The estimator theory of life and mind

How agency and consciousness can emerge

Hans van Hateren

Contents

Preface

The way in which scientific results are presented is often quite different from the way in which they were obtained. This is very much true of the work presented in this book. It is presented as if simple models of basic properties of life naturally led to complex models of phenomena such as consciousness and language. However, that is not how the work was done. In reality, it is the result of a grand detour, starting and ending with questions about mind and consciousness. While touring, the results about evolution, life and philosophy were produced as spin-offs.

The detour was unanticipated when I decided, more than a decade ago, to broaden my earlier work on the neural basis of vision towards the field of conscious perception. I started this work with an extensive literature study on consciousness and language-like communication. I was intrigued by studies that compared the capabilities of apes and human infants, such as the work on shared intentionality by Tomasello and Carpenter (2007). But gradually it became clear to me that consciousness is not the only enigma. At least as puzzling is agency, the capacity of organisms to act in meaningful ways and to initiate novel behaviour. Agency seems to conflict with the regular chains of cause-and-effect that one is used to in the natural sciences. Moreover, evolution apparently has produced agency at a much earlier time than it produced consciousness.

When I traced the biological literature to earlier and simpler forms of life, I came across a line of research that studies the genetic variability in unicellular organisms in response to how much physiological stress they endure (Galhardo et al. 2007). Such organisms contain various mechanisms that assess life-threatening conditions and that subsequently utilize that assessment to drive mechanisms that promote or suppress genetic variation. When I built quantitative models to simulate this in simplified systems, I found that this can indeed be beneficial from an evolutionary point of view. Moreover, I found that variants of the mechanism that do not involve genetics but rather behavioural changes—made and retained only within the lifetime of an organism—were beneficial as well.

I soon realized that the mechanism is quite remarkable, because it is neither deterministic nor purely random. It produces real goal-directedness and agency. Moreover, it can only work when there is sustained evolution by natural selection. The basic ideas were published in van Hateren (2015a), and were further elaborated in subsequent publications. However, at that time I had only a vague (and, with hindsight, only partially correct) notion of how it might be related to consciousness. Only over the years, these ideas have matured and have resulted in full-blown theories of consciousness and intentionality (which is the cognitive capacity that is used when thoughts refer to things).

This book gives an overview of what has been done so far. It includes updated versions of several published articles, but also a fair amount of new, unpublished material. Although some of the original publications depend on equations and quantitative simulations, these are absent from this book. It is intended for readers with a general academic background or interest, but not necessarily with the skills to read equations easily. Everything here is explained in words. Nevertheless, the theory and topics covered are not simple, and the explanations often assume that the reader has at least some intuition for the dynamics of change.

Because the theory has an extraordinarily wide scope, affecting many different fields of knowledge, I have published it deliberately in journals that serve different segments of the academic community. Publishing outside one's own specialization is quite difficult in general, because one's grasp of the literature is inevitably limited (often depending primarily on review articles), which may annoy specialist reviewers[1]. But in this particular case, the endeavour was even more difficult, because the properties that are claimed for the proposed mechanisms often seem to conflict with conventional views. I therefore thank those reviewers and editors who showed the combination of stamina and out-of-the-box thinking that is required for appreciating this line of research. I also thank those reviewers and editors who were less appreciative but who nonetheless provided helpful comments for improving the explanation and presentation of these studies. Last but not least, I thank the colleagues and friends who commented on these ideas.

J.H.v.H.
University of Groningen
June 2022

[1] References in this book are primarily given as useful entry points to the literature; there is no claim to be complete or balanced: no scholar could hope to achieve that nowadays, given the extraordinarily wide extent of the fields covered here and the enormous size of any field's literature.

PART I: GROUNDWORK

Chapter 1

Introduction

The main topic of this book is the question of how things are caused in nature, particularly in those parts of nature that are alive. The scientific revolution that started around the time of Galileo (1564–1642) and Newton (1643–1727) was based on the idea that everything in nature happens according to fixed and quantitative laws that could be discovered through experiments. These laws of nature then describe what causes what, with unlimited accuracy and precision. Laplace formulated this idea explicitly at the beginning of the 19th century. He stated that full knowledge of the state of the universe at any particular time would allow one to use the laws for calculating the state of the universe at any other time, in the future as well as in the past. Such full determinism implies that living organisms cannot behave in any other way than dictated by the laws of nature. In other words, there would be neither agency nor free will, because all future behaviour would be predetermined and unchangeable.

Full determinism of cause and effect became gradually less useful and less tenable. In the late 19th century, it became understood that the physical laws regarding the temperature and pressure of a volume of gas were statistical in nature. The random movements of large numbers of molecules in a gas could explain macroscopically observed properties and laws. However, this type of randomness is not fundamental, but merely a practical consequence of not being able to measure the positions and velocities of, say, 10^{23} particles. It is an apparent randomness that can be attributed to a lack of knowledge. Despite this lack of detailed knowledge, statistics can then still produce useful results, such as accurate macroscopic laws.

More fundamental problems for determinism arrived with quantum physics in the first half of the 20th century. Microscopic particles, such as electrons and photons, do not have deterministic dynamics. They can still be described by laws, but merely in terms of chance. Particular outcomes of single measurements are not certain—not even in principle—before the act of measurement; they only come with computable probabilities. These probabilities show up explicitly only when an identical measurement is repeated many times. In contrast to the 19th century case of statistical physics, chance in quantum physics appears to be fundamental rather than attributable to a lack of knowledge.

Even if there is fundamental randomness at a microscopic scale, one might think that such randomness would average out when going to macroscopic scales, such as those of everyday life. Then the macroscopic dynamics would still be deterministic, at least for all practical purposes. However, this expectation was undermined, in the second half of the 20th century, by the discovery that chaos and unstable dynamics are widespread in nature. Many macroscopic systems have a dynamics that is deterministic in principle, but that is at the same time sensitive to even the slightest microscopic disturbances. Small microscopic indeterminacies are then amplified to large macroscopic indeterminacies. When microscopic randomness invades such unstable dynamics, at the start as well as continually during the time in which the dynamics is observed, the state of a system may become largely indeterminate over time. This happens not because of a lack of knowledge but rather

fundamentally, because the microscopic randomness itself is fundamental. Many macroscopic systems contain at least some of this fundamental randomness, in addition to having an overall dynamics that is describable by deterministic laws.

However, randomness would do no better than determinism in producing agency and free will. Behaviour that is random is perhaps even less worth wanting than behaviour that is predetermined (Dennett 1984). Random behaviour is meaningless, by definition. It is widely believed and claimed that chance and determinacy are the only two fundamental possibilities here. This perceived dichotomy probably stems from the idea that the dynamics of the world can be described as a process that progresses through time instantaneously, in infinitesimally small steps of time. Such a description conforms to the way Newton used differential equations of time, and it is still the standard way to model physical reality today. However, behaviour that shows agency and free will is never instantaneous, but extends over macroscopic time, in the order of seconds or considerably more. In that case one might suspect that certain combinations of determinacy and randomness could exist that depend on the statistics of randomness rather than on single random events. If such a combination would exist and would have the right properties, it might have causal consequences that would not comply with the above dichotomy. As it turns out, such a combination is possible. A major task of this book is to explain this and to explore its consequences.

The basic mechanism is explained in the next few chapters. Surprisingly, it produces not only agency, but also a range of other poorly understood properties of living organisms, such as intentionality (the referring power of minds), goal-directedness, values, and, for organisms that have the capacity to communicate intentionality, consciousness. It affects basic features of life (Part II of the book) and of mind (Part III), and enlightens several long-standing issues within the field philosophy (Part IV). Most importantly, it shows that things in nature can be caused by agency, rather than exclusively through determinism and chance.

Chapter 2

The basic mechanism

This chapter and the next few will introduce the basic mechanism, which is subsequently applied to specific topics in subsequent chapters of the book. The explanation and figures are adapted from earlier publications, in particular from van Hateren (2015d, 2017, 2019). The notation of variables and processes has been updated and unified such that it is suitable for the wide range of topics discussed in this book. The specific mechanisms discussed below are all based on a single dynamical principle. But they can be grouped according to timescale and scope of evolutionary fitness, and they can be explained in two distinctly different ways. These subdivisions will now be introduced briefly.

The mechanism can be realized on two vastly different timescales. The first one is the long timescale of evolutionary change. On this timescale, changes do not occur within a single organism but through hereditary change along a line of descending organisms (Section 2.1). The second timescale is the much shorter one that spans the lifetime of an individual organism. Here, changes occur only within—and limited to—each individual, in particular as changes in its behavioural dispositions (Section 2.2). The latter mechanism can be generalized such that it can be applied to social and cultural species, but for that it is necessary to define evolutionary fitness with an extended scope (Chapter 3).

Explaining the mechanisms can be done in two rather different—but ultimately equivalent—ways. The first way of explaining depicts the mechanisms primarily in terms of cyclical dynamics. Such dynamics result in semi-random trajectories through an abstract, high-dimensional space (such as a space of hereditary forms or a space of forms with behavioural dispositions). This explanation is best suited for understanding a phenomenon such as agency, which provides an organism with some behavioural freedom. The second way of explaining depicts the mechanisms primarily as diffusion processes that produce clustering of forms. It is best suited for understanding a phenomenon such as intentionality, which lets an organism assign meaning to the world. Because both depictions are equally valid and provide insight in different ways, I will present them both.

Before explaining the basic mechanism, I will first make a few general remarks about causation in nature. For the present purpose, the term ‘causation’ is used in a common-sense way (see Section 14.2.2 for more discussion). It refers to the relationship between a cause and its subsequent effect, both understood as changes in time. Varying a cause, such as by changing its strength or by arranging it to be present or not, will then modify the effect in a systematic way. Broadly speaking, there are two fundamental forms of causation in physical nature. The first form, deterministic causation, is illustrated in Fig. 1a. The graph shows the change of a variable, such as a state or some property of a system. This change is caused by other variables (left arrow), and it subsequently causes changes in downstream variables, either in the same system or in other systems (right arrow). Causes can be multiple and complex, but the crucial property of a deterministic system is that the change of state remains fully determinate, in a similar way as the state of a clock changes in a determinate way through the motions of its cogwheels. In principle, one could predict how the system’s state changes through time, to arbitrary accuracy. In practice, there are limits to this

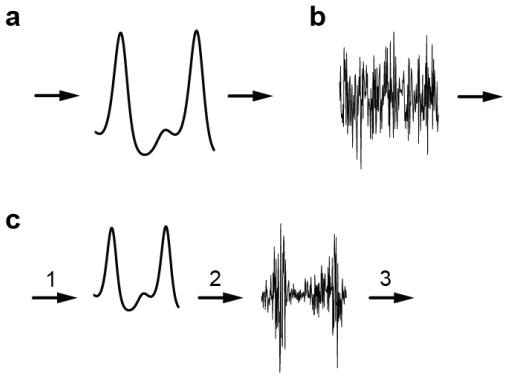

Fig. 1. Various forms of causation. (**a**) In deterministic causation, a time-varying variable (representing a system state or property) is caused by (left arrow) and causes (right arrow) other variables. (**b**) In random causation, a random variable can start new chains of causation (arrow). (**c**) In modulated random causation, a non-negative deterministic variable (left curve) drives the variance of a random variable (right curve).

predictability, because real systems always display some noise. Then 'deterministic' should be understood as 'primarily deterministic'.

The second fundamental form of causation originates from pure randomness, as is illustrated in Fig. 1b. A random process produces changes over time that are not caused by upstream factors but arise spontaneously. For example, atoms in a heated gas may emit photons spontaneously, and radioactive atomic nuclei may decay and emit a particle spontaneously. Such random events can then become the starting points of novel downstream causal chains (arrow to the right of Fig. 1b). In practice, random causation can have various origins: it may originate from thermal or quantum noise, from untraceable external disturbances of a system, and from unstable dynamics that amplifies microscopic indeterminacies (as in chaos). Random causation implies unpredictability. Randomness is ubiquitous in nature in general and in living organisms in particular, from the molecular to the behavioural level (Faisal et al. 2008; Kiviet et al. 2014). Sometimes randomness is only apparent because one has limited knowledge about a system. But the type of randomness meant here is taken to be real and fundamentally present (see further Section 14.2.1).

A very specific combination of deterministic and random causation is illustrated in Fig. 1c. It can be called 'modulated random causation', and plays a major role in the mechanisms to be explained below. In this form of causation, one variable (left curve) is caused deterministically by upstream factors (arrow 1). This variable, which is assumed to be non-negative, subsequently modulates the variance of a second, random variable (right curve). Subsequently, this random variable causes changes in downstream factors (arrow 3). For the purpose of presentation, the deterministic variable is shown here as changing slowly and the random variable as changing fast, but this is not required. This type of causation still occurs when the two variables have similar temporal properties, even though it would then be difficult to visualize in a simple graph. Modulated random causation is neither completely determinate (because of the randomness), nor completely indeterminate (because the variance of the random variable changes in a deterministic way). Nevertheless, it is merely the product of two factors that correspond to the standard forms of causation. Because it would be straightforward to separate these factors, modulated random causation is not a fundamental form of causation. Moreover, it is, in its pure form, rather special and

therefore likely to be unstable and short-lived. But below we will see that it can become stable when it is part of a highly specific mechanism, if that is incorporated in living organisms that are subject to sustained evolution by natural selection.

2.1 The mechanism on an evolutionary timescale

The theory conjectures that all living organisms contain an internal process, called X below, that estimates the evolutionary fitness of the organism itself. This process subsequently modulates the variability of the organism in such a way that the actual fitness is likely to increase, on average. Below, I explain the theory qualitatively (for quantitative studies see van Hateren 2015a, c, f and summaries in Appendix A). First, I explain how fitness is defined here, second, how X can be understood, third, how X is thought to affect the organism, and, finally, why X produces a new form of causation that does not conform to the two standard forms—deterministic and random causation—that were discussed above.

A major feature of any biological organism is its evolutionary fitness. Depending on the application, fitness is defined and used in various ways in biology. Often it is used as a purely statistical concept (as in population genetics), but alternatively it can be defined in a more mechanistic way, as a property of individual organisms. The latter is chosen here. Fitness, in its most basic form, is then understood as an organism's propensity (i.e., capacity and tendency) to survive and reproduce. It is then quantifiable by a suitable combination of the expected lifetime of an organism and its rate of reproduction. Thus, fitness is used here as a concurrent measure of an organism's likely evolutionary success. It is not used as a post-hoc measure—made with hindsight—of an organism's actually realized success. More generally, it quantifies—as a statistical expectation—how effectively an organism may transfer its features to other organisms, in particular to those of subsequent generations. This leads to generalizations of fitness that include fitness effects produced by kinship and by social and cultural transfer of properties (Chapter 3). Importantly, fitness, as used throughout this book, is a forward-looking, probabilistic measure; actually realized survival and reproduction subsequently vary randomly around the expected value. Moreover, fitness is understood to change from moment to moment. For example, fitness is lower at times when food is scarce, because such scarcity decreases the organism's chances of surviving and reproducing. Internal factors, such as malfunctioning internal organs, have similar effects. But fitness can recover when conditions improve. It becomes zero when the organism dies.

Evolution by natural selection occurs when organisms in a population vary with respect to their typical fitness, on the assumption that at least part of that fitness is produced by heritable traits. The precise form that fitness takes is not crucial for the mechanisms discussed here, as long as it is an adequate measure of likely evolutionary success. Fitness is produced by a large range of factors that originate from the environment and from within the organism. All factors that affect fitness can be conceived of as forming a highly complex fitness process, F. F is the totality of influences and processes that actually produce fitness. The latter is a single number, the outcome of the process F, and it is denoted by f. In its simplest form—in asexually reproducing organisms with a fixed lifespan—f can be interpreted as a reproductive rate. This rate equals the number of offspring of an organism that is expected, on average, over its lifespan. When the mean fitness f of the organisms in a population equals one, the population size will remain stable (apart from statistical

fluctuations). A mean f that is larger than one results in exponential growth of the population size, whereas an f that is smaller than one leads to decline and eventual extinction of the population.

Each individual in a population is assumed to produce offspring that has similar but slightly varied hereditary traits that subsequently influence the offspring's fitness. If the fitness of an individual is sufficiently large, it has a good chance of staying alive and reproducing. But when fitness becomes too low, the individual may not contribute much (in terms of hereditary traits) to the future population: the individual may have few surviving offspring or may even die before reproducing. By this process of differential reproduction, individuals vary with respect to how effectively they transfer their hereditary traits to future populations. This gradually changes the likelihood that specific traits occur in future organisms, that is, it changes the likelihood that organisms with such traits are present; equivalently, the distribution of traits over a population of organisms gradually changes across time. This process of fitness-driven change is called evolution by natural selection—which is, essentially, evolution by differential reproduction. This was Darwin's great insight, and it is symbolized by the loop 'D' in Fig. 2.

The fitness process F depends not only on external circumstances, but also on the internal state and structure of the organism itself. The state and structure—to the extent that they affect fitness—are together called here the (biological) form of the organism. When circumstances change, fitness f may change as well. If it decreases and such a decrease is indirectly detected by the organism (such as when food becomes scarce), then this usually engages compensating mechanisms. For example, the organism may switch to other food sources or may lower its metabolic rates. Such compensating mechanisms can be viewed as forms of phenotypic plasticity (Nussey et al. 2007). The phenotype of an organism is the totality of its properties, as interacting with environment and other organisms. Phenotypic plasticity then refers to systematic changes of an organism's form during its lifetime, which includes, for example, changes in behavioural dispositions. Compensating mechanisms may either be fully inherited (when they originate from previous evolution) or not or partially inherited (such as when they are mostly established by previous learning by a particular organism). In either case, they respond to a problem that has occurred before, presumably

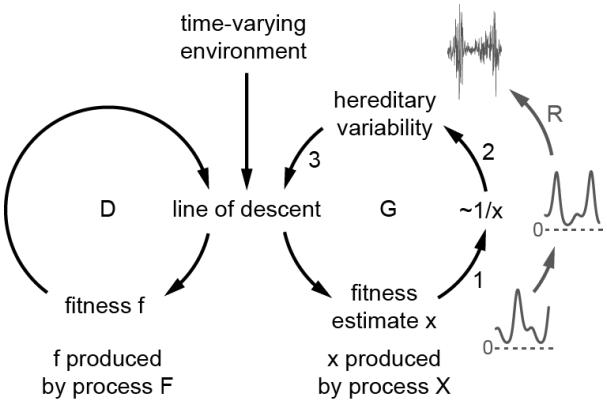

Fig. 2. The mechanism on an evolutionary timescale. A reproductive cycle D produces basic Darwinian evolution by natural selection, based on the fitness f produced by a process F. A cycle G randomly changes the hereditary properties passed on to an organism's offspring, with the on-average-expected amount of change being modulated by an internal fitness estimate x produced by a process X. The numbers correspond to those in Fig. 1c.

many times. Inherited or learned compensating mechanisms are not further considered here, but are merely acknowledged as an established baseline. The mechanism discussed below and the one discussed in Section 2.2 are taken to work on top of this baseline.

For the evolutionary mechanism discussed in this section, we will only consider hereditary change. Changes then occur in lines of descent, that is, lines of descending organisms. Such lines may split into many branches (when an organism in a line gets multiple offspring) or a line may die out (when the final organism in that line dies without having offspring). The survival of a line of descent thus depends on how its fitness varies over time, which means that it depends on the fitness of the organism that represents the line at a particular point in time. In order to keep the formulations short, 'line of descent' and 'line' below mostly stand for 'the organism representing a line of descent at a particular point in time'.

When circumstances change in an unexpected way, such that no ready-to-go compensating mechanisms are available to a line of descent, it may still need to respond. If it would not respond when fitness is low, it may die out. Without the availability of established compensating mechanisms, any response can only be random and undirected. Specifically, the response can only consist of random and undirected variations of traits as they are passed on to the next organism in the line of descent. Yet, even if it cannot be known in advance which direction of the response is best, this is not true of the mean magnitude of the response. The following qualitative considerations make this plausible. When the fitness f of a line of descent becomes large as a result of changing circumstances, then there is little reason to change the form of the line of descent (i.e., by changing the biological form of the offspring of the current organism). The line is already performing well, and even improving. On the other hand, when f becomes small as a result of changing circumstances, not changing a line's form may soon result in extinction. Then, it is better to change its form, in any direction. Although this may initially lead to even lower f and may thus increase the chance of extinction, it also increases the chance that a form with higher f is found—perhaps after continued change. On average, taking this chance is still better than not changing at all and waiting for almost certain extinction (this is supported by quantitative simulations; van Hateren 2015a and Appendix A).

Thus, the variability of changing a line's form should be a decreasing function of f: large variability when f is small ('desperate times call for desperate measures', if desperate includes undirected) and small variability when f is large ('never change a winning team', or at least not much). Note that changes are made in a random direction, and that only the statistics of their magnitude (i.e., the variance) is modulated. This means that the mechanism acts in a slow, gradual and stochastic (i.e., random) way, not unlike the process of diffusion. The random changes let the surviving branches of a line of descent drift through an abstract, high-dimensional space of forms, drifting faster where fitness is small and slower where fitness is large. In effect, it lets the forms of the surviving branches of a line of descent move away from forms with low fitness (as a result of the high variability there) and lets them stay close to forms with high fitness (as a result of the low variability there).

Although fitness is a feature of any organism, it is a factor that cannot be observed directly. The only way by which a line of descent can benefit from the above mechanism is when each of its organisms contains an internal process that makes an estimate of its own fitness. Such an estimate is evolvable, because it is part of a mechanism that increases fitness (see Appendix A). Moreover, it is under selection pressure to become and remain

adequate as a predictor of evolutionary success. The estimate is called x below, and the process that produces it, X, is called an estimator. This corresponds to the modern, statistical use of that term: an estimator is a procedure (here X) that produces an estimate (here x) of the value of a variable (here f). Then x is the outcome of a complex physiological or neuro-physiological process, X, occurring within each organism. How x affects the line of descent is symbolized by the loop 'G' at the right of Fig. 2. It runs in synchrony with the reproductive D-loop. Each cycle through the loops corresponds to the transition to a subsequent organism in a line of descent. Hereditary variability is made to depend on the fitness estimate. The '~1/x' in the figure symbolizes the requirement that large x (when fitness f is estimated to be high) should produce low variability, whereas small x (when f is estimated to be low) should produce high variability. The numbers at the arrows correspond to the numbers at the modulated random causation that is illustrated in Fig. 1c.

One way to realize a modulation of variability is by changing a rate R of random micro-changes (i.e., R is the number of micro-changes per unit of time). Such micro-changes in biological organisms are typically produced by random molecular motion (i.e., thermal noise). Because the micro-changes are random, one expects large variability of the accumulated change per unit of time when R is large, and low variability when R is small. The traces and arrows to the right of Fig. 2 illustrate this: where the fitness estimate x is low (lower trace), the desired variability (middle trace) and thus the rate R should be high, which then results in a high realized variability (upper trace).

It is important to understand that x is not a kind of fitness, but a fitness estimate. The value of x should at least roughly reflect the value of f, similarly to how the reading of a thermometer should roughly reflect the actual temperature of the medium measured. The reading is an estimate, but it is not itself a kind of temperature. Estimates need not be direct measurements, as they could also result from simulation, for example when a temperature is estimated by a computer program that is running a simulation of the weather. Again, the computed temperature estimate is then not a temperature itself. As with any estimate, the quality of the estimate x—how accurately its value tends to correspond to the value of f—could vary from poor to excellent. Finally, it is important to understand that also the processes X and F are very different entities, in the same sense that a weather simulation (made through observation and computation) is qualitatively different from the weather itself.

Both X and x are taken to be distributed throughout the organism, analogously to how that happens in an artificial neural network. Moreover, the resulting structural changes that are produced (see below) are assumed to be similarly distributed. The physiological realization of X depends on the species. In unicellular organisms (e.g., bacteria), it has to be fully realized by intracellular processes, such as those involved in sensing, computing and acting. In multicellular organisms without extended nervous systems (e.g., plants), the process also involves physiological mechanisms for intercellular communication and regulation. In organisms with brains, much of X is thought to be realized by sensory and neural processing.

The existence of X is a theoretical conjecture for which there is currently no direct empirical evidence. However, it is plausible that an X process can be present, given current knowledge of (neuro)physiology. Organisms routinely monitor many internal and external variables that affect their fitness. For example, a unicellular organism monitors the presence of nutrients surrounding it. Organisms contain physiological or neural circuits that can

respond to adverse or beneficial conditions if these are indicated by such monitoring. For example, an organism may respond to a lack of a specific nutrient by moving to a different place or by switching to a different kind of nutrient. Such responses are typically made in primarily deterministic ways, as part of conventional cybernetic control circuits (not unlike the ones used in systems engineering and robotics). However, the circuits that detect adverse or beneficial conditions can play a dual role by also participating in the X process. The response produced by this process is not deterministic at all, but purely in the form of modulating random variability. Nevertheless, X does not need much additional circuitry for being present, because it can piggyback on existing molecular, cellular and neural circuitry. Metaphorically speaking, it would be a fuzzy, stochastic mechanism that is interwoven with the more easily observed deterministic mechanisms. The term 'stochastic mechanism' is used here and below to denote a mechanism with a causal structure that depends at least partly on randomness. The G-loop of Fig. 2 is a stochastic mechanism.

The main effect of X, modulation of randomness, is a plausible mechanism as well. Physiology and neurophysiology are based on molecular processes, which are intrinsically highly variable (mainly because of the thermal variability that is inevitable when the number of molecules is small). Such variability is detrimental for the working of many biological subsystems. Thus, a large range of mechanisms exist that specifically reduce variability (e.g., DNA proofreading and repair, intracellular molecular amplification, and averaging over time and space by sensory and neural processes; see, e.g., Faisal et al. 2008). Varying the engagement of such variability-reducing mechanisms readily produces the type of modulation of variability required by the theory explained here. In other words, variability is typically controlled already, and modulating variability just requires controlling the control.

There are two ways of explaining the mechanism, through dynamical trajectories and through statistical clustering. Both help to comprehend how the mechanism works and what that implies. The first way closely follows the dynamics of the loops in Fig. 2. When an organism in a line of descent encounters a situation where the actual fitness f is high, it is likely to make a fitness estimate x that is high as well. We assume here that the organism has already acquired, through previous evolution, an X process that performs well in this respect. Large x means low hereditary variability, thus offspring (i.e., subsequent organisms in the multiple lines of descent that result when fitness is high) will remain similar. Then offspring retain high f if circumstances remain similar. But circumstances are assumed to change continually. This may happen to drive f even higher (with even lower subsequent variability), but, more likely, it may reduce f and thus lower its estimate x. Then hereditary variability increases. Because the hereditary changes are undirected, many of the subsequent lines of descent are likely to have reduced fitness, and will eventually perish. But occasionally, fitness may increase sufficiently much such that many new lines of descent can arise. The exponential growth in numbers associated with high fitness can then more than compensate for the low likelihood of obtaining high fitness. Overall, this can be a better strategy than keeping variability at a constant level, provided that the control of variability is well tuned to the variability of the environment. In computational simulations, a population of organisms that each follow this strategy outcompetes a population of which the organisms have a fixed (but optimized) variability of hereditary change (van Hateren 2015a and Appendix A). In other words, the strategy increases the fitness of lines of descent, or, equivalently, increases the fitness of each organism incorporating the strategy, at least

on average.

The organisms belonging to the surviving branches of a line of descent follow a trajectory through an abstract, high-dimensional space of organismal forms. This trajectory is driven by the G-loop of Fig. 2 to the extent that the organismal forms are shaped by heredity. The abstract space of forms is abbreviated to 'form-space' below. A trajectory moves to subsequent positions in form-space by making undirected and random steps, because each subsequent change in heredity (as realized in offspring) is undirected and random. However, the magnitude of each change is not fully random: each magnitude belongs to a probability distribution of which the mean is modulated by the fitness estimate x. As a result, the trajectory as a whole is not fully random either.

The G-loop of Fig. 2 is in fact a rather complex feedback loop, with a dynamics than can be understood as follows. Suppose we start with a particular organism with a form that, combined with the environment, produces a particular x. Then the form of the next organism in the line of descent depends, in a probabilistic way, on the hereditary variability that is modulated by this x. The resulting form then produces a new fitness f and a new fitness estimate x, which again drives subsequent hereditary variability and the form of the next organism, and so on. Thus, the trajectory is partly driven by x, because x modulates the statistically expected magnitude of the random hereditary steps. Each cycle through the G-loop of Fig. 2 further entangles two distinct factors: a random one and a determinate one (x). For a long trajectory, it is impossible to disentangle these two factors: in contrast to Fig. 1c, the deterministic and random causation can now not be separated as two multiplied factors. This means that the trajectory is, in effect, caused by a factor that is intermediate between the two fundamental types of causation of Fig. 1a (deterministic) and Fig. 1b (random). In other words, it is a form of causation that must be regarded as a third fundamental type, which is on an equal footing with the other two. It can be shown that it represents a distinct, strongly emergent form of causation (see Chapter 14). It is realized in Fig. 2 by a special mechanism that depends on life and sustained evolution by natural selection.

The second way of explaining the mechanism of Fig. 2 does not focus on the dynamics of specific trajectories, but on the statistics of clustering in form-space. There are in fact two clustering processes depicted in Fig. 2, one produced by the D-loop and another one by the G-loop. This can be understood as follows. The D-loop leads to differential reproduction of biological forms. Organisms with a form that produces high fitness f get more offspring, on average, than organisms with a form that produces low f. This means that there will be more organisms at positions in form-space that produce high fitness—in a given environment—than at positions that produce low fitness. In other words, organisms will cluster at and around high-fitness positions in form-space because of a high rate of reproduction there. In contrast, low-fitness positions will be only sparsely occupied by organisms. Thus, differential reproduction leads to clustering in form-space. A condition for such clustering to occur is that the environment is sufficiently stable to provide more or less stable fitness values within form-space, at least long enough to enable clustering. Environments are assumed to vary over time, but the slow temporal components in their time course should be sufficiently strong such that clustering—and thus natural selection—can work effectively (the simulations of van Hateren 2015a use a scale-free, power-law temporal environment, which means that it contains variation across many timescales; see also Bell 2010).

In addition to clustering by the D-loop, the G-loop of Fig. 2 leads to clustering as well, but through a completely different mechanism. Organisms with a form that produces large x will give small hereditary variation to their offspring, on average. Thus, they tend to stay close to the ancestral form: they seem to stick around in form-space. In contrast, organisms with a form that produces small x will give large hereditary variation to their offspring, on average. Thus, they tend to move away (in form-space) from the ancestral form: they seem to be repelled from low-x positions in form-space. Because of the variability, it is a statistical process that can be viewed as position-dependent diffusion in form-space. Forms diffuse away quickly from regions in form-space with low x, whereas they diffuse away only slowly from regions in form-space with high x. In effect, forms will then cluster at and around high-x positions in form-space. An analogy may help to explain this. Suppose one lets a drop of ink diffuse in a container of water, and suppose that the temperature of the water is kept inhomogeneous. Zones with low temperature water are intermittent with zones with high temperature water. Low temperature water implies a lower diffusion speed of the ink particles (because they are hit less often and less vigorously by the water molecules) than high temperature. Then ink particles will be expelled more quickly from the zones with high temperature than from the zones with low temperature. At any point in time, ink particles are thus more likely to be in the latter zones. In other words, the ink particles tend to cluster in the low temperature zones.

The two clustering processes discussed above will align when x is indeed an estimate of f, as is assumed in Fig. 2. The statistical clustering produced by x will then help to keep the organisms close to the points of highest fitness, while still allowing fast change when the positions of high fitness move around in form-space because of environmental change. In effect, the fitness that results from the alignment will be higher than when the G-loop would be absent (van Hateren 2015a). However, the resulting fitness is only obtained gradually and slowly, because it depends on a statistical (diffusion-like) clustering mechanism. In order to stress this, the gradually resulting fitness will be denoted by ‘fitness-to-be’ (symbolized by f_+). Current fitness is then still called f.

2.2 The mechanism on an individual timescale

For the mechanism discussed in this section, we will only consider changes that occur within an individual organism over the course of its lifespan. Such variability of behaviour and of behavioural dispositions affects the organism, but the resulting change is usually not directly inherited and is assumed here not to affect subsequent organisms in the organism’s line of descent. The term ‘behaviour’ should be interpreted very broadly here. It includes development, learning and phenotypic plasticity in its widest sense. It also includes internal physiological changes within unicellular organisms and plants. As before, inherited or learned compensating mechanisms in response to changing circumstances are not considered here, but are merely acknowledged as an established baseline.

When circumstances change in unexpected ways within the lifetime of an organism, and when the organism has no established mechanisms for dealing with those changes, it may still need to respond. Not responding risks sustained low fitness and eventual death. Because of the above assumptions, such a response must consist of random and undirected changes in behaviour and in behavioural dispositions. But similarly as before, the variability of the response should depend on the fitness estimate x. The G-loop of Fig. 3 depicts this. As

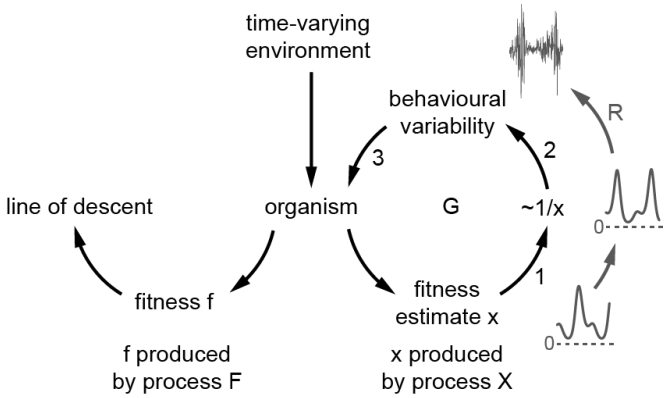

Fig. 3. The mechanism on an individual timescale. An organism participates in the evolutionary process based on its fitness f as produced by a process F. A cycle G continually updates an organism's behavioural dispositions during its lifetime, with the amount of change being modulated by an internal estimate of fitness x produced by a process X.

before, the variability can be modulated by changing a rate R of random micro-changes. These can correspond, for example, to random molecular changes in the cellular or neural circuits responsible for behaviour and behavioural dispositions. The G-loop cycles here at a much faster rate than the corresponding loop of Fig. 2, with each cycle taking only a fraction of the lifespan of the organism. Moreover, the changes made to the behaviour of the organism are assumed not to be transferred to offspring (which is the typical biological case, but see Chapter 3 for exceptions involving social and cultural transfer). Nevertheless, the G-loop of Fig. 3 helps to increase fitness. When fitness f is high, the estimate x is likely to be high as well. It is important to recall here that both f and x are changing continually in time, as produced by continuous processes F and X, respectively. Thus, f and x can vary considerably within the lifetime of a particular organism, depending on its time-varying circumstances and behaviour. When the fitness estimate x is high, the organism is likely to do well, and there is no reason to change much. Then its behavioural variability should be low (which is at the points where ~1/x is low, see the traces to the right of Fig. 3). On the other hand, when fitness is estimated to be low, the organism should change more in order to avoid deterioration and eventual death. Large variability then lets the organism quickly explore other behaviours. Most of these may produce low fitness, thus inducing low x and further change. But eventually, the mechanism is likely to hit upon behaviour with a large estimated fitness. Then subsequent variability is reduced, and the form of the organism—in terms of behavioural dispositions—stabilizes to some degree. New behaviours are then still explored, but with smaller changes. Simulations show that this strategy is evolvable under the right conditions: a population with organisms having this mechanism outperforms a population with a non-modulated (but still optimized) variability (van Hateren 2015a and Appendix A).

Although this mechanism does not directly depend on evolution by natural selection (because acquired behaviour is not inherited along a line of descent), it still increases x. Increasing x is a sustainable strategy from an evolutionary point of view because increasing x is covarying—in a statistical sense—with increasing f (since x estimates f) and because increasing f is sustainable (because of natural selection). Thus, the mechanism depends indirectly on evolution by natural selection, and it is evolvable.

The mechanism of Fig. 3 is easiest to understand when x has a simple, one-dimensional form, where it drives a single behaviour and is evaluated in a simple way from the state of the environment and the properties of the organism (see the computational models in van Hateren 2015a). In more realistic cases, x would depend on a range of different inputs (to X), and it would need to drive (via X) the variability of a range of different behaviours. Then the partial fitness effects of each input and each behavioural output would need to be taken into account and properly weighted. This will quickly become highly complex in realistic cases, where the form of X is expected to be highly intricate. X would have complex dynamics, involving nonlinearities and memory, and the number of inputs and outputs of X would be large and interdependent (even as the mechanism would still depend on how well the distributed variable x estimates f). But it is plausible that an X with a proper association of input and output factors can readily evolve, because it increases fitness. The mechanism is presented here in its simplest form in order to explain, in a comprehensible way, a range of otherwise puzzling phenomena in the realms of life and mind. It would require further elaboration before it could be a blueprint for a comprehensive quantitative model of the mechanism in a specific species.

Again, there are two ways of explaining how the mechanism works. First, the G-loop of Fig. 3 produces trajectories, now in an abstract space of forms with behavioural dispositions. Such behavioural dispositions will be partly inherited, but on top of that they can be varied by the changes produced by the G-loop. Starting at a particular position in this form-space, the organism will produce an estimate x of its fitness, depending on the current form of the organism and the current environmental circumstances. This x will subsequently modulate the variability of the changes to the organism's behavioural form. Small x means more change, on average, than large x. The new form of the organism, with new behaviours, will then affect f and produce a new x as an estimate of f, which then drives further changes in form, and thus further changes in x, and so forth. In qualitative explanations it is convenient to describe each complete cycle through the loop as a discrete event, but in reality the loop acts continuously, producing a continually changing trajectory through form-space. The trajectory will tend to remain close to positions in form-space where x is large (because of low variability there) and thus where f is likely to be large. Simulations show that this increases an organism's fitness under the right conditions (van Hateren 2015a and Appendix A). The trajectory through form-space that results is shaped equally much by random variation as by a deterministic variable (x). Cycling through the G-loop intermingles randomness and determinism in an inseparable way, and produces behaviour that has both some freedom and an effective goal (high x). Some behavioural freedom and a specific goal are the signatures of agency and goal-directedness, as will be discussed further in Chapters 9 and 15.

The second way of explaining focusses again on clustering. The form-space through which the organism's form moves is partly determined by inherited traits, and partly by traits modified during the organism's lifetime. Where fitness is high, organisms in a population (or, equivalently, organisms in a large set of lines of descent) tend to cluster. However, the current analysis considers an individual organism rather than a population or a set of lines of descent. For an individual organism, clustering can still be defined, but only in a probabilistic way. One can say that a single organism clusters at positions in form-space where fitness is high, by having a high likelihood to be at that position (because its line of descent has a high likelihood to be at that position). Thus, probabilistically, a single

organism clusters at the positions with high f, because those are the most likely positions where the organism was produced.

Apart from this reproductive clustering, there is again a second, independent mechanism of clustering, namely through the G-loop. The organism continually moves through form-space because of the x-driven behavioural variability. On average, it will spend more time at positions in form-space where x is large than at positions where x is small (because the latter produce more variability). In effect, it has a higher probability to be at large-x positions than at small-x positions. In other words, in a probabilistic sense it clusters at positions with large x. This second (statistical) way of clustering will align with the first (reproductive) way of clustering if x is indeed an estimate of f. This results in enhanced clustering and subsequently an increase of fitness. The latter is again called fitness-to-be (or f_+) because it is produced slowly and gradually, in a statistical way.

Chapter 3

Inclusive and extensive fitness

Fitness was described above, in its basic form, as an organism's propensity to survive and reproduce. Although this direct form of fitness (represented in Fig. 4 as pathway 1) may be valid for some species, fitness is often more complex. A major extension of fitness occurs when organisms help closely related organisms. If the reproductive success of a helped organism increases as a result, this can indirectly increase the fitness of the helping organism. This is so, because the helping organism shares many genes with the offspring of the helped organism. Thus, the helping organism indirectly promotes dissemination of its own traits. If this fitness benefit outweighs the cost of helping, then it is a worthwhile strategy from an evolutionary point of view. Fitness that includes this extension is known as inclusive fitness (Hamilton 1964; Fig. 4, pathways 1 and 2). Inclusive fitness is still a property of each individual organism. It is to be taken, along with the benefits it can produce, in a statistical, probabilistic sense. Benefits need not always occur, but they are expected, on average. In this more general case, fitness f then refers to inclusive fitness. The mechanisms of Figs. 2 and 3 still work as before, as can be understood as follows.

The explanations in Chapter 2 show that the mechanisms can be interpreted as aligning two clustering processes. The first process requires an x-modulated rate of micro-changes (R), and the second process requires differences of fitness. The latter clustering was explained above in terms of direct fitness, but it works for inclusive fitness as well. The reason is that kin are likely to be close in form-space, that is, to cluster. When kin help kin to survive and reproduce, this increases the likelihood that the forms in a cluster reproduce. Thus, the social component of inclusive fitness enhances reproductive clustering. This implies that alignment with the other, statistical clustering process is optimal when R is driven by a redefined x. This x must then estimate the redefined f (i.e., it must estimate inclusive fitness). The resulting fitness-to-be then also refers to inclusive fitness.

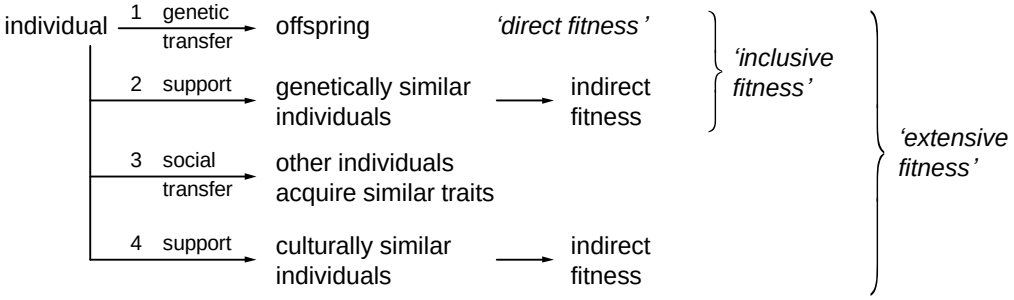

Fig. 4. Various forms of fitness. Direct fitness (pathway 1) is, roughly, the expected rate of producing offspring. Inclusive fitness combines direct fitness with indirect fitness (pathway 2) produced by helping genetically related individuals. Extensive fitness depends on the presence of an X process; it combines inclusive fitness with fitness produced socially, either directly by transferring similarity (pathway 3) or indirectly by helping others that are already similar (pathway 4).

Interestingly, this analysis suggests that there is a further way to enhance clustering, but only for the mechanism acting on an individual timescale (Section 2.2 and Fig. 3). Forms that cluster at a particular point in form-space (because of small R and high f) need not be kin. This is particularly true in species that can easily vary their form during their lifetime, by readily varying their behavioural dispositions. Then most of the individuals that display similar behaviour may be unrelated and genetically dissimilar. Such individuals then have similar forms (i.e., similar in terms of behavioural dispositions) that cluster at a particular point in form-space. As is explained in the next paragraph, they can enhance clustering by helping other individuals in the cluster, regardless of whether those individuals are kin or not. The only criterion for helping is then similarity of form.

Helping enhances the fitness f of the individuals in a form-cluster, which means that their x increases as well (because x estimates f). Increasing x lowers R, and thus reduces the likelihood that they drift away to other forms. Moreover, other individuals that happen to acquire that particular form in form-space get the same lowered R, and thus tend to keep that form. In other words, that particular form functions as an attractor in form-space. Therefore, helping individuals with a similar form enhances not only fitness, but also clustering. Both f and x need to be redefined once more, in order to include the effects of helping individuals with a similar form. For this redefined form of f, the term ‘extensive fitness’ was coined in van Hateren (2015c; see Fig. 4). Extensive fitness includes both the effects of helping individuals with a similar form (pathway 4) and the effects of inducing others to become similar in form (pathway 3). Simulations show that this extended clustering mechanism is indeed evolvable under the right conditions. Organisms that also help organisms with a similar form then outcompete organisms that help only kin (van Hateren 2015c, summarized in Appendix A). Similarity of form as such becomes heritable because the clustering establishes attractor forms in the population. In effect, attractor forms recruit new organisms by inducing them to change their form to become similar to the attractor form. This type of heredity is, thus, not an intrinsic property of specific individuals, but a property that is induced in contingent individuals by the structure of the population in form-space. This structure can remain quite stable and evolve gradually over many generations. It should be noted that this bears similarity to ideas about cultural evolution (Boyd et al. 2011) and about cultural attractors (Claidière et al. 2014). However, a major difference with these and similar theories is that they do not incorporate the mechanism of Fig. 3. Therefore, they cannot produce the special properties that are associated with this mechanism, such as agency, intentionality and consciousness (as discussed in later chapters).

There are several conditions that need to be fulfilled for the proposed mechanism to work. First, the clustering process based on a G-loop with an x and R must be present, because the fact that a form can become an attractor is based on reducing R. Second, only species that can flexibly and strongly change their behavioural dispositions during their lifetime can produce significant clustering that is unrelated to kinship. And third, helping other individuals based on the form associated with behavioural dispositions requires reliable recognition of such dispositions. Therefore, it requires considerable cognitive resources. The combination of these three conditions suggests that the mechanism may be developed fully only in humans.

The clustering proposed here depends on helping other individuals who are similar, but who can easily change their behavioural dispositions. The latter induces the risk that the

forms of the individuals in a cluster could drift apart, even when R is small. This would then decrease the efficacy of helping. Stability is, thus, a potential problem. Reciprocal communication between two individuals is an effective way to synchronize and stabilize the behavioural dispositions of those two individuals. A public system of communication can perform a similar role for large numbers of individuals, such as occur in clusters. Thus, a public language is presumably evolvable because it can stabilize clustering (see further Chapters 10 and 11). It should be noted that this is not necessarily a mechanism that makes R small. R could still be large enough to allow fast responses to environmental change. A public system of communication only ensures that the clustering remains intact, by allowing the individuals belonging to a cluster to change their forms synchronously and consistently with each other.

