists have started to distinguish, in practice, internal from external evolutionary causes in other fields as well (inner-populational dynamics and self-regulation in a multilevel-selectionist framework). Processes of frequency-dependent selection may yield an evolutionary outcome (potentially understandable as an ‘attractor’) that is mainly defined by the frequency-distribution of genes within the population rather than by the external environment alone (see, e.g., Nowak and Sigmund 2005). If one regards the initial frequency distribution of genes as an internal structure of the population and a population as a irreducible unit of evolution (perhaps due to such frequency dependent processes or other reasons), then the outcome of natural selection with regard to such populations may not only be determined by the external environment alone of such populations, but by their internal structure. Likewise, sexual selection, which in the gene-Darwinian tradition has been in

principle assimilated to natural selection, on a population level may likewise be interpreted as a process internal to a population that leads to some independence of its environment (von Sydow 2012).

If sensitivity to the tautology problem increases, one can be optimistic that evolutionary biologists will develop much more refined concepts for a proper internal-external classification than exist at present. If one accepts distinguishing internal from external 'selection' (or auto- from heteroselection: von Sydow 2012), it should be noted that the term 'natural selection' should be restricted to external selection only; otherwise the hypothesis ceases to be testable.[5] If at least *natural* selection is to be defined as testable theory in reference to an externally given environment, as proposed here, the Darwinian idea 'survival of the externally fittest', although still a major breakthrough, may turn out not at all to be a general truth, nor even to characterize appropriately the overall process of evolution (von Sydow 2001/2012). The only alternative appears to be to use 'natural selection' as a metaphysical claim that provides a framework in which one may discuss interesting empirical questions (such as the unit-of-selection debate or internal vs. external selection), without being able to test the framework itself. Although it seems to be no trivial task to turn the postulated blindness of mutations as well as the externality of selection into testable aspects of the definition of 'survival of the fittest', it seems the only way to appropriately resolve the problem of tautology.

The Problem of Tautological Formulations of Reinforcement

In this last section, it is outlined that in psychology the important theory 'trial-and-error learning' is actually beset with an analogical problem of testability and potential tautological formulation. Despite many differences, trial-and-error learning has been framed as a Darwinian process, involving blind variation and selection (Skinner 1981; Hull et al. 2001). Based on Edward Lee Thorndike's (1911) law of effect, Burrus Frederick Skinner (1904–1990), the father of operant conditioning and influential advocate of behaviourism, posited a principle of reinforcement that may be formulated as 'if a behaviour or response is followed by a reinforcing stimulus, its occurrence becomes more frequent', or more briefly:

A response increases if it is reinforced (8)

[5] Please note as well, 'selection' as such becomes a tautological term or principle referring to Proposition 2, as long as it is not defined in a testable way by one of the other discussed criteria, each with its specific problems.

Proposition 8, similarly to Proposition 1, establishes a relationship between an explanandum – 'the occurrence of a behaviour becomes more frequent' (a response increases) – and an explanans – 'a behaviour is followed by a reinforcing stimulus' (it is reinforced). The posited relationship between explanandum and explanans is normally interpreted as empirical generalization or empirical hypothesis. Skinner, however, actually proposed to define a reinforcing stimulus (a reinforcer) as any stimulus which, when presented after a response, leads to an increase in the future rate of that response. By this he avoided terms like 'pleasure', still used by Thorndike.

Early on, however, some authors noticed that on that basis the law of effect becomes a tautological claim (e.g., Postman 1947; Westmeyer 1973):

A response increases if a response increases (9)

Proposition 9, in analogy to Proposition 2, shows that the theory of reinforcement becomes an empty truism, or at best a metaphysical principle, if one follows Skinner in defining a reinforcer based on the response-rate alone.

Largely similar to the above discussion of fitness, one may object, for instance, that reinforcement has a common-sense meaning (cf. Proposition 3 and 4 in the survival-of-the-fittest debate). Postman (1947) was one of the first who pointed out the potentially circular interpretation and demanded that reinforcers be identified with effects of pleasure as originally assumed by Thorndike. Interestingly, Skinner's (1984: 220) redefinition of feelings, 'we should speak of feelings only when what is felt is reinforcing', would even lead to a tautological interpretation if he had accepted the hedonic re-definition by Postman. Here Skinner seems to have built a second immunizing protective belt. Another formulation, also going beyond a Skinnerian notion of a reinforcer, links reinforcement to reduction of a deprivation-state (such as lack of food). Again, one may question whether it is in fact possible to define states that deprive (the explanans) independently from rates of behaviour (the explanandum). Yet even if a partly independent definition is possible, deprivation in the strictest sense is understood biologically and hence this definition would tend to rule out existing social needs and secondary reinforcers, presumably playing an essential role for human beings (Westmeyer 1973). Thus these definitions either become tautological once again or they disprove the theory of reinforcement under important classes of conditions.

In a second analogy to our discussion of the tautology problem in evolutionary biology, there are also many *specific* theories in the field of reinforcement learning that are, of course, testable. One may

test specific theories about what stimuli count as reinforcers; for instance, with respect to people or situations. Second, one may test specific theories of learning, such as the Rescorla-Wagner law or one of a great number of more recent models. Although such specific models of learning often use several free parameters, most of them are clearly testable. Nonetheless, even the falsification of such theories will normally only lead to a replacement of such more specific theories without having to put the general law of effect into question.

One may claim, however, that reinforcers need to be trans-situational, trans-reactional, trans-personal and trans-temporal. The combination of the law of effect with such auxiliary hypotheses results in testable compound hypotheses (cf. Westmeyer 1973). In contradiction, studies of so-called 'biological preparedness' have in fact shown that certain reinforcers seem to be linked to particular plausible situations only. In such a situation one might still immunize the law of effect by giving up only a specific auxiliary hypothesis concerning the trans-situational applicability of reinforcers while retaining the law of effect (Westmeyer 1973: 55f.). Yet the above result has actually been seen as critical for the generality of operant conditioning. Thus the scientific community here seems to tend to interpret these finding as a restriction of the domain of application of the law of effect or of the auxiliary hypothesis that is may be seen to be actually essential to this law.

Another analogy to tautology debate in evolutionary biology is the proposal of a propensity definition of a reinforcer as a cure against the tautological understanding of reinforcement learning. It appears that this account can be described and challenged along similar lines as described above using a probabilistic formulation not only on the side of the explanans but also on that of the explanandum (cf. Proposition 7). This again appears would analogously lead to a decision between a true tautological and a false testable formulation.

The most reasonable alternative, as similarly proposed for natural selection, seems to link additional criteria associated with behaviouristic trial-and-error learning explicitly to the definition of reinforcement learning. Like for natural selection, the most reasonable proposal for the trial-and-error learning uses the postulated blindness of trials (blindness) and the exogenous character of reinforcement (externalism) as serious criteria for a testable formulation of reinforcement learning.

It has already been seen that it is no trivial task to operationalize the idea of strict blindness within biology. Nonetheless, without the blindness-criterion, sudden intelligent insight (Eureka-effects) and

Wolfgang Köhler's classical findings with chimpanzees could be reinterpreted without problem, but in a historically inappropriate way, as instances of operant conditioning. Yet conditioning has historically always been opposed to processes of sudden insight, just as Darwinism has been opposed to saltationism. Additionally, like in evolutionary biology it may be inappropriate to test blindness only against omniscience (direct insight) and not against partially directed trials. But even the preliminary historical distinction (blindness versus insight) makes plausible that the idea of blind trial-and-error learning can be formulated in a testable way. Taken as a testable theory, however, the claim of blindness, as some authors have argued, may well be empirically not universally valid (Sternberg 1998; cf. Hull et al. 2001).