Crucially, the mechanism can only work if there is an intrinsic X and a G-loop, because it requires that fitness is evaluated continually and thus drives clustering. This makes the causal structure of the mechanism fundamentally different from mechanisms of social learning and cultural evolution that are merely driven by inclusive fitness. Moreover, the mechanism should not be confused with group selection (i.e., evolution through competition between and selection of groups). Although group membership confers benefits on individuals, evolution in the present theory still happens at the level of individual organisms, not at the level of groups.

The four pathways depicted in Fig. 4 have different characteristics. In practice, all four pathways must play a role in humans, where pathways 1 and 2 may act to stabilize pathways 3 and 4. Just as pathway 2 piggybacks on pathway 1 (and could not exist without it), pathways 3 and 4 piggyback on pathways 1 and 2 (and could not exist without those). Nevertheless, pathways 3 and 4 are potentially more powerful than pathways 1 and 2, because the former make it possible to respond quickly to changing circumstances, through considerable behavioural change. But they are also more vulnerable. This is so because phenotypic helping makes it relatively easy for cheaters and freeloaders to take advantage of others. The pathways require high phenotypic (behavioural) flexibility, as well as sufficient mental capacities to recognize phenotypic similarity in a reliable way. Guarding against cheaters requires a sophisticated Theory of Mind that can assess the intentions of others. The ways by which cheating and freeloading may be suppressed is an active area of research (Rand and Nowak 2013). For the present purpose, we assume that these suppressive mechanisms are sufficiently powerful, such that helping unrelated others is an evolutionarily stable strategy.

The relative strengths of the pathways determine how much of fitness is related to competition and how much to cooperation. Pathways 1 and 3 imply competition at the individual level, either competition in terms of direct reproductive success (pathway 1) or competition in terms of being more effective than others in socially transferring one’s traits (pathway 3). In contrast, pathways 2 and 4 imply cooperation between individuals, either cooperation between genetically related individuals (pathway 2) or cooperation between individuals with similar behavioural dispositions (pathway 4). Particularly the latter form of fitness is expected to enable the cooperative forms of communication that are a prerequisite for human language (see Chapter 11).

Chapter 4

Components of F and X

The mechanisms of Chapter 2 use an internally generated variable x that estimates the organism's own fitness f. The variables x and f are produced by complex processes, X and F, respectively. The structure of these processes cannot be fully isomorphic (i.e., with an identical form), because F is orders of magnitude more complex than X could ever be. F includes a large number of factors that influence the fitness of an organism. These factors originate from within the organism itself, from its environment and from other organisms. X, on the other hand, is an approximate simulation of how the major factors affect fitness. X occurs fully within the organism. It is limited by the available processing power as well as by what the senses can tell the organism about itself and its environment.

Nevertheless, even as the structures of X and F cannot be identical, they must have similarities. The reason is that X has evolved as a means to produce an x that estimates f in many different circumstances. If circumstances change, not only f may change, but also the composition and structure of F. Then X and x must change as well, such as through evolution and learning, if the organism is to remain competitive. Changes in the structure of F typically involve coherent and correlated changes of different parts of F. For example, when food becomes scarce or when an organism migrates to another environment, this changes many parts of F at the same time. Because F is a process, the parts of F can be regarded as subprocesses. Subprocesses of F that typically change coherently are called F-components below. F-components should be roughly reflected in the structure of X, because this facilitates change of X, both evolutionary change and within-lifetime change. When an F-component changes, only the corresponding X-component (i.e., the corresponding subprocess of X) needs to change then as well. This is far more feasible than changing many disconnected parts of X at the same time, which would be required if X would lack distinct components. Therefore, organisms are likely to have evolved an X that includes not only distinct components that reflect those of F, but also the capability to develop and learn such components.

X-components that roughly correspond to F-components estimate those components, including their role in producing f. This is a more complex version of estimation than before, because components are subprocesses rather than single numbers (such as x and f). In weather terms, it is analogous to estimating an extended weather system (e.g., the course and properties of a hurricane) rather than just a single variable of the weather (e.g., the temperature at a particular place). Estimating extended processes may involve estimating many variables at once, as well as estimating the dynamics and coherence of components of the process. Estimating need not be done in a literal, isomorphic way. For example, a detailed computational simulation of the weather may be fairly isomorphic, but an experienced meteorologist interpreting a weather chart may use abstract conceptual short-cuts, and a farmer reading the sky for a short-term weather forecast may use mere rules of thumb.

We have seen above that the increase of fitness-to-be produced by the mechanisms of Figs. 2 and 3 depends on how well x estimates f. This remains true when X is parsed into

X-components. The increase of fitness-to-be is, then, not directly dependent on how well an X-component estimates an F-component, but only indirectly. The causal efficacy of an X-component depends on how it contributes to the X process as a whole, that is, to x. If it estimates the corresponding F-component and its role accurately, it is expected to contribute positively to how well the resulting x estimates f. This depends not only on how well the X-component estimates the F-component, but also on how the X-component is integrated in the X process, that is, it depends on whether its role in that process is sufficiently similar to the role of the F-component in the F process.

There are several complications that need to be mentioned. A first complication is that X, not F, determines how F is parsed. This follows from the fact that X is the source of the causal efficacy produced by parsing and estimating. Irrespective of the question whether F might have an autonomous parsing, F is necessarily parsed by X when X forms distinct components based on the available correlational structure of F. Nevertheless, the latter structure is objectively present. Therefore, there is presumably only limited scope for variations in how X can effectively parse the part of reality that is incorporated in F.

A second complication is that X-components may not always correspond to specific F-components. X is unlikely to be flawless, because it is the result of trial and error. It may contain components that have no counterpart in F, that estimate a component in a mistaken way, or that estimate the wrong component. Furthermore, X is likely to lack counterparts of many potential F-components. Such errors and omissions lower the accuracy by which x estimates f. However, in variable environments, the detrimental effect on fitness may be too small to be counteracted by evolution or learning. Small differences of fitness produce effects only slowly, if at all, because evolution as well as learning by trial and error are statistical processes. In variable environments, small fitness differences may not persist long enough to produce appropriate changes in X. Moreover, small fitness differences may drown in statistical noise when population sizes are small. And finally, correcting errors and omissions may simply be too complex or too costly for a specific species. A related complication is that the accuracy by which X-components estimate F-components may vary from poor to excellent. Poor estimates may be all that can be accomplished given the available means. Yet, poor but veridical estimates may still be better than no estimate at all.

Chapter 5

The consequences: a preview

All living organisms are conjectured here to incorporate an internal process X that makes an estimate x of an organism's own fitness f, which is produced by an external process F. This conjecture has consequences for a variety of topics that are related to life and mind. Several key topics are analysed in detail in subsequent chapters. These chapters will be previewed below, but before doing so it may be helpful to first consider the general consequences of the mechanisms of Figs. 2 and 3. These general consequences are stated here without much explanation; detailed explanations and arguments are provided in the later chapters.

The mechanisms discussed above have remarkable consequences, because they produce not only a new form of causation, but also goal-directedness and value, as strongly emergent entities (summarized in Chapter 15). The new form of causation is neither deterministic nor random, but constitutes a distinct third way of causing. Trajectories through a hereditary and behavioural space of forms are produced by an inseparable mixture of randomness and determinacy. When the mechanism affects heredity, it affects the causal structure of the evolutionary process (van Hateren 2015e). When it affects behaviour, it produces agency, the organism's capacity to act with some freedom (Chapter 9). When it affects behaviour in organisms capable of advanced forms of consciousness, it produces free will.

Genuine goal-directedness does not occur in those parts of nature that are not somehow involved in life (abbreviated in this book as 'abiotic nature'). But the mechanisms produce true goal-directedness, because high x must be viewed as an intrinsic goal of any organism. This is even true when one takes agency and free will into account. Neither of these could overrule X and x, because they are themselves produced by X and x. When agency or free will affects behavioural dispositions and behaviour, X and x are implicitly modified such that high x always remains the organism's overall goal.

The mechanisms produce estimation, because they evolve such that x tends to become as similar to f as possible. This means that components of the X process must evolve to be estimates of components of the F process. Estimation is an evolutionary invention, as it is absent from abiotic nature. It is one-sided: x estimates f, but f does not estimate x. An X-component is an internal process within an organism that strives to give an accurate account of how the corresponding F-component functions within the F process that produces f. This F-component is external to the organism. The internal X-component must be interpreted, then, as assigning meaning to the external F-component. The fact that the X-component is about the F-component can be regarded as a minimal form of intentionality (in the sense of 'aboutness', see Chapter 10).

Finally, the mechanisms produce strong emergence. A careful argument that shows this for the simplest possible variant of the mechanism is presented in Chapter 14. Briefly, one can use a variable C to denote the accuracy by which x estimates f. The better this accuracy is, thus the larger C is, the higher the resulting fitness-to-be (f_+) will become. Increasing or decreasing C produces a corresponding change in f_+. In other words, C is a cause of f_+ (on a long timescale, because the effect on f_+ is statistical and takes time). But C is not a regular

material cause, because it denotes estimation of f (on a short timescale, separate from the one on which it acts as a cause of f_+). Importantly, much of the causal efficacy of C comes from randomness. The indispensable contribution of randomness means that the causal efficacy of C cannot be reduced to—that is, is not completely replaceable by—the material causes that produce X and F. C is therefore a cause with a distinct, novel and partly autonomous efficacy. In other words, C is strongly emergent. An autonomous cause must exist as a distinct, autonomous entity. Thus, C is a distinct entity. Detailed analysis of what happens with sophisticated forms of communication between organisms leads to a plausible account of how consciousness arises and why it is experienced (Chapters 12 and 16, and van Hateren 2019).

The parts of the book that follow below focus, respectively, on life, mind and philosophy. The chapters on life (Part II) depend on the fact that the mechanism, in all its incarnations, produces a special form of causation and a strongly emergent cause C of f_+. This has two major consequences. First, it provides a novel criterion for demarcating life from non-life (Chapter 6). Second, it means that living organisms have an intrinsic goal-directedness, as well as agency, through the mechanism of Fig. 3. It has long been recognized that humans and many other species do indeed have agency and goals. The mechanism of Fig. 3 implies that such agency and goals are not merely apparent, but genuine. Moreover, and perhaps surprisingly, minimal forms of agency and meaning must be present in any organism, even a very simple one, that incorporates the mechanism (Chapter 7). The goal-directedness produced by the X process can be used to construct a theory of biological functions that has a broader explanatory scope than previous theories of function (Chapter 8).

The chapters on mind (Part III) focus on the mechanism of Fig. 3 in combination with forms of fitness that require advanced nervous systems. The new form of causation associated with the mechanism, combined with genuine goal-directedness, lead to agency (Chapter 9) and free will (Chapter 12). The estimation and meaning produced by the mechanism lead to intentionality, which is the power of minds to refer to something external to the mind (Chapter 10). It is the prime condition for the existence of human language (Chapter 11). When intentionality is prepared to be communicated between organisms, it gives rise to an additional strongly emergent cause. This emergent entity has properties that are consistent with those of consciousness, and it is plausible felt by the organism itself (Chapter 12). The properties of X then lead naturally to a theory of the human self (Chapter 13).

The final chapters (Part IV) are grouped under the title Philosophy. They contain several topics that are traditionally studied in that field. The case for strong emergence is made in Chapter 14. It argues that not all causes in nature can be reduced to constituent causes, and that not all causes can be classified as material. A major topic in the philosophy of science is epistemology, the study of knowledge and how it can be acquired. Related questions are how epistemology relates to ontology (what is ‘out there’), metaphysics (ontology plus how it changes), and ethics. The theory can provide some perspective on these topics (Chapter 15). Finally, a series of philosophical conundrums with respect to consciousness are discussed in Chapter 16.

The Epilogue contains a short discussion of the implications of the theory for the possibility of machine intentionality and consciousness. It concludes that obtaining such properties by applying the theory would be quite difficult and risky. It remains to be seen whether it is possible to overcome such problems.

PART II: LIFE

Chapter 6

What qualifies as life?[2]

It seems an intuitive truth that living organisms are qualitatively different from non-living systems, even complex ones. But it has proven difficult to formulate which differences between the two are essential. Still, having general criteria for demarcating life from non-life is important for several reasons (Cleland and Chyba 2002). First, such criteria could help to recognize life if it were discovered elsewhere in the universe, even if it were radically different from life on Earth. Second, they would help to evaluate to what extent efforts to produce artificial life in the laboratory are successful. And third, they might help to understand the origin and evolution of life.

There is currently no consensus on what would constitute sufficient and necessary criteria for establishing that a system lives (Bedau 2007; Tsokolov 2009; Benner 2010). It is clear that various properties are important, such as material and physical requirements, requirements with respect to heredity and information, and requirements with respect to system integrity and autonomy. An example of a physical requirement is that some form of metabolism is needed such that free energy can be harnessed from the environment. Free energy is needed for building and sustaining life's structures and processes (Lineweaver and Egan 2008). Hereditary requirements are, first, that some form of structural memory (such as RNA or DNA) is present in order to enable replication and reproduction (Pross 2004), and, second, that heredity can change in such a way that forms of life can adapt to changing circumstances (Darwin 1859). Heredity and physiological structure are closely related to information, which suggests that the particular ways by which living systems accumulate and use information can be used to define life (Walker and Davies 2012; Michel 2013). Furthermore, living systems are characterized by structural integrity and by their capacity to maintain themselves and to function autonomously (Varela et al. 1974; Ruiz-Mirazo et al. 2004; Di Paolo 2005). Kauffman (2000, 2003) has argued that the agency of autonomous systems—their ability to act on their own behalf in an environment—appears to be the defining characteristic of life.

Not all criteria mentioned above have a clear proposed implementation. Moreover, no single criterion appears to be sufficient. Most criteria have exceptions, and are thus not even necessary. In this section I propose, as a new criterion, that the transition from non-life to life is accompanied by a transition of causality, from the standard forms of causation of non-living physicochemical systems to a form of causation that is—at the behavioural level—equivalent to a form of agency. It therefore largely conforms to the views of Kauffman (2000, 2003) and Di Paolo (2005) that agency is a defining characteristic of life. However, it reaches this conclusion not by taking the autonomy of organisms as a starting point, but by using the mechanism that is explained in Chapter 2. It thus depends on the hypothesis that all life forms contain at least some version of this mechanism.

There can be versions of the basic mechanism on two different timescales. The first version works at the level of hereditary change, where it affects the changes along lines of

[2] This chapter is partly based on van Hateren (2013).

descending organisms (Section 2.1). The second version works at the level of organismal change, where it affects the changes within an organism during its lifetime (Section 2.2). The term ‘agency’ is normally reserved for the behaviour of single organisms and it would be a bit odd to use it for lines of descent. Therefore, I will use here the term ‘active causation’ (van Hateren 2015a) to indicate both versions at once, for the sake of brevity.

Having active causation is equivalent to having the mechanisms of Fig. 2 or Fig. 3 of Chapter 2. Both contain a G-loop where an estimate of fitness is used to modulate the variability of structural changes of the organism. As was explained in Chapter 2, this produces a special form of causation that is a distinct intermediate between deterministic and random causation. This type of causation can even be shown to be emergent in a strong sense (see Chapter 14). A strongly emergent cause is partly autonomous, because it is not fully produced by the micro-causes attributable to its constituents. The reason for this partial autonomy is the indispensable role that randomness plays in the mechanism. Randomness is understood here to be fundamental (that is, it occurs spontaneously and is not related to a lack of knowledge about underlying factors).

The presence or absence of the capacity for active causation can serve as a criterion for classifying a system as belonging to life or not. As it turns out, this criterion works well for cases that pose problems for some of the other demarcation criteria. Such criteria may require life to be able to reproduce and evolve, thus seemingly excluding non-reproducing organisms such as mules. But according to the criterion proposed here, mules would be classified as life, because they utilize active causation at the behavioural level, even if they cannot reproduce. A similar conclusion holds for a living cell that has stopped reproducing, for example because it belongs to a multicellular organism. Such a cell presumable still uses active causation to adjust its behavioural dispositions in response to changing conditions in its immediate environment.

Entities that appear to maintain themselves and reproduce, such as flames and growing crystals, might be erroneously classified as life by some criteria. However, the currently proposed criterion correctly classifies them as non-life, because they lack the capacity for active causation. Such entities do not have an X process that produces an estimate of their own capacity for reproduction and self-maintenance.

Dormant life forms such as spores and dehydrated eggs are classified as life because they have the capacity for active causation, even if they are not using active causation right now. They belong to life, although they are not alive (because they are not living at the present moment). In contrast, viruses must presumably be classified as not belonging to life, because they do not appear to use active causation themselves. In theory, they might be able to hijack their host’s X process such that viral genetic variability is driven by an estimate of viral fitness. But most likely, the X process is too strongly integrated with the host to allow for that possibility. Without an appropriate X process, viruses are merely chemical systems that are capable of reproduction (by utilizing their host).

A colony of social insects might or might not be classified as a living entity. That would depend on whether it is possible to define a proper fitness for the colony as a whole, whether such a fitness is estimated by the colony and whether such an estimate drives either the colony’s hereditary variability or its collective behavioural variability. In either case, hereditary or behavioural, the mechanism must be under control of evolution by natural selection, and therefore must increase fitness and have a hereditary component. There is

currently no indication that all of these conditions are fulfilled. Nevertheless, that is, ultimately, an empirical question.

Active causation fits fairly well with intuitions about what makes a system living. A primary phenomenological property of living systems is that such systems have agency, in the sense that they have some level of autonomy and that they act on their own behalf (Kauffman 2003). They are unpredictable to some extent, and appear to be goal-directed and self-serving. Finally, living systems can die, which is an essential requirement for the applicability of fitness and evolution by natural selection. The mechanism of active causation gives these intuitions a solid basis. Importantly, it is a basis that is accessible to scientific analysis, such as by identifying the underlying control loops and their physicochemical realizations.

6.1 Relationship with autonomy, replication and information

The current approach is related to several long-standing traditions that attempt to characterize the nature and origin of life through specific concepts, of which I will discuss here autonomy, replication and information. Central to the present approach is the concept of fitness and its estimation. High fitness is indeed associated with autonomy, faithful replication and the acquisition of new information, as is discussed below.

Firstly, autonomy is required by high fitness, because it provides the stability and time needed for effective reproduction. Autonomy in the sense of self-maintenance and homeostasis is central to the idea of autopoiesis (‘self-production’, Varela et al. 1974; see also Thompson 2007), and it was recently extended with adaptivity and agency (or adaptive self-regulation, Di Paolo 2005). The concept of active causation (AC) as proposed here resembles, but is not identical to adaptive self-regulation (ASR). There are systems that have AC but no ASR (such as the hereditary G-loop in Fig. 2, where organisms modify their offspring, but not themselves) and systems that have ASR but no AC (such as an adaptive extension of a conventional autopoietic system that is, by default, purely deterministic; such systems would lack genuine agency and would not be alive according to the criterion proposed here). Kauffman (2000, 2003) defines an autonomous agent as a system that can act on its own behalf in an environment. But these studies explain agency only in a definitional sense, by invoking thermodynamic work cycles. Agency as a form of active causation solves this problem: it is a highly specific mechanism that directly explains the causal freedom of agents in terms of underlying physical processes.

Secondly, faithful replication is required by high fitness, because otherwise fitness-promoting properties that were previously acquired in evolution would quickly deteriorate (Eigen 1971; Szathmáry and Maynard Smith 1995). However, too faithful replication would hamper the rate of adaptating to a time-varying environment. It is proposed here that a controlled modulation of the hereditary variability (as in Fig. 2), although presumably selected initially for its survival value (Galhardo et al. 2007), has produced active causation as a spin-off.

Finally, it has long been recognized that information appears to play a crucial role in the origins and functioning of life, in particular when adapting to new conditions and thereby retaining or increasing fitness (Szathmáry and Maynard Smith 1995; Maynard Smith 2000; Nurse 2008; Walker and Davies 2012). However, information is a rather elusive concept (for an exhaustive overview of how differently it has been defined and used throughout

science see Burgin 2010). It is quite useful for interpreting biological processes, but it should not be assumed to be a fundamental part of nature (see also van Hateren 2015g). Importantly, functional information depends on intentionality (in the sense of 'aboutness', the fact that some things can be thought to refer to other things). Information is necessarily about something. Intentionality is absent from the abiotic parts of nature, but can arise, in living organisms, through the mechanisms that produce active causation (see Chapter 10). This means that one should not use 'information' as a given ingredient in order to explain life, because that would be circular. The mechanisms of Chapter 2 are needed first, before one can define information that is functional to the organism.

6.2 A definition of life

Giving a definition of life may be a somewhat futile endeavour (Cleland 2012), primarily because single sentences are inevitably somewhat vague and open to different interpretations. I will nevertheless attempt to give one here, so that the present proposal can be readily identified in future discussions. It is a variant of NASA's working definition ("Life is a self-sustaining chemical system capable of undergoing Darwinian evolution"), reading: "Life is a material system capable of active causation". Active causation depends on estimating fitness and on the fact that modulating randomness can increase fitness itself. Darwinian evolution, i.e., evolution by natural selection, is the only mechanism currently known that will, in the long run, consistently promote high fitness. Therefore, Darwinian evolution is presumably required for maintaining the long-term stability of active causation. The term 'material system' is used to indicate that life is defined here as a phenomenon of the real world, and not of a purely symbolic system (such as a program running inside a computer).

Taking active causation as the primary criterion for distinguishing life from non-life implies that any system that completely lacks active causation is classified as non-life. In particular, a self-replicating system that is subject to merely the D-loop of Fig. 2, thus with a fixed mutation rate, is not considered to be life (in contrast to the NASA definition), unless it utilizes active causation on a shorter, behavioural timescale (as in Fig. 3). This does not pose a problem for defining current life, if one assumes that all current species incorporate a G-loop. However, G-loops have presumably evolved in organisms with only a D-loop. This may be viewed as (mostly) coinciding with the transition from protolife to life. Because this transition is bound to be gradual anyway, this should not be taken as a major issue for the definition. Moreover, organisms that lack a G-loop would presumably not last for long in a variable environment, because they would be outperformed by organisms with modulated variability (Ram and Hadany 2012; van Hateren 2015a). Finally, it may be argued that a system that evolves with merely a D-loop should be regarded as a self-replicating chemical system, rather than a living entity displaying some degree of agency (at the behavioural timescale) or at least a special form of causation (as in Fig. 2).

Proving that a newly discovered system uses active causation would presumably require a detailed molecular analysis or an extensive analysis of its behaviour and evolution. However, several indicators for the presence of active causation might be easy to observe qualitatively: initiative, causal autonomy (i.e., partial independence of external causes), agency, goal-directedness and self-interest.

Chapter 7

Biological meaning[3]

A biological organism may be seen as a purely material system that is driven by environmental factors and by the organism's genetic and physiological structure. But it may also be seen as an individual with agency and goals. A basic question that has been haunting biological thinking for a long time is whether the second view is a mere consequence of the first view, or whether it adds something extra. The 'mere consequence' idea implies that it is enough to study an organism's structure and physiology in as much detail as possible. Such a detailed analysis will then eventually show that agency and goals are not real but only apparent, in an 'as if' kind of way. On the other hand, the 'adds something extra' idea seems to require ingredients that have no counterpart in the non-living parts of the material world. Introducing such ingredients on an ad hoc basis is an unattractive proposition.

A way out for the 'adds something extra' view may be the concept of emergence, the idea that new properties may arise from specific configurations of matter. For example, certain spherical objects with sufficient hardness obtain the property that they can roll on a plain, and the property of rollability may then be seen as emergent. However, that would be a property that is fully predictable once the properties of the material and the configuration are specified, and rollability is not radically different from other mechanical properties that are known to exist. The problem with agency and goals is that they do seem to be radically different from anything else in nature. If agency and goals are really emergent, it needs to be shown in which specific way they can emerge and why it is plausible that they arise in the radically new form they do.

The theory discussed in Chapter 2 can indeed let agency and goals emerge from components that lack those properties. Here, I specifically put this theory within the context of the field of biosemiotics, which addresses similar issues, and show that it matches quite well with the main ideas of that field. Moreover, I argue that the emerging properties are fundamentally new and cannot be reduced to (or replaced by) a description of components and their configuration. A more rigorous argument on strong emergence can be found in Chapter 14, and a discussion of agency and goal-directedness in Chapters 9 and 15.

7.1 Extending the Darwinian approach

The approach taken here is closely associated with the original Darwinian vision of understanding evolution as a result of the differential reproductive success of organisms, a succes that depends on their phenotype[4]. This vision has often been perceived as implying a materialistic, gene-centred and deterministic view of life, which excludes genuine agency

[3] This chapter is a shortened version of van Hateren (2015d).

[4] A phenotype is the actual form (the totality of its characteristics) through which an organism interacts with the world.

and meaning[5]. I will argue below that such an implication is unwarranted, because a subtle but far-reaching extension of the basic Darwinian theory can include agency and meaning.

However, it is important to clarify from the beginning how this approach is related to other modern extensions of the theory of evolution. Modern extensions include interactions between development and evolution, phenotypic plasticity, niche construction, gene-culture coevolution, and a range of sophisticated hereditary mechanisms such as epigenetics and other forms of enhanced evolvability (Laland et al. 2011) and adaptability (Sharov 2014). These factors are specifically considered within what has been called the 'extended evolutionary synthesis' (Pigliucci and Müller 2010). Much of this extension is data driven, as more complex evolutionary mechanisms are gradually uncovered. But it is also driven by an implicit concern that the conventional evolutionary view stresses genetic causes too much, to the detriment of other causes that originate from development and behaviour. This apparently motivated Laland et al. (2014) to call it "a struggle for the very soul of the discipline".

Unfortunately, this approach appears to be misfiring—at least to the extent that it is an attempt to advocate agency as arising from the organism. Elsewhere (van Hateren 2015e) I argue that causes that seem to originate from the organism do not produce agency if they are merely a result of complex causal loops that are primarily deterministic – with any randomness regarded as noise. None of the cogwheels in a clockwork can be a source of agency and meaning, nor can any combination of cogwheels, no matter how complex. The problem with regard to agency is not the apparent origin of causes, but the assumption of determinism. The new, modern mechanisms of evolutionary change can therefore only contribute to agency if they include randomness in their causal scheme in a highly specific way (van Hateren 2015e). Below I will focus on the simplest evolutionary mechanism for the emergence of agency, the one that is easiest to understand. However, this does not imply that other processes could not be involved if they similarly entangle deterministic and random forms of causation. I also do not intend to imply that the Darwinian mechanism is the only one producing evolution. But I do claim that the Darwinian mechanism with its extension as explained below is the only one currently known that is—at least in principle—capable of generating agency and meaning. More complex forms of agency and meaning then all derive from and depend on this origin.

The discussion below will focus on the behavioural mechanism of Fig. 3, which involves modulated random causation operating in a cyclical causal loop G. This G-loop produces goals and agency, as will be argued now. The form of the X process is defined by which environmental and internal variables an organism uses for producing x, and how X does so. Here x is an estimate of the fitness f of the organism itself. It drives, via X, the variability of behavioural dispositions. The form of X thus determines which areas of behavioural space—where such areas define the possible behavioural repertoire—are associated with low behavioural variability and which areas with high variability. This association is already sufficient, purely for statistical reasons, to drive the behaviour towards the areas with low variability. The word 'towards' should not be interpreted too literally here, because the behaviour is not changed into a specific direction – all behavioural changes are taken to be random, apart from their variance. But probabilistically, behaviour will diffuse away from

[5] Throughout this chapter, the term 'meaning' is used in a general sense as in 'the meaning of an action', rather than in the more specific sense as in 'the meaning of a word'.

areas with high variability more quickly than from areas with low variability, and thus it will tend to stay in areas with low variability. Therefore, it appears to be driven towards such areas. Because low variability is associated with high x, high x must then be seen as a genuine goal of an organism[6].

Note that this reasoning does not depend on what exactly x represents. It could represent an arbitrary goal. But arbitrary goals would not be evolvable through the basic Darwinian mechanism, because they do not specifically promote fitness; most likely, they even reduce fitness, because striving for goals generally carries processing costs. It can be readily understood that the optimal goal for promoting fitness is in fact fitness itself. Organisms with high fitness as goal would outcompete organisms with any other goal. In other words, the only goal that is evolvable and stable in the long run is high fitness, that is, high f. Consequently, x must be an estimate of f, because otherwise the mechanism would not have evolved. There is no guarantee that x will keep a value close to that of f when circumstances change, but a mismatch would lead to a disadvantage relative to other organisms with a better form of X. Thus, a persistent mismatch would presumably lead to extinction, and would have done so in the past. It is therefore likely that x has evolved to become fairly robust against common disturbances.

Although striving for high x is thus the overall goal of an organism, in practice this goal will consist of a large number of sub-goals. Such sub-goals can be seen as resulting from a partitioning of the X process, that is, a partitioning of X into subprocesses (Chapter 4). Together, these subprocesses and the sub-goals they represent serve the general goal of high x. Partitioning of X into effective and coherent subprocesses is likely to facilitate improving the form of X, through evolution or learning, and is therefore likely to be evolvable.

Apart from establishing x as a genuine goal, the G-loop also produces agency, because the causation that results from cycling through the loop is rather special. The modulated random causation already intermingles deterministic and random factors (x and the behavioural variability, respectively), but the loop strongly amplifies this effect. Each time the loop is traversed (which happens continually), x and the randomness become further entangled. First, the value of x determines the behavioural variability and the random outcome determines a new behaviour; then, the new behaviour leads to a new value of f and therefore to a new value of x. In the next pass through the G-loop, the new value of x again determines behavioural variability, and so on and so forth. Eventually, there is no way to separate causation into deterministic and random components. The details of the behavioural trajectory are unpredictable because of the randomness, but the overall direction of the trajectory depends on the goal, namely high x. The behaviour therefore combines a certain spontaneity (in the form of randomness) with a certain deliberateness (in the form of striving for high x). Such a combination is the signature of agency, at least an elementary form of agency. The behavioural trajectory is driven by an internal goal (high x), but the trajectory is not fully determined, for two reasons. First, because of the randomness in the G-loop, as discussed above. Second, the form of X is not fixed, neither in evolution nor within the lifetime of an individual organism. This is because there are many different forms of X that are approximately equivalent in terms of how well they can

[6] A robust support for this claim requires the concept of strong emergence (Chapter 14), which is applied explicitly to goal-directedness in Chapters 9 and 15.

estimate the value of f. Such different forms and their improvements are evolvable and learnable as well. They may be accessible through hereditary and behavioural variability, but also more deliberately through a dialogue between or within organisms (van Hateren 2015b, 2019).

In effect, the organism internalizes the external fitness f as an internal fitness estimate x. The G-loop then utilizes this internalized measure of fitness and provides the organism with a genuine goal and genuine agency. Having a goal and agency implies that the goal is important to the organism and thereby assigns value to the goal. In other words, the behaviour becomes meaningful. This meaning is generated within the organism and is thus a form of intrinsic meaning. The emergence of meaning suggests that the current theory can be interpreted in terms of semiotics.

7.2 Interpretation in terms of biosemiotics

7.2.1 The semiotic triad

Biosemiotics involves the study of meaning in biological systems, and amongst its intellectual roots is semiotics (the study of signs and meaning). One of the most popular systems for describing signs and their meaning is the triadic one promoted by Peirce (2010). This system is often used for analysing meaning in a linguistic context (e.g., Chandler 2007), but it can also be applied to meaning in biology (e.g., Hoffmeyer 2012). My purpose here is to show that the meaning-generating theory described above can be represented as a triad.

The basic Peircean triad represents signification, the overall process of producing meaning. It consists of three elements that become mutually related. The sign (or sign vehicle) is called representamen by Peirce, because it represents. It is connected to an object (the semiotic object to which the sign vehicle refers) by the interpretant. The interpretant produces the interpretation of the sign and thereby, more generally, the meaning of the overall process. A typical example of a sign is smoke that is connected to its object, fire, through an interpretant that consists of the idea that smoke usually indicates fire. Smoke is then a sign of fire.

The mechanism of Fig. 3 can be tentatively interpreted as a (primordial) semiotic triad. The subject-generated x refers to the external f through the meaning-generating G-loop. The G-loop implicitly interprets X, and by doing so enhances the organism's fitness. This loop is the primary generator of meaning, and because of the dynamical and stochastic nature of the mechanism, it generates agency as well. The organism then gets the role of semiotic agent, which in effect uses the semiotic triad. The three entities constituting the triad are far from simple. The G-loop is an unusual stochastic feedback process, and x and f are produced by complex processes (X and F, respectively). These processes keep changing because of the variability that is utilized in the causal loop and because of changes in the organism and its environment.

As argued specifically in Chapter 10, the relation between x and f can also be seen as a primordial form of intentionality ('aboutness', the capacity to stand for or refer to something else; x is about f). In a sense, the form of X represents all that the organism knows about its situation (as objectively represented by the form of F), which is similar to the concept of knowledge as discussed in Kull (2009). Both X and F are complex processes with many

inputs and at least some of their components are likely to be related. This is likely, because only with related components, x can estimate f across a wide range of circumstances. An example of a related component is glucose surrounding a bacterium. Its presence may partially determine the fitness f, and the sensing of glucose by the bacterium may partially determine the fitness estimate x. Such related components are part of a derived semiotic triad by themselves, with as interpretant the fractional role glucose plays in the G-loop. This more detailed level of semiosis is more readily amenable to Peircean analysis than the rather abstract general level of x and f. Specific sub-goals form the bulk of specific meanings as studied in biosemiotics, for example when assigning meaning to certain molecular processes that serve an organism (Barbieri 2008).

7.2.2 Concordances and discordances with eight theses of biosemiotics

In Kull et al. (2011), the conceptual basis and basic principles of the field of biosemiotics are summarized in the form of eight theses. It is therefore interesting to see to what extent the approach presented here is consistent with these principles. This is discussed below (theses I-VIII are all cited from Kull et al. 2011).

"I. The semiosic/non-semiosic distinction is co-extensive with life/non-life distinction, i.e., with the domain of general biology." This is consistent with the argument presented in Section 7.1 that the G-loop is responsible for producing the agency and goal-directedness of life. Agency and goal-directedness together imply meaning (in the general, non-linguistic sense). Moreover, the thesis is consistent with the life/non-life distinction that is proposed in Chapter 6.

"II. Biology is incomplete as a science in the absence of explicit semiotic grounding." This is consistent with the thesis of Section 7.1 that all life has at least a minimal form of agency. As this is conjectured to require a G-loop, it automatically involves meaning.

"III. The predictive power of biology is embedded in the functional aspect and cannot be based on chemistry alone." When all organisms have agency and intrinsic meaning, prediction must utilize their implicit goal-directedness as one of the three primary causal factors (along with the conventional factors environment and heredity/physiology). In Chapter 14 it is shown that the G-loop indeed produces an autonomous and distinct causal factor. Sometimes the conventional factors (e.g., a harsh winter or genetic disease) may determine biological outcomes without also being caused by the organism's agency and goals. But usually, biological outcomes also depend on (and are partly caused by) agency and goals, for example, when an animal deliberately migrates to a new territory. Although x is ultimately produced by a physiological process X, that process can only be interpreted if it is understood as a key component of the stochastic mechanism from which agency, goals and meaning emerge. The intention to migrate is then a real phenomenon that must be used for a complete explanation of why the animal migrates, as well as for predictions of such behaviour.

"IV. Differences in methodology distinguish a semiotic biology from the non-semiotic one." The current approach does not specifically address methodology, but it is at least compatible with this thesis. Meaning is often implicitly used for analysing living systems in terms of using and processing information. Examples are cases where genetic information

is interpreted (for a review of biosemiotic interpretation see El-Hani et al. 2006) and where sensory and neural processing is viewed as a form of information processing. The specificity of the current theory may help to distinguish information that is meaningful to the organism itself from information that is merely used as an analysis tool by the investigator (and therefore may be only meaningful to the investigator rather than to the organism; see also van Hateren 2015g).

"V. Function is intrinsically related to organization, signification, and the concept of an autonomous agent or self." This thesis is closely related to the thesis of autopoietic theory (e.g., Thompson 2007) that autonomy and self-maintenance as such represent meaning. I am critical of this viewpoint, because self-maintenance may be purely deterministic (or have randomness without utilizing a G-loop) and thus may fail to produce agency. Self-organization is sometimes seen as the source of autonomy, but self-organization is quite common in nature, occurring whenever systems have unstable and self-reinforcing dynamics (e.g., spontaneously generated tornadoes). Furthermore, maintaining the self as an autonomous unit can only be regarded as normative (implying goals and meaning) when the additional (tacit) assumption is made that existing is better than not existing. Such an assumption is unwarranted (see also Davies 2009, pp. 86–87), unless there is already a G-loop. I also do not agree with the thesis "Evolution presupposes function, rather than vice versa" (Kull et al. 2011, p. 32) if the term 'function' is regarded as normative (see further Chapter 8). The basic Darwinian theory of evolution by natural selection could, in principle, work without the extension with a G-loop. It would lead to self-reproducing systems without agency and meaning, and could not produce systems with consciousness (see Chapter 12). Nevertheless, this is a hypothetical case, because the extension provides an evolutionary advantage and presumably evolved very early on. Moreover, it is conceivable (but nearly impossible to prove) that without enhanced fitness-driven selection—enhanced because x amplifies f—the overall drive would be too weak, in practice, to let proto-life get off the ground or to prevent it from becoming extinct at an early stage.

"VI. The grounding of general semiotics has to use biosemiotic tools." This thesis is consistent with the idea that complex forms of meaning, such as associated with human consciousness and language (see Chapters 9–12), emerge from more basic forms of meaning that are also present in non-human species. The term 'grounding' acknowledges the possibility of emergence and the subsequent necessity to use novel concepts (e.g., in the social sciences and humanities).

"VII. Semiosis is a central concept for biology – however, it requires a more exact definition." The G-loop and its elaborations can be seen as a defining, prototypical model, as a valid proxy for a verbal definition. It incorporates several of the seven specific criteria mentioned by Kull et al. (2011, pp. 36–38), in particular agency, normativity, teleo-functionality, form generation (as through the G-loop) and inheritance of relations (as in the structure of X). Categorization is not specifically included, but is consistent with how high-level symbolic systems may arise from the basic theory (see Chapter 10). I believe that there is discordance with the final criterion, namely that a sign vehicle must be insulated from the dynamics that it constrains. This is similar to the notion that the controlling system must be separated from the controlled system (Pattee 2008). However, this requirement of a strict separation of initial conditions (doing the controlling) and laws (subsequently determining the fate of the controlled system) implicitly assumes systems described in a deterministic

manner. When the actual physical system is not deterministic, but partly stochastic in the specific way of the G-loop, it is no problem to have controller and controlled being part of the same dynamics. A key point here is that agency and meaning are not instantaneous, but only gradually build up statistical significance. This implies an entanglement between determinacy and randomness that makes it impossible to separate controller and controlled.

"VIII. Organisms create their umwelten." The Umwelt is a concept that comes from von Uexküll (1982), who suggested that organisms perceive and interpret the world in which they are embedded by generating internal meanings. The concept of Umwelt is closely associated with the form of X, the means through which an organism attaches meaning to everything it implicitly takes to be relevant for its F and f. The organism actively interacts with its world, modifying it and being modified by it. In effect, the organism lives in a semiotic niche (Hoffmeyer 2008a) that depends on the organism's own interpretations and that coexists with the ecological niche. However, the semiotic niche is still strongly connected to the ecological niche, because X is tied to F. Therefore, the word 'create' in thesis VIII should not be interpreted as 'freely construct', that is, the construction of an Umwelt is neither completely free nor completely determined.

The conclusion from the above discussion is that there is clearly a considerable overlap between the theory explained in this book and standard biosemiotic notions. Apart from a minor discordance with part of thesis VII, there is a stronger discordance with thesis V with respect to the origin of agency and meaning. The current theory partially agrees with thesis V to the extent that it also requires that organisms have enough autonomy such that the fitness process F takes a form that enables evolution. But such autonomy is only necessary for normative functions, not sufficient. Normativity and intrinsic goal-directedness are proposed here to emerge from X and the stochastic mechanism of the G-loop, which, in addition, produces agency. Agency as understood here is in fact largely consistent with its typical use in biosemiotics (Tønnesen 2015), where the "core attributes of an agent include goal-directedness, self-governed activity, processing of semiosis and choice of action" (see also Chapter 9). For most species, the expression 'choice of action' is probably a bit too strong, because choosing seems to presuppose sharp categorization. I rather prefer to call it 'some behavioural freedom', where behaviour is interpreted broadly to include also processes within plants and unicellular organisms. But apart from wording, it points to a similar concept.

The causation produced by the G-loop belongs exclusively to life. It is an elementary form of agency, closely related to what is elsewhere called 'semiotic causation' (Hulswit 2002; Hoffmeyer 2008b), i.e., the bringing about of effects through interpretation. The new form of causation has emerged from the highly specific combination of deterministic and random causation as occurring in the G-loop. It is a form of strong emergence (Chapter 14). Once it has emerged, it can no longer be described purely in physical terms. It depends on goal-directedness, meaning and agency, which are phenomena that are not present in the abiotic parts of the physical world. As a result, changes in the world of life can only be understood from three rather than two basic forms of causation: deterministic, random and active/semiotic. The latter form can subsequently evolve into increasingly complex forms of agency (Chapters 9 and 12; van Hateren 2015b).

Chapter 8

Biological functions[7]

Many of the parts and processes of biological organisms appear to have functions. For example, pumping blood appears to be the primary function of the heart, and enabling vision appears to be the primary function of the eye. The concept of function has several interpretations (Wright 1973), but at least some of these seem to imply an implicit goal-directedness. The heart is expected to pump blood and it has properties that are well suited to that end. There is often also a valuative, normative aspect to functions, because a properly functioning heart seems good for an organism and a malfunctioning one seems bad. Both goal-directedness and normativity are puzzling, because they do not occur in the non-living parts of nature. One may therefore wonder if and how they can arise in living organisms.

In this chapter, I will analyse biological functions from a naturalistic perspective. Thus, I assume that they can be understood as being produced by basic, physico-chemical processes. I will show that functions can be autonomous causal factors, not depending on human understanding. This also applies to their goal-directedness and normativity. I will not perform a detailed conceptual analysis of the term 'function'—neither an analysis of how it is typically used in natural languages, nor of how it is typically used by biologists studying functions. Approximate agreement between the concept of function developed here and typical usage is expected, but it is not a specific requirement or goal. The goal is to explain the ontology of functions, including their goal-directedness and normativity.

This chapter focusses on biological functions in non-human species. The reason for this restriction is that the analysis of biological functions in humans is complicated by the dual role humans have. They are biological organisms with functions of their own, but they are also the ones doing the interpretation of functions. Human sociality further complicates matters, because goals may become widely shared with others, which diffuses the benefits of a particular function. Although it is possible to extend the present approach to human functions, this is left to a future study. The same goes for an extension to the function of artefacts.

8.1 Are functions epistemological constructs or ontological causal factors?

It is clear that the material structures that perform a function, for example the heart and its muscles and valves, are ontological causal factors, or at least are fully composed of such factors. These material structures produce their effects in the standard way of any physico-chemical process. However, it is less clear what causal status one should assign to the function as such, for example, the function of pumping blood. If the function as such has no causal efficacy beyond that of its material realization, then it should be regarded as an epistemological construct. It may be real (pumping blood is real), but the function ascription would not need to be included in a complete and sufficient causal inventory of the world. Including the material realization of the function would suffice for that. On the other hand,

[7] This chapter is a shortened version of van Hateren (2017).

if a function as such has causal efficacy that goes beyond that of its material realization, then it should be regarded as an ontological causal factor. A causal inventory of the world would not be complete without it.

This distinction between the ontology and epistemology of functions is used extensively below. Functions that possess autonomous causal efficacy are denoted by the term 'ontic-causal'. 'Ontic' is meant here to denote that such functions exist independently of whether human intellect (or equivalent) exists. 'Causal' denotes that they are embedded in the causal dynamics of the world and that they form an autonomous and indispensable part of that dynamics. Functions that lack autonomous causal efficacy are denoted by the term 'epistemic-real'. 'Epistemic' means here that the perception of humans (or other life forms) is required for noting the material structure associated with such functions. 'Real' denotes that this structure is still objective. It is neither subjective, nor disputable, nor dependent on the attitude of observers.

A standard physicalist view assumes that all material processes are completely defined by the underlying, fundamental physical processes. In that view, biological functions would be epistemic-real only, by definition. Moreover, their apparent goal-directedness and normativity would be epistemic-real as well. However, theoretical and computational work (Chapter 2; van Hateren 2015a; Appendix A) has shown that goal-directedness is not necessarily epistemic-real. It can become ontic-causal through a subtle combination of deterministic and random processes, if this combination is subject to sustained evolution by natural selection. The structure of this theory is such that it can explain how ontic-causal functions can arise. In the next section, the theory is explained and applied to biological functions. Subsequently, other theories of biological function are discussed with respect to the question whether they produce epistemic-real or ontic-causal functions. It is argued that these theories produce epistemic-real functions only. Nevertheless, many of the key properties of these theories transfer to the new theory, which can thus be seen as a unifying one. Finally, it is shown that the theory is consistent with an existing list of intuitions about functions (Wouters 2005).