With regard to externalism, reinforcement learning like natural selection is an opportunistic process of adaptation to the present, externally given environment. Similar to the biological internal-external distinction, it seems historically appropriate in psychology to distinguish between internal cognitive causes and an external environment. Skinner (1981: 503), for instance, explicitly advocated that in his theory 'there is no place for the initiating agent'. Nonetheless, again it is no mean task to operationalize this distinction. Although trial-and-error learning and behaviourism in general have explicitly stressed the opportunistic responses to an environment, and opposed the relevance of internal causes, it should be noted that one actually implicitly assumes internal changes, for instance, when accounting for secondary reinforcers. By taking the previous learning history of an organism into account, the internal-external distinction may become blurred and externality ceases to be a truly testable criterion. If the questions surrounding the operationalization of externality are not resolved, this poses a problem for both advocates and critics of reinforcement-learning. Although a more principled operationalization of externality would be helpful, common sense and history of science again do provide at least a preliminary understanding of the externality criterion. Historically, conditioning and behaviourism have been opposed to insightful inner restructuring, to learning based on inferences in internal representations – such as mental maps or causal relationships (for the latter see, e.g., Cheng 1997; Hagmayer et al. 2011; Waldmann and Hagmayer, in press) –, and to learning without external reinforcers. The historic studies by Edward Chase Tolman, for example, have often been interpreted as critical of strict reinforcement learning (without internally mediating variables). Tolman, for instance, studied the learning of rats in a maze, and showed that learning may take place without external reinforcement and without a rat having exhibited similar responses previously (reasoning in mental maps). Such findings were mostly seen as

problematic for strict classical conditioning and behaviourism, since they required the causal relevance of internal representations (such as reasoning with mental maps and expectations). Furthermore, Bayesian approaches in cognitive psychology suggest that humans test for instance logical hypothesis in a more intelligent, directed and inductively informed way than would be expected based on Popper's Darwinian approach of (blind) conjectures and (external) falsifications (e.g., Oaksford and Chater 2007; Kruschke 2008; cf. Sober 2008). Such examples may be taken to violate the externality assumption (and/or the blindness assumption), and even though further work is needed to operationalize the internal-external distinction (as well as the blind-seeing distinction), the examples show that the externality criterion (and the blindness criterion) for reinforcement learning may provide testable formulations. One might speak of *internal reinforcers* or perhaps, more accurately, of internal causes or reasons of behaviour as well. But the notion of 'internal reinforcers' would abandon the criterion of externality and-if defined in this way – the testability of the law of effect with regard to this criterion. If one gives up the criteria, a 'reinforcer' would be nothing but some cause or reason for showing a behaviour, whether in the past, in the present or the future, whether caused by the environment or the organism, whether based on observation or on reasoning The law of effect would become an empty tautological principle or, more positively put, a metaphysical framework to produce and investigate interesting empirical claims, without being empirically testable itself.

Conclusion

First, it was shown that tautological interpretations of 'survival of the fittest', based on defining the explanandum by the explanans, are surprisingly stable against several modifications of the meaning of the term 'fitness'. Simultaneous shifts on the sides of both the explanans and the explanandum were shown to be possible. If natural selection in biology is defined in this circular way it can never be refuted and is at best a metaphysical principle. Second, it was argued that one may nonetheless provide testable definitions of natural selection, based explicitly on using the concepts of *blind* variation (blindness) and *environmental* selection (externalism) in its definition. Although taking the tautology problem seriously, and pointing out that even the criteria cited are not trivial to operationalize, it was defended that 'survival of the fittest' may be formulated in a testable way. The testable formulations, however, may actually lead to a falsification of natural selection or to restricting its domain of application. Finally, it was argued that

in reinforcement learning, a Darwinian process analogous to natural selection, the problem of tautology can be discussed in an analogous way as well. Again much care is needed to disentangle tautological from testable aspects. Only then can one obtain a truly empirical theory that may indeed turn out to be false or at least incomplete. Alternatively, one may of course treat these theories as non-empirical metaphysical frameworks only, generating empirical hypotheses and contributing to a larger Darwinian metaphysics without being testable. Nevertheless, an implicit shifting between a testable and an untestable interpretation can be an illicit tactic to immunize natural selection or reinforcement learning while conveying the impression that one is concerned with testable hypotheses.

Bibliography

Brooke, J.H. and Cantor, G. 2000 [1998]. *Reconstructing Nature: The Engagement of Science and Religion*. Oxford: Oxford University Press.

Campbell, D.T. 1960. Blind Variation and Selective Retention in Creative Thought as in Other Knowledge Processes. *Psychological Review*, 67: 380–400.

Cheng, P.W. 1997. From Covariation to Causation: A Causal Power Theory. *Psychological Review*, 104: 367–405.

Darwin, C.R. 1985–. *The Correspondence of Charles Darwin,* edited by F. Burkhard et al. Cambridge: Cambridge University Press.

Darwin, C.R. 1986–1989. *The Works of Charles Darwin,* edited by P.H. Barrett and R.B. Freeman. London: Pickering.

Darwin, C.R. 1871. *The Descent of Man, and Selection in Relation to Sex*. 1st ed. (2 vols.) London: John Murray.

Dawkins, R. 1989. *The Selfish Gene*. Oxford: Oxford University Press.

Dawkins, R. 1983. Universal Darwinism, in *Evolution from Molecules to Men*, edited by D.S. Bendall. Cambridge: Cambridge University Press, 403–25.

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

Dawkins, R. 1989 [1982]. *The Extended Phenotype*. Oxford: Oxford University Press.

Dawkins, R. 2006. *The God Delusion*. London: Bantam.

Dennett, D.C. 1995. *Darwin's Dangerous Idea*. London: Penguin Press.

Dennett, D.C. 2006. *Breaking the Spell: Religion as a Natural Phenomenon*. London: Penguin Books.

Depew, D.J. and Weber, B.H. 1995. *Darwinism Evolving: Systems Dynamics and the Genealogy of Natural Selection*. Cambridge, MA: MIT Press.

Desmond, A. and Moore, J. 1992 [1991]. *Darwin*. London: Penguin.

Eichenlaub, M.P. and Ettwiller, L. 2011. De Novo Genesis of Enhancers in Vertebrates, *PLoS. Biol.* 9(11), 1–11.

Fehr, E., Fischbacher, U., Gächter, S. 2002. Strong reciprocity, human cooperation and the enforcement of social norms. *Human Nature 13*, 1–25.

Fodor, J. and Piattelli-Palmarini, M. 2010. *What Darwin Got Wrong*. London: Profile books.

Friedman, M. 1953. *Essays in Positive Economics*. Chicago: University of Chicago Press.

Gould, S.J. 2002. *The Structure of Evolutionary Theory*. Cambridge, MA: Belknap Press.

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 of London*, Series B, 205(1161), 581–98.

Gould, S.J. and Vrba, E. S. 1998. Exaptation – A Missing Term in the Science of Form, in. *The Philosophy of Biology*, edited by D.L. Hull and M. Ruse. Oxford: Oxford University Press.

Greene, J.C. 1981. *Science, Ideology and World View: Essays in the History of Evolutionary Ideas*. Berkeley, CA: University of California Press.

Hagmayer, Y., Meder, B., von Sydow, M. and Waldmann, M.R., 2011. Category Transfer in Sequential Causal Learning: The Unbroken Mechanism Hypothesis. *Cognitive Science*, 35, 842–73.

Hodgson, G.M. 1993. *Economics and Evolution: Bringing Life Back into Economics*. Cambridge, Oxford: Polity Press.

Hodgson, G.M. 2002. Darwinism in Economics: from Analogy to Ontology. *Journal of Evolutionary Economics*, 12, 259–81.

Hofstadter, R. 1955. *Social Darwinism in American Thought*. Boston, MA: Beacon Press.

Hull, D.L. 1981. Units of Evolution: A Metaphysical Essay, in *The Philosophy of Evolution*, edited by U.J. Jensen and R. Harré. Brighton: The Harvester Press, 23–44.

Hull, D.L., Langman, R., Glenn, S. 2001. A General Account of Selection: Biology, Immunology and Behavior. *Behavioral and Brain Sciences*, 24(3), 511–73.

Knight, D. 2004. *Science and Spirituality: The Volatile Connection*. London: Routledge.

Khalil, E.L. 1992. Neo-classical Economics and Neo-Darwinism: Clearing the Way for Historical Thinking, in *Economics as Worldly Philosophy: Essays in Political and Historical Economics in Honour of Robert L. Heilbroner*, edited by R. Blackwell, J. Chatha and E.J. Nell. London: Macmillan Press, 22–72.

Knudsen, T. 2002. Economic Selection Theory. *Journal of Evolutionary Economics*, 12, 443–70.

Kruschke, J. K. (2008). Bayesian Approaches to Associative Learning: From Passive to Active Learning. *Learning & Behavior, 36*, 210-226.

Lipton, P. and Thompson, N.S. 1988. Comparative Psychology and the Recursive Structure of Filter Explanations. *International Journal of Comparative Psychology*, 1, 215–29.Nowak, M.A. and Sigmund, K. 2005. Evolution of Indirect Reciprocity. *Nature*, 437, 1291–8.