8.2 Explanation of the new theory of functions

The new theory of biological function is based on the conjecture that all living organisms contain an internal process X that makes an estimate x of the evolutionary fitness of the organism itself, as explained in Chapter 2. All factors that affect fitness can be conceived of as forming a highly complex fitness process, F. F is the totality of influences and processes that actually produce fitness (which is denoted by f, the organism's tendency to survive and reproduce). It is important to understand that both F and f are epistemic-real constructs. The process F is just a standard physico-chemical process, and thus is causally effective only through the microscopic factors of which it is composed. Neither F nor f have autonomous causal efficacy, that is, causal efficacy that goes beyond that of their composing factors (including how they interact). Another point that should be noted is that fitness as used here focusses on the organism, as the natural reproductive unit. However, the approach is not committed to a particular level of selection. Fitness depends on the entire process F producing the organism's tendency to survive and reproduce. F includes organismal factors and factors arising from the physical environment. But it also includes population-level feedbacks, such as the Malthusian factor. This factor reduces the fitness of all organisms in

a population when the population size approaches the environmental carrying capacity (e.g., when food or space becomes scarce). Frequency dependent effects, such as those occurring in mimicry, are automatically included in F as well. Factors at a level below that of the organism, such as developmental and genetic ones, are also included. The approach is therefore, intrinsically, a multi-level one with respect to natural selection (i.e., differential reproduction). It does not assume, a priori, that any level of selection is more important than another one. This also applies to the mechanisms that can sustain traits across evolutionary time. Evolution by natural selection depends on the existence of such mechanisms. Although the most obvious mechanism is genetic, there are significant additional ones (e.g., epigenetics, the retention of cellular structures, niche construction and social transmission).

The fitness f of an organism is conjectured to be estimated by a variable x that is produced by a process X within the organism. Thus, X is part of the organism (called 'agent' below) and it typically has both a hereditary part (as formed by previous evolution) and a behavioural part (as formed during the lifetime of a particular organism). Through the particular mechanism explained in Chapter 2, where x modulates the variability of random structural change of the organism, fitness can be increased. This can happen both at the timescale of evolution (Section 2.1 and Fig. 2) and at the timescale of an organism's lifetime (Section 2.2 and Fig. 3). Importantly, the ultimate causal efficacy of X depends on the condition that x estimates the fitness f. This estimating relationship between x and f (i.e., the fact that x is an estimate of f) is in fact an emergent factor with autonomous causal efficacy (see Chapter 14). It has causal efficacy in addition to the direct (proximate) causal efficacy of the material parts of X. In particular, the model implies that the material parts of X can only affect fitness if the non-material relation between x and f is present as well. The latter is partly independent of X, because the relation not only depends on x, but also on F and f (which can vary autonomously and, to some extent, randomly). Therefore, both causal aspects of X are needed in conjunction, and they can be regarded as complementary. They produce neither epiphenomenalism, nor causal overdetermination.

The autonomous causal efficacy of the relation between x and f gives an ontic-causal status to x. Its relation with f needs to be included in a complete and minimal causal inventory of the world. As stated above, f itself is an epistemic-real construct that is fully defined by its microscopic constituents and their interactions. Readers may be puzzled by the fact that x obtains ontic-causal status by being related to an epistemic-real f. However, one should realize that the relation between x and f is not based on a regular physico-chemical connection. Rather, it is an estimating relationship that cannot be defined in terms of physico-chemical constituents. Properties of f do not transfer to x, just like the properties of the weather (e.g., that it is wet, hot, cold, or windy) do not physically transfer to a weather simulation. The weather and its simulation belong to different categories.

The estimating relationship between x and f is an emergent, non-material factor with causal efficacy. The drive towards high x must be regarded, then, as the implicit goal of the agent (see also Chapter 9). The agent combines this goal-directedness with the behavioural freedom provided by agency (for the fast timescale of Section 2.2). Agency makes it possible that the agent changes its behaviour in a direction away from the goal (i.e., towards lower x), even though changing in a direction towards the goal remains more likely. The strength of attraction towards the goal must be equated, then, to the value that the agent implicitly attaches to the goal. The goal of high x is implicitly normative, for the agent itself (Chapter 7). The agent is expected to strive for high x, intrinsically. It is supposed to strive

for high x not from the point of view of any external agent, but from the point of view of the agent itself. Thus, the G-loop of Figs. 2 and 3 produces primordial forms of goal-directedness and normativity, as emergent factors. Moreover, it also produces a primordial form of causally effective reference, because X is causally effective only because x implicitly refers to f (in the form of an estimating relationship). Whereas reference plays no causal role in abiotic nature, it is present in systems if (and probably only if) these contain an X process. Because X presupposes evolution, such systems must be living organisms.

As argued above, high x must be regarded as the overall goal of an agent. But in practice, the process X is decomposed into subprocesses that serve specific sub-goals, such as having a well-functioning heart, finding food and finding mates. Together, these subprocesses and sub-goals contribute to X and x. The intrinsic goals of the agent are completely defined by X. New goals are, by definition, incorporated into an accordingly changed X. Because X has a non-material causal aspect (in the form of the relation between x and f), also its subprocesses have a non-material causal aspect (in the form of the relation between their sub-goals and the corresponding parts of F). Subprocesses that monitor specific functions then produce a causal efficacy that goes beyond that of the material realization of the functions themselves.

Similarly, the way in which X modulates variability (as based on x) is also decomposed into subprocesses affecting different parts of the agent differentially. If x is low because a specific trait is malfunctioning, variability need not (and will not in general) be redirected to that specific trait. How variability is redirected and distributed in specific organisms is likely to be quite complex, depending on the particulars of the organism and its habitat. However, the way in which X distributes variability is readily evolvable through standard evolutionary mechanisms, because it affects f. It is therefore likely to be adequate, on average. As an example of how variability may be redirected, we can consider the function of haemoglobin in vertebrates. It has the function of enhancing oxygen transport, according to existing theories of biological function. The new theory ascribes this function to haemoglobin as well, as follows. If haemoglobin starts to work less effectively, such as in the presence of interfering chemicals, then this is detected by control circuits regulating the oxygen levels in an organism. Compensatory changes (e.g., to respiration) are then made through standard feedback control, primarily in a deterministic way. The new theory conjectures that a deficient oxygen level produces, in addition, effects through X. This is done in a stochastic way and is based on estimating the organism's overall fitness. The oxygen level is one of the factors likely to be used for producing such an overall fitness estimate, because this level is highly significant for the actual fitness. Thus, X has likely evolved to include it, because that improves the adequacy of x as an estimator of fitness. Therefore, a poor performance of haemoglobin reduces x, and thus, indirectly, drives more variability anywhere in the organism. For example, it may result in behavioural variations that eventually result in the organism finding a less energetic lifestyle. Such a lifestyle can enable it to survive, despite suboptimal oxygen levels. The new lifestyle can become fixed (through a reduction of behavioural variability), because X subsequently indicates that the expected (i.e., estimated) fitness has become fairly high again.

In conclusion, biological functions can acquire ontic-causal status as follows. If a trait, process, or behaviour is of evolutionary significance to an agent, for example the pumping of blood by the heart, then it is likely to be represented in X. This is likely, because X would need to monitor the blood circulation in order to produce an x that is a reasonable estimate

of f. A poorly working blood circulation should be reflected in a decreased x. A reasonable estimation of f by x is required for obtaining high fitness (through the mechanism of the G-loop). It is therefore under positive selection pressure. We have seen above that subprocesses of X have autonomous causal efficacy, that is, they are ontic-causal. Therefore, the function as such is also ontic-causal. It has a non-material causal aspect (through X) that occurs in addition to the material realization of the function itself (such as realized by the heart and its muscles).

In order to decide whether a trait or process is functional in the ontic-causal sense, one needs to determine whether it is represented in X, that is, whether it is monitored by X (and thus used for producing x and for modulating organismal variability). Whether a trait or process is monitored by X is ultimately an empirical question. X is just a physiological or neural process that can be identified and modelled, including if and how it tracks the performance of specific traits or processes. If X exists (as conjectured here), it must be included in any adequate model of the organism. When a good model of X is established, then this also establishes what is represented in X and what not.

Until such empirical and modelling studies are available, common sense arguments may be used to evaluate the proposal made in this chapter (see Section 8.5). The key notion here is that X itself has evolved and is subject to continuing selection pressure. If x estimates f well, it gives the organism an evolutionary advantage. But like any biological process, X is costly (e.g., in terms of energy and material use), thus it will typically acquire parts that are useful and, eventually, loose parts that have become useless. Moreover, useless parts may even reduce how well x estimates f. Such a reduction would decrease the organism's evolutionary advantage, because it would decrease how well the G-loop works. Useless parts in X would, then, be specifically selected against. Thus, one can use the usual evolutionary reasoning to make plausible arguments as to what is included in X and what not.

A provisional definition that may be useful for such common-sense arguments is that "the working of a biological trait or process has an ontic-causal function if and only if its performance is monitored by X—where how X implements the sign of the trait's contribution to x determines how one should formulate the function". It is important to note that monitoring as such is neutral with respect to the question whether the effects of a trait or process in specific cases contribute positively or negatively to x. The mere fact of being included in X is already sufficient for having an ontic-causal function. Therefore, a malfunctioning heart still has the function of pumping blood, because its performance continues to be monitored by the X process. Nevertheless, the implemented sign of the contribution to x is important for how one should, linguistically, formulate the function. Saying that the function of the heart is 'to pump blood' is correct, because 'pumping blood' is implemented in X in such a way that it contributes positively to x (and thus is an implicit goal). One might perhaps interpret 'monitoring pumping blood' alternatively as 'monitoring not pumping blood'. But saying that the function of the heart is 'not to pump blood' is incorrect, because 'not pumping blood' contributes negatively to x (and thus is not a goal, but something to be avoided). The definition explicitly includes 'ontic-causal', because one is free, of course, to define biological functions more broadly, i.e., in an epistemic-real sense. A broadly defined concept of function may be convenient when used metaphorically in certain scientific explanations, even if it assigns functions to processes that have no autonomous ontic-causal status.

Ideally, functional goals represented in X would always serve f, because x is under selection pressure to estimate f as well as possible. However, this is not guaranteed, and agents may therefore have goals that are not in their best interest. Such goals can only be transient, because they are selected against or found to be disadvantageous through learning, eventually. Therefore, x tends to be well aligned with f.

8.3 Other theories of functions

Broadly speaking, there are two main traditions for explaining biological functions. The Causal Role (CR) school (Cummins 1975, 2002) characterizes functions by their current causal role in accomplishing assumed capacities of a containing system. In biological organisms, such capacities may take the form of specific goals, e.g., survival and reproduction (Boorse 1976). In contrast, the Selected Effects (SE) school (Millikan 1984, 1989; Neander 1991) looks at the historical, evolutionary causes of biological functions. Although some approaches incorporate elements of both schools (e.g., Walsh and Ariew 1996; Buller 1998) and there are alternative approaches, I use a clean dichotomy here for explanatory purposes. This clearly exposes the problems that arise if one seeks to assign ontic-causal status to biological functions.

8.3.1 Selected Effects functions

The upper diagram in the left part of Fig. 5 illustrates the basic idea of the SE explanation. This explanation is also known as etiological, that is, with the explanation provided by a chain of historical causes. A particular agent has functions that are active, or at least potentially active, in the present or future (black dots and arrows). The SE approach assumes

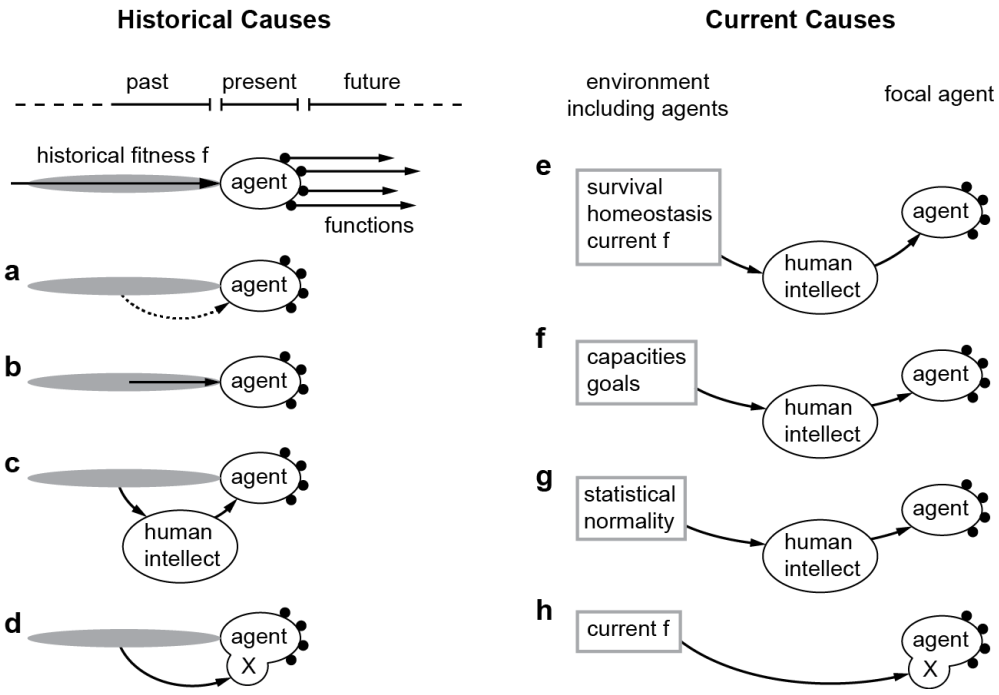

Fig. 5. Theories of biological function are typically based on historical, evolutionary causes (left) or current causes (right). See the main text for further explanation.

that these functions can be explained by their origin, through natural selection, in the evolutionary past of an agent (Millikan 1984, 1989; Neander 1991). Alternatively, such selection can be formulated in terms of fitness (Griffiths 1993; Buller 1998), by requiring that functions have contributed positively to the fitness of the agent's ancestors. In the figure, this is symbolized by the historical fitness f. Either way, natural selection and the effects of fitness occurred in a distributed way over time, which is symbolized by the grey area.

As stated above, we seek to assign ontic-causal status to functions. Functions exist in the present. If they are ontic-causal partly because of a historical process (historical fitness), then the question arises how this historical process is connected to the current entity. Some causal connection must be present if functions are to be ontic-causal. Without such a connection, the functions could only be epistemic-real. The main possibilities I can think of that might produce such a causal connection are depicted schematically in Fig. 5a–d.

The first possibility (Fig. 5a) assumes that there is a causal connection through immaterial (e.g., Platonic) means (dotted arrow). For example, one may assume that a historical process consists of objective facts, that it exists in its own right, and that it extends its existence across time (similar to a Platonic circle, which could be seen as timeless). It can then connect to the present function. However, such an immaterial explanation has no clear naturalistic interpretation. It seems too implausible to be considered further here.

The second possibility of a causal connection (Fig. 5b) is the standard way by which causal influences are thought to be connected to one another in physico-chemical processes. Such processes are fully defined by an instantaneous state, at each moment in time, that proceeds to the next state, at the next moment in time. Importantly, such processes do not contain explicit information about earlier states. This lack of historical information implies that the mechanism of Fig. 5b cannot directly connect the relevant parts of the fitness history to the present. At most, it only transfers information about the state immediately preceding the present one. Everything before is 'forgotten' and irrelevant from a physico-chemical point of view, because physico-chemical states unfold locally in time. There is no way to tell, purely from the state, how the system got to that state. Its history can only be reconstructed by using specific background information. But that would be epistemic inference. Relying on it would only produce functions with epistemic-real status.

A variant of the causal connection of Fig. 5b was proposed by Millikan (1984). Lines of descending organisms are connected by an uninterrupted chain of reproduction. Reproduction thus transfers the effects of natural selection (or of fitness) across time. However, reproduction has no special status from a naturalistic point of view. It is just a physico-chemical process that is completely defined by processes unfolding locally in time. In other words, nothing is transferred beyond the immediate physico-chemical state.

One might think that developmental processes in an organism can solve the problem of causally connecting the present function to the evolutionary past, because they construct a trait as homologous to ancestral traits. However, such an explanation depends on epistemic interpretation. It requires human perception to note the structural correlation that is associated with 'homologous'. Such a correlation has no autonomous causal efficacy. It can be a factor in scientific explanations, but it does not belong to the fundamental causal inventory of the world. Thus, functions explained in this way are only epistemic-real.

Similarly, nothing is solved if one would invoke DNA as a carrier of historical information. Biological functionality is used already when one interprets DNA as a form of

memory. Memory presupposes biological functionality, because it assumes that it is possible to refer across time. Conventional physico-chemical processes cannot refer across time or space, because all interactions are strictly local in time and space. In contrast, the theory explained above can produce non-local causation because of the non-local reference that x makes to f. However, this already requires agents that are subject to selection pressure and that possess an X system (this is formalized mathematically in van Hateren 2015f). Fundamentally, non-living physico-chemical processes lack memory (the memory in machines is a macroscopic phenomenon that presupposes human interpretation; at the microscopic level, machines do not utilize memory). Using memory for explaining the ontic-causal status of biological functions would be circular, unless one first explains non-local reference across time (by introducing X; see also Chapters 6 and 10, and van Hateren 2015f, g).

The scope of memory is evolutionary in the case of DNA, but the problem remains for faster forms of memory. For example, Garson (2012) proposes a generalized selected effects theory for functions in neural systems, utilizing selective (but non-evolutionary) processes acting on synapses, neurons, or neural groups (e.g., through development and learning). However, selected neural functions are formed at an earlier moment than when they are typically used. In other words, the causation would depend on memory and would be epistemic-real again. Therefore, it would fail to give ontic-causal status to functions.

The current argument is similar to the intuition inherent in well-known counterexamples against SE theory. Such counterexamples involve organisms that are identical to actual ones but with a completely different history, such as hypothetical instant organisms (e.g., Swampman) that are produced spontaneously (Boorse 1976, p. 74; Neander 1996; McLaughlin 2001, pp. 108–113). If functions have an ontic-causal status and naturalism is true, identical organisms must have identical functions. But according to basic SE theory, different histories would imply different functions. Therefore, basic SE theory must be amended if one seeks to assign ontic-causal status to functions (see below).

The third possibility of causally connecting history with present functions is sketched in Fig. 5c. It involves human intellect interpreting the fitness history of a specific agent and assigning functions to the appropriate processes. The historical information that was lacking in Fig. 5b is now implicitly present in human intellect and memory. Human intellect thus connects historical fitness to the present agent. However, intellect already presupposes biological functionality, because it depends on memory, agency, goal-directedness, and non-local reference in general. The possibility of Fig. 5c is perfectly legitimate and is standardly used for scientific inference. But it only produces functions that are epistemic-real. The function ascription is objective and real for the human (in the sense of the real patterns of Dennett 1991a), but it does not produce an ontic-causal function in the agent.

The final possibility of producing a causal connection between fitness history and functions (Fig. 5d) assumes a special process in the agent, X. As explained above, this can indeed produce ontic-causal functions. Information on the fitness history is implicitly stored in the structure of X. Part of the theory can be seen as an amended version of SE theory, where x, rather than f, is utilized (see below).

8.3.2 Causal Role functions

The right half of Fig. 5 illustrates several variants of the CR explanation. This theory focusses on the present and investigates the causal role that functions have for the current capacities of an agent. Such capacities are typically relative to the agent's internal state and to its environment, including other agents. Clearly important to an agent are the capacities to survive, to maintain homeostasis, and to obtain a high fitness f. However, these are compound factors, which do not have causal efficacy beyond that of their constituent factors. For example, the fitness of a bacterium is produced by a multitude of physical factors (temperature, presence of nutrients, absence of antibiotics, and so on). Only these factors directly influence the bacterium and its chances of survival and reproduction. In contrast, fitness itself is an epistemic-real factor. Fitness is objective and real, and plays an important role in human scientific theories. But it has no autonomous causal efficacy beyond that of its constituents and their interactions, and it cannot make functions ontic-causal. Therefore, functions acquire mere epistemic-real status if they are explained by their role for survival, fitness and homeostasis. Human intellect is then required (Fig. 5e).

Capacities (Cummins 1975) or goals (Boorse 1976) are explicitly assigned during human analysis of a system (Fig. 5f). They depend on the causal organization of the system. However, 'organization' is an epistemic-real phenomenon, not an ontic-causal one. Inferring organization is part of human functionality. Organization in abiotic systems never has causal efficacy of its own, even if such systems are complex. For example, there appears to be structure and organization, in the form of non-local correlations, in the atmospheric system that produces weather and climate. Scientific theories about the atmosphere depend on specifying this structure. They may use complex explanatory factors in the form of correlated aggregates, such as clouds, tornadoes, seasons and ice ages. But such structure has arisen gradually and naturally from the history of system states, without structure itself participating in the causal dynamics. The actual causation is purely local, through local pressure, local radiation, local mass transport, and so on. Only those local factors are needed in the fundamental causal inventory of the world.

According to the standard naturalistic view of living organisms (i.e., without conjecturing an X process), they are also just physico-chemical systems, albeit highly complex ones. They may be more complex than most abiotic systems, but they are still fully driven by the standard local causation of any physical and chemical process. Nevertheless, living organisms appear special, because they have a cyclically closed organization. This forms the basis of organizational accounts of function (Mossio et al. 2009; Moreno and Mossio 2015). In a closed organization, the system specifically produces products and conditions that are required for sustaining the working of the system itself. This also happens in some simple abiotic systems, such as a candle flame (which sustains itself by drawing in its own fuel and oxygen). But living organisms do this in ways that are far more differentiated and complex. However, one can still completely define the dynamics of a complex cyclical system in terms of the local processes and local interactions of which the system is composed. Its complexity does not make it fundamentally different from the atmospheric system. One could specify all molecular components and interactions of a metabolic system in a similar way as those of the atmosphere, and readily simulate either system. In other words, 'organization' need not be included in a fundamental causal inventory of the world. It has no autonomous causal efficacy, neither in a candle flame, nor

in a standard (X-lacking) living organism. It cannot give ontic-causal status to biological functions.

In contrast, living organisms that contain an X process do have an additional causal factor that goes beyond the standard causation of abiotic systems. The presence of X in a G-loop introduces a relation as a causal factor, namely the estimating relationship between x and f. This relation cannot be reduced to local processes and local interactions (see Chapter 14). Moreover, X and x integrate processes across the organism, both by affecting and by being affected. This provides the organism with a form of unity that is lacking in abiotic processes. In abiotic processes, one can always eliminate structure as a causal factor, as in the weather and climate example given above. But this eliminative strategy does not work in the case of living organisms that contain an X process. Elimination would leave no room for the relation between x and f. It would thereby neglect an essential, ontic-causal part of how living organisms work. Living organisms are, therefore, intrinsically distinct, non-epiphenomenal entities, in contrast to, e.g., a tornado. X is evolvable, and could gradually emerge from systems lacking X. Therefore, the theory does not assume a property (distinctness) in order to explain that property. The explanation involves gradual change through time, which makes the explanation cyclical rather than circular. It is thereby perfectly legitimate. Finally, it should be recognised that both organismal unity and causally efficacious relations are key notions of the organizational theory of functions (see, e.g., Moreno and Mossio 2015, Ch. 2). The current theory may be viewed as providing a naturalistic grounding of such notions.

One way to detect what functions are typically doing is to observe the distribution of their properties in a population (Fig. 5g). This yields an estimate of statistical normality (Boorse 1977). The distribution of properties in a population approximately reflects the evolutionary history of the function, in that it is likely to be concentrated at properties that contribute positively to fitness. Therefore, current statistical normality can be regarded as the population version of the historical SE approach. However, distributions of properties have no autonomous causal efficacy and cannot directly influence organisms. Such distributions are epistemic-real entities, not ontic-causal ones. This approach, therefore, produces epistemic-real functions, depending on human intellect (Fig. 5g). The recent modal theory of Nanay (2010) also requires human intellect, because it depends on inferring the effect of functions in ‘relatively close’ possible worlds. Possible worlds are entities that cannot exert direct causal influence, and thus can only be used for explaining functions as epistemic-real.

As before, the only way to avoid human intellect is through an internal process X within the agent (Fig. 5h). X refers, implicitly, to the relevant factors in environment and agent. Functions become ontic-causal because of the causal efficacy of the relation between x and f.

8.4 Unification of theories of biological function

As argued above, the ontic-causal efficacy of functions derives from the fact that x estimates f. X is itself an evolved physiological or neural process. Therefore, the history of f has shaped the way in which X lets x estimate f. Thus, the structure of X depends on that history. It is, therefore, closely associated with the Selected Effects theory of functions. When X is used for explaining functions, the history of f is used as well, albeit only implicitly and

indirectly (Fig. 5d). The implicit memory of X that appears to be present here does not presuppose biological functionality (in contrast to when one would directly invoke developmental or genetic memory). It has emerged naturally from the evolved property of X that x estimates f (and that parts of X estimate corresponding parts of F).

In addition to the part of X that focusses on heredity and fitness history, there is also a behavioural component in X. This component modifies the organism during its lifetime, through phenotypic plasticity and similar processes. Again, these modifications depend on the requirement that x estimate f. There is no certain way for the organism to verify, on the spot, the correctness of this estimation. But effective mechanisms to that end must have evolved over evolutionary time. For example, learning strategies must have evolved that are likely to produce adequate estimations, on average. The behavioural part of X has no direct selected effects explanation (unless the concept of selection is stretched, as in Garson 2012). It is particularly associated with the Causal Role theories of function (Fig. 5h), because it specifically attempts to track real-time changes in F and f. The behavioural part of X is continually adjusted during the lifetime of an organism. Capacities and goals can thus become part of X.

X is causally effective because of two different causal aspects that are both necessary, as was explained above. First, a non-conventional causal aspect in the form of an estimating relationship between x and f. Second, a conventional material causal aspect in the form of the physico-chemical realization of X. The latter is a conventional process that monitors the condition of the organism and affects its variability. Functions are fully defined by how X monitors. In other words, functions do not depend on the etiology of X, but only on the current structure of X. Organisms that arise spontaneously (e.g., Swampman) have exactly the same X as identical evolved organisms, and they have therefore exactly the same functions. Neither does the causal efficacy of X depend on its etiology. Given identical organisms in identical circumstances (now and in the future), F and f will be identical, as well as the relation between x and f. X will then have the same effects on Swampman as on its natural counterpart. Nevertheless, etiology is still needed for understanding how X and its structure could arise.

The above considerations suggest that replacing f by x (and F by X) in existing theories of function has two major consequences. First, it aligns these theories with specific aspects of the new theory. Second, the existing theories will then actually produce ontic-causal rather than epistemic-real functions (Fig. 5d, h). This follows from the fact that X has autonomous causal efficacy, whereas f has not. One way to state the novelty of the present proposal is by noting that earlier accounts only consider the direct material realizations of functions (e.g., how they work or how they have been formed by natural selection). In the new account, natural selection works, in addition, on the X process. The X process monitors, but it does not directly (i.e., immediately and proximately) participate in the working of functions. X only indirectly affects functions, by modulating how much they can vary (and thus how fast they can change, potentially). The material realizations of functions do not require relations as causal factors (similarly to the fact that the weather does not require relations). In contrast, the ultimate causal efficacy of X does require relations (similarly to a weather simulation, which depends on relations with the actual weather if it is to be accurate and useful).

A taxonomy of existing theories of biological function is provided by Perlman (2004, 2009). The three main branches of that taxonomy are non-naturalistic theories (Platonic and

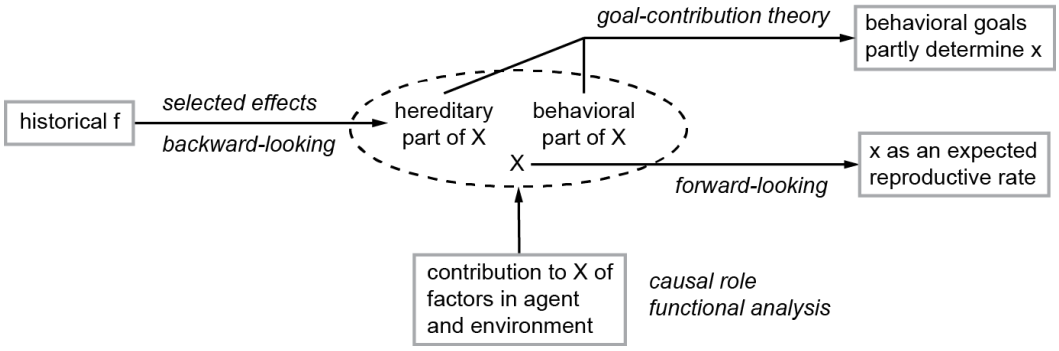

Fig. 6. The internalized process estimating fitness, X, can serve as an anchor point for amended versions of most previous theories of biological function.

religious), quasi-naturalistic theories that depend on the notion of emergence, and naturalistic theories. The latter theories are subdivided into conventionalism and theories that are primarily backward-looking, present-looking, or forward-looking. I will focus here on the latter three. Figure 6 illustrates that they can be viewed as representing different aspects of the new theory, by reformulating them within the new framework (by using X and x, rather than F and f).

The reformulation of the Selected Effects theory focusses on functions with goals related to the hereditary part of X. This part is formed by the evolutionary history of f in an agent's lineage (leftmost arrow). That part of X can be regarded as backward-looking (in accordance with Perlman's classification), because the structure of X implicitly refers to the evolutionary history. The reformulations of Causal Role theories (present-looking in Perlman's classification) specify how the factors of agent and environment contribute to X and its sub-goals (Fig. 6, upward pointing arrow). Formally, biological functions can then be regarded as capacities that are expected to realize the present sub-goals of X. This realization involves mechanisms using factors in agent and environment, as sensed by the organism in the present.

Goal-contribution theories (e.g., Boorse 1976) depend on current goals of an agent. Perlman classifies them as backward-looking to the recent past. Such theories can also be reformulated within the new framework. The current goals may then have been established recently in the hereditary part of X. Alternatively, they can belong to the behavioural part of X when they are acquired during the lifetime of an agent, such as through learning. The upper rightmost arrow, originating from both parts of X, symbolizes the rationale of these theories. Finally, forward-looking approaches (e.g., Bigelow and Pargetter 1987) focus on the overall goal of obtaining high f. When reformulated within the present framework, they focus instead on the overall goal of obtaining high x (which is in fact a true goal, in contrast to obtaining high f, which is only an 'as if' goal).

Figure 6 shows that these previous theories can be positioned in the new theory, although always with an essential and obligatory switch from f to x. The new theory unifies the earlier ones, and adds their explanatory power (see the next section). All causation in the theory is based on well-understood forms of causation, either primarily deterministic, primarily random, or combinations. The theory is therefore fully naturalistic. The required mechanisms are evolvable through standard natural selection (van Hateren 2015a and Appendix A). Nevertheless, the special, non-deterministic G-loop, as depicted in Figs. 2 and 3, produces a unique, emergent goal-directedness. This arises from the unusual fact that a

relation, namely the one between x and f, has acquired autonomous causal efficacy.

8.5 Intuitions about functions

Based on an extensive literature review, Wouters (2005) compiled a list of 15 intuitions about functions with which a theory of functions should ideally comply. He concluded that no existing theory could handle them all. Below they are discussed from the perspective of the new theory (all quotations are from Wouters 2005, pp. 133–134). The arguments rely on the fact that X itself has evolved, and that it continues to change on evolutionary and behavioural timescales. It gives the organism an evolutionary advantage only if x is a reasonable estimate of f. Therefore, X will typically contain and acquire components that are useful for such an estimate, and loose those that have become useless or detrimental.

1. "A theory of function should distinguish between activities that are functions (such as the beating of the heart) and activities that are side-effects of functional organs (such as heart sounds and pulses)." Side-effects are not included in the hereditary part of X (as they played no role in evolving X) and are therefore not automatically functional. However, when a side-effect is incorporated into the behavioural part of X, through learning, it may become functional.
2. "A theory of function should not allow one to ascribe functions to parts of systems that are not believed to have parts with functions (such as our solar system)." The solar system is not a living organism. It has neither f nor X, and therefore no parts with functions.
3. "A theory of function should allow for maladapted functions." The fur of a polar bear has as its primary function the reduction of heat loss. This function is determined by the hereditary part of the bear's X (as heat loss is of such importance for fitness that X must have evolved to utilize it for making x an adequate estimate of f). However, when the bear lives in a zoo in the tropics, f deviates from x (and the corresponding parts of F deviate from the corresponding parts of X). The fur is then maladaptive because it lowers f, but it is still a function for the bear because it remains incorporated in the bear's X.
4. "A theory of function should not depict the use other organisms make of the items of a certain organism as functions of those items. It is, for example, not a function of a dog's long hair to harbor fleas." For the dog, using its long hair for harbouring fleas is not a function, because it is not incorporated in the dog's X as a goal, i.e., as a factor that increases x. For the flea, living in the long hair of a dog is likely to be incorporated in the flea's X as a goal.
5. "A theory of function should distinguish between effects that are functions and effects that are accidentally useful. Although belt buckles occasionally save their wearers' life by deflecting bullets, it is not a function of belt buckles to deflect bullets." Accidentally useful effects just happen to contribute to f. But they are not incorporated in X (as they played no role in evolving X) and they are therefore not functions.
6. "A theory of function should not depict the systematic use humans make of existing items for new purposes as functions of those items. It is, for example, not the function of the human nose to support eyeglasses." It is not the default, biological

function of the nose, because it is not included in the hereditary part of X (as eyeglasses played no role in the evolution of X). Only when X is adjusted through learning, the nose may acquire an additional (though learned rather than biological) function for an agent.

7. "A theory of function should allow one to attribute functions to traits that currently do not vary in the population." The theory only requires that traits are expected to contribute positively to x and thereby probably to f. A positive contribution to fitness may not be observable in population variability. For example, some functions may play such a fundamental role for cellular functioning that any genetic variation in them would be lethal. Such variations are nevertheless bound to happen (for molecular reasons), but would not produce viable cells. They would therefore not be observable as phenotypic variation in a population.
8. "A theory of function should distinguish currently functional items from vestiges (like vestigial eyes in cave dwellers)." Vestigial eyes in cave dwellers are likely to have lost their representation in X, because if they would still be included then that would lower the accuracy by which x estimates f. Thus, it would have been selected against in previous evolution. Without representation in X, such eyes have no function for cave dwellers.
9. "A theory of function should allow one to attribute functions to the parts and behaviors of so-called 'instant organisms', hypothetical organisms that have no evolutionary history." Instant organisms are created including their X. X is just a concurrent physiological process. Those parts and behaviours that it monitors are functional. This is the same in an instant organism as in an identical organism with another history. Thus, the former has the same functions as the latter.
10. "A theory of function should enable us to attribute functions to items that do not actually perform it (most sperm cells will never fertilize an egg cell and mating displays quite often do not have the intended effect)." Functions correspond to sub-goals of X, which are, like X itself, to be understood in a probabilistic sense. They are expected to contribute, on average, to x and therefore, probably, to f. Sperm cells are indeed likely to contribute to f, statistically. Most do not, but the few ones that do are highly significant for fitness.
11. "A theory of function should enable us to attribute functions to items such as malformed hearts that are incapable of performing their function." A malformed heart influences only f, not the inclusion of its functional goal in X (which was established when X evolved). Therefore, it retains its function, even when X and x indicate it is malfunctioning. The same applies to the case when epidemics and major disasters reduce f in an entire population. Functions only depend on the form of X and they are therefore not changed by epidemics.
12. "A theory of function should allow one to attribute functions to the parts and behaviors of sterile organisms such as mules." Mules have a normal X and thus have the usual functions.
13. "A theory of function should not allow one to attribute functions to organisms as a whole." Organisms as a whole could only have a function if they are part of a larger system that has f and X. In that case, they would have a function for that larger system, not for themselves. One possible candidate for such a larger system is an ecological system. But such a system does not have a clear reproductive rate

(required for f), and there are no indications that anything resembling X and a G-loop could be present in an ecological system. A larger system that perhaps might have f and X, is a colony of social insects (briefly discussed in Chapter 6). Animal husbandry is a clear case where organisms as a whole can indeed have a function, e.g., when keeping sheep for their wool is incorporated into the behavioural part of human X. But sheep are then merely functional for humans, not for themselves.

14. "A theory of function should not allow one to attribute functions to such things as junk DNA, selfish DNA, and segregation distorter genes." Junk DNA and other forms of DNA that do not contribute to f are unlikely to have their working monitored by subprocesses of X. If X would implicitly attribute x-enhancing effects to such forms of DNA, the estimation of f by x would be less accurate. Therefore, it would be selected against.
15. "A theory of function should allow one to attribute functions to traits that are selected against." Circumstances may have changed such that not having a specific evolved trait, or having another trait, produces higher f. The trait is then selected against. But it may still be relevant for producing an x that estimates f, and therefore still be monitored by X (and thus be functional). There will be growing selection pressure on X to stop monitoring a trait if the trait gradually disappears or becomes irrelevant for f.

It is clear that the new theory performs very well. All intuitions are aligned with the explanations of the theory. Yet, the original theory (van Hateren 2015a) was not explicitly intended for explaining intuitions about biological functions. In that sense, the correspondence shown above is a successful prediction of the theory.

8.6 Discussion and conclusion

The analysis in this chapter makes it plausible that biological functions can indeed have an ontic-causal status. This requires a physiological process X within an agent that produces an estimate of the agent's actual fitness, f. The intrinsic X participates in a causal loop that is evolvable and sustainable by conventional evolutionary mechanisms. The loop produces genuine agency and goal-directedness in living organisms and makes the goal-directedness and normativity of functions ontic-causal as well. This ontic-causal status requires that functions in an agent be represented in X. Processes contributing to f without being monitored by X might be perceived by an observer as adaptations. They could be perceived as functional in the sense of objectively contributing to the agent's fitness f. However, such functionality would only be epistemic-real. It would only play a role for human scientific understanding. The agent itself is only directly connected to X, not to f and its history. Therefore, only functions that are included in X are ontic-causal. Only those functions strictly exist as autonomous, goal-directed parts of the causal dynamics of the agent.

Functions based on X combine the historical view of Selected Effects theories with the ahistorical view of Causal Role theories. The reason is that X forms, in effect, an implicit memory of previous evolutionary outcomes. In addition, it is adjustable in the present through learning and phenotypic plasticity. On the one hand, it is backward-looking to the distant and recent past. On the other hand, it is present- and forward-looking, because fitness is associated with the current likelihood of surviving and reproducing.

The theory presented here is new and largely conjectural. Nevertheless, there are strong reasons to think it is a plausible one. First, there are computational reasons, second, theoretical reasons, third, it can explain and unify a wide range of phenomena, and fourth, it is consistent with mounting evidence for the role of randomness in living organisms. Computationally, simple models show that the mechanism presented in Chapter 2 not only works, but also is evolvable for a range of conditions and models (van Hateren 2015a). The mechanism is advantageous, is quite simple in simple organisms, and requires only a slight variation on existing mechanisms (see Section 2.1). It is therefore plausible that evolution has produced it, at or close to the origin of life. The mechanism uses modulated randomness as an essential causal factor. The proposed system critically depends on and is evolvable through evolution by natural selection. This makes it understandable why the specific properties that it produces can only be observed in living systems.

As shown in the previous section, the theory is quite successful in explaining intuitions about biological functions. Moreover, it largely matches with the concept of meaning in biological systems that has been developed in the field of biosemiotics (Chapter 7). It explains why life seems to be characterized by having agency (Chapters 6 and 9). Other examples could be added. Many of these applications of the theory concern topics where alternative theories are absent, problematic, or only partially successful. A theory that can integrate wide, seemingly disconnected parts of reality in a well-defined way has intrinsic plausibility. Even if its components have not yet been shown explicitly, the fact that the theory has considerable explanatory power adds to the likelihood that such components actually exist.

Finally, there is mounting empirical evidence for the importance of functional randomness in living systems (Faisal et al. 2008; Brembs 2011; Kiviet et al. 2014). Several studies provide circumstantial evidence for the specific mechanisms of Figs. 2 and 3. At the subcellular level, mutation rates are known to be modulated in proportion to cellular stress (Galhardo et al. 2007), with stress presumably inversely related to cellular x. At the cellular level, the run-and-tumble behaviour of the bacterium *E. coli* (Macnab and Koshland 1972) provides an example of randomness modulated by the availability of nutrients, also associated with fitness. At the neural level, a similar modulation of turning rates and randomness has been shown in the nematode worm *C. elegans* (Gray et al. 2005; Gordus et al. 2015). In the context of foraging behaviour, switching from local search to a wider search area when the yield of food patches becomes low, appears to follow a similar pattern in many species (Hills 2006). Neural plasticity as controlled by how dopamine depends on reward prediction errors (Glimcher 2011) seems to conform as well. The dopaminergic system may thus contribute to X, at least partly.

However, all such examples may have alternative explanations, and their precise role for fitness is not clear. Ultimately, only targeted experiments with associated theoretical modelling can provide conclusive evidence for the theory. X is conjectured to integrate information about much of what is going on in an organism, and to produce effects throughout the organism. Therefore, a comprehensive system-theoretic understanding of the entire organism is required. Quantitative evaluation is probably only practicable, then, in very simple organisms. Nevertheless, there is no reason why empirical testing could not be performed, even though it would require considerable effort.

PART III: MIND

Chapter 9

Minimal agency, goal-directedness and value[8]

Agency is defined here as the capacity of an organism to initiate and generate behaviour directed towards a goal that the organism gauges, implicitly or explicitly, as meaningful and valuable. Thus, having agency provides the organism with some degree of behavioural freedom. From a fundamental, naturalistic perspective, agency has been difficult to understand, as it does not yield easily to mechanistic explanations. One would expect that meaningful behaviour should follow from certain criteria and rules. But rules suggest a deterministic mechanism, which, by its nature, could not initiate anything truly novel. The capacity to initiate novel behaviour suggests a mechanism that uses the indeterminacy and novelty of randomness. But that might only produce behaviour that is random rather than meaningful. This chapter proposes a mechanism that avoids this conundrum. It will argue that the mechanism of Fig. 7 (similar to Fig. 3, reproduced here for convenience) is sufficient to produce a minimal form of agency. It entails the presence of a goal with implicit value. Combining agency with consciousness can subsequently produce explicit goals and free will (see Chapter 12).

The G-loop of Fig. 7 lets a particular organism follow trajectories through an abstract and high-dimensional space of forms, with forms varying in behavioural dispositions (see Section 2.2). When traversed, the loop produces a sequence of random changes in behavioural dispositions. The resulting sequence of consecutive behaviours will be called a behavioural trajectory below. Each particular change in this trajectory appears to be fully random, but is in fact not completely so. The non-random part of each change is hidden in

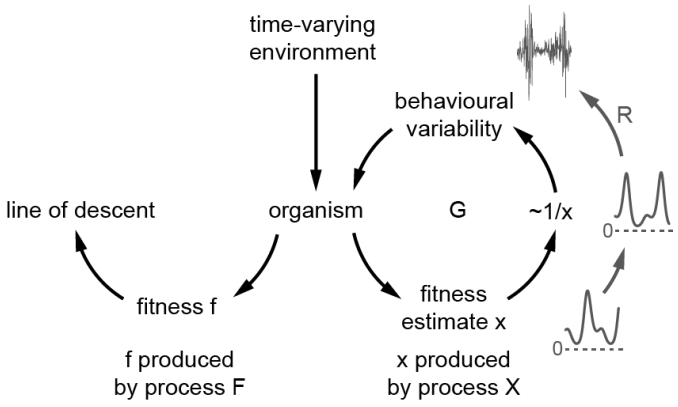

Fig. 7. Origin of agency and goal-directedness. Basic evolution by natural selection depends on the fitness f of each organism. Within each organism, a G-loop generates agency and intrinsic goal-directedness. The loop continually updates an organism's structure and behaviour in a random way, with the mean amount of change being modulated by an internally made estimate of fitness, x. The figure is slightly adapted from Fig. 3.

[8] This chapter is partly based on van Hateren (2015b); see also van Hateren (2022).

the on-average-expected magnitude of each change. This magnitude depends on an internally made estimate of fitness, x, which drives the variability of changes. The effect of the non-random parts gradually accumulates along a trajectory, the more so with each time the loop is traversed. It results in a behavioural trajectory that—in a statistical sense—gradually becomes strongly dependent on x (van Hateren 2015a). This dependence applies to the trajectory as a whole, whereas the trajectory is mostly random in its details. In effect, the G-loop produces a behavioural trajectory that is an inseparable compound of determinacy and randomness. It may be helpful to elaborate here on the main reasons for this, which have to do with a multiplicative interaction, cyclical causation, mixing of unknown external factors, and structural change that persists across time.

A determinate causal factor (say, d) can interact with a random one (say, r) in various ways. We will consider here the simple interactions of addition and multiplication. First, let us assume that the interaction between d and r is additive. Then the result would be d+r. When either d or r dominates, the result would be primarily deterministic or primarily random, respectively. For example, with d=249 and r=7, one might neglect r in 249+7 and regard d as the dominant factor. Alternatively, we can assume that the interaction between d and r is multiplicative, with d×r as result. Now there is no dominant factor. It would make no sense to neglect 7 in 249×7. Both factors are indispensable for the result. In other words, a multiplicative interaction, such as how ~1/x modulates variability in Fig. 7, puts the determinate and random factors on an equal footing right from the start.

This equality is further established by the continuous cycling of the G-loop. Cycling through the G-loop accumulates structural changes in the organism (in the form of physiologically engrained behavioural dispositions), originating from both d and r. Had the interaction between d and r been additive, then one would expect that changes attributable to r would gradually average out, after many cycles through the G-loop. Then particular values of r would typically affect a trajectory for only a limited time. Only d would remain as the factor that dominates the ultimate result. But such averaging out does not happen when the interaction between d and r is multiplicative. Each random value of r then puts the result on a significantly different trajectory, typically affecting the result for an unlimited time. Which particular values of r are randomly realized then influence the trajectory as strongly as the specific way how d (i.e., ~1/x and thus x) develops over time, over the time course of the trajectory. There is no way to view the resulting trajectory as some combination of the time courses of x and r during that trajectory. The determinacy and randomness of the trajectory have become inseparable.