Nowak, M. A., Tarnita, C. E., and Wilson, E. O. 2010. The Evolution of Eusociality. *Nature 466*, 1057-1062.

Nelson, R.R. 2007. Universal Darwinism and Evolutionary Social Science. *Biology and Philosophy*, 22, 73–94.

Mills, S.K. and Beatty, J.H. 1979, The Propensity Interpretation of Fitness. *Philosophy of Science*, 46, 263-288.

Oaksford, M. & Chater, N. (2007). *Bayesian Rationality. The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Okasha, S. 2006. *Evolution and the Levels of Selection*. Oxford University Press.Ospovat, D. 1995 [1981]. *The Development of Darwin's Theory: Natural History, Natural Theology, and Natural Selection, 1838–1859*. Cambridge: Cambridge University Press.

Paley, W. 2005 [1802]. *Natural Theology*, edited by M.D. Eddy, D.M. Knight. Oxford: Oxford University Press.

Plotkin, H.C. 1994. *Darwin Machines and the Nature of Knowledge: Concerning Adaptations, Instinct and the Evolution of Intelligence*. Harmondsworth: Penguin.

Popper, K.R. 1972. *Objective Knowledge: An Evolutionary Approach*. Oxford: Oxford University Press, 1972.

Postman, L. 1947. The History and Present Status of the Law of Effect. *Psychological Bulletin*, 44, 489–563.

Rosenberg, A. 1983. Fitness. *Journal of Philosophy*, 80, 457–73.

Rosenberg, A. and Bouchard, F. 2008. Fitness, in *The Stanford Encyclopedia of Philosophy*, edited by E.N. Zalta. [Online] Available at: http://plato.stanford.edu/ [Accessed 2 February 2013].

Scriven, M. 1959. Explanation and Prediction in Evolutionary Theory, *Science*, 130, 477–82.

Skinner, B.F. 1981. Selection by Consequences, *Science*, 213, 501–4.

Skinner, B.F. 1984. The Evolution of Behavior. *Journal of the Experimental Analysis of Behavior*, 41, 217–21.

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

Sober, E. 2008. *Evidence and Evolution: The Logic Behind the Science*. Cambridge: Cambridge University Press.

Sober, E. 2011. *Did Darwin Write the Origin Backwards: Philosophical Essays on Darwin's Theory*. New York: Prometheus Books.

Sternberg, R.J. 1998. Cognitive Mechanisms in Human Creativity: Is Variation Blind or Sighted? *Journal of Creative Behavior*, 32(3), 159–76.

von Sydow, M. 2005. Charles Darwin: A Christian Undermining Christianity? On Self-Undermining Dynamics of Ideas Between Belief and Science, in S*cience and Beliefs: From Natural Philosophy to Natural Science, 1700–1900*, edited by D.M. Knight and M.D. Eddy. Burlington: Ashgate, 141–56.

von Sydow, M. 2012. *From Darwinian Metaphysics towards Understanding the Evolution of Evolutionary Mechanisms.* Göttingen: Universitätsverlag.

von Sydow, M. 2013. Darwinian Metaphysics (pp. 1306-1314), in *Encyclopedia of Sciences and Religions*, edited by A. Runehov and L. Oviedo,. Heidelberg: Springer.

Waldmann, M.R. and Hagmayer, Y. (in press) Causal Reasoning, in *Oxford Handbook of Cognitive Psychology*, edited by D. Reisberg. New York: Oxford University Press.

Westmeyer, H. 1973. *Kritik der psychologischen Unvernunft. Probleme der Psychologie als Wissenschaft*. Stuttgart: Kohlhammer.

Williams, G.C. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton, NJ: Princeton University Press.

Wilson, D.S. and Wilson, E.O. 2007. Rethinking the Theoretical Foundation of Sociobiology, *The Quarterly Review of Biology*, 82(4), 327–48.

Wilson, E.O. 1976 [1975]. *Sociobiology: The New Synthesis*, 4th ed. Cambridge, MA, London: Harvard University Press.

Wilson, E.O. 2005. Kin Selection as the Key to Altruism: Its Rise and Fall. *Social Research, 72, 159-166.*