A further feature of the G-loop of Fig. 7 is that x depends on a time-varying environment. This is so, because the actual fitness f depends on environmental circumstances (such as the availability of food or the presence of predators). These circumstances are part of the fitness process F that produces f. Then the process X within the organism should take such circumstances into account if it is to produce an accurate estimate x of f. Information about such circumstances may be obtained through the organism's sensors and its capacity to infer. Changes in environmental circumstances and their effects may be predictable, and then lead to behavioural change based on evolved or learned adaptive mechanisms. However, such mechanisms are not the ones considered here (see Chapter 2). Here, the unpredictable part of environmental change is considered. In other words, the environmental change in Fig. 7 is assumed to be partly random. This randomness affects f and thus x, and adds to the variability that is being modulated by x. Because the resulting

random structural changes affect the form of the organism, and therefore the resulting f and x, it has lasting effects on future randomness, through x. The value of x and thus the variability of structural change at a particular point in time then depend on the entire history of x and on the entire history of random changes. This follows from the fact that the G-loop entangles these factors and that the result is stored in persistent structural change. The changing structure thus establishes, in effect, a form of memory that is often permanent within the lifetime of individual organisms.

Unfading structural memory means that the G-loop follows ever-changing, newly created trajectories through newly created parts of behavioural form-space. Consequently, it produces nonstationary and non-ergodic dynamics, with future form-spaces that cannot be known or defined in the present. 'Nonstationary' means here that the statistically expected properties of an individual trajectory change across time. 'Non-ergodic' means here that the statistically expected properties of an individual trajectory taken across time differ from the statistically expected properties of the trajectories of a population of individuals taken at a single point in time. Stationarity and especially ergodicity are typically assumed when one uses standard statistical methods, an assumption that is not valid for what is produced by the mechanism of Fig. 7.

The above discussion argues that the behavioural trajectory is an inseparable compound of determinacy and randomness. In other words, the behaviour as it manifests itself on a longer timescale is neither random nor deterministic, but something in between. In effect, it provides the organism with some behavioural freedom and thus establishes a minimal form of agency. New behaviour can be initiated because of the randomness that participates in producing the trajectory. Behaviour is meaningful because it is shaped—through x—by the internal X process, which incorporates meaning (see Chapter 10). Moreover, behaviour is non-ergodic because the G-loop produces non-ergodic dynamics. This means that future behaviour and its statistics cannot be anticipated based on the current state of the organism or on the current states that are present in a population of organisms. Form-space itself is not fixed across time.

We have argued above that the G-loop produces a minimal form of agency. A related feature of the G-loop is that it produces a genuine goal in the organism, in the form of implicitly striving for a high x, that is, a high self-estimated fitness. Before explaining this, it should be noted that the basic process of evolution by natural selection itself does not involve any goal-directedness. The evolutionary process has no foresight or goal. When it produces organisms with adaptations that are matched to their current environment, it is because such adaptations happened to promote fitness in previous environments. Although adaptations may be perceived, post hoc, as goal-directed ones produced by a goal-directed process, either goal-directedness is only an apparent, an 'as if' one. In the present, such adaptations are just regular physicochemical processes with no more intrinsic goal-directedness than any other such process. Similarly, if a line of descent results in an organism with high fitness f, then this is just an observation after the fact. It is not produced by any intrinsic goal.

However, this is different for the mechanism that utilizes x. The G-loop produces behavioural freedom and agency, as it replaces the standard causation involving F and f by the special form of causation involving X and x. Given this agency, high x should then be viewed as a genuine goal of the organism. Chapter 5 introduced a factor C for denoting the fact that x estimates f and for quantifying the accuracy of this estimate. C does not only

quantify estimation, but can be shown to be, in addition, an emergent cause, namely of fitness-to-be (i.e., the fitness that gradually results, in a statistical way, from the action of the G-loop; see Section 2.2). Moreover, in Chapter 14 it is shown that this causal aspect of C is strongly emergent. The causal aspect of C cannot be reduced to the causal efficacy of a configuration of material constituents—in essence, because of the special way fundamental randomness is involved. C depends on x, which is produced by the X process. Thus, the causal efficacy of the X process (on fitness-to-be) cannot be reduced to that of a configuration of material constituents. Then striving for high x, as implied by the mechanism of Fig. 7, means striving for a goal that cannot be reduced to a configuration of material constituents. In other words, the goal-directedness of the G-loop is a distinct, irreducible phenomenon (see also Chapter 15). It is different from the reducible, 'as if'-kind of goal-directedness one might perceive in standard physical processes (such as when the water in a river seems to be heading towards to the sea, as if that were the water's goal).

A final consequence of the G-loop is that it produces genuine value. It was argued above that the G-loop provides the organism with agency and a goal. Having behavioural freedom combined with having a goal implies that the goal is important to the organism. This assigns value to the goal: the behaviour becomes meaningful to the organism itself. Because striving for large x is the overall goal, and x is produced by the process X, more specific meaning is attached to subprocesses of X. The way in which such subprocesses refer to subprocesses of the fitness process F can be interpreted in terms of intentionality. This is the topic of the next chapter.

Chapter 10

Intentionality and meaning[9]

The terms ‘intentionality’ and ‘intentional’ are used below in their technical, philosophical sense (see, e.g., Jacob 2014). They designate the power of minds to be directed towards something, for example when forming thoughts about objects or events. The terms are not used in their colloquial sense of having to do with intentions (in the sense of aims and purposes). ‘Intentional behaviour’ in this chapter does not mean behaviour that is done on purpose. Instead, it means behaviour that is based on processes that are about something. Thus, intentionality is used in the sense of ‘aboutness’.

Intentionality seems to be absent from those parts of nature that are not somehow involved in life. Such parts may causally affect each other, but they are not, by themselves, about each other. Intentionality is quite puzzling from a causal point of view, because a thought can be about a non-existing object (e.g., a unicorn) or about events that never happened (e.g., those in a novel). It is not clear, then, how intentionality might be explained in a naturalistic way. One possibility is to refrain from explaining it explicitly and directly, by assuming that it is fundamental itself, or that it depends on consciousness, which might be fundamental or at least must be explained first (Searle 1983; Strawson 2008, pp. 281–305; Kriegel 2013). However, that is not the approach taken here. Here we aim to derive intentionality from basic processes that may occur within living organisms, thus providing a direct naturalistic explanation of intentionality.

In recent decades, several theories for naturalizing intentionality have been proposed (reviewed in Shea 2013; Mendelovici and Bourget 2014; Hutto and Satne 2015). The main issue is how external entities, such as objects and processes, can be connected to internal processes of the mind. Tracking theories of intentionality assume that external entities are tracked (i.e., indicated) by internal processes, through a causal, correlational, or informational connection (Dretske 1981; Fodor 1990). However, such theories have difficulty explaining cases where the objects to be tracked do not exist (such as unicorns). Teleosemantic theories of intentionality (Millikan 1984; Neander 2017) assume that external entities produce the causal dispositions of internal processes in an indirect way, through an organism’s etiology (i.e., causal history) of evolution by natural selection. However, such theories have difficulty explaining cases where this history is deviant or does not exist, for example when an organism is synthesized or arises purely by chance. The explanation would ascribe a deviant or non-existent intentionality to such an organism, despite the fact that it would be identical to the normal one and would go through identical states. Other theories, such as based on functional learning, explanatory ascriptions of intentionality (Dennett 1989, 2009), and social constructions of intentionality (e.g., Brandom 2008) suffer from problems as well, typically because they implicitly depend on elementary forms of intentionality. Given these persistent problems, one may find it implausible that intentionality could ever be naturalized. How could it possibly work? The main purpose of this chapter is to offer such a possibility. It proposes and explains a

[9] This chapter is a slightly modified version of van Hateren (2021a).

biological process, that, if it exists, could provide a naturalistic explanation of intentionality. The current proposal thus takes a different approach than extant ones, by depending on a process for which there is no independent evidence yet.

The theory to be presented here superficially resembles correlational and etiological theories, but it has in fact a radically different causal structure. It is based on a conjectured internal process within each organism that estimates the organism's own evolutionary fitness (including causal constituents of fitness; see Chapters 2 and 4). The theory might be called an estimator theory. The term 'estimator' has here its modern statistical meaning of a method or procedure that produces an estimate of the value of a variable. Estimation is fundamentally different from causal, correlational, or informational tracking, because it is one-sided (see Section 10.2.1). But estimation is not a standard part of nature, as it is usually regarded as belonging to human epistemic practice. Because epistemic practice depends on intentionality, it would be circular to assume estimation in order to explain intentionality. What is first of all needed is a naturalistic theory of how estimation can arise in nature, without involving humans or any other source of intentionality. Section 10.2 shows that this is indeed possible. The result is a bare minimum, loosely called 'minimal intentionality' (reminiscent of the Ur-intentionality proposed by Hutto and Satne 2015). It does not require a human mind, not even a mind at all—strictly speaking, it thus falls short of the concept of intentionality as defined above. Sections 10.3–10.5 then use this minimum to build a construct that approaches the conventional, human kind of full-blown intentionality. However, the chapter only sketches the contours of the latter. Human language, a major means of human intentionality, is addressed only briefly (but see Chapter 11).

Human intentionality is closely associated with consciousness and agency. Such phenomena can be tentatively explained with variants of the theory presented here (see Chapters 9, 12 and 13). This implies that the current theory of intentionality is embedded in a much wider theoretical context. This blocks several potential objections to the theory. In particular, the estimating process explained below, X, is a process that fully integrates agency (and fully integrates consciousness in organisms capable of consciousness). Thus, agency and consciousness cannot be used to override X.

As stated above, an important caveat of this study is that the existence of the internal estimating process X is a conjecture. The process is evolvable and its existence appears quite plausible given what is currently known about (neuro)physiology (see discussions in Chapter 8 and in van Hateren 2019), but whether it is actually present or not has not yet been established. Hence, the process and its role have the status of a working hypothesis, for the time being.

10.1 Desiderata for a theory of intentionality

A naturalistic theory of intentionality should generate all of the presumed properties of intentionality. The list below contains properties that are commonly assumed.

(a) Directedness. An intentional component of an intentional process is directed towards something, points towards something, refers to something and is about something. The entity towards which it points may or may not exist, may be vague and may not be consciously perceived. But in any case, entities towards which intentionality points do not automatically point back. Intentionality is, thus, fundamentally one-sided. This is

different from the standard properties of a relation: if A is related to B, then B is related to A (though often in a different way); moreover, the existence of a relation between A and B presupposes that both A and B exist. Intentionality has neither of these properties, and is, strictly speaking, not a proper relation (Brentano, discussed in Kriegel 2016).

(b) Capability to make contingent errors. An intentional component may happen to point in the wrong direction, that is, it may point towards another entity than—implicitly or explicitly—assumed within the intentional process to which the component belongs. For example, an intentional system may perceive a predator where there is actually only a bush.

(c) Capability to make systematic errors. An intentional component may misrepresent, that is, it may always point to another entity than assumed within the intentional process to which the component belongs. Systematic errors can be related to ignorance. For example, one may not know that hoverflies (which commonly look like wasps) are flies rather than wasps. Then referring to a hoverfly as a 'wasp' is a misrepresentation: the actual target (a hoverfly) is different from the intentional target (a wasp). The term 'intentional target' is used here and below as short for 'the target of an intentional component that is assumed by the intentional process to which the component belongs'. The intentional target may or may not correspond to the 'actual target' (i.e., the entity that is actually targeted, if it exists). The capability to make systematic errors means that there is no disjunction ('or') problem (Fodor 1990): referring to a hoverfly as a 'wasp' is an error, not an indication that the term 'wasp' actually means [wasp or wasp-like hoverfly].

(d) Capability to point to non-existent entities. An intentional component may point to an entity that does not exist, a fact that may or may not be known to the intentional process. The former case corresponds, for example, to imagining a unicorn. The latter is a special case of making an error, as in (b) or (c).

(e) Capability to point to abstract entities. An example of a purely abstract entity is a mathematical object, such as the number π.

(f) Capability to point rigidly to some entities (Kripke 1980). For example, proper names of entities (e.g., the nearby star called 'the Sun') can have a unique and unambiguous reference.

(g) Directedness can be many-to-one. A single entity may be the target of many different intentional components at once. For example, an intentional process (e.g., a thought) may characterize a single object (e.g., an apple) by many different properties (such as colour, shape, taste and texture), which each correspond to a different intentional component. Such components may interact and overlap in complex ways and may not be fully separable.

(h) Directedness can be one-to-many. A single intentional component may target many different entities at once. For example, it may target 'all red objects present in the room'. In extreme cases, the number of entities targeted may become indefinite or unlimited (e.g., 'anything in the future that will be red'). An intentional component (e.g., the one associated with the word 'jade') may even be directed towards two different materials at once, regardless of whether this is known to the intentional system or not.

(i) Capability to target a single entity in different ways with different meanings. This is related to the distinction, made by Frege (1892), between reference ('Bedeutung', used by Frege for the actual entity that is targeted) and sense ('Sinn', the way in which the

entity is targeted, that is, the meaning or 'content' of the intentional component). One consequence of different senses is that an intentional component may target an entity A and not target an entity B, even if, unknown to the intentional system, A and B are the same entity. This is Frege's puzzle: one may refer to the morning star (target A) as if it were different from the evening star (target B), whereas in reality they are both the same planet (Venus).

(j) Perspective and grain. Intentional components have a perspectival or fine-grained nature. Many different perspectives are possible for a single intentional target. For example, the interpretation of the same visual scene may change depending on one's knowledge. Similarly, the meaning of words shifts depending on surrounding text, on context and even on the backgrounds of speaker and listener. It may appear, then, that meaning is indeterminate and that reference is inscrutable (Quine 1960). However, any indeterminacy and inscrutability are quite limited in practice (Searle 1987; Horgan and Graham 2012). Intentionality has, at least approximately, determinate content.

It is clear that intentionality is a complex phenomenon that requires a complex theory. Before explaining the theory in detail, it may be helpful to provide a rough sketch of how it works. The key innovation is the introduction, by conjecture, of a specific internal process (X) within each organism. This process continually evaluates how well the organism is likely to fare in terms of its evolutionary fitness. This includes both the organism's present performance and predicted future success (thus deviating from teleosemantic theories, which focus on the past). Crucially, the internal process then drives structural changes in the organism by combining random and determinate processes (a mechanism that can be shown to gradually increase fitness). Because of the randomness, the causal link between internal process and eventual increase of fitness occurs only slowly and indirectly. The better the internal process mimics the external world (as relevant for fitness), the higher the eventual increase of fitness that results. Because the mechanism is indirect, it avoids the too close causal coupling—between parts of the internal process and parts of the external world—one finds in tracking theories. The mechanism results in the one-sided directedness of estimating (Section 10.2.1). The internal estimation of fitness in different species should mimic their actual fitness, which may involve complex factors in some species (including social and cultural factors). Complex aspects of intentionality can then be inferred by subsequently analysing increasingly complex variants of the fitness estimator (Sections 10.3 to 10.5). Examples of how to apply the theory can be found for fairly simple cases (depending on the explanations up to Section 10.3) in Appendix B and for more complex cases in Section 10.5.2.

The sections below gradually develop the theory in detail. Section 10.2 starts with an explanation of the most fundamental property of intentionality: one-sided directedness. It is the conceptually hardest part of the theory, because it depends on a subtle, evolvable combination of determinacy and randomness (see also Chapter 2).

10.2 The evolvability of minimal intentionality

(1) Assume a variable environment in which organisms of various forms are evolving by natural selection, that is, by differential reproduction: some forms tend to reproduce more than others. The tendency to survive and reproduce of each individual organism is given by

its fitness f. It is defined here as a time-varying variable that quantifies to what extent the organism may transfer its traits to the next generation. Thus defined, high fitness usually requires both a good chance of not dying (per unit of time) and a good chance of reproducing (per unit of time).

Fitness f is assumed to be the instantaneous outcome of a highly complex physicochemical process F, which includes all factors of organism and environment that affect f. F unambiguously combines within-lifetime and evolutionary aspects of fitness. It includes within-lifetime aspects, because f changes instantly when circumstances deteriorate or improve (e.g., f decreases when there is a drought or epidemic, because these decrease the chance of surviving and reproducing). It includes evolutionary aspects, because f is a forward-looking measure of (statistically expected) evolutionary success. Note that f is probabilistic and prospective, and is thus immune to the issue that, in retrospect, actually realized short-term success sometimes conflicts with actually realized evolutionary success.

The totality of organismal factors that participate in the F of a particular organism is abbreviated below as the 'form' of that organism. Which parts of the organism compose its form, and how they do so, is well-defined, because F is assumed to be well-defined at any point in time. However, F changes over time, because environment and organisms change. The form of organisms is assumed to change continually, both within the lifetime of a particular organism (such as through development and learning) and across generations (through hereditary change across a line of descending organisms). An organism that typically has a high f over its lifetime is more likely to transfer its hereditary properties to offspring than an organism that typically has a low f. As a result, the distribution of properties over a population of organisms usually changes gradually, particularly in response to environmental change. Equivalently, the probability of finding specific properties in an organism changes, as well as the probability of finding specific forms of the organism. Thus, the typical form of organisms evolves.

(2) Item (1) describes a basic version of evolution by natural selection. Importantly, it defines f for each individual organism, that is, fitness is here not defined as a property of populations, nor as a property of specific traits. Moreover, it takes f as forward-looking, probabilistic and time-varying. Natural selection depends on differential reproduction as a result of variation of the forms of organisms. Hereditary changes to the form of an organism are assumed to be random and undirected. It is assumed here, in addition, that non-hereditary changes to the form of an organism that occur during its lifetime consist of micro-changes that are random and undirected, too. The latter assumption is made in order to keep the explanation below simple. However, it is not essential. The presence of directed changes (as produced by, e.g., phenotypic plasticity or learning), occurring along with undirected ones, would not change the conclusion of the argument below.

Intentionality is a feature of individual organisms and it occurs within their lifetime. Therefore, we will focus here on changes to the form of individual organisms that occur within their lifetime. Let us call the number of micro-changes per unit of time R (i.e., R is a rate of change). The source of such micro-changes in biological organisms is typically thermal noise (i.e., random motion of molecules). Cellular and neurophysiological processes are usually based on small, fluctuating numbers of molecules. Inevitably, such processes are partly random (Faisal et al. 2008).

When unfamiliar environmental change challenges an organism, a series of micro-changes enable it to explore novel forms that might meet those challenges, i.e., that might restore or increase fitness. However, the value of R needs to be set carefully, because it should be neither too low, nor too high. If R is too low, an organism could not change its form fast enough to keep pace with environmental change. The result would be low fitness and the prospect of death. On the other hand, if R is too high, the form of an organism changes strongly per unit of time, in a random direction (as the net result of a large number of random micro-changes). The forms that would result from strong changes are likely to function poorly in current and imminent environments, because such changes are likely to overshoot environmental change. This would produce low fitness as well. Thus, the rate of micro-changes R should be well matched to the rate of environmental change. In statistically variable environments, it could be advantageous to have an adjustable rate, that is, a controlled R. This is elaborated on next.

(3) The main conjecture made here is that, as a means to control R, an internal process X has evolved within the organisms. X has a time-varying output value x that modulates R (more on that later). Both X and x are assumed to be distributed throughout the organism, in an analogous way as how that happens in a neural network. In humans, most of X is assumed to reside in the brain. Modulation of R by x is accomplished through conventional causal mechanisms. For example, x might modulate the rate by which behavioural dispositions change. This can be done by facilitating or suppressing the effects that molecular randomness has on forming and modifying the cellular or neuronal structures that generate behaviour. Because X is part of the organism, its form can be modified as well. Such variations then happen within the lifetime of the organism as modifications of X on top of the basic form of X that was inherited (and that is modified only on an evolutionary timescale). The major question is now which form of X would maximize fitness. At first sight, this may seem like an intractable problem. Yet, it has a unique and simple solution, explained below and in items (4) and (5).

The key notion is that the rate R results in a diffusion-like process and that a variable rate can produce structure in the distribution of organismal forms. R lets the form of an organism migrate through an abstract and high-dimensional space of possible forms (abbreviated to ‘form-space’ below). Migration through form-space is analogous to molecular diffusion, because random micro-steps continually change the organism’s form in random directions in form-space. This is similar to the random walk of molecules (produced by random inter-molecular collisions) that results in molecular diffusion (e.g., of ink particles in water). The speed of diffusion (i.e., the average speed that results from the statistics) depends on how many micro-steps are taken per unit of time. Thus, it depends on the rate R. When R is small, the form migrates slowly, that is, it changes little per unit of time, on average. The organism then tends to linger close to its current form. On average, the form gradually moves away (in form-space), but only slowly. Therefore, forms that contain an X that produces small R appear sticky: organisms that happen to acquire such a form tend to stick around (i.e., stay similar to this form for a while). In contrast, when R is large, the form of an organism changes fast, on average. It quickly migrates away from such a form. Therefore, forms that contain an X that produces large R appear repellent: organisms that happen to acquire such a form seem to be repulsed and move away quickly (in form-space).

It should be noted that stickiness and repulsion change dynamically depending on the internal dynamics of X as well as on structural changes of X. The form of X can change from moment to moment, because it depends not only on heredity, but also on changes made within the organism's lifetime. In addition, environmental variations can change the output x (and thus R) for a given X.

(4) The modulated diffusion process explained in (3) tends to let organisms cluster around forms that have an X that produces small R. This is true of an individual organism in a probabilistic sense: it spends more time while having such forms. Conversely, it spends less time while having forms that produce large R. In effect, the probability that the organism has specific forms is clustered (i.e., is high) at forms with small R. Because this clustering applies to each organism, a population of organisms displays clustering as well. A population clusters in the sense that there is an increased density (in form-space) of organisms that have forms with small R at any particular time, on average. This directly follows from the fact that individual organisms spend more time close to such points in form-space. Thus, (3) can be regarded as a mechanism that produces clustering of forms.

Importantly, there is a second clustering process present. When an organism reproduces, it produces a new organism that is partially the same (that is, the hereditary part of that organism is partly similar). Therefore, differential reproduction tends to form clusters of similar forms as well. A population clusters in the sense that there is an increased density (in form-space) of organisms that have a form that produces high fitness. The density at forms that produce low fitness is low (because of a low rate of reproduction). An individual organism clusters in a probabilistic sense: the probability of producing a similar form is clustered (i.e., is high) at forms with high fitness.

We conclude, then, that there are two independent clustering processes. The first is based on a differential rate of micro-changes and the second is based on a differential rate of reproduction. Would it be possible, then, to align these two clustering processes? And if so, what would be the consequences? The two clustering processes can indeed be aligned by requiring that R is small when fitness f is large (and that R is large when f is small, with intermediate values of f and R covarying in an appropriate way). Then the (stochastic) clustering produced by small R coincides with the (reproductive) clustering produced by high fitness.

According to (3), R is assumed to be modulated by x (in a still to be specified way). Therefore, the simplest way to produce alignment is when x is made similar to f and when x then modulates R in an inverse manner (i.e., small x gives large R and large x gives small R). Because x and f are quantified by single numbers, similarity of x and f just means that these two numbers are similar, including how they change over time. The system produces enhanced clustering, because the clustering produced by high f is now automatically aligned with the clustering produced by small R. Small R results here from high x, which obtains because high f implies high x (as x is similar to f). The latter condition (i.e., that x is similar to f) is introduced here as an assumption, but it is shown to be evolvable in (5).

(5) Aligning the two clustering processes has two major consequences. First, it increases the fitness of organisms that utilize this mechanism. The reason is that when fitness is high, R is small (because x is high, as implied by high fitness). This means that such forms stick around in form-space. If they stick around, the organism that has such a form gets ample opportunity to take advantage of the fact that its form has high fitness. Thus, its survival

and reproduction are facilitated, that is, its time-averaged fitness is increased. On the other hand, when fitness is low, R is large (because x is low). This means that such forms change quickly, and move away (in form-space) from their low-fitness form. An organism may then have to move through forms with even lower fitness. But it might survive and eventually migrate to forms with high fitness (and then automatically stick around there). On average, this is still better than staying at a low-fitness form and waiting for certain death. Computational simulations (summarized in Appendix A) show that this mechanism is indeed one that enhances fitness when environments are variable. Organisms that modulate R in this way outcompete organisms that have an optimized, but unchanging R. In other words, alignment of the two clustering mechanisms is evolvable and it is sustainable by continued selection pressure. The effect on fitness is slow and gradual (as it depends on stochastic clustering). In order to emphasize this, the resulting fitness will be called fitness-to-be below. The current fitness is still denoted by f.

(6) The second major consequence of aligning the two clustering processes is even more interesting. Alignment requires that x becomes similar to f. It is hard to overstate the significance and the extraordinary novelty of such a similarity. One should realize that f and x are unrelated, intrinsically. The fitness f is the result of a complex process in nature, F. It objectively describes the tendency of an organism to survive and reproduce. In contrast, x is the output of an internal process X that has, in principle, nothing to do with fitness—it does not participate directly in F. If X evolves (through trial and error) in such a way that x tends to mimic f, then that produces, fundamentally, an arbitrary correspondence. It is a correspondence that is evolvable, according to (5), but there is no intrinsic, pre-existing connection between x and f (or between X and F). The best way to describe what x does is that it estimates f (in the theoretical sense as used in estimation theory). Because X is the process that produces the estimate x, X is properly called an estimator. An estimator is a procedure (here realized in the form of the process X) that yields an estimate (here x) of the value of a variable (here f).

It is important to understand that X (and x) are categorically different from F (and f). F is a regular physicochemical process, in the same category as, for example, the atmospheric processes that produce the weather. In contrast, X is an internal estimating process, in a similar category as a process that simulates the weather (through observation and computation). In other words, the evolvability of mechanism (4) produces estimation as a categorically novel factor. It should be stressed that this estimation is intrinsic to each organism: it is fully made within the organism, by process X. It has autonomous causal efficacy (on fitness-to-be) and it does not depend on human interpretation (and thus differs from a weather simulation in these respects). Moreover, it is a true evolutionary innovation, because estimation does not occur in those parts of nature that are unrelated to life.

(7) We have seen above that X is likely to evolve such that its output x estimates f. However, we have not specified how well x must estimate f. Perfect estimation is unattainable, because F usually includes complex physicochemical processes as well as complex other organisms. However, even poor or mediocre estimation produces some alignment of the two clustering processes, and can therefore already enhance fitness-to-be. The better the estimation becomes, the higher the fitness-to-be can become. Therefore, there is selection pressure on organisms to improve the estimation, given the means available to specific species and given the benefits (in terms of increasing fitness-to-be) compared with

the costs (in terms of decreasing fitness, because of the energy, materials, learning time, and hereditary resources that are consumed by X).

10.2.1 Intermediate evaluation

Minimal intentionality has property (a), directedness, because one can say that x estimates f, but it would make no sense to say that f estimates x. The reason is that x and f have quite different causal properties. Although x modulates R by conventional causal mechanisms, this modulation only increases fitness-to-be when x and f are similar. Without this similarity, the two clustering processes would not be aligned and there would be no effect on fitness-to-be. Thus, x acquires an additional causal efficacy (on fitness-to-be) when x and f are similar. In contrast, f does *not* acquire an additional causal efficacy when x and f are similar. The fitness f still quantifies expected evolutionary success, irrespective of whether there is an X process or not. This causal difference between x and f implies that x points to f, but that f does not point back in any meaningful way, that is, in a way that has causal consequences for the organism itself. This conforms to the fact that intentionality is one-sided. Roughly speaking, x is about f, but f is not about x.

Minimal intentionality has properties (b) and (c), contingent and systematic errors, only in a weak sense, as associated with the inevitable limits to how accurately x can estimate f. Any inaccuracy may be viewed as indicating errors in the estimator. However, in order to make this more explicit and more convincing, it is necessary to parse the processes that produce x and f, that is, to parse X and F (see below). Property (d), the capability to point to non-existent entities, is not realized, because f must exist. Moreover, f is not abstract, thus (e) is not realized either. All other properties depend on multiple components in the intentional process (X) and in its target process (F), and, thus, depend on parsing X and F.

10.3 The parsing of minimal intentionality

Section 10.2 showed that organisms can evolve an internally generated variable x that estimates the organism's own fitness f. How the variables x and f can be parsed was already explained in Chapter 4, of which the relevant parts are reproduced here, for convenience. The variables x and f are produced by complex processes, X and F, respectively. The structure of these processes cannot be fully isomorphic, because F is orders of magnitude more complex than X could ever be. F includes a large number of factors that influence the fitness of an organism. These factors originate from within the organism itself, from its environment and from other organisms. X, on the other hand, is an approximate simulation of how the major factors affect fitness. X occurs fully within the organism; it is limited by the available processing power as well as by what the senses can tell the organism about itself and its environment.

Nevertheless, even if the structures of X and F are not identical, they must have similarities. The reason is that X has evolved as a means to produce an x that estimates f in many different circumstances. If circumstances change, not only f may change, but also the composition and structure of F. Then X and x must change as well, through evolution and learning, if the organism is to remain competitive. Changes in the structure of F typically involve coherent and correlated changes of different parts of F. For example, when food becomes scarce, or when an organism migrates to another environment, this changes many

parts of F at the same time. Because F is a process, parts of F can be regarded as subprocesses. Subprocesses of F that typically change coherently are called F-components below. F-components should be roughly reflected in the structure of X, because this facilitates change of X, both evolutionary change and within-lifetime change. When an F-component changes, only the corresponding X-component (i.e., the corresponding subprocess of X) needs to change then as well. This is far more feasible than changing many disconnected parts of X at the same time, which would be required if X would lack distinct components. Therefore, organisms are likely to have evolved an X that includes not only distinct components that reflect those of F, but also the capability to develop and learn such components.

X-components that roughly correspond to F-components estimate those components, including their role in producing f. This is a more complex version of estimation than before, because components are subprocesses rather than single numbers (such as x and f). In weather terms, it is analogous to estimating an extended weather system (e.g., the course and properties of a hurricane) rather than just a single variable of the weather (e.g., the temperature at a particular place). Estimating extended processes may involve estimating many variables at once, as well as estimating the dynamics and coherence of components of the process. Estimating need not be done in a literal, isomorphic way. For example, a detailed computational simulation of the weather may be fairly isomorphic, but an experienced meteorologist interpreting a weather chart may use abstract conceptual short-cuts, and a farmer reading the sky for a short-term weather forecast may use mere rules of thumb.

Estimating complex components is, as before, fundamentally one-sided. The causal efficacy of an X-component depends not only on the actions and interactions of its micro-parts, but also—and crucially—on how it contributes to the X process as a whole, that is, to x. Thus, X-components obtain their causal efficacy (on fitness-to-be) from that of x. In contrast, the causal efficacy of an F-component depends fully on the actions and interactions of its micro-parts. Therefore, X-components estimate corresponding F-components, but the reverse is not true (because the latter would lack causal efficacy).

However, there are several complications. A first complication is that X, not F, determines how F is parsed. This follows from the fact that X is the source of the causal efficacy produced by parsing and estimating. Irrespective of the question whether F might have an autonomous parsing, F is necessarily parsed by X when X forms distinct components based on the available correlational structure of F. Nevertheless, the latter structure is objectively present. Therefore, there is presumably only limited scope for variations in how X can effectively parse the part of reality that is incorporated in F.

A second complication is that X-components may not always correspond to specific F-components. X is unlikely to be flawless, because it is the result of trial and error. It may contain components that have no counterpart in F, that estimate a component in a mistaken way, or that estimate the wrong component. Furthermore, X is likely to lack counterparts of many potential F-components. Such errors and omissions lower the accuracy by which x estimates f. However, in variable environments the detrimental effect on fitness may be too small to be counteracted by evolution or learning. Small differences of fitness produce effects only slowly, if at all, because evolution as well as learning by trial and error are statistical processes. In variable environments, small fitness differences may not persist long enough to produce appropriate changes in X. Moreover, small fitness differences may

drown in statistical noise when population sizes are small. And finally, correcting errors and omissions may simply be too complex or too costly for a specific species.

A related complication is that the accuracy by which X-components estimate F-components may vary from poor to excellent. Poor estimates may be all that can be accomplished given the available means. Yet, poor but veridical estimates may still be better than no estimate at all. A final complication is that clusters of X-components may be used to estimate clusters of F-components, including many-to-one and one-to-many mappings. Such clusters may have a complex internal structure, with complex interactions between the components of the cluster. Many-to-one and one-to-many mappings are likely to depend on context, because context affects both X and F. Therefore, context affects how clusters can best be formed.

10.3.1 Intermediate evaluation

The parsed form of intentionality has property (a), directedness, because an X-component has causal efficacy (on fitness-to-be) only because it estimates an F-component. The causal efficacy of an X-component occurs regardless of whether it estimates an F-component well and regardless of whether it participates in X in a veridical way (i.e., in a way that improves x as an estimate of f, on average). The only condition for being causally efficacious is that an X-component actually contributes to the X process and thus affects the way by which x estimates f. Even if that estimate deteriorates as a result, and thus decreases fitness-to-be, the X-component remains a directed intentional component of the intentional process.

Property (b), the capability to make contingent errors, is present, because X changes dynamically and can temporarily produce an X-component that points to the wrong F-component. If such errors remain in X for a long time, it produces (c), the capability to make systematic errors, as well as (d), when pointing to a non-existent F-component. However, (e), the capability to point to abstract entities, is not yet realized, because F is assumed to be a concrete process. Then its parsed components are not abstract either, because they fully consist of concrete micro-parts.

Property (f), the capability to point rigidly, is realized when the accuracy of x (as an estimate of f) strongly requires that a particular X-component points rigidly to a particular F-component. For example, an organism that would not reliably (i.e., rigidly) recognize specific mates or specific sources of food would have an unsustainable (i.e., lethal or infertile) version of X. More abstract versions of (f) require the extensions of intentionality discussed in Sections 10.4 and 10.5.

Properties (g) and (h), that directedness can be many-to-one and one-to-many, are realized when X combines components. Different versions of clustered X-components may point to different versions of clustered F-components. Again, abstract versions of (g) and (h) require more elaborate versions of intentionality. This also applies to (i) and (j), but they are already present in primordial form. Two X-components X_A (a subprocess about target A) and X_B (a subprocess about target B) may point to different assumed F-components, F_A (an assumed target A) and F_B (an assumed target B), even if there is in reality only a single F-component F_C (the actual target C). In contrast to Frege's use, 'reference' has to be interpreted here as the intentional target (A or B), not as the actual target (C). Within X, X_A may have a role in producing x that is independent of X_B's role in producing x. Frege's 'sense' (or 'meaning' or 'content') can be identified with each of these roles, which are

estimates of the conjectured roles of F_A and F_B in producing f. X may produce a reasonably accurate x even if it does not incorporate the fact that X_A and X_B estimate the same actual entity; nevertheless, incorporating such a fact (thus equating F_A and F_B) is likely to improve x on average, across a wider range of circumstances. Because X can change dynamically, roles can change dynamically as well, which leads to primordial forms of being fine-grained (j).

In conclusion, all desiderata of Section 10.1 have at least a minimal incarnation, with the exception of pointing to abstract entities. The theory up to this point is applied to several examples of minimal intentionality in Appendix B. However, the focus of this chapter is not on minimal intentionality, but on full-blown intentionality. Several applications of the latter are presented in Section 10.5.2. The required extension to abstract entities is the topic of Sections 10.4 and 10.5.

10.4 The extension of intentionality to other organisms with intentionality

Above, F is viewed as a fully physicochemical process. However, this is not true any more if the environment contains organisms with an X process. Each X process affects fitness by using intentional components that obtain causal efficacy through estimation. Although estimation is realized by a physicochemical process, it is an overlay on such a process. Estimation itself is not physicochemical—roughly in the same way as one can say that a machine that computes a weather forecast is a physicochemical process, but that the forecast itself (particularly the fact that it is about the real weather) is not physicochemical. Thus, the incorporation of other organisms makes F not fully physicochemical.

An organism may benefit from taking this into account. It can do that by utilizing components in its own X process that point to X-components of organisms in its environment. Thus, this requires intentional components pointing to intentional components (in a similar way as in Dennett 1989). However, this is complex and difficult, especially because intentional components cannot be directly observed. They need to be inferred from observed behaviour. Therefore, it is only worthwhile for an organism to have the appropriate inferential means if the intentional behaviour of other organisms is highly significant for the fitness of that organism. Moreover, it could be accomplished only by organisms that have access to sufficient resources to maintain a sophisticated X. Intentionality pointing to intentionality is related to the idea that some animals may utilize a Theory of Mind in order to predict the behaviour of other creatures (e.g., Call and Tomasello 2008).

Targeting an X-component in another organism is semi-abstract, because such a component is only partly concrete. Its physiological implementation is a physicochemical process, but the fact that it estimates an F-component is not. As before, an X-component pointing to another X-component remains an intentional component, regardless of whether it characterizes its target well and regardless of whether its target exists at all. If two organisms share mutual interests, they may benefit from producing behaviour that explicitly displays the content of their X-components. In this way, they can more easily infer each other's X-components. Such reciprocal intentionality creates the possibility of intentional communication, intentional cooperation and intentional deception. It may even involve X-components that point to X-components that point to X-components. Then organism 1 could estimate how organism 2 assesses the X-components of organism 1. However,

constructions along these lines cannot become too complex, because the amount of processing required of X would quickly rise, as well as the uncertainty in the estimates. Therefore, complex constructions can evolve only if the fitness benefits are considerable.

In conclusion, having X-components that point to X-components adds some abstraction, but not yet the full abstraction that can occur in human language and mathematics. That requires a further extension of intentionality.

10.5 The human extension of intentionality

Above, fitness f was defined as an organism's tendency to survive and reproduce. This may be adequate for some species, but fitness is often more complex than individual survival and reproduction. For example, social organisms may help their kin. This can indirectly increase the likelihood that their properties are transferred to subsequent generations, if those properties are hereditary (and thus similar in kin). Such transfer increases an organism's fitness, even if the organism does not reproduce itself. Fitness that includes these indirect effects is known as inclusive fitness (Hamilton 1964). Hence, f has to be redefined accordingly.

Section 10.2 showed that a minimal form of intentionality is produced by aligning two clustering processes. The first process requires a modulated rate of micro-changes (R), and the second process requires differences of fitness. In Chapter 3 it was explained that alignment continues to works for various forms of fitness. The relevant paragraphs of that explanation are mostly reproduced below, for convenience.

The fitness-based clustering was explained above in terms of individual fitness, but it works for inclusive fitness as well. The reason is that kin are likely to be close in form-space, that is, to cluster. When kin help kin to survive and reproduce, this increases the likelihood that the forms in a cluster reproduce. Thus, the social component of inclusive fitness enhances reproductive clustering. This implies that alignment with the other, statistical clustering process is optimal when R is driven by a redefined x. This x must then estimate the redefined f (i.e., it must estimate inclusive fitness). The resulting fitness-to-be then also refers to inclusive fitness.

Interestingly, this analysis suggests that there is a further way to enhance clustering (van Hateren 2015c). Forms that cluster at a particular point in form-space (because of small R and high f) need not be kin. This is particularly true in species that can easily vary their form during their lifetime, by readily varying their behavioural dispositions. Then most of the individuals that display similar behaviour may be unrelated and genetically dissimilar. Such individuals then have similar forms (i.e., similar in terms of behavioural dispositions) that cluster at a particular point in form-space. As is explained in the next paragraph, they can enhance clustering by helping other individuals in the cluster, regardless of whether those individuals are kin or not. The only criterion for helping is then similarity of form.

Helping enhances the fitness f of the individuals in a form-cluster, which means that their x increases as well (because x estimates f). Increasing x lowers R, and thus reduces the likelihood that they drift away to other forms. Moreover, other individuals that happen to acquire that particular form in form-space get the same lowered R, and thus tend to keep that form. In other words, that particular form functions as an attractor in form-space. Therefore, helping individuals with a similar form enhances not only fitness, but also clustering. Both f and x need to be redefined once more, in order to include the effects of

helping individuals with a similar form. Simulations (summarized in Appendix A) show that this mechanism is indeed evolvable under the right conditions. Organisms that help organisms with a similar form then outcompete organisms that help only kin. Similarity of form as such becomes heritable because the clustering establishes attractor forms in the population. In effect, attractor forms recruit new organisms by inducing them to change their form to become similar to the attractor form. This type of heredity is, thus, not an intrinsic property of specific individuals, but a property that is induced in contingent individuals by the structure of the population in form-space. This structure can remain quite stable and evolve gradually over many generations. It should be noted that this bears similarity to ideas about cultural evolution (Boyd et al. 2011) and about cultural attractors (Claidière et al. 2014). However, intentionality is either not used in these and similar theories, or is implicitly assumed. Therefore, these theories fall outside the topic of naturalizing intentionality.

There are several conditions that need to be fulfilled for the proposed mechanism to work. First, the clustering process based on x and R must be present, because the fact that a form can become an attractor is based on reducing R. This implies that the mechanism can work only if there is intentionality of the kind explained above. Second, only species that can flexibly and strongly change their behavioural dispositions during their lifetime can produce significant clustering that is unrelated to kinship. And third, helping other individuals based on the form associated with behavioural dispositions requires reliable recognition of such dispositions. Therefore, it requires considerable cognitive resources. The combination of these three conditions suggests that the mechanism may be fully developed only in humans.

The clustering proposed here depends on helping other individuals who are similar, but who can easily change their behavioural dispositions. The latter induces the risk that the forms of the individuals in a cluster could drift apart, even when R is small. This would then decrease the efficacy of helping. Stability is, thus, a potential problem. Reciprocal communication between two intentional systems (mentioned in Section 10.4) is an effective way to synchronize and stabilize the behavioural dispositions of two individuals. A public system of communication can perform a similar role for large numbers of individuals, such as occur in clusters. Thus, a public language is presumably evolvable because it can stabilize clustering. It should be noted that this is not necessarily a mechanism that makes R small. R could still be large enough to allow fast responses to environmental change. The mechanism only ensures that the clustering remains intact, by allowing the individuals belonging to a cluster to change their forms synchronously and consistently with each other.

So how does this lead to abstract entities? Section 10.4 argued that an X-component pointed to by another X-component is only semi-abstract, because X-components are partly concrete. However, once there is a public language, there must be X-components that are shared by all individuals that use that language. Each individual then has a version of such a component. Such versions need not be fully identical, but should at least be sufficiently similar to allow effective communication—ineffective communication would decrease clustering and fitness. Let us call $\langle X_A \rangle$ the average, public version of an X-component X_A that targets an F-component F_A. Then, $\langle X_A \rangle$ is the version that an individual variant of X_A should approximate if it is to function effectively in public communication. The proper way to let an individual variant of X_A approximate $\langle X_A \rangle$ is to let $\langle X_A \rangle$ be a secondary intentional target of X_A. Thus, X_A estimates both F_A and $\langle X_A \rangle$. This utilizes property (h), that

directedness can be one-to-many. For example, when referring to a specific tree, using the word ‘tree’ (which produces the subprocess X_A) points not only to the tree (F_A), but also to how the word is used in the language community ($\langle X_A \rangle$). This applies to both speaker and listener(s).

It is clear that $\langle X_A \rangle$ is not a concrete process. It involves X processes across a large and variable population of individuals. Moreover, $\langle X_A \rangle$ could be derived partly from individuals of previous generations, and it could be documented. Therefore, $\langle X_A \rangle$ should be regarded as fully abstract. Now suppose that there exist specific public components $\langle X_A \rangle$ that, by themselves, partly determine (or are assumed to determine) the fitness of individuals, by being assumed parts of F. For example, a specific $\langle X_A \rangle$ may point to the number π, and one might conjecture that having mathematical abilities can contribute to an attractor in form-space. Then one could have an X_A that points only to this $\langle X_A \rangle$ (and thus to π), a fully abstract public entity. This establishes a pure example of (e), the capability to point to abstract entities.

10.5.1 Concluding evaluation

Intermediate evaluations above have already discussed and explained most desiderata of the list in Section 10.1, which will not be repeated here. The evaluation here focusses on how the addition of a public language makes several properties more distinct. Property (c), the capability to make systematic errors, can now acquire a nearly discrete, binary status (i.e., formulated in terms of true and false, with true interpreted as ‘beyond a reasonable doubt’) rather than a continuous one (i.e., lying somewhere on a scale ranging from very accurate to very inaccurate). Using the public word ‘wasp’ when one refers to a hoverfly is false, not just very inaccurate. The reason is that the use of the word is stabilized by public knowledge. Such stabilization can even become absolute (and hence truth and falsehood can become absolute) within abstract symbolic systems (such as mathematics and logic).

An estimate that is either true or false (i.e., that has a truth value) can be regarded as a representation in the full, symbolic sense. Therefore, the X-component that obtains when one uses the word ‘wasp’ is a representation of a wasp. Saying that the word ‘wasp’ represents a wasp is short for saying that the symbol ‘wasp’ (such as in the form of an ink pattern, sound pattern or memory trace) produces an X-component that estimates a wasp—usually truthfully, but sometimes falsely so, such as when it actually points to a hoverfly.

Property (d), the capability to refer to non-existing entities, is facilitated, in particular in the form of deliberately imagining a non-existent entity. Its non-existence is stabilized by public knowledge (such as that unicorns do not exist; in normal individuals, this is maintained as independent of, and therefore not destabilized by, privately fantasizing about unicorns). The capability to point rigidly, (f), becomes more pronounced as well. Public knowledge fixes the reference to ‘the Sun’. One-to-many directedness, (h), is facilitated by the fact that publicly supported reference enables abstract generalizations (such as ‘all entities that could have a colour’).

Property (i), that the same reference can have different senses, can occur within individuals (see Section 10.3.1), but also between individuals or between groups of individuals. The latter happens when individuals or groups use idiosyncratic versions of X_A for an otherwise publicly fixed reference $\langle X_A \rangle$. Then X_A depends on the perspective of an

individual or group. It is fine-grained and may change rather quickly. However, $\langle X_A \rangle$ is determinate at any point in time and is expected to be quite stable across time. Thus, there is no significant problem with indeterminacy and inscrutability, as required by (j). Nevertheless, $\langle X_A \rangle$ can gradually change across historical time, for example when there are changes in the meaning of specific words.

10.5.2 Applications

Applying the theory of intentionality requires the following steps: first, decide which F-component is involved, and then assess the presence and structure of the corresponding X-component. We will first analyse a case that produces problems for previous theories based on evolutionary arguments, but not for the current one. Subsequently, we will analyse a case that produces problems for conventional tracking theories, but again not for the current one. Finally, the theory is applied to a case that is challenging to naturalistic theories in general.

Suppose that an organism with full-blown intentionality is copied, either artificially or by a lucky coincidence (such as 'swampman', e.g., McLaughlin 2001, pp. 108–113). Such an organism has no conventional evolutionary history, nor a conventional life history. This means that etiological ('causal-historical') theories of intentionality must ascribe intentionality to such a copy that is different from that of the original (or that is even non-existent). Such ascription is problematic, because original and copy are indistinguishable, including any memory they might have of their own history. The current theory has no problems with this case. When a copy is made and placed in the same environment, one gets an organism with exactly the same F and exactly the same X as the original. Intentionality is then exactly the same as well, because it is produced by X-components estimating F-components. The main reason why the theory works well here is that it defines fitness as forward-looking, as the—statistically expected—tendency to survive and reproduce. When the present is given, past fitness is irrelevant for future fitness: original and copy will have the same chances of surviving and reproducing (given identical current environments, and assuming that any future environments and contingencies would happen to be the same for both). The past is only relevant if one wants to explain how original and copy came into existence, but different explanations do not lead to different futures (given current identity). It is clear that the current theory, while partly based on evolutionary arguments, is not etiological at all (see also Chapter 8). It depends on an internal estimate of (statistically) expected future evolution, not on past evolution.

Tracking theories usually suffer from the disjunction problem. For example, on a dark night, viewing a horse may give rise to the same sensory impressions as viewing a cow (example from Fodor 1990). One might think that this necessarily leads to a confusion or collapse of the mental representations of horses and cows, but this is not what happens in practice. Such representations remain separate. The current theory readily explains that. Sensory impressions on a dark night involve F and F-components. In contrast, mental representations involve X-components. The latter will remain separate for horses and cows, independent of current lighting conditions. If an F-component involving a cow on a dark night happens to be best estimated by an X-component associated with horses, then that is just an error made by the X-components associated with horses and cows. There is no reason to adjust the content of horse-related X-components or cow-related X-components based

on such an isolated error. Adjustment is only justified when horses and cows are consistently confused across many different viewing conditions and over a considerable period of time, and when most people in the individual's language community are confused too. The main reason why the theory works well here is that the process of estimation separates the intentional components of the X process from the immediate sensory impressions and immediate causation that belong to F.

As a final example, we consider Putnam's thought experiment 'brain in a vat' (BIV). Suppose that an evil scientist removes a person's brain and puts it in a vat with the right nutrients to keep it alive. The brain is connected to a computer that simulates the normal input to the brain as well as the effects of the output of the brain. It is assumed that the simulation is perfect, such that the brain does not notice anything abnormal. What can we say about intentionality in this case? The assumption of the thought experiment is that the external parts of F are replaced by a computer simulation. But X is still (mostly) the same, because most of it resides in the brain. It still estimates its assumed version of F in the same way. Therefore, intentionality is initially not changed, despite the fact that the estimation is severely flawed (because F has been replaced by a completely different physical process). However, intentionality cannot be maintained indefinitely in this way. Intentionality depends on the causal efficacy (on fitness-to-be) of x estimating f; this efficacy requires genuine fitness (i.e., physical survival and reproduction). Such fitness is abnormal in the BIV (because it has no physical body, has no physical relatives in the simulated environment, and is not part of a physical community). Crucially, fitness is fully lacking in organisms simulated by a computer because of lack of embodiment. Inside a computer there is no genuine fitness, that is, no physical survival and reproduction. Hence, the simulated organisms have no intentionality, and the original assumption (perfect simulation) is necessarily false according to the theory proposed here. When the BIV tries to bond and communicate with simulated people, it will soon find out that their intentionality is fake. As a result, the BIV is likely to become very confused and to develop erratic forms of intentionality. Eventually, the persistent lack of the prospect of meaningful dialogues will destroy the BIV's consciousness according to the theory of van Hateren (2019; see also Chapters 12 and 16).

10.6 Discussion and conclusion

The explanations and evaluations above show that it is possible to construct intentionality in a naturalistic way if one conjectures that an X process exists. The crucial step is the alignment of two clustering processes, one associated with a differential rate of reproduction and the other associated with a differential rate of micro-changes. The alignment enhances fitness, and is thus evolvable through regular evolutionary mechanisms. It necessarily produces estimation. However, estimation is not part of the causes that standard evolutionary theory utilizes. Therefore, the mechanism explained above can explain intentionality, whereas standard evolutionary considerations fail. This failure is often summarized by stating that natural selection cares only about reproductive success, not about truth. However, this is only true of F, but not of X. X cares about truthfully estimating F, and thus cares about truthfully estimating the processes in Nature that participate in producing F. This is so despite the fact that X itself has evolved as a means to improve reproductive success.

Another objection to evolutionary theories of intentionality is that natural selection cannot distinguish between two different external entities that have exactly the same effect on fitness, and that have always had so in the past (Fodor 1990). Again, this is only true of F, but not of X. The implicit task of X is to estimate components of F reliably in as many different circumstances as possible. This includes circumstances that have not yet occurred. Therefore, if X obtains indications that the two external entities are not identical, it should represent them separately, because their effects on fitness may differ in future circumstances. In specific cases this may not produce an evolutionary advantage, or at least not immediately, but the strategy as such is evolvable. A type of X that systematically follows this strategy is likely to outcompete a type of X that does not. Versions of X that are capable of social and cultural communication are particularly likely to produce the strategy, because they can quickly and flexibly change the way by which they parse F.

An extensive critique of previous theories of intentionality is beyond the scope of this chapter. The primary purpose here is to explain the new theory in sufficient detail such that its explanatory power can be understood. Nevertheless, it is useful to briefly state key characteristics of the theory in which it differs from some or most extant theories. Specifically, the theory implies that there is no hard connection to the actual target of an intentional component, that meaning is internal, not external, that the evolutionary past is irrelevant for intentionality in the present, and that intentionality is associated with a process, not with a state.

Many theories assume that there is a hard connection to the actual object towards which a thought is directed, or to the actual object to which a word or expression refers. In Frege, reference denotes the actual target, and the assumption is common in recent studies as well (such as in tracking theories of various kinds). Hard connections produce all kinds of problems, such as the disjunction problem (see Section 10.5.2) and troubles when the actual target is absent or imaginary. In contrast, estimator theory holds that such hard connections do not exist. The actual target does not directly drive the intentional system, but is inferred by the intentional system. Although this is often based on sensory data induced by an actual target, the intentional system is not controlled by such data, but specifically acquires them for the purpose of estimating. The accuracy of intentional components affects the intentional system only indirectly, through the stochastic clustering mechanism explained above. The internal realm of thoughts (part of the X process) is only softly coupled to external reality (part of the F process) because of the randomness utilized by X.

The lack of hard connections to external entities implies that estimator theory denies common claims that meaning and mental content are ‘not in the head’. Meaning is produced by the estimator X and its components, which are processes in the head (and presumably somewhat in the body too). The process of estimating does not necessarily depend on the actual identity of the entity estimated (e.g., to take the example of Putnam 1975, it is irrelevant for the meaning of ‘water’ whether the actual target is H_2O or XYZ, if X and the associated language community cannot distinguish these). The theory implies that embodiment is important (see the discussion of the ‘brain in a vat’ thought experiment in Section 10.5.2), but does not view the mind as extended into the outside world.

Although the theory depends on evolutionary arguments, these concern estimated future evolutionary success, not past evolution. In that sense it differs strongly from teleosemantics. As discussed in Section 10.5.2, only the present form and circumstances of an organism (including the form of its X process) determine its intentionality, not how it

obtained that form (through evolution, learning, or otherwise). A dependence on past evolution is circumvented by the presence of X, which has the kind of structure that is needed for producing reasonable estimates of future evolutionary success.

Finally, intentionality is associated with a process and should not be ascribed to a static state. The effect of X and its components crucially depends on time. Clustering is not an instantaneous phenomenon, but arises gradually, by accumulating sufficient statistics over time. Therefore, intentionality can be ascribed to a mental process, but not to a mental state (if that is interpreted as static, such as dispositional; thus, a 'belief' has intentionality only during the time it is a real-time 'believing'). It is, according to the theory, not correct to ascribe intentionality literally to neural memory traces or to a book. Such entities are intentionality aids, not forms of intentionality. They merely assist in producing meaning at the moment when they are utilized as input to a real-time subprocess of X (such as a thought). Saying that a book is meaningful is then metaphorical: it is short for saying that the book produces meaningful intentional components when read.

Chapter 11

Language

Having language is clearly a defining characteristic of the human species. It is crucial for understanding the human mind and human social behaviour, and instrumental in sustaining the evolutionary success of humans. Structural aspects of languages, such as their phonetics and syntax, are well amenable to study and are quite well understood. A more enigmatic aspect of language is its meaning, that is, how language empowers the mind to deal effectively with the physical and social world. It is not well understood how reference works, the apparent capacity of words to refer to objects and events—not just to near-by physical objects, but also to objects distant in space and time, to non-existent objects and to abstract ideas. Nor is it well understood how language could evolve and why it has evolved to a full-blown form only in humans. The purpose of this chapter is to provide tentative answers to these questions by elaborating on the theory of intentionality presented in the previous chapter.

Spoken language can fulfil several functions at the same time. It can help the speaker to assert social agency, that is, to initiate actions coordinated with other individuals. It often serves to communicate the intentions and goals of the speaker. Most importantly, it lets the speaker communicate about objects or topics that may or may not be nearby. The term ‘aboutness’ (or ‘directedness’) can be used here for the latter property of language, the property that it is about something. Agency, goal-directedness and aboutness are, thus, important aspects of language. How these properties could arise in biological evolution has not been easy to explain. Standard evolutionary theory seems to incorporate an implicit goal, in the form of a drive towards survival and high reproductive success. But such a goal is mere appearance, merely ‘as if’. Only with hindsight, one can say that the surviving organisms were successful. Prospectively, most organisms will eventually turn out to be unsuccessful, which contradicts any foresight and goal-directedness. Similarly, agency and aboutness are rather enigmatic, and are indeed absent from those parts of nature that are unrelated to life.

The theory presented in this book offers solutions to these enigmas. The preceding Chapters 9 and 10 specifically explain how agency, goal-directedness and aboutness can arise in nature, when organisms evolve a mechanism such as the one in Fig. 7. This mechanism depends on an internal process X within each organism that produces an estimate x of the organism’s own fitness f (which is produced by a process F). This fitness estimate can subsequently enhance eventual fitness (denoted by fitness-to-be) by modulating the variability of the (neuro)physiological structures that produce behaviour. The elaborations in Chapter 10 show how increasingly complex forms of aboutness can arise when fitness becomes more complex (such as when helping kin or peers is included) and when communication acquires a sophisticated form.

This chapter will specifically focus on such sophisticated communication in humans. Tomasello and Carpenter (2007) argue that a distinct property of humans is their well-developed ability to share aboutness, for example when forming shared goals and when focussing with shared attention on an object or task. The term typically used for this is

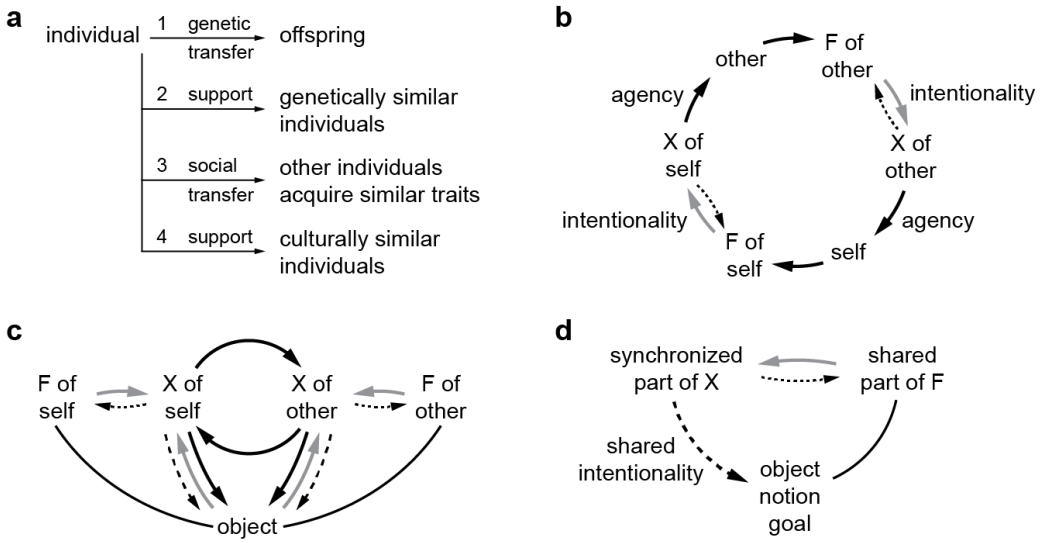

Fig. 8. The origin of shared intentionality and shared meaning. (**a**) Human extensive fitness consists of four pathways. (**b**) Cooperation enables a behavioural cycle involving the X of two individuals. (**c**) A specialized communicative channel (inner loop) synchronizes the intentionality of two individuals. (**d**) Shared intentionality. In all panels, solid arrows indicate conventional causation, dashed arrows intentionality, and grey arrows changes implied by intentionality.

'shared intentionality'. The term intentionality is used in philosophy for what is called aboutness and directedness above, and it should not be confused with intentionality in the sense of having intentions (see also Chapter 10). It is often used specifically for conscious, mental directedness. It is broader than having intentions, because it can also involve attention, or, in general, any mental activity that is about or directed towards something outside one's own mind.

Shared intentionality can be understood by extending the scope of fitness (Chapter 3). The latter is summarized in Fig. 8a (simplified from Fig. 4), which shows the four basic forms of human fitness (on the assumption that there is an X process). Several of the pathways in Fig. 8a can benefit from communication, which is assumed here to be between partners that are typically inclined to cooperate (because of shared interests). The simplest form this can take is illustrated in Fig. 8b, with two individuals, denoted by the self and the other. They can engage in a cooperative behavioural loop that involves agency and intentionality as follows. First, the agency associated with the X process of the self (left) produces a behaviour that affects the other (upper-left). Consequently, this changes the fitness process F of the other (upper-right), which is implicitly perceived by the X of the other (right) through intentionality (denoted by the dashed arrow). This intentionality is part of the X process, which thus changes along with F (grey arrow). The latter arrow is rendered grey in order to indicate that it is not a regular causal connection, but a consequence of the fact that X-components (i.e., subprocesses of X) estimate F-components (i.e., subprocesses of F). In response to the change of the X of the other, the associated agency of the other produces a behaviour that affects the self (lower-right) and thereby the F of the self (lower-left). Finally, this affects the X of the self (left), starting the next cycle of the loop. Note that this loop does not require an explicit form of communication. There is only an implicit form

of communication as implied by how the behaviour of the self affects the other, and vice versa.

Over evolutionary time, the sequence of Fig. 8b may gradually evolve into that of Fig. 8c. Now there is a direct link between the X process of the self and the X process of the other, through a specialized behavioural channel (the upper two central arrows). This channel produces explicit communication, but no direct action beyond what is needed for communicating. Actions affecting the self and the other are still produced by the agency associated with their X. Suppose that such actions primarily affect an object that is of interest to both the self and the other. These actions are denoted by the solid arrows from each X to the object. Because the object is part of the F of the self, as well as of the F of the other (solid lines), changes in the object affect the X of both individuals (grey arrows) via intentionality (dashed arrows). For the sake of clarity, the grey and dashed arrows towards the object are shown separately, but they are in fact part of the general connection between X and F. We see here that an entity in the outside world, such as a physical object, is connected to the self and to the other in two different ways, both through X. First, it is influenced by agency, and second it affects X through intentionality. The intentionality is similar for the self and the other, and it is indirectly connected through the communication channel that connects their X processes.

If the two partners have, on average, cooperative intentions, then the similar way in which their X is directed towards the object will produce similar changes in their F, on average. Figure 8d then shows Fig. 8c in a different way. The part of X of the self and the other that is engaged with the object can be regarded to be synchronized between the self and the other, through their communication. It is associated with a shared part of the F of the two individuals. The synchronized part of their X is directed towards the object as a shared intentionality. The object can be a simple physical object, but also more general entities, such as notions, goals, and the intentionality of other individuals (see Sections 10.4 and 10.5).

The communication channel that is needed for shared intentionality (Fig. 8c, d) might be a private channel, with codes that are shared only by the two partners. This would suffice for species where most cooperative behaviour would benefit close relatives (pathway 2 in Fig. 8a). For example, offspring benefits from communication during parental care and pair bonding. But in humans, much of cooperation is based on pathway 4 in Fig. 8a, involving large groups with similar phenotypes. Then the full benefits of communication can only be reached when communication can be shared, in real time, amongst more than two individuals. Which particular individuals communicate may vary from case to case. Effective communication then requires a standardized means of communication, that is, a public language. Once a standard has been established, this further facilitates communication between just two individuals as well. A public language requires symbols that represent entities, including abstract ones (see Section 10.5). Mentally processing such symbols can subsequently also enhance individual thought, such as in reasoning.

Discussing the detailed forms that human language can take is beyond the scope of this chapter. Rather, we will focus here on several general topics, from the perspective of the theory explained above. In particular, the sections below will discuss how language is related to time, meaning, truth and consciousness.

11.1 Language and time

Ultimately, the aboutness of language depends on the G-loop of Fig. 7. This loop is part of a dynamic process involving modulated randomness. This implies that intentionality does not exist as an instantaneous property. It needs to be generated and accumulated over time, by a continual cycling through the G-loop. The same is true of the other phenomena generated by the G-loop, such as agency and goal-directedness. We will assume here that meaning is indeed fully generated by the G-loop and its more elaborate versions (van Hateren 2015a, b). Then language as a means of transferring meaning between individuals is, fundamentally, a dynamic process. It cannot be frozen in time.

Nevertheless, human language is often associated with static entities as well. For example, learning a language leads to structural changes in the brain that are more or less static, in the sense that they are changing more slowly than the timescale of speech. Furthermore, the public nature of human language has led to static records in the form of written sentences, books and stored audio recordings. Often, one ascribes meaning to such static records. However, this is only a convenient manner of speaking. In reality, a written sentence only acquires meaning at the moment it is read, written, spoken, or heard. There has to be a human brain actually producing its meaning, in real time. When one says that a particular sentence contains meaning, this is short for saying that it is expected to produce meaning when a suitable person would hear or read it.

Producing meaning from stored records can be viewed as reproducing the meaning originally produced by the author of the records. This is not fundamentally different for records that are stored as memory traces in the brain, records that consist of scribbles written for oneself, or records that are written for communicating with others. But particularly in the latter case, the type of recording is standardized, because public accessibility requires socially shared symbols that are conventionally understood.

The importance of time for language and meaning also implies that one should be cautious when talking about mental ‘states’ and mental ‘content’. If states and content are interpreted as static, that is, as properties of the system at a specific point in time—frozen in time, so to say—then this means that they cannot have intentionality and meaning. Mental phenomena are processes, not states. Memory traces in the brain, such as those relevant for implicit beliefs and desires, do not have intentionality. At most one can say that their presence makes it likely that the meaning associated with specific beliefs and desires is produced in certain appropriate situations of thinking or acting. Such beliefs and desires are only involved in producing intentionality during the time that their meaning is actualized as a dynamic process. Of course, as a manner of speaking, one may still say that a person holds certain beliefs and desires, as long as it is understood that this does not produce intentionality, most of the time.

11.2 Language and meaning

The term ‘meaning’ has two major denotations in the English language. The first is when it is used to denote the significance or purpose of something, in expressions such as ‘the meaning of an action’ and ‘the meaning of life’. The second is the common denotation with respect to the meaning of language, such as in ‘the meaning of a word’ and ‘the meaning of

a sentence'. The two uses are related, because we have seen above that the G-loop generates agency and goal-directedness along with intentionality. In the pragmatics of everyday dialogue, these aspects of meaning always coincide (Grice 1957; Sperber and Wilson 1995). A meaningful dialogue presupposes that the words spoken are not spoken inadvertently, that is, they involve agency by the speaker. It also presupposes that the words are spoken for a reason, that is, they are significant and goal-directed, even if the goal might not be immediately clear. And finally, it presupposes that the words are about something that the speaker and listener could share in an intentional way (e.g., physical affairs, social situations and emotional states).

However, agency, goal-directedness and intentionality are strongly dependent on context and on which partners are communicating. Agency and goal-directedness are often hardly relevant for language as a formal, public system of communication. Semantic meaning, such as recorded in a dictionary, is essentially an averaged form of pragmatic meaning. It gives the meanings one can typically expect in a neutral context, i.e., 'unmarked' in linguistic terms. It is understood that the actual, pragmatic meaning may shift depending on context and speakers. For example, a sentence such as 'that tree will be cut down' is quite neutral from a semantic point of view. But it acquires different aspects of goal-directedness and agency when uttered by a lumberjack, a gardener, or a conservationist. In contrast, other sentences, such as 'that war criminal should be sentenced to life in prison', have a semantic meaning that is already loaded with goal-directedness, agency and normativity (in the sense of socially shared and fixed values). Thus, even when regarded as a formal semantic system, language is not completely free from implied goals and agency. However, for many isolated sentences the plausible contexts are so varied that their formal, averaged meaning carries almost no goal-directedness.

11.3 Language and truth

Traditionally, the meaning of language has often been associated with truth conditions. Particularly for a proposition, meaning then requires an understanding of the conditions under which the proposition is true. For example, the meaning of 'that tree has been cut down' is then clear if one could decide on its truth, that is, on the question to which facts in reality it would correspond. The theory presented here and in Chapter 10 has implications for how the concept of truth can be understood. Developing this into a full theory of knowledge is beyond the scope of this chapter (but see Chapter 15 for a slight elaboration). Nevertheless, a brief sketch will be given here because of the importance of truth for traditional theories of language.

Theories of truth typically focus on either correspondence, pragmatics or coherence. The theory that is presented here has aspects of all three: correspondence of some kind because X-components estimate F-components, pragmatics of some kind because estimation can be judged by how well it works (ultimately for increasing fitness), and coherence to some degree because X is expected to work best if its structure has a coherence that matches the implicit structural coherence of the F process. The F process is part of objective reality, and the F-components can then be viewed as components of reality.

Clearly, the correspondence, pragmatism and coherence of X cannot be perfect, because X can only approximate F. Nevertheless, there is evolutionary pressure[10] towards more truthful, that is, more accurate X. In most species, this drive is slow. It requires either evolutionary change of X, or learning and social change of X that is either not transferred across generations or transferred in a limited way. In contrast, the human X is strongly dependent on language, which enables fast cultural change of X. Moreover, language and culture are readily maintained across generations. The flexibility of language, including formal variants such as mathematics, has enabled human collective knowledge to quickly expand beyond that of other species.

It should be clear from the above discussion that the concept of truth produced by the theory is usually not the logical truth that one can have in purely formal systems, such as in mathematics and logic. Natural language is viewed here not as a formal system, but as a tool for successfully interacting with reality. Its truth resembles the truth of the empirical sciences, not the truth of mathematics. It is closely associated with accuracy and likelihood. Truth then at most means very likely true, perhaps so likely that it is true for all practical purposes and beyond a reasonable doubt.

This probabilistic, tentative nature of knowledge is related to the probabilistic, correlative nature of the dashed arrows in Fig. 8. Establishing a logical truth or falsehood about reality would require a direct connecting line to reality. No such lines exist (see also Section 10.6). Something resembling such lines only exist internally within constructed, purely formal systems, which are detached from the world that is ‘out there’. Word meanings always require a mapping to reality, which is inevitably somewhat vague, with exceptions and boundary cases. Even mathematical statements, such as ‘1+1=2’, are not necessarily true when applied to reality (e.g., 1 litre + 1 litre < 2 litres, when adding a litre of water and a litre of alcohol).

11.4 Language and consciousness

Creatures can be conscious without having a shared symbolic language. Nevertheless, introspection suggests that language has an important role for the consciousness of adult humans. In order to understand this, it is necessary to develop a theory of consciousness first. It is indeed possible to develop a theory of consciousness based on the proposed mechanism that produces intentionality. This theory is explained in detail elsewhere (Chapters 12 and 16, and van Hateren 2019). Very briefly, it conjectures that consciousness is produced when intentionality is prepared to be communicated, either externally to others or internally to different parts of the individual’s brain. The transformation that is required for such a preparation can be shown to be a strongly emergent cause, which is plausibly sensed as the feeling of consciousness. Importantly, this basic form of consciousness is not

[10] ‘Evolutionary pressure’ is used here as a generalization of ‘selection pressure’. The latter term usually refers to the genetically mediated effects of natural selection on fitness. Evolutionary pressure includes, in addition, the positive effects on fitness that occur—during an individual’s lifetime—when X (through x) estimates F (through f) more accurately. This accuracy can have a genetic component, but it can also be accomplished by cultural means.

perceptive (as when perceiving a visual scene) but communicative. All other forms of consciousness, including conscious perception, are then derivable from the basic form.

If consciousness is indeed a consequence of intentionality being communicable, it can be understood why language plays an important role for human consciousness. While there are many nonverbal ways to communicate intentionality (e.g., through touch, facial expressions, postures, and sounds such as laughing and crying), verbal communication is tantamount in humans. Components of intentionality that are prepared for verbal communication can lead to spoken language, but may also be used within the brain as further input to the parts that produce intentionality. The resulting intentional components can then be transformed once more, and so on. This sets up a continual cycle of intentionality and transformed intentionality (van Hateren 2019). This cycle progresses and changes continuously, because the X process depends on other inputs as well, such as input from memory and from the senses. An internal processing loop of this kind can be useful for preparing to communicate externally, perhaps at a much later time. But language-based internal processing can be useful for other reasons as well, because abstraction and symbols enable reasoning of a kind that may not be easily realized otherwise. Ultimately, these forms of thinking depend on the presence of shared intentionality and of extended forms of fitness (pathways 3 and 4 of Fig. 8a).

11.5 Discussion

Investigating the nature of meaning and language has a very long history, perhaps longer than any other intellectual endeavour. A comprehensive discussion seems quite infeasible. Therefore, I will focus here on a selection of topics that have remained particularly intractable over time. Yet, they have a clear interpretation when analysed from the perspective of the present theory.

11.5.1 Reference, extension, sense and intension

Words and linguistic expressions appear to have at least two different aspects. First, they are about something, that is, they relate and refer to something external to the mind. And second, they carry an intrinsic meaning that goes beyond this aboutness. Frege (1892) captured the relatedness by the term ‘reference’ (Bedeutung, used by Frege as the actual entity pointed to) and the intrinsic meaning by the term ‘sense’ (Sinn). In modern linguistic analyses, one often finds ‘extension’ and ‘intension’ (with an s, different from both ‘intention’ as related to aboutness and ‘intention’ as related to purposes and goals). There are subtle differences in how these terms are used (Chalmers 2002). But both reference and extension are normally used to relate a word to an object in reality. Superficially, they seem to resemble the dashed arrows in Fig. 8. However, it is crucial to understand that they are actually quite different from these arrows, and from the aboutness these represent. Reference and extension are typically not used as a direction (as analogous to a force vector), but as a connecting line, which connects word and object. They attach the object to the word, that is, to the meaning of the word. How such very different entities could be literally connected is typically not explained.

However, the present theory does not have that problem. There is no literal connection between meaning and object. The dashed arrows in Fig. 8 are attached to X, but they do not

actually connect to their intended objects. They just point in a certain direction. Evolutionary pressure stabilizes them against drifting away from their target, on average. Such pressure is taken to act continually, through stabilizing mechanisms in the individual's neurophysiology. For example, evolved or culturally established learning strategies may evaluate mistakes in X and improve its subsequent use and structure. The pressure keeps X and its components directed towards F and its components, on average.

Meaning is thus only indirectly coupled to its target, through a stabilized estimation. The coupling is essentially produced by the distant, tacit and implicit threat of death and extinction. Therefore, it is strong enough to prevent the disconnect from reality that could easily result from purely ideational systems of meaning (i.e., systems defining ideas purely in terms of other ideas). At the same time, the coupling is loose enough so that it can explain why meaning can be about abstract or non-existing entities.

The dashed arrows in Fig. 8 only represent one aspect of X. In addition, X generates agency and goal-directedness, on the individual level as well as on a socially shared level (Fig. 8c, d). For language, this produces those aspects of meaning that go beyond directedness, that is, aspects associated with sense and intension as mentioned above. The meaning of a language utterance is not completely given by the fact that it is about something, but also by how the utterance serves the agency and goals of speaker and audience. This is most clear for specific situations, where pragmatic meaning applies. It is less clear when semantic meaning is studied, as detached from a specific context. But as discussed above, seemingly neutral meaning is always implicitly understood as a provisional average, valid for a hypothetical, neutral situation. Such neutral meaning may be modified strongly—in terms of sense and intension—when a specific context is present.

11.5.2 Original and derived intentionality

Searle (1983) argues that humans have original intentionality, whereas artefacts such as books have mere derived intentionality. According to the present theory, the term 'derived intentionality' is potentially confusing, because a book does not have intentionality at all, derived or not. It is just a shared physical tool that can help its reader to produce meaning. It is a delayed, one-way communication channel, which, in a sense, channels intentionality from the author to the reader. But there is no intentionality in the channel itself, because the channel has no X process and no G-loop. The channel itself is purely physical, but coded in such a way by the author that the reader can translate it into real-time intentionality, while reading. Saying that a book has derived intentionality is analogous to saying, pointing to a sweater lying on a shelf, 'that is a warm sweater, it must have derived warmth.' The sweater has neither original nor derived warmth. It is just a physical tool that can help its wearer to keep the body warm. Books and sweaters do not possess intentionality and warmth, but can enhance intentionality and warmth that is present elsewhere.

Dennett, on the other hand, thinks that humans have no original intentionality. He states (Dennett 1989, p. 318) 'We may call our own intentionality real, but we must recognize that it is derived from the intentionality of natural selection, which is just as real.' This view implies that our own intentionality is just as much 'as if' as the one we may perceive in designed artefacts. Ascribing intentionality to entities or processes is, according to Dennett, primarily an effective way to describe and predict their behaviour. In contrast, the theory presented in this chapter implies a different view. It ascribes real intentionality to humans

(as well as intentionality or minimal intentionality to other species, see Chapter 10), but no intentionality to human artefacts. Moreover, it does not ascribe intentionality to natural selection, neither original nor derived intentionality. Rather, it proposes that evolution by natural selection has produced, through its regular, non-intentional mechanisms, organisms with an X process and the G-loop of Fig. 7, as well as more elaborate versions (van Hateren 2015b). Because of the special stochastic structure of this loop, it produces a primordial form of intentionality. Intentionality has, thus, been a novel evolutionary invention. But the basic evolutionary process itself has no intentionality, because it has no fitness itself, let alone an X process and a G-loop.

11.5.3 Internalism and externalism

A recurring theme in the study of mind and language is the question how much of meaning is internal ('inside the head') and how much is external, produced by or present in the physical and social environment. Putnam (1975) gave an influential argument for externalism. It uses a Twin Earth and XYZ for water rather than H_2O, but a simple variant of this argument goes like this. Suppose that Alice has a goldfish, Jeremy. When Alice talks about Jeremy, the reference and extension of her words connect to Jeremy. Unknown to Alice, someone replaces Jeremy by another goldfish, Bob. Bob is in all respects similar to Jeremy, so Alice does not notice any difference. When Alice subsequently talks about Jeremy, Alice's mental processes are the same as before. But the actual reference is now not to Jeremy, but to Bob. Again, Alice does not know this. In other words, one can apparently change reference and extension (to Bob now, but to Jeremy before) without changing the mental processes inside Alice's head. According to Putnam, meaning is therefore not exclusively in the head, but must be partly external.

From the present perspective, Putnam's argument falls apart. It is based on the assumption that meaning requires reference and extension as literal connections between internal mental processes (thinking about Jeremy) and external physical facts (the goldfish is either Jeremy or Bob). But the dashed arrows in Fig. 8 do not form literal connections. They are determined by the X of Alice, irrespective of whether they are actually pointing in the direction of Jeremy or in the direction of Bob. Therefore, when Alice talks about Jeremy, the meaning for Alice (i.e., as incorporated in her X and its intentionality) is independent of whether Jeremy has been replaced by Bob or not. Things only change when Alice finds out about it, which would change her X. Similarly, how others understand talk about Jeremy depends on whether they know about the switch or not.

Are all meanings then generated internally, all in one's own head? Not always. The process X and the G-loop are indeed entirely in the head, or perhaps in the head plus body. But the inputs into the X process have, at least partly, an external origin. Because X is not fixed but flexible, it will change depending on external facts (e.g., when finding out that someone has kidnapped Jeremy). The crucial point here is how the timescale of internally generated meaning compares with the timescale of externally driven change in X. Meaning corresponds to neural processes that take some time. Typically, short but meaningful verbal utterances or thoughts take times in the order of seconds. Only the order of magnitude is meant here, essentially the notion that sentences lose their meaning when their time frame is not right (say, a sentence compressed into one millisecond, or expanded to be evenly spread out over one day).

If external influences on X change X much more slowly than the timescale of verbal utterances, one can say that the meaning associated with these verbal utterances is indeed generated in the head. But this is different in a dialogue as the one depicted in Fig. 8c. Verbal dialogues involve quick exchanges of meaning. Levinson (2006) calls this the human Interaction Engine. Gaps between the turn takings of two persons in conversation are shorter (typically 0.2 s, Levinson and Holler 2014, p. 2) than the time it takes to plan a response (at least 0.6 s). This means that each person tracks as well as anticipates the meaning produced by the other. Presumably, this is accomplished by maintaining and updating a model of each other's X to the extent that this is relevant for the conversation. Because such simulated models are part of each person's X, each person's X is quickly modified during a verbal dialogue.

Therefore, in this situation it would be wrong to state that meaning is entirely generated in one's head. It is externally shared with and produced by the X of others as well. Mutually shared meanings then become inextricably entangled in a real-time process. The important point here is that the interaction is in real time and happens between partners that each have their own X process and G-loop. Then meaning is not entirely generated in one's own head, but rather in a collection of heads. If the interaction would not be in real time (e.g., when reading a book previously produced through the X process of its author), or if it would involve just one X process (e.g., when only interacting with physical objects), then meaning is indeed fully generated in a single head. External input is then used, but the essential dynamics of the process, the stochastic G-loop that changes X, is internal. When one reads a book, there is no genuine shared intentionality with the author, because the author has no agency at the time when the book is read. At most there is an 'as if' kind of shared intentionality, with a dialogue (with the author) that is simulated entirely inside the reader's mind.

11.5.4 Representation, symbols and communication

The intentionality that is generated in the brains of higher animals is often not well captured by the term 'representation', if a representation is defined as something that symbolically stands for something else, and 'representing' as the associated process. X-components can be directed towards assumed F-components without being used as if they represent the real thing. They are just used within the X process, without having an explicit symbolic role. However, this is different when some of the factors are internally generated, or processed, in a form suitable for communicating with many others. Symbolic factors shared with a language community then represent what they stand for, again not literally, but in the directed sense of the dashed arrows in Fig. 8. Representation, therefore, requires at least shared intentionality (see also Section 10.5.1). Because symbolic representation enables symbolic manipulation, having a symbolic language enhances the brain's capacity for reasoning. Symbolic processing makes the human X inherently complex. But it also makes it possible to model and estimate the X of others in an effective and efficient way. In addition, it enables more accurate modelling of those parts of F that are primarily concerned with the physical world. The symbolic system that has been established is particularly powerful, because language is essentially open-ended by using hierarchical syntactic structures (such as Chomskyan recursion).

It is possible to view language as a tool for thinking as well as a tool for communicating. Most likely, these functions have coevolved. The sequence that seems to make most sense from an evolutionary point of view is, first, the evolution of phenotypic helping (pathway 4 of Fig. 8a), followed by—and coevolving with—enhanced communication and shared intentionality, and, finally, followed by symbolic communication and symbolic thought. The pivotal role of shared intentionality for the development of language is consistent with the arguments and empirical evidence given by Tomasello et al. (2005) and Tomasello and Carpenter (2007).

Chapter 12

Consciousness[11]

The terms ‘consciousness’ and ‘conscious’ have various meanings. They may refer to the state of being awake (as in ‘regaining consciousness’), the process of gaining access to certain facts as they affect the senses or are retrieved from memory (as in ‘becoming conscious of something’), and the subjective sensation associated with experiencing (e.g., when feeling pain or joy and when undergoing a visual experience). The primary topic of this chapter is the latter meaning, sometimes referred to as phenomenal consciousness (Block 1995). The main purpose here is to explain why consciousness is *felt*. Nevertheless, the explanation given below has implications for the first two meanings as well.

The science of consciousness is making considerable progress by studying the neural correlates of consciousness (Dehaene 2014; Koch et al. 2016). However, these studies primarily aim to identify which particular neural circuits are involved in consciousness, but not how and why exactly such neural mechanisms would produce subjective experience. Theories that explicitly address the latter typically focus on a specific neural, cognitive or informational process, which is then hypothesized to be accompanied by consciousness. There is no shortage of such proposals, more than could be mentioned here. Some representative examples are: a narrative that the brain compiles from competing micro-narratives (Dennett 1991b); a regular, but unspecified physiological process (Searle 2013); broadcasting messages to a widely accessible global workspace within the brain (Baars 1988); neuronal broadcasts to a global neuronal workspace (Dehaene et al. 2003); having representations in the form of trajectories in activity space (Fekete and Edelman 2011); the self perceiving its own emotional state (Damasio 1999); having representations about representations (Lau and Rosenthal 2011); attending to representations (Prinz 2012); perceiving socially observed attention (Graziano and Kastner 2011); recurrent neuronal processing (Lamme and Roelfsema 2000); having a dynamic core of functional neural clusters (Edelman and Tononi 2000); having the capacity to integrate information (Oizumi et al. 2014); and having unified internal sensory maps (Feinberg and Mallatt 2016).

Only some of these theories are closely associated with neurobiological measurements. In particular, the global neuronal workspace theory (Dehaene et al. 2003) assumes that, when perceiving, the brain first engages in a nonconscious processing stage, which can (but need not) lead to a global second stage (a brain-wide ‘ignition’). There is empirical support for consciousness arising in the second stage, which is taken to provide a globally accessible workspace for the results of the first stage, as well as for subsequent processing. Another well-known neural theory (Lamme and Roelfsema 2000) also assumes two subsequent stages. When a visual stimulus is presented, there is first a fast forward sweep of processing, proceeding through the cortex. This first stage is nonconscious. Only a second stage of recurrent processing, when earlier parts of the visual cortex are activated once more by feedback from later parts, is taken to be conscious (with, again, empirical support). Finally, Edelman and Tononi (2000) propose that a loop connecting thalamus and cortex forms a

[11] The first and second parts of this chapter are partly based on van Hateren (2019, 2015b), respectively.

dynamic core of functional neural clusters, varying over time. This core is assumed to integrate and differentiate information in such a way that consciousness results (for an elaborate theory along these lines see Oizumi et al. 2014).

The theory of consciousness that is explained in this chapter takes a somewhat unusual approach, as it first constructs a stochastic causal mechanism that plausibly produces something distinct that may be experienced. Only then does it conjecture which neural circuits in the brain are good candidates for the mechanism's implementation, and how that could be tested. However, discussions of detailed system diagrams and neural circuits are beyond the scope of this book. For such specifics, the reader is referred to van Hateren (2019). Here, I will merely attempt to give an accessible overview of the theory. The first part, Section 12.1, aims to explain consciousness in its primordial form, that is, the form in which it presumably was first established in evolution and in which it presumably still exists today—close to the transition between non-conscious and conscious life forms, as well as close to the beginning of consciousness during the development of any conscious organism. Primordial consciousness is argued to be communicative rather than perceptual. Elaborate forms of consciousness, such as occur in adult humans, are discussed in Section 12.2.

12.1 Primordial consciousness

The explanation below consists of a series of increasingly detailed theoretical elaborations, which may, at first reading, seem unconnected to consciousness. I will therefore briefly state here where the argument leads to and why the elaborations are needed. The argument concludes with the conjecture that consciousness is a transient and distinct cause that is produced when an individual prepares to communicate—externally or internally—a particular class of internal variables. The particular class of communicated internal variables is rather special, because they estimate components of the evolutionary fitness of the individual itself. We have seen in Chapter 10 that such internal variables are related to intentionality (in the sense that such variables are about external things, with the term 'variable' taken here broadly to include complex subprocesses). How to prepare these internal variables for communication is subsequently explained.

As proposed in Chapter 10, intentionality depends on an X-component estimating an F-component. An X-component is part of the X process within an organism that produces x, the internal estimate of the organism's own fitness f. An F-component is part of the fitness process F that produces f. The fact that x estimates f is crucial for the functioning of the G-loop in the mechanism of Fig. 7. The estimative aspect of x can be denoted and quantified by a factor C. The larger C, the more accurately x estimates f. In Chapter 14 it is shown, for a maximally simplified case, that C must be considered as a cause of the increased fitness that will eventually result from the presence of the G-loop. Because the G-loop functions in a statistical way, by utilizing randomness, it affects fitness only slowly and gradually. In order to make clear that fitness does not increase instantly when C is increased, but only slowly, the resulting fitness is called fitness-to-be (or f_+) below. Then C is a cause of fitness-to-be: when C is increased, f_+ increases as well, and when C is decreased, f_+ decreases. This corresponds to the standard way cause-and-effect relationships are understood. If a cause is changed, the effect should subsequently change as well (at least in a statistically reliable way). It should be noted that C has a double role here: first, it denotes

estimation (of f by x, on a fast timescale), and, second, it is a cause (of f_+, on a slow timescale). This is further discussed in Chapter 14.

It can be shown that C is a strongly emergent cause (Chapter 14). Briefly, this means that its causal efficacy is not fully produced by any set of micro-causes (as associated with its material parts). Simply put, this is so because its causal efficacy depends on uncaused randomness in an indispensable way. As a result of being strongly emergent, C exists in a literal sense, as a distinct and autonomous entity. When the X process is parsed into X-components that estimate F-components (see Chapters 4 and 10), one can define factors c that denote how accurately each X-component estimates its corresponding F-component (including the role it has in the F process). Such c-factors contribute to producing fitness-to-be. The causal efficacy of c-factors ultimately depends on how much they contribute to C. In other words, each c-factor is strongly emergent as well, because its causal efficacy is not produced by any set of micro-causes (because its causal efficacy depends on a strongly emergent C). Below, each X-component will be called an 'intentional component'. Each intentional component is a part of an individual's intentionality (Chapter 10) and has autonomous causal efficacy—and is, therefore, a distinct, strongly emergent entity. Further below it will be argued that the intentionality associated with C and the c-factors is not well localized, and would not qualify as an entity that can be equated with consciousness. Remarkably, another strongly emergent cause can arise that is—in contrast to C—well localized, and that does qualify as an entity that can be equated with consciousness. This entity emerges when intentional components are prepared to be communicated, as is explained next.

In its simplest form, fitness quantifies how well an organism can survive and reproduce. However, fitness is more complex in many species, such as when organisms help closely related organisms. This was explained already in Chapter 3, of which the relevant parts are reproduced here, for convenience. If the reproductive success of a helped organism increases as a result of being helped by a related organism, this can indirectly increase the fitness of the helping organism. This is so, because the helping organism shares many genes with the offspring of the helped organism. Thus, the helping organism indirectly promotes disseminating its own properties. If this fitness benefit outweighs the cost of helping, then it is a worthwhile strategy from an evolutionary point of view. Fitness that includes this extension is known as inclusive fitness (Hamilton 1964). Inclusive fitness is still a property of each individual organism. It is to be taken, along with the benefits it can produce, in a statistical, probabilistic sense. Benefits need not always occur, but they are expected, on average. Below, fitness f refers to inclusive fitness, and x is then an estimate of inclusive fitness. For the explanation of primordial consciousness, we will assume here the simple case of two closely related individuals that can mutually benefit from this mechanism, by cooperating with each other.

Cooperation often relies on communication between the cooperating individuals. Cooperative benefits then depend on exchanging useful messages, such as about the environment or about behavioural dispositions. One possibility for such communication is that it is hardwired or otherwise ingrained in the organism's physiology (e.g., through learning). Then communicative behaviour is either fixed or can only be learned within the narrow margins of fixed constraints. Moreover, the information that is transferred is then fixed as well, and it does not involve an X process that drives random, undirected changes of an organism's form. Examples of this type of hardwired communication are quorum

sensing in bacteria and the food-pointing waggle dances of honey bees. Such stereotyped transfer of information is not further discussed here.

We will assume here that all non-stereotyped communication is based on communicating intentional components (i.e., X-components). When such communication is performed in a cooperative setting, the fitness of both sender and receiver is likely to benefit, as is argued now. An intentional component estimates a corresponding F-component, which is a factor in the world that is relevant to the organism (because it participates in producing its fitness). An F-component depends on the state of the world as well as on how this affects the sender. Communicating an estimate of this is likely to enhance fitness for two reasons. First, it may directly increase fitness, in a similar way as a stereotyped transfer of useful information can increase fitness. Second, it is likely to increase the accuracy by which the receiver can estimate, through its own X process, those parts of the sender's X process that are relevant to the cooperation (i.e., parts of X to which the intentional component belongs). As a result, the receiver's x becomes more accurate as an estimate of its f, because parts of F are determined by the cooperation. A more accurate x subsequently and gradually increases the receiver's fitness through the G-loop mechanism. The fitness of the sender is then likely to increase as well, because of the cooperation. This mutual effect is further enhanced when sender and receiver engage in a dialogue, as will typically happen (where 'dialogue' is taken in this section as nonverbal, because of the focus on primordial consciousness). We conclude that X-based communication is likely to enhance fitness. Hence, it is evolvable and assumed here to be present. Examples of cooperative settings that support this type of communication in its most primordial form are mother-infant bonds in mammals and pair bonds in breeding birds. On average, fitness is increased either directly (as for infants) or indirectly (because offspring is supported).

Importantly, an intentional component is an internal factor of the sender. From here on, we will assume that it is internal to the sender's brain. The reason for this assumption is that the theory that is explained below requires quite complex transformations (such as inversion of a complex process). These transformations are at a level of complexity that is presumably only realizable in advanced nervous systems (and not in multicellular organisms without a nervous system, nor through processing within unicellular organisms). A sender communicating one of its intentional components to a receiver requires that the receiver obtains access to a factor that should be, in effect, similar to this intentional component. The question is, then, how the sender can communicate its intentional component in such a way that it will produce a similar factor in the brain of the receiver. Specific hardwired solutions, such as in the case of honey bee dancing, are not viable here, for two reasons. First, the X process is assumed to change in an unpredictable way across time and thus cannot be anticipated in detail. And second, the number of intentional components that are potential candidates for communication may be huge.

A viable way to communicate an intentional component, at least approximately, is the following. It is reasonable to assume that the receiver has an X process that is similar to the X process of the sender. This assumption is reasonable, because the cooperative setting presupposes that the two individuals are similar, that they share similar circumstances and possibly a similar past, and that they share goals to which the communication is instrumental. As can be shown (van Hateren 2019), the sender can then communicate its intentional component by utilizing a process that is approximately the inverse of the process that produced the intentional component in the first place. When the sender communicates

the inverted intentional component (abbreviated to ‘communicable component’ below), the receiver can decode it through its already present X process, as if it were a regular part of the world (abbreviated to ‘regular component’ below). Explaining this in more detail requires equations and system diagrams (see van Hateren 2019), which will be avoided here.

One concern here is whether the sender would be able to produce the communicable component in a form that resembles a regular component, with the means of expression available. We will assume here simple cases where that is possible. For example, if the regular component is a facial expression (such as a smile) observed in another individual, then this could be produced equally well as a communicable component (e.g., reciprocally by mother and infant). The same goes for sounds, touches, and expressions of emotions (such as laughter and crying) that are produced by other individuals. However, if the regular component is part of the physical world, such as a general visual scene, the communicable component may have to be produced by indirect means, using learned conventions that are understood by both sender and receiver. This applies also to other complex communications, for instance when social situations or language are involved. How such conventions may be gradually learned is discussed in Section 12.2.

A further concern is whether an intentional component has an inverse at all. In general, this is not guaranteed. If an exact inverse does not exist, then it is often possible to define an inverse that at least minimizes the distance between the communicable and the regular component. However, the more serious problem is not whether there is a mathematical solution, but rather whether there is a plausible and realistic neurobiological mechanism that could invert an intentional component. This is a major issue, because intentional components are unpredictable and not known in advance. Yet, there is a possible solution, both how it may be realized as a computational mechanism and how it may be realized in the neurobiology of the brain (van Hateren 2019). It depends on a well-known feedback mechanism that can invert transformations in real time. However, for the present purpose it is not necessary to address the details of these proposals. It suffices to assume that a solution exists. We will now argue that the required transformation constitutes a partly autonomous cause, and that this cause is a second distinct entity that is related to, but different from, C.

The fitness in the sender that ultimately arises from the communication will again be denoted by fitness-to-be. Because of the cooperative setting, the positive effect on fitness-to-be depends on how accurately the intentional components of the sender are communicated to the receiver. If the accuracy is low, the effect on fitness-to-be may be negligible, or even negative. If the accuracy is high, the effect on fitness-to-be is expected to be positive, on average. However, the sender can only directly influence its own part of the communication. If the receiver does not pay attention or external circumstances interfere with the communication, then there is little the sender can do directly (other than subsequently try to remedy the situation). But the sender should, in any circumstance, try to optimize communication by translating the intentional component in such a way that it has a good chance of being transferred accurately. The major bottleneck here is the stage where the intentional component is transformed and prepared such that it is ready to be communicated. This is a stage prior to the actual communication; the latter typically requires muscular movements, such as for producing sounds, postures and gestures. We can denote the adequacy of the internal transformation stage (from an intentional component to a component that is ready to be communicated) by T, which can be quantified by a non-negative number. T is small when the transformation is such that the resulting accuracy of

communication can only be low. T is large when the transformation is such that the resulting accuracy of communication is potentially high (though still depending on receiver and circumstances). T has the characteristics of a cause, with the fitness-to-be of the sender as its effect. Increasing T will increase fitness-to-be, on average, whereas decreasing T will decrease fitness-to-be, on average.

Above we saw that the cause C is produced in a complex way, because it depends on how the intentional components (i.e., the X-components) contribute to the process X that produces x. Similarly, T is produced in a complex way, because it is produced by transformed intentional components that the sender might communicate. But T is still a single unit, because the effect of T on fitness-to-be ultimately depends on which role the communicable components have for producing x (which is a single unit that inseparably integrates the intentional components contributing to the X process). Within the sender, intentional components act as partly autonomous causes (by contributing to C and by being an integrated part of X as producing x and C). The communication is successful if similar intentional components are replicated within the X of the receiver, acting similarly as partly autonomous causes. In effect, distinct entities are transferred from sender to receiver—not by a literal physical movement, but by replication. Being capable of transferring such distinct entities is an essential part of T being a cause, because its potential effect depends on this transfer. This implies that T itself is partly autonomous: it is a cause of which the causal efficacy cannot be explained by any set of micro-causes, because intentional components are not micro-causes. As T is a partly autonomous cause, it must be a distinct entity. Hence, organisms that are capable of communicating X in this way have two emergent, distinct entities: not only C, but also T. However, the characteristics of these two entities are quite different with respect to spatial extent and ownership, as is discussed next. Here we will keep focussing on C and T, for the sake of simplicity. A more detailed discussion can be found in van Hateren (2019), which considers the causal role of specific X-components (as denoted here by c-factors) and that of the corresponding communicable components (as denoted here by t-factors).

C denotes and quantifies how well x estimates f. The process that produces x is fully realized within the brain, thus x is localized to the brain (though somewhat diffusely, because x is assumed to be distributed throughout the brain). However, the process that produces f is not localized to the brain, because any part of the world may, in principle, contribute to fitness. For example, if current atmospheric conditions produce a life-threatening drought, then this decreases fitness (because the tendency to survive and reproduce is affected negatively). Thus, atmospheric conditions contribute to the process that produces f. Although the organism may take the drought into account as a relevant factor for compiling x, the required processing is all done within its brain (such as based on sensory data and on memory traces of earlier, similar weather conditions). Because C depends on both x and f, and f is not localized to the brain, C is not localized to the brain either. Moreover, the cause C is not fully owned by the individual, because f is not owned by the individual (even if x is). In other words, C is a distinct entity, but it is neither well localized nor a proper part of the individual. Because C is not well localized, it is unlikely to be clearly sensed. Thus, intentionality and its components are not good candidates to be equated to consciousness and its internal structure.

The situation is quite different for T. T denotes and quantifies how well x is transformed and prepared for communication. Both the process that produces x and the process that

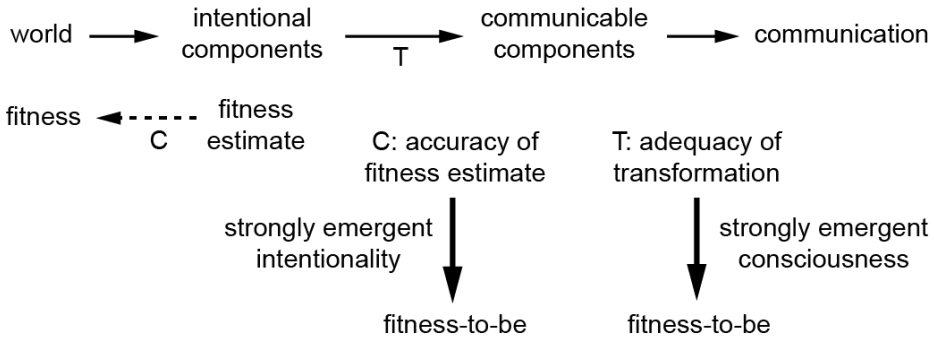

Fig. 9. Summary of the theory of consciousness. Intentional components are part of an internal process within the brain that estimates fitness and its components (which form part of the world). C denotes how well fitness is estimated. It is a cause of fitness-to-be, the fitness that eventually results from a stochastic optimization process. C can be shown to be a strongly emergent cause, making intentionality strongly emergent. Intentional components can be prepared for cooperative communication through a transformation with accuracy T. T can be shown to be a strongly emergent cause of fitness-to-be as well. In contrast to C, it is an entity that is localized to the brain. It is plausibly felt as the feeling of consciousness.

performs the transformation are fully realized within the brain. Therefore, T is fully localized to the brain (though, again, somewhat diffusely). Moreover, the cause T is fully owned by the individual, because both x and the transformed result are present in the individual's brain. In other words, T is a distinct entity that is localized to the brain and that is a proper part of the individual. Normally, spatiotemporally localized entities correspond to material objects (such as a rock) or localized forms of energy or energized material (such as lightning). Here, T is not only spatiotemporally localized, but also distinct and internal to the individual's brain. It is plausible, then, that T is sensed by the individual. It is analogous to an object that would continually change and materialize inside the brain, in a strongly emergent way. The only difference is that T as such is not a form of matter or energy. It is conjectured here that sensing the presence of T is equal to feeling conscious. The content of consciousness at a particular moment then depends on which particular intentional components are being prepared, at that moment, for communication.

Figure 9 summarizes this theory of primordial consciousness. Parts of the world that are relevant for an organism's fitness can be used to make an internal estimate x of this fitness. This estimate is produced via a process X, of which the subprocesses are the components of intentionality. This estimate, with an accuracy quantified by C, subsequently causes a slow increase of fitness (denoted by fitness-to-be) via a stochastic process. C can be shown to be a strongly emergent cause of fitness-to-be, thus C is a distinct entity. Intentional components are subsequently transformed into communicable components, intended for dialogue with kin in a cooperative setting. The adequacy of this transformation for the communication is quantified by T, which is another strongly emergent cause of fitness-to-be. In contrast to C, T is localized to and fully owned by the organism. T is a distinct entity of which the presence is plausibly sensed by the organism as the feeling of being conscious.

12.2 Elaborated forms of consciousness

The previous section described a primordial form of consciousness that arises when an organism uses a transformation (with an accuracy denoted by T) for preparing intentional components to be communicated to a related organism. We assume here that the communicative setting is cooperative on average, which means that preparing to communicate tends to increase inclusive fitness-to-be. Preparing is sufficient when the communication is realized at least part of the time, because fitness is a forward-looking, probabilistic variable. The reasonable likelihood that a cooperative communication will be realized is already sufficient to increase fitness (because it concerns expected evolutionary success). There are several ways in which this primordial form of consciousness can be extended, as is discussed below.

Rather than engaging in a dialogue with a communicative partner, an individual may engage in an internal dialogue. This is produced when communicable components are used as input to the individual's own X process. Because the X process gradually changes over time, intentional and communicable components do so as well. It should be noted that the capacity to perform internal dialogues would be unlikely to evolve if it were exclusively stand-alone (i.e., if it would never involve external dialogue). This is so, because then there would be no benefits obtained from producing communicable components. Any processing that might increase fitness could be performed at the level of X, without going to the trouble of performing a transformation T. Thus, internal dialogue is only useful if it is, at some point at least, combined with a regular dialogue between partners. For example, internal dialogue may prepare intentional components for more adequate future interactions in a regular dialogue. More adequate means here having a better chance of increasing fitness, at least on average. The presence of internal dialogue can explain why not only senders are conscious, but receivers as well, and why non-communicative stimuli (such as a general visual scene) can be consciously perceived. In both cases, the stimuli may induce an internal dialogue, and thus produce consciousness—again, with the ultimate prospect of communicating and engaging in a dialogue with a partner.

Communication could be improved if the sender would not transform the sender's intentional components, but rather a prediction of the intentional components that the receiver is likely to use when receiving the communication (and then the receiver should do something similar in return). This requires that the communicating partners maintain a model of the relevant parts of each other's X, and that part of their communication is used to keep these models up to date. Elaborations along these lines and references to similar ideas in the literature can be found in van Hateren (2015b). Below, a few examples are given of how more complex forms of consciousness might be constructed from the basic one. They should primarily be seen as draft proposals. The order and specific paths presented here may be different and are likely to be more complex in actual organisms.

Figure 10a shows at (1) a basic pair-bond in a symbolic way. The double-headed arrow stands for a dialogue (i.e., reciprocally communicated intentionality) between two subjects, S_1 and S_2. As the prototypical pair-bond we will consider the human mother-infant bond, which has been particularly well studied (e.g., Trevarthen and Aitken 2001; Reddy 2003). In Fig. 10, S_1 stands for the infant and S_2 for the adult. We assume that S_2 has full-blown consciousness, and we will focus here on changes in S_1. The dialogue that is enabled by the basic bond at (1) produces a simple form of consciousness in S_1, firstly, when S_1 acts as a

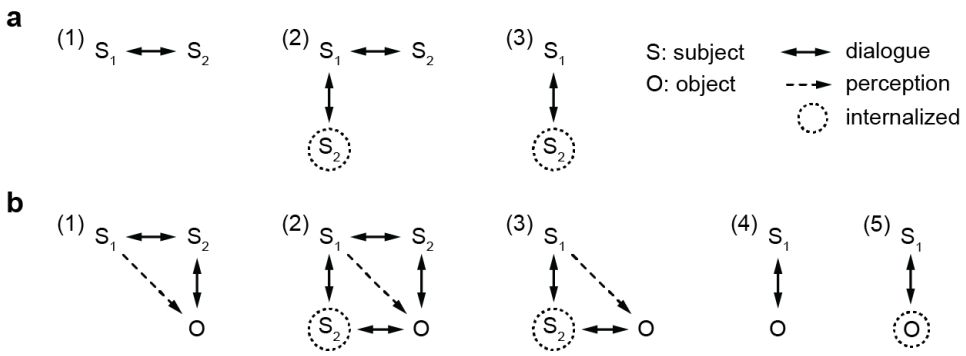

Fig. 10. Origin of various forms of consciousness. (**a**) Basic mother-infant bond (1), with the double arrow denoting a dialogue (continued communication of intentionality, implying subjective experience) between infant S_1 and adult S_2. S_1 gradually (2) internalizes S_2 (circled-S_2), where the internalized dialogue with circled-S_2 produces consciousness even when S_2 is absent (3). (**b**) Consciousness of the natural world arises when objects O are first perceived in interaction with S_2 (1), and S_2 and the interaction are internalized (2), also when S_2 becomes absent (3); this finally results in an internal dialogue within S_1 that involves O (4) or even only its internalized version (5).

sender, and secondly, when an internal dialogue is triggered in S_1 when S_1 receives intentional components from S_2. The subjective experience in S_1 is about S_2, particularly about how S_2 relates to S_1. At (2) the basic bond is used to gradually acquire an internalized version of S_2, symbolized by the dashed circle. This can be acquired through modifying the X process within S_1, because it is likely to help dialogue with S_2. When this internal model of S_2 becomes sufficiently realistic, it can be used to produce an internal dialogue between S_1 and the internalized version of S_2, thus enabling S_1 to be conscious of S_2 without the presence of S_2 (3).

Although the basic pair-bond is used here to derive more complex forms of consciousness below, it is based on an even simpler form of subjective experience. Subjective experience is assumed to occur already whenever intentional components are prepared for communication. A dialogue is not strictly necessary, although it is implicitly anticipated. Even when S_2 is absent, a new-born infant who is crying (or intending to cry) in response to a painful stimulus will have the corresponding subjective experience. This is so, according to the present theory, because the crying is the result of preparing and executing communication of intentionality (even if there happens to be no other person around).

The natural world can become part of subjective experience in the way that is illustrated in Fig. 10b. The basic S_1-S_2 bond is extended with an interaction of S_2 with an object O. This is in the form of an internal dialogue within S_2 that involves O, because S_2, the adult, has already formed an internalized model of O. Initially, S_1 lacks such an internal model suitable for dialogue, and perceives O without subjective experience (1), as an intentional component (dashed arrow). However, the internalized version of S_2 (as formed according to Fig. 10a) can be gradually extended with the interaction with O (as used in internal dialogues by S_2), as shown in (2). Once this has become sufficiently engrained in S_1, S_2 need not be present any more (3). Finally, the role of the internalized version of S_2 can fade away, and S_1 can directly interact with O by using an internal dialogue that involves O (4) or only the internalized version of O when O is absent (5). In either case, the interaction

involves an internal dialogue, where the transformation of intentional components is accompanied by subjective experience.

Consciousness is proposed here to occur whenever intentional components are prepared for communication. In Chapter 10 it is argued that intentional components can become abstract when fitness is generalized to extensive fitness, which includes social and cultural forms of transmission that utilize the X process. This requires that there is a widely shared communal means of communication. Consciousness that is based on abstract intentional components then enables communicating through a symbolic language. It also enables abstract thought through internal dialogues. Moreover, an X process that uses abstract intentional components and consciousness will produce full-blown human agency (recall that agency is produced by the X process and the G-loop, see Chapter 9). Because full-blown agency involves some degree of behavioural freedom, goal-directedness and value, it is equivalent to free will. Since the effects of the G-loop are only established slowly and gradually, actions based on free will require some time to be formed and executed. Thus, free will does not imply the possibility of instant decisions (for further discussion see van Hateren 2015b). Many of the conundrums with respect to free will that one can find in the neuroscience literature are produced by not properly taking the slowness of free will into consideration. Moreover, it is usually not acknowledged that agency and consciousness are likely to depend on strong emergence (see further Chapters 14 and 15).

Chapter 13

The human self

The self is arguably central to human psychology. It is fundamental for understanding how humans perceive themselves as individuals, how they position themselves within their social niche, how they change during development, and how problems with the self affect psychological functioning and well-being. Within social psychology, there is a range of approaches that aim at identifying and explaining aspects of the self and its dynamics (Leary and Tangney 2010). But what the self is and if it really exists as a unitary and continuous entity is not so clear. The purpose of this chapter is to present a theory that explains the human self as an evolved construction, combining biological as well as social mechanisms. It implies that the self is indeed real, unitary and continuous. Moreover, it explains why the self involves agency, goal-directedness and meaning. The theory is primarily intended as a meta-theory. It does not intend to replace existing theories of the self, but rather provides an evolutionary framework for interpreting and connecting more specific theories.

The theory takes an evolutionary perspective, but it does not use the standard approach of the field of evolutionary psychology. In particular, it uses the proposed additions to evolutionary theory that were discussed in Chapters 2 and 3. The first addition explains the agency and intrinsic goal-directedness of living organisms. Agency is defined here as the capacity to initiate and generate behaviour that is meaningful to the organism itself. Throughout this chapter, the terms ‘goal-directedness’ and ‘goal’ are used in a general biological sense, not necessarily associated with human goals and human motivations, and not necessarily perceived consciously. The second addition to evolutionary theory that is used below explains the strong human reliance on social and cultural processes. In addition, the theory depends on the conjecture that consciousness is produced when communication between individuals involves meaning itself (Chapter 12). Taken together, these mechanisms are sufficient to provide a basic theory of the human self.

The present theory mixes elements of social and evolutionary psychology by combining a short-term and long-term perspective. Living organisms and their properties can always be explained at two rather different levels, proximate and ultimate (Mayr 1961). At the proximate level, one studies the mechanisms as they are functioning right now or at least within the lifetime of an individual. For example, one can investigate how physiological mechanisms, psychological motivations and social processes influence how the self functions and develops throughout life. The ultimate level of explanation, the evolutionary one, arises from the fact that life is the result of evolution by natural selection, that is, evolution driven by differential reproductive success. Physiological and social mechanisms that systematically interfere with survival and reproduction will not endure on an evolutionary timescale. The ones that endure are therefore likely to have an evolutionary interpretation, that is, an interpretation that transgresses the lifetime of the individual and its proximate processes. This is essentially the perspective of evolutionary psychology.

The theory presented here falls neither in the proximate nor in the ultimate tradition, but combines the two approaches in a novel way. As will be explained below, it conjectures an internal compound drive that utilizes an internally produced estimate of an individual’s own

evolutionary fitness. This drive continuously functions within the individual, thus acting as a proximate mechanism relevant within the individual's lifetime. But at the same time, this drive has an evolutionary role as a proxy for the true evolutionary fitness. Therefore, it also acts as a mechanism that has a direct ultimate interpretation, as relevant for the timescale of evolution.

The basic elements of the theory are summarized in Section 13.1, and the core of the theory of the self is explained in Section 13.2. In the first sections of the Discussion, this is connected with existing concepts and theories, in particular with self-esteem, sociometer theory, self-determination theory, terror management theory and evolutionary psychology. Finally, the unity, continuity and stability of the self are discussed, and how the theory can be tested empirically.

13.1 Agency, goal-directedness, consciousness, self-awareness and extensive fitness

An important aspect of the self is that it is a prime source of agency, which is taken here as an individual's capacity to initiate and construct novel behaviour that is significant for that individual. If we assume that all living organisms have some version of the mechanism of Fig. 7 (Chapter 9), then they must all have some form of agency. Chapter 9 discusses how a minimal form of agency can be produced. It assumes that each organism has an internal process X that produces a value x as an estimate of the organism's fitness f (which is produced by an external process F). Both X and x are assumed to be distributed in a similar way as processes and values are distributed in a neural network. How X is realized depends on the species; it ranges from purely intracellular circuits (in unicellular organisms) to neuronal networks (such as in humans). The fitness estimate x is assumed to drive the variability of structural changes within the organism, thus affecting behavioural dispositions. A sequence of behaviours and behavioural dispositions then acquires agency because of the specific way the mechanism of Fig. 7 combines determinacy and randomness.

Chapter 9 also shows that high x must be regarded as the overall intrinsic goal of the organism (see also Chapter 15). In practice, it consists of a large set of sub-goals that are all expected to contribute to the overall goal. Such contributions should typically contribute to fitness f as well, but they are not guaranteed to do so, because x can only approximate f. For practical reasons, the goal-directedness one can observe in biological organisms is normally studied through the sub-goals and how they are related (e.g., Carver and Scheier 1982, 2002). The theory implies that the structure formed by all sub-goals together corresponds to the structure of the process X, including its dynamics. Sub-goals can only be fully understood, then, from their role in constituting the structure of X.

In a minimal sense, the process X as discussed above can be viewed, for any living organism, as at least a proto-self. The structure of X defines what the organism implicitly takes as important for its own survival and reproduction. The self-related identity of the organism can be equated to the form of X, that is, to which internal and external factors the organism takes into account for X, and how. It is important to stress that only X—and not the fitness process F—can produce a self, because X is an essential part of the agency-producing G-loop of Fig. 7. However, the concepts of self and identity as used in the context of psychology require not only agency and goal-directedness, but also subjective awareness

and awareness of the self. Organisms such as bacteria, worms and insects may have minimal forms of agency and goal-directedness according to the theory presented here. But it seems reasonable to assume that they do not have subjective awareness and a self in any psychological sense. Below I will briefly summarize how the theory is thought to produce subjective awareness.

Meaning, goals and values are not present explicitly within most species, but are merely contained implicitly in the structure and dynamics of an organism's X process. But this must change when organisms engage in reciprocal communication of meaning, that is, in a (usually nonverbal) dialogue. Then internal goals and values have to be transformed into regular physical signals—such as touch, posture, gestures and sounds—that can be interpreted by the partner in the dialogue. In Chapter 12, it is conjectured that consciousness is produced by a critical step required for communicating these aspects of X. Merely preparing for dialogue is then already sufficient to produce consciousness. Why this is plausible has been explained elsewhere (Chapter 12 and van Hateren 2019), but it boils down to the fact that preparing meaning for communication produces a strongly emergent cause that has the right kind of properties for consciousness. A strongly emergent cause can be equated to a distinct, partly autonomous entity (because it is not completely caused by its constituents, in this particular case because of the indispensable role randomness plays in the proposed mechanism). Because this distinct entity is owned by and localized to the brain, it is plausible that it can be sensed as the feeling of consciousness. Presumably, the kind of (nonverbal) dialogue that produces consciousness first evolved in organisms with advanced nervous systems and social lifestyles, such as mammals and birds. It may have its origins in dialogue with particularly strong fitness consequences, such as dialogue that occurs within mother-infant and pair bonds.

In its simplest form, awareness then occurs when two subjects communicate in such a dialogue. This produces a subjective sense of awareness, roughly corresponding to the minimal or core self as proposed by Zahavi (2005, p. 106). However, subjective awareness is not the same as awareness of others, of objects and of the self. Figure 11 sketches how more complex forms of awareness may arise during development. This is analogous to what was previously suggested by Mead (1934) and Vygotsky (1986), and it is related to work in developmental psychology (Tomasello 1993; Trevarthen and Aitken 2001). It is shown in Fig. 11 specifically for the self, but similar schemas can be made for awareness of others, of objects (see Fig. 10 in Chapter 12) and of groups of others. The diagram is not meant as a detailed theoretical proposal, but merely as a minimalist tool for explaining the general

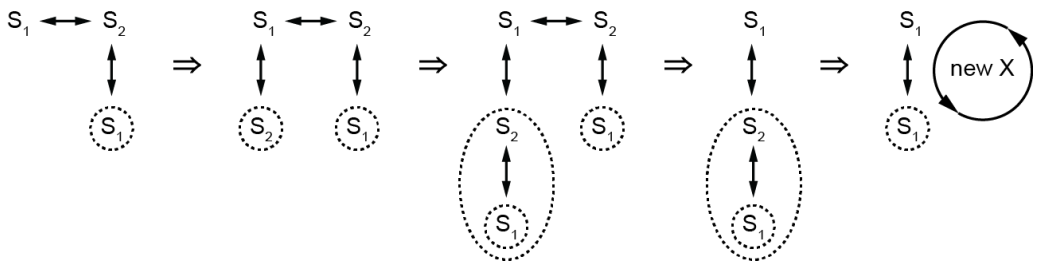

Fig. 11. Development of the self as a perceived object. Subject S_1, while communicating with subject S_2, gradually develops a continually updated model of the self (final diagram). Dialogues are symbolized by double-headed arrows, models are encircled by dashed lines. See the main text for further explanation.

ideas needed here. S_1 stands initially for an infant and S_2 for an adult. The nonverbal dialogue between S_1 and S_2 then provides both individuals with subjective awareness. But the awareness of the adult S_2 is considerably more complex, because it includes an implicit model of S_1. Such an implicit model may take the form of a simulation or it may have a symbolic form. Internalized models are denoted in the figure by dashed circles. When S_1 and S_2 interact over an extended time, S_1 will gradually develop an internal model of S_2 (second diagram). A vertical double-headed arrow connects S_1 and this model. This arrow indicates that S_1 can engage in a simulated dialogue with the internal model of S_2. The engagement produces awareness of a purely internal nature.

Initially, S_1's model of S_2 will be simple, but gradually S_1 will learn that S_2 communicates partly based on an internal model of S_1. Subsequently, S_1's model of S_2 is gradually extended accordingly (third diagram of Fig. 11). Optionally, the actual dialogue with S_2 can then be replaced by a purely internal dialogue (fourth diagram). The model of S_2 contains a model of S_1 themselves. In a final stage (last diagram), the modelled S_2 is not needed any more. Then S_1 can directly engage with their own modelled self and be aware of their own self. The final stage thus represents a primary form of awareness of the self.

Because the model can contribute to the X of S_1, the dialogue between S_1 and the model of S_1 is in fact a dynamic cycle. The changing model affects X, which subsequently may induce further changes in the model, and so forth. The cycle drawn at the far right emphasizes the dynamic nature of this interaction. The continual dialogue associated with this cycle is conjectured to be accompanied by subjective experience, as discussed above. The explanation applies to any kind of internal model, including nonverbal ones with only non-symbolic simulation. In human development, the diagram of Fig. 11 is presumably executed several times, probably in overlapping, continuous, and increasingly complex ways. This then results in increasingly sophisticated internal models of the self (Reddy 2003), as a form of Theory of Mind. Below, it is explained how symbolic (language-based) and social (communal) layers (Tomasello 1993; Tomasello and Carpenter 2007) can be added. A more detailed proposal of how abstraction and symbols can arise can be found in Section 10.5.

The above theory of agency and awareness may be adequate for understanding such phenomena in nearly all species where they occur. X and x then estimate F and f in the form of the standard evolutionary fitness, that is, inclusive fitness (see Fig. 12, reproduced here

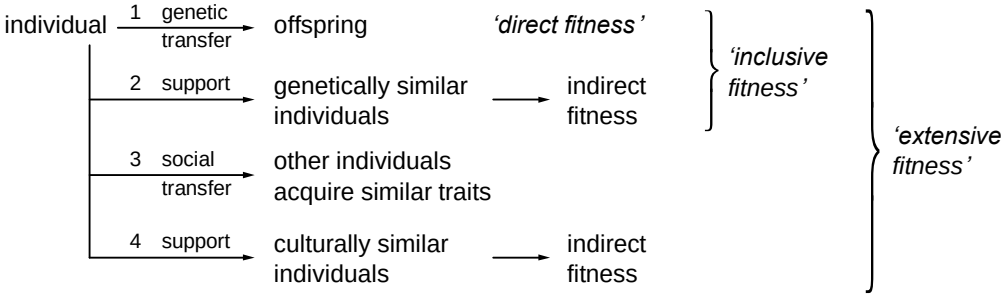

Fig. 12. Various forms of fitness. Direct fitness (pathway 1) is the expected rate of producing offspring. Inclusive fitness combines direct fitness with indirect fitness (pathway 2) produced by helping genetically related individuals. Extensive fitness combines inclusive fitness with fitness produced socially, either directly by transferring similarity (pathway 3) or indirectly by helping others that are already similar (pathway 4).

from Fig. 4 of Chapter 3, for convenience). However, in particular for humans, the standard fitness requires elaboration in order to include social factors that are not included in inclusive fitness. The way this is done here is different from previous accounts of social and cultural influences on fitness, because it extends fitness itself rather than adding factors that enhance inclusive fitness. This extension is necessary when there is an X process. The extended form of fitness, dubbed extensive fitness (see Chapters 3 and 10, and van Hateren 2015c), is still a form of biological fitness. It is part of the fitness process F and it is taken into account by the fitness estimator X. It belongs to individual organisms, that is, it concerns organismal fitness, not the fitness of cultural traits in a population of organisms, nor how easily cultural memes can spread. The extension can arise when organisms are capable of helping their con-specifics on the basis on phenotypic similarity rather than genetic relatedness. The extended form of fitness requires that individuals can flexibly change their phenotype throughout their lifetime, such as through learning and imitating. Within a population, sub-groups with similar phenotypes can then readily become larger than sub-groups with similar genes. Individual benefits from helping typically increase with group size, because that increases the probability of helping. Helping based on phenotype may then outperform helping based on genetic relatedness, because of differences in sub-group sizes (Appendix A, which summarizes computational simulations in van Hateren 2015c).

Increasing one's extensive fitness by supporting similar others can enable those others to become supportive in return, and thereby further increase their own extensive fitness. Formation of in-groups strongly amplifies this effect. The mechanism is therefore analogous to direct and indirect forms of reciprocity that have been proposed as explanations for human cooperation. But in contrast to existing proposals, the current theory does not regard such forms of reciprocity as necessarily produced by adaptations (selected traits) serving inclusive fitness. Rather, it reinterprets them as partly produced by a specialized, human form of fitness that goes beyond inclusive fitness.

Fitness itself is not an adaptation, because it is a core property of the evolutionary process: it is not facultative (Bell 2008, pp. 5–6). The reinterpretation is therefore not a trivial one. It is essential if one wants to understand the social aspects of agency, goals and meaning. These factors depend, for humans, not merely on self-estimated inclusive fitness (when x estimates the f of pathways 1 and 2 in Fig. 12), but on self-estimated extensive fitness (when x estimates the f of all four pathways together). The G-loop of Fig. 7 functions as before, thus the process X is still producing agency, goal-directedness and meaning. These factors then automatically acquire social, non-genetic aspects through X. Importantly, the extended form of X is necessary for understanding the origin and nature of the human self (see below).

Helping based on phenotypic similarity makes the structure of X and F considerably more complex. It requires phenotypic flexibility, with as side-effect that phenotypes can be faked easily. In other words, reliably recognizing cheating and free-loading becomes important. The internal structure of X must reflect how the contributions of the four basic pathways are balanced within the individual. Although these contributions may be aligned (as mentioned above), they can also produce internal conflicts. This is well known for inclusive fitness (pathways 1 and 2), for example in parent-offspring and sibling-sibling conflicts (e.g., Schlomer et al. 2011). But the two additional pathways multiply the possibilities for tension within the structure of X. For example, genetic alliances may

conflict with phenotypic alliances, different group alliances may conflict with each other, and direct transfer of one's traits (pathway 3) may conflict with supporting similar others (pathway 4).

Apart from tension within the structure of the X of specific individuals, the properties of X can also produce tension and conflict between individuals and between groups. In general, pathways 1 and 3 imply competition between individuals. For example, competition arises when there is mate selection and when raising children depends on shared resources that may be scarce. Moreover, pathway 3 typically requires competition, because a population has only a limited capacity to absorb socially transferred traits. Thus, individuals must compete with other individuals. In contrast, pathways 2 and 4 imply cooperation between the individuals of the relevant groups. However, these pathways may subsequently lead to conflict between different clans or in-groups, again because of limited resources and a limited capacity for cultural absorbance. The balance between prosocial and antisocial behaviour then depends, first, on the details of how an individual engages pathways 1 to 4, second, on the specific in-groups to which the individual belongs, and third, on how these in-groups overlap or have conflicting interests.

13.2 The human self

We are now in a position to formulate the key thesis of this chapter. This thesis is that the human self is defined by the structure of the process X, because it is X that produces human agency, subjective awareness, and sense of meaning and purpose. Because the process X is, formally, an estimator that produces an estimate x of f, the theory may be called the estimator theory of the self. As discussed above, the process X is an evolved, layered phenomenon, which is summarized in Fig. 13a. It can be conceptualized along two major dimensions. The first dimension is the depth of awareness of the self. This can be interpreted as a dimension of qualitative experience. It ranges from the non-experienced proto-self that only requires agency, through the aware self as presumably present in many species of higher animals, to the symbolic self that is present in humans. The psychological concept of the human self then corresponds to the aware and symbolic self, with the proto-self

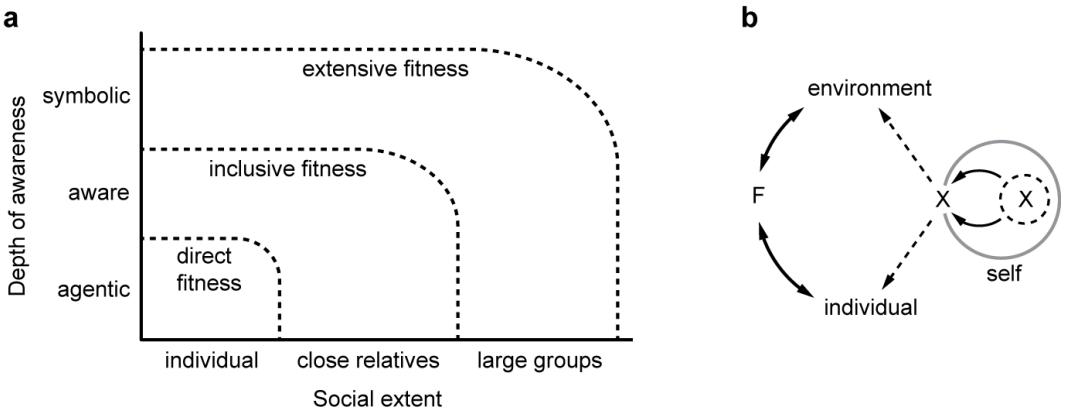

Fig. 13. The estimator theory of the self. (**a**) The self has evolved along two dimensions, depth of awareness and social extent. (**b**) The self is aware of itself in a continual cycle involving the process X and its model. It incorporates parts of individual and environment, and it depends on and modifies X. X is an estimator that produces x as an estimate of f.

merely providing the basis for agency. The second dimension, the social extent of the self, is a scale for the social aspects of fitness. It ranges from the individual (as in direct fitness), through genetically related groups of individuals (as added in inclusive fitness), to large groups of individuals with socially formed traits (as added in extensive fitness). Human extensive fitness is conjectured to be a weighted amalgam of the various aspects of fitness along these dimensions. The two dimensions are usually highly correlated, because the aware and symbolic self co-develop with more complex interactions with the social environment (e.g., Reddy 2003). But the correspondence may be lower in specific individuals, for example when the symbolic self is more strongly developed than the social self.

There are two complications that need to be discussed. First, the individual can only be aware of the self, as an object of awareness, to the extent that X contains an intentional component—a subprocess of X—that points to X. This model of X is depicted in Fig. 13b by a dashed circle around an X. It is shown as separate from X for the sake of clarity, but is in fact an internal part of X. The model continually changes when its target X changes, and itself continually evaluates and modifies X in return. The self is thus changing dynamically. This is shown in Fig. 13b as a cycle represented by the arrows connecting X and its internal model. It is equivalent to the cycle at the far right in Fig. 11. The modelled self is part of X proper. The border between the self as producer of subjective awareness (X), the self as intentional component (X within a dashed circle) and the self as object of awareness (the intentional target X towards which the encircled X points) is therefore fluid. X, including its internal model, forms the self, as indicated by the grey circle labelled ‘self’.

The second complication is that the self may seem to incorporate parts of the social and physical environment, as well as parts of the individual’s body. The self as perceived object is the process that estimates one’s fitness. This process (X) includes aspects of the physical and social environment that are judged to be vital for one’s goals. For example, individuals living in a country may regard its territory and culture as vital for their X. Components of X that are directed towards the territory and culture will then be perceived as part of the self. Similarly, political and religious ideas may become a significant part of X. The perception of the individual’s own body parts will often be incorporated into the self as well. At first sight, one might think that all parts of the individual automatically belong to the self, but that is not so. For example, even if one’s liver is part of one’s body and is essential for survival (i.e., essential for the fitness process F), it is usually not part of the aware and symbolic parts of one’s X. But some other parts of an individual may typically be involved in X. For example, one’s bodily features and general appearance are usually taken as important parts of the self. It should be noted that parts of environment and body become parts of the self not in a literal sense, but rather by how parts of the X process refer to them (symbolized by the dashed arrows in Fig. 13b).

13.3 Interpretation of X in psychological terms

X is assumed to be an internal, neurophysiological process within the organism. It is interesting to see how the human X can be interpreted in psychological terms. The parts of X that are situated in the lower-left corner of Fig. 13a presumably affect human behaviour through nonconscious processes. Such behaviours are still not automatic, because X inherently produces agency (through the G-loop of Fig. 7). In psychological terms, such

nonconscious agency can be viewed as behaviour influenced by drives or needs. The dashed lines in Fig. 13a are only soft demarcations, because also agentic drives may be influenced by kin and social groups. Drives and needs have been established by previous evolution, and usually contribute to X in a way that makes the resulting x align well with fitness f. The result is agency that increases fitness, on average. For example, basic nutritional and sexual drives belong in this category. However, x only estimates f, and agency is non-deterministic. Thus, there is no guarantee that drives and needs work out well in specific cases.

The aware parts of X (Fig. 13a) lead to agency associated with consciousness, that is, they lead to volition. Motives influencing behaviour contribute to these parts of X. Motives are typically flexible and partly formed through learning. A motive adjusts the structure of X in such a way that x is decreased when perceived circumstances indicate that there is a decreased likelihood that the implicit goals associated with the motive could be reached. Similarly, x is increased when goals appear more attainable. The result is agency that, on average, enhances the chances of attaining goals. Whether specific motives and their associated goals are indeed helpful to increase the actual fitness f as well is a different matter. Formation of motives and goals must be subject to evolved and learned high-level constraints that increase the likelihood that motives turn out to be beneficial. However, such high-level constraints have to balance two conflicting demands. On the one hand, too tight constraints will limit agency, and thus decrease the likelihood of finding and producing beneficial motives. On the other hand, too loose constraints will increase the risk that motives take an adverse form.

The symbolic parts of X lead to agency that is strongly dependent on social, communicated factors (see Chapters 10, 11 and 12) and on internal reasoning. Explicitly formulated goals pursued by individuals and groups, and agreed-upon policies by organizations and societies, are all represented in this part of X. Such goals and policies are the result of inter-individual communication (horizontally in Fig. 13a) combined with intra-individual drives, motives and goals (vertically in Fig. 13a).

13.4 Discussion

The general model of the human self as presented here can be related to specific existing approaches in the psychological literature on the self. Baumeister (2010) distinguishes three basic roots of selfhood: self-knowledge, the interpersonal self, and the self as agent. These roots are also embodied in the present theory. Self-knowledge corresponds to the interaction of X and its model (Fig. 13b). The model (X within dashed lines) represents self-knowledge, through its structure and memory, which is used and modified by the agency of X and the G-loop (Fig. 7). The interpersonal self corresponds to the interpersonal aspects of extensive fitness (particularly pathways 3 and 4 of Fig. 12), where X and the social environment interact. Finally, the self as agent corresponds to agency generated by X and the G-loop.

Leary and Tangney (2012) argue that the three most appropriate uses of self as a psychological term are the attentional self, the cognitive self and the executive self. The attentional self corresponds to the awareness that X produces of its model (Fig. 11). The cognitive self corresponds to the modelled X as interacting with X (Fig. 13b), in particular when the interaction involves symbolic awareness (Fig. 13a). The executive self corresponds to how X and its model produce agency.

Several other well-known approaches to the self can be similarly related to particular aspects of the current theory (which is abbreviated below as ETS, the Estimator Theory of the Self). I will discuss here in some detail how ETS relates to self-esteem, self-determination theory, terror management theory and evolutionary psychology. Subsequently, I will discuss how explanations based on X differ from those based on F, and how ETS understands the self with respect to unity, continuity and stability. As stated in the first paragraph of this chapter, ETS is intended as a meta-theory, and not as a replacement of more specific and detailed theories. It provides the infrastructure for connecting theories, but highly detailed predictions should not be expected from the broad, evolutionary considerations on which ETS is based. Nevertheless, empirical testing should be possible, as is discussed in Section 13.4.8 below.

13.4.1 Self-esteem

Explicit self-esteem presumably corresponds to the subject's perception of that part of X that estimates the subject's role in producing extensive fitness—in particular how much the subject contributes to that or may come to contribute to that. It is, then, an evaluation of one's self and self-worth (Heppner and Kernis 2011). Not all parts of X are produced with a direct, active role for the subject. For example, external factors, such as disease and war, can strongly affect X. They thereby affect general psychological well-being, but not self-esteem, at least not directly. However, in reasonably favourable circumstances, X is likely to be strongly determined by agency and by how the individual is functioning within the social environment. Because the process X has a complex, heterogeneous structure, self-esteem is heterogeneous as well (Heppner and Kernis 2011). If the dynamics of X is unstable, that is, easily swung by minor variations in input, self-esteem is fragile. Whereas explicit self-esteem concerns the aware and symbolic parts of X (Fig. 13a), implicit self-esteem (Heppner and Kernis 2011) corresponds to nonconscious, agentic parts of X.

The social aspects of self-esteem are stressed by sociometer theory (Leary 1999). Self-esteem is then regarded as a gauge that indicates how well the individual is socially accepted. This overlaps with the concept of self-esteem proposed here, although it is not completely identical. Social acceptance is a prerequisite for producing extensive fitness through pathways 3 and 4 (Fig. 12). It is difficult to transfer one's traits, for example by acting as a role model, if one is not socially accepted. Moreover, there would then be few opportunities to support others (pathway 4). Conversely, being supported is also less likely, because others will not perceive the individual as similar. Being supported would improve the individual's general circumstances, and thus would indirectly make it easier for the individual to acquire extensive fitness (through any pathway).

However, self-esteem as proposed here can also derive from individual goals rather than from social acceptance, if the individual regards such goals as important for obtaining a high value of x. Such a high value may lie in an envisioned future; for example, a future with hoped-for success as an artist, an athlete, or an entrepreneur. Current social acceptance may be low, but the individual may still assess their current agentic role as favourable, and therefore have high self-esteem. Nevertheless, X always integrates individual and social factors (as it is unitary, see below). Recent studies across a range of different cultures (Scalas et al. 2014; Becker et al. 2014) indicate that self-esteem is primarily determined by socially shared values rather than by individually held ones. Socially shared values are

indeed expected to strongly affect X, because many fitness pathways (Fig. 12) have a social component. Socially held values are therefore vitally important for F (and thus for X). If the individual does not endorse the values of the group, this will make pathway 4 more difficult (because it decreases perceived similarity, either way). On the other hand, divergent personal values can, potentially, make competing based on pathway 3 more effective, thereby boosting self-esteem.

Striving for high self-esteem need not be beneficial to an individual, because it can produce detrimental side-effects when it becomes obsessive and socially negligent (e.g., Crocker and Park 2004). From the perspective of ETS, it is important to note that high x does not necessarily reflect high f. X and x could be a poor, distorted estimate of F and f. X is about an unknown and uncertain F, including how that might change as a result of the agency of the individual and of others. Individuals might therefore form an X, and goals associated with its structure, that are, objectively, not in the best interest of themselves and of those involved in the prosocial parts of their X. High-level constraints—evolved, learned, or provided by social institutions—should help to avoid this problem, at least on average. But there is no guarantee that they can do so in specific cases.

The concept of self-esteem as proposed here shares with sociometer theory the notion that high self-esteem is not a direct goal. It is merely a means to induce behavioural change that leads to the actual goal. The actual goal is social acceptance in sociometer theory and high x in ETS. An example of a theory that views high self-esteem as a goal in itself is terror management theory (see also below). A drive towards high self-esteem is then primarily seen as a way to create an emotional buffer against anxiety that is produced by being aware of one's mortality. Moreover, it is regarded as a cultural construction (Pyszczynski et al. 2004, pp. 436–437). For ETS, striving for high self-esteem may be culturally shaped, but it has firm biological roots. High self-esteem is correlated with high x, high x is correlated with high f, and high f is necessary for sustaining life.

13.4.2 Relationship to self-determination theory

Self-determination theory (SDT) is particularly concerned with the various intrinsic and extrinsic factors that motivate people, and how that affects their functioning and well-being (Ryan and Deci 2000; Deci and Ryan 2000). It poses that the self primarily depends on three intrinsic, innate factors: the needs for autonomy, relatedness and competence. In a similar vein, Swann and Bosson (2010) distil three commonly posed motives from an overview of the literature. These three motives are the ones for agency, communion and coherence.

The three factors identified by these theories are consistent with major components of ETS. The need for autonomy is implied by the agency of the self combined with the goal of obtaining high x. In the G-loop of Fig. 7, it is the freedom to act—as called agency here and autonomy in SDT—that enables enhancing x and the subsequently resulting fitness. This freedom is therefore intrinsically desirable. When agency is strongly constrained by external factors (controlled behaviour in terms of SDT), then the possibilities to enhance X are constrained as well. Such constraints are typically less optimal and thus less desirable than more freedom (autonomous behaviour in terms of SDT).

The need for relatedness to others (Baumeister and Leary 1995), or a motive for communion, is consistent with the other-directed (interpersonal, communal and societal) aspects of extensive fitness (Figs. 12 and 13a). In order to obtain high f and x, people need

to relate to others. Prosociality in the form of pathways 2 and 4 (Fig. 12) is an intrinsic part of F and X. Finally, the need for competence can be understood from the fact that behaviour driven by X will typically only result in actual high f when it is carried out competently. Similarly, X is only likely to be successful if it is coherent. Because the fitness process F is inherently coherent (and f unitary and continuous, see below), an incoherent X is likely to approximate F and f less than adequate. It would then be maladaptive. Incompetently performed behaviour based on a coherent X would be maladaptive as well.

More generally, ETS is consistent with the basic notion of SDT that the self does not primarily strive for equilibrium. Rather, the self actively seeks out change in order to explore new possibilities in its interactions with the physical and social environment. This is very much in the spirit of biological evolution, where organisms that have more (cryptic) variability are better prepared for adapting to new circumstances (Masel and Trotter 2010). Such preparedness is indeed enhanced by the G-loop of Fig. 7.

13.4.3 Relationship to terror management theory

Terror management theory (TMT) assumes that the capacity of human beings to understand the inevitability of their own future death induces an existential anxiety. This existential anxiety is then a major factor that drives the self (Solomon et al. 2004; Landau et al. 2007; Pyszczynski et al. 2012). In particular, existential anxiety has led to the construction of cultural systems that give value and meaning to life. These systems thus enable individuals to transcend (or deny) their own death, by affiliating with such systems. There is empirical support for the theory from experiments that increase an individual's awareness of their own death. Increasing mortality salience will generally induce them to defend their cultural worldviews more strongly. It also induces them to invest more strongly in self-esteem, which can then act as an emotional buffer against anxiety.

The empirical results of TMT are largely consistent with ETS. However, TMT assumes different primary causes of the self and of meaning than ETS. In ETS, the main factor driving the self is the need to get or keep a high x. Because the value of x is an internal estimate of the rate by which an individual induces others to become similar, x goes to zero (along with f) when the individual dies. Avoiding death is therefore an absolute condition for maintaining a positive x. However, avoiding death is not the primary goal. It is a derived goal that supports the primary goal of high extensive fitness. For example, an individual can risk or choose death when the corresponding action is expected to let x peak strongly. Such a peak can occur through one or more of the pathways depicted in Fig. 12, for example when protecting one's children or defending a community. If the peak is high enough, it can accumulate more extensive fitness than would have resulted from staying safe, or from staying alive with low to moderate x in the remaining lifespan.

Nevertheless, increasing mortality salience is clearly a particularly powerful way to let individuals feel that their x may be too low. A way to compensate for a possibly low x is to invest more in some of the pathways of Fig. 12, such as by giving more weight to the views of the in-group or by investing more in self-esteem. A belief in immortality, such as life after death, can be an effective way to increase x as well, even if such a belief does not correspond to the reality of f. The reason is that x is merely an estimate of f. It needs to be reasonably close—but not perfect or optimal—if it is to be adaptive.

13.4.4 Relationship to evolutionary psychology

ETS relies on evolutionary arguments and stresses the importance of evolutionary fitness for human psychology, in particular through the internalized estimate of fitness. Fitness is evaluated continuously, which makes the approach compatible with developmental evolutionary psychology (Lickliter and Honeycutt 2013) as well as with ecological, Gibsonian approaches to psychology (Heft 2013). By emphasizing the evolutionary context, ETS is clearly related to evolutionary psychology (EP, Tooby and Cosmides 1992; Maner and Kenrick 2010). But there are important differences with conventional EP that need to be understood. In EP, human psychological mechanisms are seen as adaptations. Such adaptations are typically assumed to have originated in response to challenges posed by environments in the human past, such as those in the Pleistocene. Tooby and Cosmides (1992, p. 54) emphasize that individuals are not fitness-maximisers in a teleological (goal-directed) sense, but rather adaptation-executors.

ETS does not conflict with the notion that much of human behaviour is produced by relying, more or less automatically, on evolved adaptations. But it claims that those parts that involve agency must rely on the G-loop of Fig. 7 and its elaborations (see also van Hateren 2015b). The G-loop does not produce behaviour that consists of executed, ready-to-go adaptations, but creates novel behaviour. Such novel behaviour will usually rely partly on existing behavioural adaptations, but these are used then as mere components. Novel behaviour is still evolutionarily constrained by the requirements that x estimates f and that both are sufficiently high. However, such constraints must be rather abstract, high-level ones. They should protect the advantages of using a highly flexible X, because that can potentially increase f, as in a self-fulfilling prophecy. But the constraints should also protect x from becoming too different from f, or from producing maladaptive forms of F. There is no direct fitness-maximization—that would be impossible, because the fitness consequences of novel behaviour are not known in advance. But there is an indirect drive towards increasing fitness, albeit only in a statistical way. Making the fitness estimate high is a genuine, innate goal, thus the mechanism is in fact teleological (i.e., goal-directed), in a weak sense. The teleology is weak, because it only exists within organisms and does not depend on an external teleology. Moreover, the G-loop is a fairly weak addition to the more basic action of fitness f. Nevertheless, this weak addition seems to be responsible for human's most precious assets, such as experiencing consciousness, agency and purpose.

A major addition to conventional EP is indeed that individuals have agency. They have more behavioural freedom than mere adaptation-executors would have. Human agency is further enhanced by symbolic reasoning (Deacon 1997) and by society and culture. The latter integrate and accumulate the agency of others and thereby usually empower an individual's agency in return. The status of X as an estimating process provides considerable freedom to individuals—and indirectly to society—to estimate F in different ways. Different forms of X subsequently affect F in a continual cycle (Fig. 13b). ETS therefore partly complies with the Standard Social Science Model that is criticized by Tooby and Cosmides (1992). It thus combines evolutionary constraints as stressed by EP and societal constraints as stressed by the social sciences. It does so without introducing biological or social determinism, because it incorporates human agency and awareness as essential components.

13.4.5 Differences between explanations based on X and those based on F

The present theory explains the self and its motives as based on X rather than on F. Evolutionary explanations of the self have been given before (e.g., Sedikides and Skowronski 1997), based only on the conventional F. Because X and x have evolved to estimate F and f, explanations based on either X or F are often fairly similar. If a particular behaviour increases f, it is likely to increase x, and vice versa. However, the form of F depends in important ways on the presence of X. Without X, the social pathways 3 and 4 (Fig. 12) would not exist as forms of fitness (van Hateren 2015c). If only conventional F existed, all fitness effects of human sociality must be explained through their effects on inclusive fitness (pathways 1 and 2). Explaining prosocial interactions with non-relatives, such as helping strangers, is then far from straightforward. In contrast, pathway 4 readily explains why it is often adaptive to help individuals that are judged to be actually or potentially similar. Such judgment is mediated by X, which can flexibly define similarity to varying degrees of inclusiveness (e.g., based on clan, region, nationality, or on just being human, or even on just being a form of life or a part of nature).

A primary problem for explanations based purely on F is that they assume that adaptive behaviour is pre-specified, or at least produced by a pre-specified system with determinate rules formed by previous evolution. Evaluation of evolutionary fitness then has necessarily happened completely in the past. That means that people must rely on tried-and-tested solutions when they encounter new challenges during their lifetime. In contrast, X offers more freedom. New behaviour, generated partly through trial-and-error, is evaluated, in real time, through X. Behaviour is then changed to varying degrees, depending on that evaluation. Although the structure of X itself must have formed partly in previous evolution, it includes high-level constraints that allow it to adapt through agentic and cultural influences (through the G-loop and its elaborations).

Explanations based on X are particularly insightful when individual, social, or cultural behaviour is clearly maladaptive when judged by conventional F. Conventional F implies that maladaptive behaviours must be understood as (unintended) errors and misfirings, perhaps as a result of evolutionary lag. In contrast, such behaviours can often be interpreted as behaviours that are intended to be adaptive—and implicitly judged to be adaptive—by the individuals and social groups displaying them. The reason is that x may differ from f, or at least the f as inferred by independent observers. In some obvious cases, the latter f may be known to be accurate. Discrepancies between x and such accurate f can explain, for example, individual and group behaviour arising from mental delusions and delusional ideologies. Then one can conclude that the X and x of the individual or group is objectively wrong, that is, different from present and forthcoming F and f. However, in other cases, F, and how it will develop into the future, may be rather uncertain. Then an unconventional X may in fact turn out to be adaptive, eventually.

13.4.6 The unitary and continuous self

It is sometimes stated that the self is less real than perceived and may be merely a convenient conceptual term for a loosely connected bundle of phenomena. For example, Dennett (1992) compares the self with the centre of gravity of a material body. Such a centre is a convenient concept for understanding the motions of a body, but it only exists in a loose sense.

Similarly, the self might be primarily interpreted as a constructed narrative (reviewed in McAdams 2001). This narrative and how it is socially constructed may be real, but the self itself should then be regarded as primarily epiphenomenal. Indeed, there are many indications that much of the self is constructed socially, as is also assumed here (e.g., part of the processes in Fig. 11). However, concluding from this that the self has no solid reality would be wrong, according to ETS. X is taken here as a real neurophysiological process with significant causal consequences that are crucial for life. Ultimately, it may decide, via its influence on f, between flourishing and withering. X may be partially constructed, but it is a constrained construction, because an x that becomes too different from f is maladaptive.

The unity of the self can be understood from the unity of fitness. The fitness f is a scalar (i.e., a single number) that quantifies the prospective evolutionary success of an individual organism. Although the fitness process F is complex and multifaceted, its outcome, f, is a unitary quantity. Because fitness is unitary, its estimate x should be unitary too. Then X has an implied unity too, because it has to produce a unitary x. In certain pathologies, the unity of X may be poorly maintained, but that would be a sign that something is wrong. It is then likely to produce low fitness because it implies a mismatch between X and F. Such a mismatch would not be sustainable for a long time. The unity of the self does not conflict with the fact that the self usually manifests itself with different identities in different contexts (e.g., the contexts of home, work, or hobby; the term ‘identity’ is used broadly here, for a fine-grained analysis see Oyserman et al. 2012). Different identities are fully consistent with a unitary self, as long as they are consistent with the structure of F. F, too, results from different aspects, depending on context and situation.

Perceiving the self as continuous is important for well-being (e.g., Smeekes and Verkuyten 2014). According to ETS, the self is expected to be continuous because the fitness process F is automatically continuous. Continuity does not imply gradualness, because abrupt changes are possible. For example, circumstances may suddenly change, or the individual may go through an abrupt personal transition. But such abrupt changes are never completely discontinuous in the sense of being unrelated to the previous self. They always follow a historical trajectory of changes. Again, this is strictly true of F, but only by implication true of X when one assumes evolutionary sustainability. Pathologies may still break the continuity of the self.

According to Vignoles et al. (2006) and Vignoles (2011), people construct their identities based on several motives, one of which is the motive to see one’s identity as continuous. The other motives concern self-esteem, distinctiveness, meaning, efficacy and belonging. Several of these motives have already been discussed above as consistent with ETS, such as self-esteem, efficacy (which combines competence and agency) and belonging (which is similar to relatedness and communion). Distinctiveness is necessary in order to be competitive through pathways 1 and 3 (Fig. 12), and distinctive traits can be useful when supporting others through pathways 2 and 4.

The meaning motive implies that people are motivated to see their lives as meaningful. That motive is hard to explain with conventional evolutionary theory. It is not clear how a sense of meaning as such can benefit fitness. Adverse behaviour that decreases fitness could also be felt as meaningful. One might assume that people feel disturbed when they think that their lives are not meaningful, and that such a feeling interferes with normal functioning. But this begs the question: feelings are proximate phenomena, the presence of which requires an evolutionary explanation in the first place. Not having feelings about

meaning would then be more adequate from the point of view of conventional fitness. If such feelings are mere side-effects of previously evolved, but outdated traits, it is hard to understand why reaching for meaning is such an important motive for people. In contrast, meaning and purpose are readily explained by ETS, because the structure of X represents the individual's goals and overall purpose. If one senses meaning in one's life, this essentially means that x indicates that f is high, or is likely to become high. This meaning is sensed even if x's indication is in error. Felt meaninglessness indicates that X needs work, along any of its pathways.

13.4.7 The stability of the self

Finally, there is the question of the stability of the self, which is taken here as its resilience to perturbations. As argued above, continuity of the self does not rule out abrupt change. Abrupt changes in X may just follow corresponding abrupt changes in F. On average, X should then remain dynamically aligned with F. Alternatively, changes in x may be produced by contingencies that do not similarly change f. This would correspond to variability of the self that generates a mismatch between X and F. Such a mismatch would affect fitness negatively, if sustained.

However, some mismatch between X and F is to be expected, as part of the regular, stochastic functioning of the G-loop of Fig. 7. Agency and goal-directedness require variability. Such variability modifies the behavioural dispositions of the individual, and thereby the subsequent X and F. In particular when the values of x and f are low, large variability and fast changes are to be expected. Then the structure of X, and thereby of the self, may change quickly, and may thus induce a quick change in F as well. For example, new coping behaviour may be tried, and when it appears to be successful, it can produce a permanent change in the behavioural repertoire. As a result, also F is changed permanently. Then the self may appear to be instable, during the transition, but it eventually settles to a new, stable structure of X and F.

If the instability continues without finding a favourable X and F, it may become maladaptive. The healthy self is thus expected to show at most transient instabilities, as a normal consequence of an adapting X. If X changes only slowly over time, this can indicate a situation where the values of both x and f are high. Then slow change is indeed adaptive: it will keep x close to their high values, while the residual change still allows exploring even better versions. However, an unchanging X can also indicate a situation where x is high, but f is low. The individual is then in trouble, but believes—nonconsciously or consciously—that everything is all right. This is likely to be maladaptive. If x is low, but f is high, the individual believes the situation is worse than it actually is. This is likely to be maladaptive as well, because as a result of the low x, behaviour would be varied more than would be optimal. As a final possibility, x and f may both be low. Suppose that external circumstances allow change, but that the individual does not manage to change, for example because of a depression. This is maladaptive, because it leaves f low, or it may cause a decrease of f even if f starts out as moderately high. Low f is then a self-fulfilling prophesy based on an x that is inaccurately low because of the depression.

13.4.8 Empirical testing of the theory

The theory is formulated in a way that is sufficiently concrete to allow empirical testing, at least in principle. However, such testing will not be easier than testing any other theory of the self, for two specific reasons. The first reason is that the central component of ETS is an implicit self-estimate of fitness. But fitness is inherently difficult to measure. Straightforward statistical measurement of f or x would require similar individuals and similar circumstances, which are difficult to realize for humans. Alternatively, a theoretical model of F or X might be developed that could be compared with experimental outcomes with variable individual properties and circumstances. But a theoretical model of F or X would be crude at best, because human traits are complex and hard to model. Moreover, the environment of humans, in particular the social environment, is highly complex as well.

The second reason why testing ETS is not simple is that there is an asymmetry when comparing it with existing theories of the self. As discussed above, ETS incorporates several of the key components of such theories. Therefore, empirical support for these theories will often also support ETS. For testing the current theory specifically, one needs predictions that distinguish it from other theories. Fortunately, such predictions exist. One general prediction is that individuals with low x should show more behavioural variability than individuals with high x, at least under stable conditions. More specific predictions follow from Figs. 12 and 13, respectively, as is detailed below.

An enhanced X implies an enhanced self-worth, as perceived by the self but partly based on how others are believed to value oneself. Fig. 12 implies that there are four major pathways to increased self-esteem. Some of these pathways are amenable to experimental manipulation, and could be specifically tested for their effectiveness and interactions. For example, an increased opportunity to teach (pathway 3) should then be able to compensate for a (properly scaled) decreased loyalty to the in-group (pathway 4), thus keeping self-esteem approximately constant (as the observable). Similarly, pathways 2 and 4 may be manipulated into opposite directions.

A specific prediction following from Fig. 13b is that the extent of the self can be manipulated into including less or more of the environment and of the individual. The internal model of X is dynamical. Therefore, it can change which aspects of the environment and of the individual are judged to be so important that they are part of one's identity. Again, this is amenable to testing. A more detailed prediction follows from the fact that the model of X is layered. The aware and symbolic layers of Fig. 13a are presumably produced by repeated internalizations as in Fig. 11. Such layers may be amenable to separate manipulation. This would then enable, for example, an experiment with conflicting non-symbolic and symbolic information. This could be scaled such that the contribution of the layers is changed, but not the self-esteem that results.

13.5 Conclusion

It is proposed here that the human self obtains its agency and goal-directedness from a stochastic cycle that incorporates an internalized process estimating evolutionary fitness, broadly defined. Awareness *by* the self occurs when this process is involved during actual or internalized dialogue. Awareness *of* the self arises when the dialogue utilizes a modelled version of the self. The fitness of an organism corresponds to the rate by which it produces

traits similar to its own in other organisms. In humans, this takes the form of extensive fitness, which consists of two genetic and two social pathways. The genetic pathways concern, first, directly producing offspring and, second, supporting individuals with similar genes. The social pathways concern, first, directly influencing others to adopt one's cultural traits and, second, supporting groups of individuals that are already culturally similar. The internalized version of extensive fitness, in particular the structure of the process producing this internal estimate, is conjectured to produce the human self. It is dynamically modified through an internal dialogue with its modelled version, and it integrates the individual with the social and physical environment. The theory explains to what extent the self is unitary, continuous and stable, and it provides an evolutionary interpretation of self-esteem. Interestingly, the theory contains core components that are also central to other theories of the self, in particular sociometer, self-determination and terror management theory. It thereby provides an evolutionary framework for understanding the foundation of these theories.

PART IV: PHILOSOPHY

Chapter 14

Strong emergence[12]

Are all causes in nature material in the sense of being ultimately equivalent to law-like interactions of matter-energy, in any combination? The success of the natural sciences over the last centuries suggests that this might be true. The material basis of life is understood in increasing detail, and the same goes for the neural basis of mental processes. Descartes's dualism, which assumed that an immaterial mental substance interacts with matter, has become untenable. Similarly, it is clear that living organisms do not contain a mysterious life force, an élan vital, which was assumed by the vitalists. There is overwhelming evidence, nowadays, that physiological and neural systems are fully composed of physico-chemical processes.

Nevertheless, there are puzzling phenomena in living organisms that seem to have a non-material flavour—phenomena such as agency, intentionality and consciousness. Are these phenomena indeed completely produced by law-like material interactions? Below it is shown, by construction, that the answer need not be affirmative. The constructed system is maximally simplified in order to make it comprehensible and explainable. It shows that an autonomous cause can emerge from a realizable system that is purely material. The construction depends on a subtle interaction of deterministic and random processes. Moreover, the emergence requires sustained evolution by natural selection. Importantly, the emergent cause is ontological (related to what exists 'out there') and not merely epistemological (related to knowledge and its limitations).

The construction can be fully realized with components at the level of chemistry and basic cellular physiology. Actions and interactions at these levels have been studied extensively, and there is little doubt about the reality of the entities involved (such as molecules and biological cells). The construction thus stays away from lower and higher levels with a more uncertain ontology, such as fundamental physics (where the ontological interpretation of quantum physics is uncertain) and the behavioural sciences (where the ontological status of mental phenomena is uncertain). The main point here is to show that well-understood entities can be combined in such a way that an ontologically novel entity emerges that has novel causal efficacy.

This chapter is organized as follows. Section 14.1 contains the main result of this chapter, by constructing and explaining a system that produces a strongly emergent cause. Section 14.2 discusses the system, in particular with respect to indeterminacy, causation, emergence and materialism. Section 14.3 states the conclusion.

14.1 Construction and explanation

This section presents and explains a minimal material system from which a novel and autonomous causal factor emerges. The system is intended as a proof-of-feasibility: the purpose here is to show that the proposed system works and indeed produces strong

[12] This chapter is a slightly modified version of van Hateren (2021b).

emergence. The system is not intended as a faithful model of any existing system. Existing, living systems are far more complex, with more complex heredity and physiology. Moreover, the X process that is assumed in point (2) below has not yet been identified and investigated empirically; hence, its actual existence is not known, even though it is physiologically realizable. Consequently, the system and its evolution have a novel causal structure that has not been studied in biology[13]. The minimal system was proposed and analysed computationally by van Hateren (2015a; see also Appendix A). However, it is analysed here specifically with respect to the question of strong emergence, and it is formulated in a fully non-mathematical, yet self-contained form.

The proposed system combines three causally complex ingredients: evolution by natural selection, random[14] structural change of which the variability is systematically modulated (which is a novel ingredient), and cyclical causation. The statements below are specifically intended to clarify the intricate dynamical feedback structure of the system[15]. They include some redundancy and comments (usually included in paragraphs starting with 'Note that'), because a minimal explanation would become incomprehensible if even a single clause were misinterpreted. The explanation of the basic dynamics of the system does not depend on a particular conception of causality, but uses whatever seems clearest. Several points require more discussion than would be consistent with a lucid presentation. Such points are deferred to Section 14.2. As a guide to the reader, Fig. 14 illustrates the overall structure of the mechanism and its causal consequences. However, a diagram of this kind cannot depict the dynamic and stochastic aspects of the system, and several details are omitted for the sake of clarity. The reader should therefore rely primarily on the explanations below.

(1) Assume a collection of systems S of various forms, embedded in a changing environment. The systems have a fixed lifetime and reproduce continuously. They are being modified continually, gradually and randomly, each at a variable rate R. The fitness f of system S is defined as its capacity and tendency to reproduce. The fitness of each system varies from moment to moment, depending on system modifications and on environmental change. The latter is assumed to happen continually and to be partly random. Fitness f can be thought to be produced by a fitness process F, which properly combines all relevant factors and dynamics within S and its environment.

Note that each individual system varies continually throughout its lifetime, that fitness varies during that time as well (i.e., as a propensity to reproduce at each point in time, and

[13] It resembles a population of organisms undergoing Darwinian natural selection, with the crucial difference that conventional natural selection depends on a rate of random change (R) that is either fixed or at least not systematically modulated by an internal estimate (x) of each organism's own evolutionary fitness.

[14] The terms 'random', 'randomness', 'determinate', 'indeterminate', 'non-determinate' and 'deterministic' are used, throughout this chapter, in an ontological sense (that is, they do not refer to epistemic uncertainty about the system under consideration, but refer to the presence or absence of intrinsic randomness within the system itself). See also Section 14.2.1.

[15] Feedback always utilizes implicit time delays, which produces a cyclical rather than a circular causal structure. Therefore, the statements should not be interpreted as a static logical system, because that is likely to produce the buzzer fallacy (Bateson 1979, pp. 58–60, shows that treating the functional description of an electromechanical buzzer as a logical system produces 'if *P*, then not *P*'). In particular, the recursion that the system seems to contain (by utilizing an estimate of fitness to produce a gradual increase of fitness) is only apparent, because of time delays and differences in timescales (see also Section 14.2).

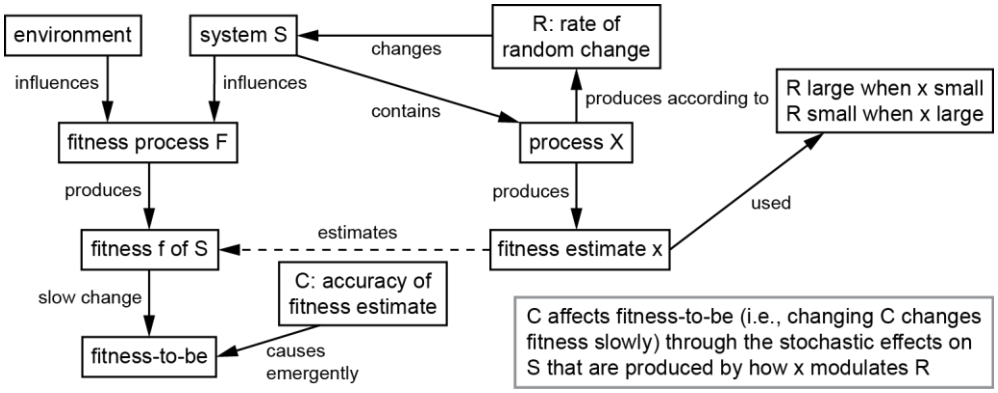

Fig. 14. Overall structure of the proposed mechanism. See the main text for explanations.

not as a post-mortem reading of actual reproduction), and that fitness is defined here as a variable that applies to an individual system (i.e., not to a population of systems and not to specific traits of a system). It is here not necessary to specify which structural modifications of a system S (compared to the parent of S) are heritable and are thus transferred to the offspring of S. However, at least some such transfer of modifications is needed in order to enable evolution by natural selection.

(2) Assume an internal process X that has gradually evolved within S. X has a time-varying output value x that modulates the rate R by which S is modified. This is accomplished through conventional causal mechanisms (e.g., by modulating the efficiency of repair mechanisms of random changes of S). Modulation is performed in such a way that R is a decreasing function of x, that is, R is high when x is small, and R is low when x is large.

Note that R quantifies the rate of random, undirected modifications, again occurring throughout the lifetime of S. This implies that large R produces large variability in the structural changes of S, and small R produces small variability. Thus, x modulates how much structural variation of S one can expect: much variation when x is small and little variation when x is large. Because X is part of S, x produces structural variation of X as well. Therefore, X varies continually during the lifetime of S.

(3) Systems drift randomly through the (high-dimensional and abstract) space of possible structural forms (abbreviated to ‘form-space’ below). This drift through form-space occurs continually, because systems are being modified continually. It follows from (2) that systems are more likely to linger at forms that produce a large x than at forms that produce a small x. This is so, because systems drift away more quickly from forms with small x (since the rate of change, R, is high there) than from forms with large x (since R is low there). Purely for statistical reasons, forms that produce large x appear sticky, whereas forms with small x appear repellent.

Note that the above mechanism produces clustering (in form-space) around forms that produce large x, at least when observed at the level of a population of systems. Individual systems only cluster in a probabilistic sense, by having an increased likelihood of acquiring and sustaining a form with large x. Because the mechanism is purely statistical, a system only gradually responds to changes of X and x (that is, changes in which parts of form-space produce large x; such changes can be produced by environmental changes of the input to process X as well as by changes of the structure of X). Thus, the relation between cause (x) and effect (probabilistic clustering) is gradual and slow: it takes time. Formally, drifting

through form-space is analogous to how molecules drift—in a random-walk-like manner—through a fluid or gas in a diffusive process (e.g., of a drop of ink in water). Such processes are not instantaneously effected, but slowly and gradually. More specifically, the current mechanism would be analogous to diffusion in a volume in which there is structure in the speed of diffusion, such as a volume of water with zones of different temperatures. Ink particles would then gather mostly in the zones with low temperature, because they are expelled more quickly from the zones with high temperature (which produce a higher speed of diffusion).

(4) Assume further that X has evolved to be, in effect, an estimator that produces x as an estimate of the fitness f of S. The term ‘estimator’ is used here in the modern statistical sense of a method (here realized as the process X) that produces an estimate (here x) of the value of a variable (here f). The process X may, for example, get input from evolved sensors of environmental factors relevant for fitness. Such factors are then used to produce an x that is likely to have, at least roughly, a similar value as f, and that is likely to mimic, at least roughly, how f varies from moment to moment. The similarity (or ‘correspondence’ in the colloquial sense) between x and f is denoted and quantified by C, that is, C quantifies how well x estimates f. Another way to state this is that C denotes the level-of-fit that x has to f. Here x and f usually vary across time because of environmental change, and C might gradually change as well because of changes in X or F. If C is small, then x is a poor estimate of f. If C is large, then x is a good estimate of f. If C is negligible (zero or close to zero), then x and f are unrelated.

Note that X and what it does are introduced here and in (2) by assumption. However, in (7) below it is argued that X and what it does are evolvable and sustainable (see also Chapter 2). This makes the presence of X in a form similar to what is assumed here consistent with evolution by natural selection, and far from implausible.

An analogy may help to understand how X and x are related to F and f. One can think of F as the weather (a dynamical process) with f then a single time-dependent variable of that process (e.g., the temperature at a particular place). Then X would be a simulation of the weather (such as running in a computer) with x the resulting estimate of the temperature at that same particular place. C then quantifies how well the simulation estimates that temperature, including how it changes over time. An accurate estimate means a large C, that is, it means a high level-of-fit that x has to f. Both the temperature f and its estimate x usually change across time. In addition, C itself might change across time as well. For example, C could increase when the simulation is improved (by refining the computer program), and it might decrease when the climate changes, which might render the assumptions that were used for the simulation less accurate.

(5) Because x estimates fitness according to (4), large x is correlated with large fitness. A system is more likely, then, to reproduce during the time that it lingers, according to (3), at forms with large x. In effect, process (3) combined with (4) biases each system to spend more time at forms with large fitness than at forms with small fitness. As a result, this enhances the time-averaged likelihood that a system reproduces, that is, its average fitness is enhanced. This effect on the fitness of individual systems is gradual and slow, because it depends on the statistics of drifting through form-space. In order to stress the fact that fitness is not immediately affected but gradually, the resulting fitness will be denoted by ‘fitness-to-be’ below. ‘Fitness’ and f still refer to current fitness.

Note that reproduction is increased here by combining two quite different mechanisms. The first mechanism, (2), utilizes conventional material causation. It increases the likelihood of having systems with large x according to (3). This mechanism would also work if x were unrelated to fitness, and it does not influence fitness-to-be by itself (i.e., the gradual increase of fitness is not a direct effect of x). The second mechanism, (4), is unconventional. It depends on x estimating f, which is denoted and quantified by C. By itself, this estimate (or its accuracy) has no direct effect on fitness-to-be either (because it lacks, by itself, a mechanism that can affect the structure of S and X). Separately, neither of the two mechanisms affect fitness-to-be. But in combination they do.

(6) The efficiency of mechanism (5) is enhanced when the strength of the similarity C between x and fitness is increased, such as may result from random modifications of X in a system. The efficiency is enhanced, because a system with increased C is even more likely to reproduce when it lingers, according to (3), at forms with large x (since large x combined with a large C implies a higher fitness than large x combined with a small C). Therefore, mechanism (5) implies that, on average, an increase of C leads to an increase of fitness-to-be, that is, a higher fitness-to-be than would occur when C would not have been increased.

(7) As already noted in (5), lingering at forms with large x would have no effect on reproduction if x were unrelated to fitness. Therefore, the similarity C between x and fitness is required for producing the slow effect on reproduction (i.e., the effect on fitness-to-be). Perturbing C tends to change the fitness-to-be that a system will reach, according to (6). Specifically, slightly increasing C will slightly increase fitness-to-be, and slightly decreasing C will slightly decrease it. Changing fitness-to-be in this way tends to change actual reproduction. Therefore, C is a factor with causal efficacy, that is, C is a cause. This argument complies with the standard (interventionist) use of causation as is discussed in Section 14.2.2 below.

More specifically, one might perturb C by changing the structure of the processes X or F, or by changing aspects of the environment that affect x (through X) and f (through F) in different ways (see further the fourth paragraph of Section 14.2.2). For example, when the structure of X is changed such that C is increased (which is analogous to improving the computer program that simulates the weather), it just means that x then better estimates f (i.e., the level-of-fit that x has to f is then higher). It does not, by itself, change f (in the same sense as improving a weather simulation will not change the weather). But when x better matches f, the statistical clustering—towards large x—that results (via R) from mechanism (3) will better align with fitness; the better aligned clustering subsequently facilitates and enhances fitness. This happens in a slow and gradual way because the clustering is slow, that is, it affects fitness-to-be. Alternatively, when the structure of X is changed such that C is decreased, alignment gradually decreases, and as a result fitness does so too.

Note that C does not directly affect fitness (nor does it directly affect X, R, or S). It affects fitness only indirectly, slowly and gradually, by enhancing the time-averaged likelihood that a system reproduces, through mechanism (5). This enhancement depends on the fact that mechanism (3) produces a basic tendency towards large x, with x being similar to fitness f because the X process has a structure that produces a large C. Mechanism (3) works through the conventional causal mechanism (2). Mechanism (2) does not directly

depend on fitness, nor does it directly change fitness. Nevertheless, it is evolvable and sustainable, because it indirectly enhances fitness-to-be through (5). Mechanism (4) is evolvable and sustainable as well: the better x estimates fitness (i.e., the larger C), the higher the resulting fitness-to-be will become according to mechanism (6). Hence, there is selection pressure on X to produce large C. Mechanisms (2) and (4) can coevolve, because they reinforce each other. Moreover, initiating their evolution is facilitated by the fact that they are beneficial even when weakly present.

(8) The major conclusion of (7) is that C is a cause. C gives system S a causal efficacy that still requires the material constitution of X, but that goes beyond the conventional physical efficacy of X as produced by its constituents (including their activities and interactions). The latter only influence R, by conventional material causation. However, an effect on fitness-to-be occurs only when a non-negligible C is present as well. C acts as a cause in system S, but it is a strongly emergent cause (in the sense of Bedau 1997, see Section 14.2.3). It produces a causal power in the system that is ontologically novel, that is, the causal power does not follow completely from the micro-physical properties of X (nor from those of X and its environment, see Section 14.2.4). Novel causal powers are indicative of strong emergence. We conclude that C is a strongly emergent cause.

(9) Perturbing C tends to slowly and gradually change the fitness of S. In particular, increasing C typically increases the fitness-to-be of S according to (7). Thus, increasing C is likely to have significant and directed material consequences (because it changes reproduction). In other words, the strongly emergent cause C can direct matter. In addition, C can be changed by material changes in X. Hence, this provides an example where a strongly emergent cause interacts with conventional material causes.

14.2 Discussion

Points (8) and (9) above show that the proposed material system produces a strongly emergent cause and that this cause can interact with conventional material ones. It is useful to state the more general reasons why this system can accomplish this. The main reasons are related to a cyclical use of non-determinism, to evolution by natural selection and to complementary causes. This is discussed further below, as an informal means to help the reader to comprehend the system explained in detail above. But before doing so, a major point is addressed that may appear confusing at first sight (that is, why C can have both an estimative and causal aspect). It concerns the basic question of how one should interpret C.

Naively, when the system is first defined and before its dynamics are analysed, C is simply a standard level-of-fit that x has to f. Although it is an objective feature of the world (in the sense of being objectively assignable by any competent observer), it does not participate in the dynamics of the world. Thus, it has not much of an ontic standing. It is not an entity in the sense that it could influence, change or cause things. But because of the mechanism that runs from X via R to randomly modifying S (see Fig. 14), C acquires a novel causal power (directed at fitness-to-be). It is its only causal power. Having this power makes C ontic: it becomes an entity in the above sense (see also the final paragraph of Section 14.2.2). C is then still a level-of-fit (and, technically, it still quantifies the accuracy by which x estimates f), but it is a non-standard level-of-fit, an ontic one with causal power.

A potential source of confusion is that C seems to have a dual role in the analysis. On the one hand, it quantifies how well x estimates fitness—it is then simply the level-of-fit that x has to f. On the other hand, Section 14.1 claims that C is, in addition, a cause, namely a factor that affects fitness-to-be. Wouldn't this amount to a category mistake, by using a single quantity, C, for two radically different notions (i.e., quantifying estimation and being a cause)? Moreover, estimation might be viewed as a merely descriptive and epistemic tool, presumably objective and real, but without causal clout in the real world. In contrast, the causal aspect of C is claimed to be ontological. Actually, there is no erroneous mix-up of categories here, but this is exactly the kind of categorical uncertainty that is to be expected of strong emergence. Prior conceptions of categories cannot suffice, because an entity with a novel causal power cannot just appear out of thin air. Novel causal power has to be acquired by something that is there already, even if this something is not yet a full-blown entity (such as C as a standard level-of-fit). A key point here is that the two aspects of C, related to estimation and causation, occur on different timescales. The estimative aspect of C occurs on a fast timescale, namely the one at which the components of X operate to produce an x that mimics f. The causal aspect of C—by which it affects fitness-to-be—occurs on a much slower timescale, namely the one at which fitness gradually increases through the statistical clustering mechanism of point (3) in Section 14.1. The two aspects are simultaneously there, but their dynamics occur at disparate timescales. The two aspects of C can thus operate alongside without significantly interfering with each other. The process X can easily let the estimate x of fitness follow the slow changes of fitness that depend on the clustering. In effect, estimation and causation are decoupled.

The objection that estimation is only a human epistemic tool does not hold either. As is argued in the final paragraph of point (7) in Section 14.1, C is an evolvable cause. Hence, there is no dependence on humans or human epistemology when C evolves and acquires objective causal clout in the world. Nevertheless, one might view the estimative aspect of C as part of a primitive epistemology that has evolved within system S itself: the fact that x is 'about' f may be regarded as a primordial form of 'knowledge' (ascribable to S) about the world. Such a minimal epistemology is evolvable, because C has a causal aspect that promotes fitness.

The proposed mechanism is non-deterministic (the term 'non-deterministic' is used here and below to denote 'not completely deterministic'; see also Section 14.2.1), because it depends in a fundamental way on utilized randomness. When systems of the kind considered here are deterministic, any emergent causation may be classified as weak emergence (Bedau 1997; see Section 14.2.3). Such a classification also applies to most non-deterministic systems, that is, systems that combine determinism with randomness. Randomness usually does not contribute to novel causal powers; thus, involving randomness is not, by itself, sufficient to produce strong emergence. However, the non-deterministic system explained here is of a rather special kind. It couples deterministic and random factors in a cyclical and sustained way, such that it is impossible to separate them and impossible to neglect the effects of randomness. The randomness is even indispensable, because without the randomness there would be no effect on fitness-to-be and C would not be a cause. Importantly, the specific way by which deterministic and random factors are coupled (i.e., by how x and R are related) tends to increase fitness. In combination with selection pressure, this results in a stable factor with a strongly emergent causal efficacy, as explained in Section 14.1.

Complementary causes are responsible for the fact that the emergent cause C can still affect matter. X has two different causal aspects, an emergent and a conventional material one. The emergent causal aspect of X is produced by the fact that x is an estimate of fitness (as denoted and quantified by C). The presence of a non-negligible C is required for producing the effect on fitness-to-be. But actually producing this effect depends crucially on the material causal aspect of X, which is the fact that X is, in addition, a regular physical process occurring within system S. This physical process modulates the rate of change R of the system, through conventional material causation (such as by facilitating or counteracting the effects of random events[16]). Yet, this material causation can only affect fitness because of the emergent causal aspect of X. Exactly the same material causation, with the same X and x, would have no effect on fitness-to-be when x would not mimic fitness any more (for example, when fitness would have been drastically changed by large random disturbances of the environment). Thus X couples the emergent cause with the material cause. These causal aspects of X are complementary, because they are both indispensable. Neither of them can do the job alone. Therefore, the complementary causes produce neither epiphenomenalism, nor causal overdetermination. The mechanism does not require or produce a dualism of the Descartian type. Nevertheless, one might perhaps view it as letting some sort of dualism emerge from a physical monism.

The proposed system is based on several fundamental assumptions, in particular that there is indeterminacy and that causation can be understood in a particular way. Moreover, it has consequences for several fundamental questions, in particular concerning the existence of strong emergence and the correctness of materialism and physicalism. These topics are discussed below.

14.2.1 Indeterminacy

A major aim of science is to uncover, systematize and explain the law-like regularities that one can find in reality. However, it is a valid question whether reality is fully produced—according to such regularities—in a determinate way (i.e., whether it is fully deterministic) or that reality is partly indeterminate (such as when some events occur randomly and without sufficient cause). For the construction explained above it is not necessary to consider this question in its most general form, pertaining to all of reality at once. It suffices to consider the constructed system combined with its causally relevant environment (i.e., the environment relevant for producing its evolutionary fitness). Let us call this combination the ‘inclusive system’. Above we assumed that there is some indeterminacy—in addition to law-like regularities—in the constructed and inclusive systems, in the form of random system modifications and (partly) random environmental change. The indeterminacy is called ‘randomness’ here, and it is assumed to be present throughout the systems. Note that ‘randomness’ refers here to temporal indeterminacy, and not to some structural property of a system (in the sense that one might call a spatial pattern less or more random). Although most engineered systems and many biological subsystems specifically tend to reduce and

[16] Such mechanisms are known in molecular biology; a simple (non-biological) way by which one can understand the concept of modulated randomness is when one would modulate the distance between a radioactive source and a biological organism, and thus modulate the mutation rate in the latter.

control randomness (then usually referred to as ‘noise’), the proposed system specifically utilizes it.

It is plausible that, in effect, any constructed or inclusive system indeed contains indeterminacy and is, thus, non-deterministic (i.e., not fully deterministic). This is plausible because no system is completely isolated from the rest of the universe. In classical physics, long-range electromagnetic and gravitational fields that originate from outside the inclusive system inevitably disturb it in an intractable way. Such influences are not negligible (for an extreme example see Berry 1988, who argues that even the unknown whereabouts of a single electron at the edge of the visible universe produces sufficient gravitational uncertainty to yield intractable molecular motion, here, within a tiny fraction of a second). In quantum physics, systems inevitably couple to their environment and subsequently decohere, producing indeterminate outcomes (as observables, irrespective of how one interprets the ontology of quantum physics). In any case, the constructed and inclusive systems contain indeterminacy that is either strictly ontic or at least indistinguishable from being strictly ontic.

Empirically, indeterminacy is commonly observed in systems. At the submicroscopic scale there are indeterminate quantum events, at the microscopic scale there is thermal noise (random movements and interactions of molecules), and at larger scales there is indeterminacy caused by nonlinear and unstable dynamics. The latter can readily amplify (sub-)microscopic indeterminacy to macroscopic scales. Such instabilities are known to be very common in dynamical systems. A dramatic example is a study by Laskar and Gastineau (2009), who show that infinitesimal causes can have runaway effects even in the planetary system. Similarly, it is clear that biological systems (ranging from the cellular to the neural and behavioural level) are non-deterministic to a considerable extent (e.g., Balázsi et al. 2011; Faisal et al. 2008). Indeterminacy is not negligible at any level in such systems.

14.2.2 Causation

The explanation of the proposed system utilizes cause-and-effect language. Such language has many uses in many different fields, and it has proven difficult to define causation in a way that covers all such uses at once (for an overview of different approaches see Illari and Russo 2014). However, the proposed system belongs to the realm of the basic (matter-oriented) natural sciences, in particular the ones that deal with complex material systems (such as occur, for example, in cellular physiology, neuroscience, and atmospheric science). Within these fields, causality is usually captured by making a mechanistic model of the investigated system. The model—such as the one presented in Section 14.1—then contains the conjectured causal structure of the mechanisms or processes involved. The importance of mechanisms and mechanistic causation in the sciences has recently become the focus of a range of studies (e.g., Glennan 1996; Machamer et al. 2000; Bechtel and Abrahamsen 2005; Craver 2007; Illari and Williamson 2012; Glennan 2017). Much of this work has a reductive flavour, that is, it is consistent with ontological reduction of causation. However, the term ‘mechanism’ is used throughout this book in a general sense, not necessarily reductively and not necessarily referring to deterministic systems.

The causal structure of mechanistic systems can be probed either by an empirical intervention (i.e., through an experiment on an actual system), or by a theoretical

intervention (i.e., by mathematical or computational analysis of a modelled system). Interventions are typically performed by slightly perturbing (altering) some part of a system. When a system is thus perturbed (or would be perturbed), and subsequently some part of it tends to respond in a statistically reproducible way (or would do so), then one can infer that a causal connection exists. This strategy (perturb and observe the consequences) was used in point (7) of Section 14.1 to show that C is a cause of fitness-to-be. The strategy is standing practice in science, being closely related to interventionist approaches to causation that are discussed in the philosophy of science (Woodward 2013) and in statistics (Pearl 2009).[17]

Yet, there are valuable alternatives to the interventionist approach. One may then wonder whether such alternatives would also classify C as a cause. A few examples are discussed here briefly. First, a counter-factual take on causation would indeed classify C as a cause: if C had been negligible (i.e., if x would have been unrelated to f), an effect on fitness-to-be would not have obtained either. Again, the effect that is meant here is the gradual and slow effect produced by the drift-related process utilized in mechanism (3) of Section 14.1. A second possibility is to equate causation with the existence of a productive causal mechanism between cause and effect (see Glennan 2017, Ch. 7, and the remarks above about capturing causation through a mechanistic model). Using this approach again leads to the conclusion that C is a cause: there is indeed a (stochastic) mechanism (as explained in Section 14.1) that produces a change of fitness-to-be when C is changed. As a final possibility one may assume that causation requires that something (e.g., energy) is transferred from cause to effect. Both in applied and fundamental physics this is typically the case (for an extended version of the transference theory see Ardourel and Guay 2018). However, it is not immediately clear what C might transfer to fitness-to-be. Given the mechanism, whatever is transferred must at least involve something statistical, such as entropy or information. But it would require further analysis to see what is going on in terms of transfer.

The argument in point (7) of Section 2, as well as several of the alternatives discussed above, establish that C is a cause. One might object that the true cause may not be C but rather the X process (through x), because C depends on x. However, such an argument would ignore the fact that C also depends on the process F (through f), since C denotes how well x mimics f. A perturbation of C could result from perturbing X or F or both. It could even be produced by an environmental change that affects X and F in different ways. The effect of such changes depends on changes in C, because mechanism (3) of Section 14.1 does nothing but letting S cluster around large values of x, which can only change fitness-to-be in relation to how similar x is to f (that is, depending on the value of C). No change of fitness-to-be will occur when C is not changed. The latter may occur even when X, F, and the environment all change but approximately compensate each other. This is possible because X and F are two separate, physically unrelated processes that can be perturbed independently (compare that to the fact that a weather simulation and the weather can be perturbed independently). The key point here is that any change

[17] One might worry that an interventionist approach would be difficult here if there is a causal loop from C to fitness and from fitness back to C. However, such a loop does not occur here. Fitness does not directly affect C (f will usually change all the time because of environmental change, but x can normally follow that, keeping C unaltered; compare this to the fact that the weather normally does not affect the accuracy of a weather simulation). In contrast, C affects fitness (but only slowly). The deeper reason for this asymmetry is that the latter change depends on a driven diffusive mechanism (from x via R to random changes of S). Such a mechanism is irreversible (for statistical reasons).

must always run via a change in C in order to affect fitness-to-be. In other words, C is the crucial, indispensable causal factor in this mechanism.

The above discussion primarily concerns how scientific knowledge about causes and causation is acquired, that is, it is primarily related to the epistemology of causation. On the assumption of realism, scientific knowledge should somehow correspond to or refer to what actually exists (i.e., to the ontology of reality). This is often unproblematic, such as when referring to well-studied objects and their characteristics (e.g., a specific planet or a specific molecule). But causation and the associated law-like regularities are more problematic, because it is not clear to what extent fundamental physical laws or fundamental causal dispositions are ontological (see, e.g., Mumford and Anjum 2011; Bird 2016). However, at non-fundamental levels (such as used for the system proposed here) this problem can be avoided. When one assumes a mechanistic framework, one can explain what a particular mechanism causes (i.e., which outcome it produces) by the activities and interactions of its components (Machamer et al. 2000; Illari and Williamson 2012; Glennan 2017). Such components and how they (inter-)act are then taken as given facts, with an ontology that need not be known in all its details. This strategy thus avoids entering a long regress of explanations at lower and lower levels, of which it is not clear if and how it bottoms out.

If a particular cause can be completely explained by the above strategy, then such a cause can be said to be amenable to ontological reduction in the sense of Van Gulick (2001). For example, it is assumed above that x—as a cause of R—is fully explained by the detailed workings of X. Then the ontology of the cause x is reducible to the ontology of the components of X (including their interactions). Similarly, it is assumed that f—if interpreted as a cause of reproduction—is fully explained by the ontological details of system S, its environment, and their interactions (i.e., it is explained by the process F). This implies that such causes are not autonomous, but rather are produced by other causes, namely those at the component level. Although such non-autonomous causes are objective and real, they have no independent causal power[18]. The causal power of a non-autonomous cause is then fully defined by the causal powers of its constituents (including their interactions).

However, the ontology of the emergent cause C—as affecting fitness-to-be—is different. Its causal power is autonomous in the sense that it is not fully defined by the causal powers of the constituents of the system and its environment. This is so, because an indispensable ingredient of the causal power of C is produced by randomness (which has no fixed ontology and does not correspond to constituents with causal powers). The randomness is indispensable, because without it C would not be a cause and there would be no effect on fitness-to-be. A cause C with autonomous causal power must be an autonomous entity. This is based on the notion that something that has the potential to influence other entities (by having some sort of causal power) has to exist, i.e., has to be an entity. Having causal autonomy (distinctive efficacy) then guarantees ontological autonomy (distinctness), by Leibniz's law (identicals are indiscernible, thus discernibles are not identical); see Wilson (2015, p. 372). In other words, the causal analysis of the proposed system shows that it produces a novel, emergent entity C with autonomous causal power.

[18] Having 'causal power' is used in this chapter in the general and weak sense of having a disposition, tendency, or propensity to produce certain effects; the use of the term here does not assume a specific ontology of powers.

14.2.3 Emergence

The mechanism explained above produces emergence. A general discussion of emergence is beyond the scope of this chapter (but see the anthologies by Bedau and Humphreys 2008 and Gibb et al. 2019, as well as Wilson 2015; Gillett 2016; Humphreys 2016; O'Connor 2020). Nevertheless, it is important to indicate where the proposed mechanism is positioned within the field studying emergence. Guay and Sartenaer (2016) usefully distinguish three dimensions along which one can analyse emergence. First, their ontological-epistemological axis distinguishes ontological emergence (which is 'out there' in the natural world) from epistemological emergence (which is just a consequence of our representation of the natural world). Second, their strong-weak axis corresponds to the extent to which emergence is present fundamentally (my take on what they mean by 'in principle') or only in practice. Finally, their synchronic-diachronic axis corresponds to whether the emergent phenomenon (such as an event, a property, or a process) occurs simultaneously with its constituents (such as events, properties, or processes) or distinctly later in time.

In other literature (e.g., Bedau 1997; Chalmers 2006; Kim 2006), the ontological-epistemological and strong-weak dimensions of Guay and Sartenaer (2016) are often combined into two broad possibilities for emergence, denoted by 'strong emergence' and 'weak emergence' (see also O'Connor 2020). Weakly emergent phenomena are then novel and surprising primarily for epistemological reasons. A system may be so complex that its properties and dynamics cannot be explained in simple terms. In principle, one could simulate the microdynamics of such a system and, thus, reproduce the emergent phenomenon (Bedau 1997, 2008). The phenomenon is then explained computationally, but it is still novel and surprising. Strongly emergent phenomena, on the other hand, are novel and surprising primarily for ontological reasons. In particular, such phenomena produce novel causal powers that are not explainable in terms of those of the components of the system and its environment (Kim 2006, 2010).

The type of emergence produced by the system that is proposed here can be unequivocally positioned on two of the three axes of Guay and Sartenaer (2016): it is ontological and present fundamentally. But the position on the synchronic-diachronic axis is more complex. A purely synchronic relation between constituents and emergent may be viewed as one of dependence or constitution (e.g., Gillett 2016). However, the present study assumes a regular matter-based system, which thus has to be consistent with fundamental physics. All of the fundamental theories of physics contain differential equations of time; hence it is difficult to see how such a system could produce emergence that is both ontological and purely synchronic (see also remarks and references on this topic in Guay and Sartenaer 2016, p. 301). Pure synchronicity depends on connecting properties (or events, or processes) at different positions at the same time, which fundamental physics cannot do (at least not in a sense relevant for the current study).

That the emergence produced by the proposed system is not purely synchronic is also implied by the fact that it focusses on causation (in the sense of producing effects), which—in matter-based systems—inevitably involves time delays. However, it would be inappropriate to characterize the emergence here as diachronic. The constituent causes (involved in how X produces x and thus modulates randomness) and the emergent cause (C) act continuously and spread out in time. The emergent cause-and-effect relation (between C and fitness-to-be) builds up statistically and gradually. It occurs on a much

slower timescale than the constituent cause-and-effect relations within X. Yet, these constituents and the emergent cause C overlap in time to a considerable degree, that is, they occur almost simultaneously. They do not occur at clearly separable, distinct moments in time (as would be required for pure diachronicity)[19]. Therefore, the emergence is best characterized as near-synchronic. In order to keep the formulations in this chapter simple, the term 'strong emergence' is used here to refer to the near-synchronic, ontological and fundamentally present emergence of a causal power. In the terminology of Van Gulick (2001), strong emergence as used here corresponds to a radical-kind emergence of a causal power.

A specific form of emergence, where an ontological transformation occurs across time, is called transformational emergence (Humphreys 2016, Ch. 2; Guay and Sartenaer 2016). One may wonder whether the proposed system could be understood in those terms, as one might think that C is transformed from a simple level-of-fit to a cause. However, nothing is transformed here, because C remains a level-of-fit and the causal power is just an emergent addition. Moreover, transformational emergence is diachronic, whereas the emergence proposed here is not: C is simultaneously a level-of-fit and a cause, even as the dynamics of these two aspects occur at mostly different timescales.

Finally, one may wonder what would be the emergence base from which C emerges. From Section 14.1 and Fig. 14 it is clear that many different parts of the world are participating in producing C: system S and its components, as well as the environment, which coproduces fitness. But a crucial part of the emergence base that is not explicitly represented in Fig. 14 is ontic randomness, without which C would not be a cause. Yet, ontic randomness is not definable in a definite way, because its concurrent outcomes are not determined before they occur. Only with hindsight one could reconstruct the randomness and its consequences. But hindsight is a form of epistemology that is not available to the mechanism itself, including during the time that C gradually causes fitness-to-be (see also below). Therefore, the concept of a definite emergence base cannot be applied to the proposed mechanism, at least not in an exact way.

14.2.4 Materialism and physicalism

According to point (8) in Section 14.1, C is a strongly emergent cause. But is C a material cause? The answer to this question depends on how one interprets the meaning of the term 'material'. As indicated by the first sentence of this chapter, we will take it as referring to being fully produced by law-like interactions of matter-energy, in any combination. Thus, it is not restricted to mere matter, but also includes, for example, forces and fields. This is consistent with how the term 'material' is commonly understood in the natural sciences (including by those working on complex material systems). The common interpretation is that a material cause is a cause that would be completely specified by its material constitution, that is, by its material constituents and their activities and interactions[20].

[19] Note that the notion of causation taken here is more general than the traditional notion of event causation (i.e., one distinct event causing another distinct event). This is necessary, because mechanisms (such as the one proposed here) often do not operate with discrete events, but rather through continuous and overlapping influences.

[20] In the context of physicalism, Hüttemann and Papineau (2005) argue for distinguishing part-whole reduction and inter-level reduction (e.g., when connecting the mental and physical levels). No such

Moreover, the material constitution must be well-defined and must be knowable, at least in principle. This implies that such a cause is amenable to ontological reduction (in the sense of Van Gulick 2001), at least in theory: if one would reassemble the full material constitution from scratch[21], then one would get exactly one's material cause and its efficacy, no less and no more. A second common interpretation of material causation is based on the notion (e.g., Wilson 2015, p. 351) that causation is first of all a cause-and-effect relation between spatiotemporally located goings-on (with 'goings-on' standing for events or processes). Then material causation can be interpreted as a cause-and-effect relation between the material events or processes to which such goings-on correspond (i.e., to which they can be reduced ontologically). If either one of these two common interpretations of material causation does not apply, then it seems best not to regard the causation as material in the conventional sense of the term. As is argued below, neither of these interpretations does in fact apply to C.

The first interpretation does not apply because there is no suitable basis for an ontological reduction of C (or, in alternative formulations, a basis on which C supervenes or a basis by which C is realized). One might think that C could be reduced to the material process (X) that is required for producing C and its effect on fitness-to-be. This might be so, because the value of x is fully produced by the material constituents of X, and the effect of C on fitness-to-be depends on how x modulates R, the rate of random change. However, such a reduction cannot work, because C depends not only on x, but also on fitness f (recall that C quantifies how well x estimates f). Fitness depends on interactions of the system S and its environment, which both extend beyond X. Therefore, X is not sufficient for specifying C, nor for specifying the effect of C on fitness-to-be.

Alternatively, one might think that C could be reduced to a broadened basis, which includes not only X and S, but also any part of the environment that is relevant for fitness. Such a reduction cannot work either, as is argued next. The effect of C on fitness-to-be is stochastic, being produced slowly and gradually by modulated random change. This randomness is assumed to be ontic, which means that each individual change is indeterminate up until the short time interval during which it is produced (this remains true also when the statistics of changes are modulated over time). The specific outcome of each change only becomes determinate during this production, but before that it is fully indeterminate: not just epistemologically (due to lack of knowledge about the system) but ontologically (irrespective of how completely the production could be characterized; see also Section 14.2.1). The effect of C on fitness-to-be is slow, thus it depends on the specific outcomes of a series of individual changes during a stretch of time. The system compounds this effect over time in such a way that the probability distribution of each subsequent change, as well as its micro-effects, depend on the results of previous changes (because these results partly determine how the system subsequently modulates randomness). This means that ontological reduction as defined above is not possible: one could not reassemble the material constitution of the broadened basis—over the stretch of time where C assembles its effect on fitness-to-be—from scratch, because randomness prevents reproducing this constitution. This is so, because randomness would produce a different

distinction would apply to the current study, because it focusses on a single, well-understood level (chemistry-based physiology, see the introduction).

[21] For open or non-deterministic systems this could include external contingencies (i.e., randomness) assumed to be independent of the form and dynamics of the system (see Bedau 1997).

material constitution each time when one would attempt such a reassembly: the constituents and their activities and interactions would not be the same. Consequently, not only the sequence of forms of X and S would be significantly different each time, but also C and its effect on fitness-to-be. Therefore, the causal efficacy of C is not fully given by its ontological basis, because that basis stretches across time and is partly indeterminate[22]. Of course, one could observe the specific realizations of the random changes of X and S over a stretch of time (after the fact) and use those for faithfully reproducing the cause C and its effect from scratch. But that would be cheating, because a reproduced series of random changes is not indeterminate, but determinate (and thus not random; recall that the terms random, indeterminate and determinate are all used here in an ontological sense). That would fundamentally change the ontology of the system that is presented here, in effect taking it to be fully deterministic. We must conclude, therefore, that C cannot be reduced ontologically, not even in a broadened basis.

The second commonly used interpretation of material causation, that it is a relation between spatiotemporally located goings-on of a material kind, does not apply either. The effect, an increase of fitness-to-be, is a going-on of the right kind, because fitness characterizes a material process (reproduction) that is fully produced—in a complex way—by the physical properties of system S and those of its environment. However, the cause of this increase, C, is not a material going-on, as is argued next. C denotes and quantifies how well x estimates fitness, and this estimation is in fact the crucial condition for C acting as a cause. An estimate is, roughly speaking, a relation of some kind[23]. The process that produces x, X, is a material going-on, and the same goes for the process that results in f. However, the fact that C is a cause (by slowly affecting fitness-to-be) is fully attributable to the relation between x and f as such (see the fourth paragraph of Section 14.2.2). This relation, as such, is not a material, spatiotemporally located going-on. Essentially, it is a similarity between two values as they evolve over time, somewhat like a correlation (but more constrained, because the specific value of x needs to be close to the specific value of f, on average). A relation between two values does not consist of material events or processes. There is no way in which a relation between two values is identical to—in the sense of being indistinguishably replaceable by—a material event or process. Relations between values belong to a different category than material events and processes. We conclude that causation here is not a relation between two material goings-on, but is a relation between a relation of some kind (the estimative aspect of C) and a material going-on (the material process characterized by fitness-to-be). Therefore, the causal aspect of C fails to comply with the second interpretation of material causation.

[22] Note that this is not an epistemological issue, but an ontological one: even the system itself could not construct the causal aspect of C before that aspect is slowly and gradually produced through a series of changes, because each change remains indeterminate even to the system itself until the short time interval in which it is produced.

[23] In a proper (two-sided) relation, A being related to B entails that B is related to A (though often in a different way). In contrast, an estimate is one-sided: the estimate points to the estimated, but the estimated does not point to the estimate. Here x points to f, but f does not point to x (only the former is causally relevant to S, which contains and utilizes x, whereas the latter would be causally irrelevant to S).

Because C does not comply with both common interpretations of material causation, we must conclude that the cause C is best regarded as non-material[24]. Hence, if the proposed system were constructed (or variants of it actually exist), it would be a counter-example against materialism. However, one should not conclude from this that the cause C is non-physical as well, as this depends on how broadly one defines the term 'physical'. One might argue that anything real that exists must be physical, by definition (Strawson 2008), that anything that occurs in the physical realm (including uncaused random events) is physical, and that anything that arises from that is automatically physical as well. Then C would be a physical cause, along with any other possible cause (if one assumes naturalism). Thus, the current proposal is consistent with physicalism, if that is broadly defined.

Nevertheless, the mechanism as proposed here is incompatible with common, more narrow conceptions of physicalism. Arguments against the existence of strong emergence often depend on the thesis that the physical realm is causally closed, such as "Every physical effect has an immediate sufficient physical cause, in so far as it has a sufficient physical cause at all" (Papineau 2009). The 'immediate' is included in order to exclude causal chains running indirectly through a non-physical cause, and can be ignored here. The thesis appears to be correct for deterministic processes, and it is silent about uncaused random events (which are not 'effects' because they are not caused). But it is not obvious that the thesis, or any variant of it, is always true of processes that are both deterministic and random. The mechanism explained above is in fact a realizable counterexample against physical causal closure[25]. It clearly has a physical effect, on fitness-to-be and thus on reproduction. It produces this effect by systematically utilizing randomness (via x and R) to produce deterministic effects (via the structure of S and X) that subsequently modulate further randomness (via x and R), and so on. This continual cycle of mixing determinism and randomness produces an effect on fitness-to-be that is of mixed origin: both caused and uncaused. The cycle is complex, nonlinear, and nonstationary, making it impossible—not only in practice, but fundamentally—to disentangle the causal effect that C has on fitness-to-be in terms of physical causes and uncaused randomness. More specifically, the cause C produces its effect on fitness-to-be by an inseparable composite of physical causes and uncaused randomness. One might think that the 'in so far as it has a sufficient physical cause at all' could deal with this, but that is not so. As the definition stands, 'in so far .. at al' seems to be equivalent to 'if', and is intended to exclude ontic randomness (Papineau 2009). It does not deal with causation of mixed origin. We might change the definition by removing the 'at all', and hope that 'in so far as' (i.e., 'to the extent that') works for the current mechanism. However, this would tacitly assume that one can separate the effect on fitness-to-be in a caused and an uncaused part, which is not the case. Thus, 'in so far as it has a sufficient physical cause' is inapplicable here; it is meaningless with respect to the proposed mechanism.

Still, some form of physical closure might be formulated by saying something like "Every physical going-on that is caused, is caused by nothing other than physical goings-on or causes that are strongly emergent from physical goings-on." The term 'physical going-on' denotes here any form of physical change, irrespective of whether it is determinate, random, or mixed. Then the physical domain is still closed in some sense, but

[24] Material causation is undermined here even if one thinks that only one of the two interpretations is convincingly discredited.

[25] At least for any physical system that is not equal to the entire universe at once; the latter is not sufficiently well understood to presume that it can be regarded as a regular closed system.

it is not necessarily causally closed (because any strongly emergent cause, such as C, would be a metaphysical novelty, though still physical if that is broadly defined).

Van Gulick (2001) states that mainstream ('atomistic') physicalism assumes two core principles, namely AP1 "The features of macro items are determined by the features of their micro parts plus their mode of combination" and AP2 "The only law-like regularities needed for the determination of macro features by micro features are those that govern the interactions of those micro features in all contexts, systemic or otherwise." The system presented here conforms to AP1, but not to AP2. For example, the process X is, as a matter-based process, fully defined by how its micro items are combined (as in AP1). However, the interactions of its micro features are not the only law-like regularities that are needed for understanding the causal efficacy of X (contra AP2). In addition, the emergent law-like regularity "The factor C, which depends on X, affects fitness-to-be" is needed. This regularity is not defined by the system, nor by the system and its context (i.e., by the inclusive system). It is a slow regularity that is established stochastically and gradually, and that depends on how well x keeps estimating f over time. Thus, it depends on how the inclusive system and the environmental statistics change over time, potentially a long time into the future. Of course, S cannot really tell the future, but has to rely here on the predictive qualities of its X process, the structure of which was gradually established—through natural selection—over time, potentially a long time into the past. This deviates from the standard assumptions of mainstream physicalism (as well as from those of a non-mainstream version as in Papineau 2008). Such standard assumptions are that the causal fate of a system is fully determined by the current system and its current environment, plus possibly some current external contingencies (for open or non-deterministic systems; see Bedau 1997). An explicit and irreducible statistical dependence on the (non-determinate) future form of the system and the (non-determinate) future environmental statistics, as applies to the system explained here, is not part of the standard view. Nevertheless, the proposed system is fully consistent with standard science.

14.3 Conclusion

The mechanism constructed above shows that strong emergence—taken here to be the near-synchronic, ontological and fundamentally present emergence of a causal power—is feasible in a universe that appears to be fully based on micro-physical laws. This is enabled by the fact that randomness escapes such laws, at least in its detailed realizations (i.e., its actual outcomes). The proposed mechanism takes advantage of this non-deterministic loophole in the apparent law-like nature of reality. An internal estimator of a system's own reproductive fitness thus produces a slow and gradual increase of fitness, by stochastic means. As a result, the estimate of fitness obtains a novel quality, namely that of being a—strongly emergent—cause of fitness-to-be. It thus provides an example of how strong causal emergence can be realized in a material system. If it actually exists in one form or another, it may explain several of the more puzzling phenomena that occur in living organisms (see Chapters 10 and 12 for specific elaborations).

Chapter 15

The real, the true and the good

Broadly speaking, philosophy consists of three major fields of study. Metaphysics studies what is ‘out there’ and how it changes, epistemology how one can know about such things, and ethics whether they are good or bad. Epistemology has a pivotal role here, because it provides data that feed metaphysical considerations as well as data to which ethical considerations can be applied. Epistemology depends on intentionality, which denotes the power of minds to be directed towards something, for example when a thought is about an object or event (reviewed in Jacob 2014). Chapter 10 proposes that intentionality is produced by an evolved form of estimation. This depends on the conjecture that each individual has an internal process that produces an estimate of the individual’s own evolutionary fitness. Such a process, as well as the fact that it estimates, can be shown to be evolvable through evolution by natural selection. The components of the internal process can be regarded as intentional components that point to the external world. All expected properties of human intentionality can subsequently be constructed by parsing fitness and by extending it to include social and cultural factors.

Here, I will argue that the basic form of this theory is consistent with the tripartite structure of philosophy. This suggests that the usefulness of this separation of fields has deep biological roots. The analysis below can be regarded as a form of metaphilosophy: it aims to provide objective reasons for studying metaphysics, epistemology and ethics as mostly separate fields. Moreover, it may help to understand the origin of the ways in which these fields are related. However, the reader should bear in mind that the relevance of this analysis depends on whether the conjectured internal fitness estimator actually exists. That is an empirical question, which cannot be answered unequivocally yet, as there is currently only circumstantial evidence for the corresponding processes. Nevertheless, their existence seems plausible, or at least quite possible (see discussions from the perspective of general biology in Chapter 8 and from the perspective of neuroscience in van Hateren 2019).

Figure 15 shows the basic diagram of Fig. 3, with a few additions. The mechanism that it depicts has been explained in various ways in previous chapters and will be summarized here only briefly. Any biological organism has an evolutionary fitness f, which is taken here as a variable that quantifies likely evolutionary success. In simple cases, it is an individual’s propensity to survive and reproduce. In complex cases, it can also involve social and cultural factors. Fitness f is understood to be produced by a highly complex process F through which organism and world interact. Both F and f are distributed throughout the world (with f in an analogous way as a variable can be represented in a distributed way throughout an artificial neural network). The higher an organism’s fitness is, the higher the chances are that the organism can pass on its traits to the next generation. Under quite general conditions, fitness differences between organisms then lead to evolution by natural selection.

Fitness is taken here as a variable that continually changes during an organism’s lifetime, because prospects tend to vary over time, for example during a time of famine. An organism

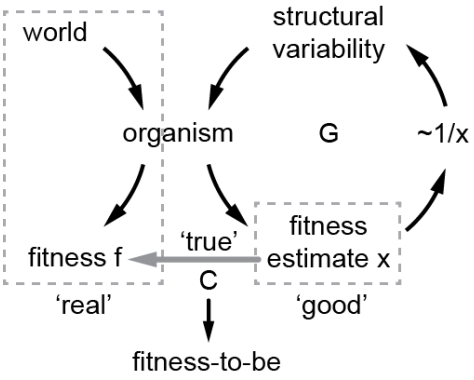

Fig. 15. Conjectured origin of the real, the true and the good. The prospective evolutionary success of an organism depends on its fitness f, resulting from a time-varying world and the organism's features. Fitness is assumed to be estimated by an internally produced variable x, which drives variation of structural properties of the organism during, and limited to, its lifetime. The scheme is evolvable if small x produces large variation and large x small variation, symbolized by ~1/x. By continually modulating random causation in this way, the loop G produces goal-directedness and agency (see the main text). The 'real' is defined by the effects of the world on f, the 'true' by how well x estimates f, and the 'good' by the goal of obtaining high x.

has many means to adjust to changing circumstances in order to keep its fitness as high as possible. Most of these means depend on primarily deterministic mechanisms. But there is one possible mechanism that is non-deterministic in an essential way. It is illustrated by the causal loop G in Fig. 15. It works by modulating the rate of random structural changes in an organism during—and limited to—its lifetime, for example by changes in neural connectivity. Such structural changes can affect behavioural dispositions, that is, they can affect which behaviours the organism is likely to produce. The structural variability (i.e., the on-average-expected variation) is modulated by an internally generated variable x, which is assumed to depend on behaviour. This variable is produced by an internal process X within the organism, again with both X and x present in a distributed way. The structural variability that is modulated by x is also assumed to be produced, via X, in a distributed way throughout the organism. The variability is modulated—through conventional physiological mechanisms—in such a way that when x is large, the variability is small, and when x is small, the variability is large. This is symbolized by ~1/x in the figure.

Because x is assumed to depend on behaviour, cycling through the G-loop will, in effect, produce behavioural dispositions that typically produce large x. This occurs purely in a statistical, diffusion-like way: when x is small, large structural variability lets the organism quickly drift away (within an abstract space of behavioural forms) from such behaviours, whereas it will tend to stay close to behaviours that produce large x (because large x produces small variability). Thus, small x seems repellent and large x seems sticky. As a result, x is likely to become large, on average. One can summarize this by stating that the G-loop is an optimizing mechanism (producing large x) that works in a stochastic way, by modulating randomness. It has a [goal], large x, that is not explicitly built into the mechanism in the way outcomes might be built into a deterministic mechanism. The [goal], large x, arises here in a stochastic, non-deterministic way. The square brackets indicate that the goal might be only apparent (in an 'as if' kind of manner). They are used below until the point where it becomes clear that they should be omitted.

In principle, the variable x might indicate any specific [goal]. However, a [goal] can only evolve if it is not detrimental to fitness. It is easy to see that the optimal [goal] is in fact high fitness f itself. If the value of x were identical to the value of f (also as these values vary over time), then the statistical outcome of the G-loop, large x, is guaranteed to produce large f. A problem here is that f is not directly observable and that it is produced (by process F) in a highly complex way. Then the best an organism can do is to produce an estimate of f. Therefore, x must have evolved to be, in effect, an estimate of fitness. The better this estimate is, the higher the fitness will eventually become through the stochastic optimization produced by the G-loop. This process takes time (as it is statistical) and thus increases fitness only slowly and gradually. In order to emphasize this point, the resulting fitness is called fitness-to-be. The accuracy of the estimate x (that is, how closely it mimics f over time) is quantified by C. Put differently, C quantifies the level-of-fit that x has to f. In Fig. 15, C is placed close to the grey arrow that points from x to f and that symbolizes estimation. An analogy (from Chapter 14) may be helpful here: one can think of F as the weather (with f a time-varying variable of the weather, such as the temperature at a specific place) and of X as a computer simulation of the weather (with x the simulated temperature). C then quantifies how accurately the temperature f is estimated by x.

It can be shown that, in addition to quantifying estimation, C is a cause, namely of fitness-to-be (Chapter 14, for a maximally simplified variant of the mechanism). Briefly, this is shown by the fact that perturbing C (by making it larger or by making it smaller) will systematically affect fitness-to-be accordingly, though only gradually and slowly. This effect occurs, because the stochastic process that produces large x becomes more effective in enhancing fitness-to-be when x becomes more similar to f (i.e., when C is made larger), and less effective when C is made smaller. Thus, C is a cause of fitness-to-be. Note that C has a double role here: on a short timescale, it quantifies estimation, and on a longer timescale, it is causal. The causal aspect of C can be shown to be strongly emergent (Chapter 14). This means that the causal efficacy of C is not completely given by the causal efficacies of its constituents (including their interactions). Very briefly, this follows from the fact that an indispensable ingredient of the causal efficacy of C consists of slowly accumulating causal effects that are attributable to randomness. Randomness is assumed here to be ontic, that is, uncaused and not a consequence of insufficient knowledge. Importantly, a strongly emergent cause must exist autonomously as a distinct entity (Chapter 14).

The fact that the level-of-fit C has thus become a strongly emergent entity with causal power has several consequences. C quantifies how well x estimates f, which means that the X process has, through x, a minimal form of intentionality: x is about f (see Chapter 10). C being strongly emergent then implies that intentionality is strongly emergent as well. Although the value of x is produced by a conventional process X, and x modulates randomness through conventional causal mechanisms, x has an additional quality—by stochastically affecting fitness-to-be via C—that does not follow completely from its constituents. The latter implies that also x is strongly emergent. With x being strongly emergent, the [goal] of high x is strongly emergent too. This means that high x can be called a proper goal and that the square brackets should be omitted. In other words, the goal-directedness of the G-loop is genuine, because it is strongly emergent. This makes it categorically different from the apparent (‘as if’) goal-directedness that one might ascribe to conventional mechanistic processes.

The behavioural trajectory that an organism follows when cycling through the G-loop depends on x. Following such a trajectory can be regarded as a minimal form of agency (see Chapter 9), which is neither fully deterministic (because of the modulated randomness) nor fully random (because of x). Agency is strongly emergent, because it depends on the strongly emergent goal of high x. In other words, minimal agency consists here of non-deterministic behaviour directed towards an autonomous goal that has emerged—in a strong sense—within the organism itself. Finally, the presence of an autonomous goal within the organism implies that reaching such a goal represents a minimal form of value to the organism. Thus, obtaining high x represents value. Value is strongly emergent as well, because it depends on a strongly emergent goal.

The arguments above claim that the G-loop of Fig. 15 produces strongly emergent, yet minimal forms of goal-directedness, intentionality, agency and value. But philosophy is traditionally not primarily interested in such minimal forms, but rather in the sophisticated, full-blown forms one can find in adult humans and in human society. The question is then how full-blown forms can arise. This can be explained by three different moves away from minimality: firstly, a parsing of the processes X and F, secondly, communicated intentionality, and thirdly, an expanded concept of fitness. These elaborations have been explained before (see Chapters 2–4, 10 and 12) and will, again, be summarized here only briefly.

The basic form of intentionality is produced by x estimating f, but in order to produce x by a process X, more detailed forms of estimation are required. X can only be efficient and effective if it takes the structure of the fitness process F into account. Therefore, the process X has to consist of subprocesses (called X-components below) that each estimate—in a usually complex form of estimation—a presumed subprocess of F (called an F-component below). Then an X-component can be regarded as a component of intentionality that is directed towards the corresponding F-component. An X-component is only useful to the extent that it participates in the overall X process for producing x, because the G-loop of Fig. 15 still uses x. The causal efficacy of an X-component then depends on how well it estimates an F-component (including its role in the F process), but only because of the causal efficacy by which x—via C—produces fitness-to-be. Because the latter causal efficacy is strongly emergent, the causal efficacy of each X-component is strongly emergent as well. In other words, components of intentionality are each strongly emergent. This also applies to the sub-goals, sub-values and meaning they might involve. It is clear that this kind of parsing of X and F will increase complexity, going beyond minimality.

The second move away from minimality involves introducing sophisticated forms of communication, in particular those performed in a cooperative setting. Such a setting makes it advantageous, on average, to communicate intentionality itself. Communicating intentional components requires, in its simplest form, a complex inverting transformation (van Hateren 2019), the accuracy of which is quantified by T. T can be shown to constitute a second strongly emergent cause, which is related to but different from C. The complexity of communicating intentional components—either for external or for internal use—presupposes a complex brain. The emergent cause T corresponds to a distinct entity that is—in contrast to C—spatially localized to the brain (van Hateren 2019 and Chapter 12). It is plausibly sensed as the feeling of consciousness. The content of consciousness at a particular point in time then corresponds to the content of the intentional components that are being prepared for communication at that point in time.

The final move away from minimality involves fitness itself. In its minimal, direct form, it denotes the prospective evolutionary success of individual organisms, in terms of individual survival and reproduction. But fitness is often more complex than that. First of all, it may include indirect effects via helping kin. Helping one's kin increases the chances that traits one shares with kin (because of genetic relatedness) are transferred to the next generation. Fitness that includes such indirect effects is known as 'inclusive fitness' (Hamilton 1964). Inclusive fitness can subsequently be affected by social and cultural factors (e.g., Boyd et al. 2011). However, for fitness that is partially produced by a G-loop as in Fig. 15, it is necessary to redefine fitness itself. This is explained in Chapter 3 and analysed computationally in van Hateren (2015c, see also Appendix A). It is called 'extensive fitness' (which includes inclusive fitness, which itself includes direct fitness). Extensive fitness depends on organisms with similar phenotypes (i.e., similar effective forms) helping each other. Under certain conditions, this can be shown to be more effective for increasing evolutionary success than when organisms with similar genotypes help each other (as in inclusive fitness). Part of heredity is then, in effect, outsourced to the cultural environment. Extensive fitness is facilitated when a population has a shared language, and it naturally leads to abstraction and symbols (Chapter 10).

The three moves summarized above appear to suffice for explaining how full-blown forms of agency, intentionality, consciousness, goal-directedness, values and meaning can emerge. Because these phenomena are all strongly emergent, there is no prospect or claim of reduction here. Although their emergence as such can be understood, detailed theories and explanations can only be developed at the emergent level.

We can now, at last, turn to the dashed boxes and grey arrow in Fig. 15, and explain the labels 'real', 'true' and 'good'. The fitness f of an individual organism is produced, through F, by how the organism confronts and is confronted by the world, and by how its internal structure and state can cope with that world. The term 'world' denotes here that part of reality that affects the individual. Because fitness is a matter of life or death, that is, of existing or not existing, it performs a fundamental reality check. The dashed box labelled 'real' in Fig. 15 can thus be viewed as demarcating that part of metaphysics that is relevant to the individual. When the individual has extensive fitness, the dashed box includes metaphysics that is relevant to the individual's social and cultural community. Because fitness is a forward-looking quantity, and anything in reality may eventually have direct or indirect consequences for fitness when cultures become scientifically and technologically advanced, the box will gradually expand towards including all of reality.

As we have seen above, the organism is driven to estimate its own fitness as accurately as possible, because a more accurate x (i.e., a larger C) will increase fitness-to-be, on average. When intentionality takes an advanced form by including consciousness and a shared symbolic language, 'as accurately as possible' can become 'as truthfully as possible' (Chapter 10). The grey arrow labelled 'true' in Fig. 15 then symbolizes the epistemology of the individual and associated community. It is interesting to connect this to existing theories of truth. Such theories typically focus on either correspondence, pragmatics or coherence (Audi 2011, pp. 286–290). A correspondence theory of truth holds that a proposition is true if it corresponds to a fact in reality. A pragmatic theory holds it as true if the proposition works and is useful when put into practice. And a coherence theory holds it as true if it is consistent with other propositions that are taken to be true. In their pure form, all such theories have problems. The facts of reality are not directly available to the human mind,

some things may happen to work but be false, and basing propositions only on other propositions leads to circularity or to an infinite regress.

Interestingly, the ‘true’ in Fig. 15 combines aspects of correspondence, usefulness and coherence. First, the G-loop only produces evolutionary benefits if X-components (which include mental propositions) correspond to F-components (which are part of reality) in a sufficiently accurate way. Second, increasing x is expected to work, on average, because it correlates with increased fitness f. A high valued and working x is properly called ‘useful’ for the organism, because the value associated with the goal of high x is strongly emergent. Finally, F is first of all a material process, which is coherent by definition: material processes cannot contain internal inconsistencies, they just happen as they do. In contrast, X is not a purely material process, because its X-components have strongly emergent intentionality. But with F coherent, the intentional structure of X must be approximately coherent too if x is to estimate f well. X should use its components in a mostly coherent way. The ‘mostly’ here indicates that the particular X of a particular individual may be partly incoherent and still happen to be effective. But incoherence presumably becomes less frequent in a community-wide system of shared X-components.

Finally, as argued above, obtaining high x is a genuine overall goal of any individual. One might worry that having high x as a fixed goal conflicts with conscious agency (i.e., free will), because the latter seems to imply that an individual could deliberately decide to strive for, say, low x. However, that implication is incorrect and there is no conflict with free will. The X process is a crucial part of agency and consciousness (Chapters 9 and 12). Although the individual is at liberty, to some degree, to decide, out of free will, to change goals and purposes, this can only happen when X changes accordingly in the course of that process. A changed X means a changed way of producing x. As a result, high x will remain the overall goal, no matter what, and high x will still represent value to the individual. Nevertheless, if a change of sub-goals lets x strongly deviate from f, then C is strongly reduced. This would not be sustainable from an evolutionary point of view, because a reduced C leads to a reduced fitness-to-be. Presumably, a range of neural mechanisms must have evolved (by variation and natural selection) that nudge individuals away from sub-goals that are likely to be detrimental in this way. Moreover, such protective mechanisms can be learned and, with extended fitness, can be produced and retained through cultural mechanisms.

The above discussion implies that the dashed box labelled ‘good’ in Fig. 15 demarcates the overall values of the individual, as present in the process X that produces x. High x will usually indicate high fitness. As explained above, human fitness does not rely exclusively on direct or inclusive fitness. The cultural components of extensive fitness can become dominant, and they rely strongly on the individual’s cultural community. Norms arising from the cultural community then produce values—in the X of most of its individual—that will include communal interests, or at least presumed communal interests. The good can, thus, gradually transform into the right, and one can start to speak of an ethics. Individual goals may then become misaligned with communal norms. However, there are many issues and caveats here, discussion of which would go far beyond the scope of this chapter. One complication that should give pause for thought is that the goals and values included in this box are strongly emergent, and, as such, become a distinct part of reality. Then one might argue that they should be included in the box ‘real’ as well, as an emergent part of metaphysics (and of F). This would make the basic picture that is presented here

considerably more complex. Although it would be quite interesting to pursue this further, it would go against a major objective of this particular chapter. The idea here is to make complex things simple first, before even considering to make simple things complex.

Chapter 16

Philosophical problems of consciousness

Philosophical analysis of consciousness has produced a rich literature on actual and potential problems of consciousness. Several major problems—partly derived from formulations by Tye (2017), Searle (2017) and Chalmers (2017)—are analysed below from the perspective of a recently proposed theory of consciousness (Chapter 12) and intentionality (in the sense of 'aboutness', Chapter 10). Section 16.1 summarizes the theory, but the reader is advised to consult the abovementioned chapters as well as the one on strong emergence (Chapter 14). Although the theory is conjectural and requires empirical investigations, I will avoid the many 'mays' and 'mights' that could litter the text below. Instead, it is written as if the theory concerns established facts. But the reader should keep in mind that such is merely a stylistic choice.

A note on terminology and notation may be helpful. The term 'intentional component' is used below for what is called X_i in van Hateren (2019) and X-component in van Hateren (2021a; see Chapter 10). Similarly, 'fitness component' corresponds to F_i (or F-component), and inverted intentional component (which is experienced) corresponds to $\bar{X}_i$. The estimate of an individual's fitness is denoted by x both here and in van Hateren (2019, 2021a), but by f_{est} and $\hat{f}$ in van Hateren (2015a, b). In the latter, the process X that produces the value x is denoted by the form of f_{est} (analogously to a mathematical function that has both a value and a form). A similar notation concerns fitness f_{true} and f (van Hateren 2015a, b), which is denoted by a process F that produces a value f in van Hateren (2019, 2021a).

16.1 Summary of the theory of consciousness

A key feature of any biological organism is its evolutionary fitness f, which is, in the simplest form, its propensity to survive and reproduce—the term 'propensity' indicates that fitness is used here as a forward-looking, predictive factor. However, fitness is often not so simple, because it can include the effects of helping related individuals (which is known as inclusive fitness), as well as social and cultural effects. Under quite general conditions (such as on heredity and variability of traits), fitness differences between individuals lead to evolution by natural selection.

Fitness as defined here acts continuously during the lifetime of any organism. Therefore, each organism typically strives to keep its fitness high during its lifetime, through various mechanisms. Usually, such mechanisms are primarily deterministic, but it is in fact possible to enhance fitness through a remarkable non-deterministic mechanism (Chapter 2). An organism then produces an implicit internal estimate, x, of its own evolutionary fitness, and utilizes that estimate when randomly varying its internal structures (with x and its effects present in a distributed way). The variation is done in the following way: when fitness is estimated to be low, structures are changed with much variability ('desperate times call for desperate measures', if desperate includes undirected), whereas a high fitness estimate

produces little variability (‘never change a winning team’, or at least not much). This mechanism is conjectural but can be shown to be evolvable (see computations in van Hateren 2015a, summarized in Appendix A). The key point here is that evolution can produce estimation as a causal factor: the better the internal estimate of fitness is, the higher the subsequent fitness (denoted by ‘fitness-to-be’) will become—slowly and gradually because the mechanism is stochastical.

Estimation does not exist as a causal factor in abiotic nature; thus, it is a purely biological novelty. It can be regarded as a minimal form of intentionality (Chapter 10). Importantly, this particular form of estimation can be shown to be a strongly emergent cause (Chapter 14). This means that its causal efficacy cannot be explained by any set of micro-causes (essentially because it depends in a cyclical way on the structural effects of noise). As a result, estimation exists in a literal sense, as a distinct and autonomous entity. However, this entity is not well localized, because what is estimated, fitness, is produced by a complex process F with components that are scattered widely throughout the world.

An entity that is well localized to the brain emerges—again in a strongly emergent way—when components of the fitness estimate (called ‘intentional components’ below, see Chapter 10) are being prepared to be communicated to a related organism. If the setting is assumed to be cooperative, the inclusive fitness-to-be of the sender will increase, on average. Preparing to communicate an intentional component by a sender requires—in its most basic form—an inversion, such that it leads to a similar intentional component in the receiver (this depends on the fact that an operation followed by its inverse produces an identity operation). Inversion can be performed by the thalamocortical feedback loop in the mammalian brain, if it is used in a switched, dual-stage way (van Hateren 2019). The first stage produces intentional components, whereas the second stage inverts them through a specific feedback mechanism. Stages are switching continually, at a rate of roughly 10 Hz in the primate brain. Inverted intentional components are either communicated to a partner or are used internally as further input to the thalamocortical loop.

Inverted intentional components are causal factors that can be shown to be both strongly emergent and spatially localized (van Hateren 2019; see also Chapter 12). They produce an entity that is autonomous, distinct, spatially localized to the brain, transient, and strongly emergent. Thus, they appear to mimic a localized material cause (i.e., as if they were a transient material object within the brain) and it is plausible that their presence is sensed, as the feeling of consciousness. Their content equals that of the corresponding intentional components. The total content of consciousness depends on which inverted intentional components are active at any point in time. The unity of consciousness is produced by the fact that all intentional components get their causal efficacy from the causal efficacy of the overall estimate x of fitness f, which are both scalars (and thus unitary).

The above summary focusses on the most basic form of consciousness, which occurs when intentional components are being prepared to be communicated. If this happens internally, it can set up an internal conscious cycle. Consciousness can be produced by perception in an indirect way, such as when a communicated intentional component or a visual scene induces an internal conscious cycle. Complex forms of consciousness can arise in a way that is similar to how complex forms of intentionality can arise (see Chapter 10). The relationship between intentionality and consciousness is indeed a close one: consciousness arises when intentionality is being prepared to be communicated.

16.2 Problems and solutions

16.2.1 Ownership (Tye 2017, p.18)

Problem. Specific subjective experiences are necessarily owned by a specific individual. This makes them different from ordinary physical things. Such things are sometimes owned, but they can still exist if they are not owned. In contrast, subjective experiences cannot exist in an unowned state. Thus, there is a problem if one assumes that phenomenal consciousness is wholly physical.

Solution. Although consciousness is produced by a physical system, it depends on strongly emergent causal factors, specifically estimation and its inversion. Inversion of estimation produces a concrete entity (in the sense of being distinct and spatially localized) that is sensed, but that is indeed not an ordinary physical thing (in the sense of consisting of matter-energy). It is attached inseparably to the neurobiological system that produces and owns it, and it cannot exist in an unowned state.

16.2.2 Perspectival subjectivity (Tye 2017, p.19)

Problem. Phenomenal conscious states are perspectival, but physical states are not. Whereas the latter can be fully understood from a complete description of state and dynamics, the former can only be fully comprehended by having the proper experiential perspective (such as when having a pain, feeling a depression, and having the visual experience of red).

Solution. A conscious state (or, more appropriately, a specific conscious process) consists of components that are the inverse of specific intentional components. Thus, the content of conscious experience depends on the content of intentionality. The latter is a form of estimation produced by the brain. It concerns an estimate of components of the individual's own evolutionary fitness, produced by the individual itself. Hence, it has a subjective perspective. It is not an ordinary physical state, because it is a strongly emergent entity.

16.2.3 Mechanism (Tye 2017, p.20)

Problem. What is the mechanism that produces the what-it's-like feeling? In the natural world, it seems that higher-level states or processes or properties are always grounded in—and are explained by—what is going on at lower neurophysiological or chemical or microphysical levels.

Solution. Ontological reduction may be generally applicable in the abiotic natural world, but not here. The key point here is evolutionary fitness, which confers causal efficacy (through affecting fitness-to-be) to an internal fitness estimate made within the individual. This mechanism depends in an indispensable way on the micro-effects produced by randomness. Estimation is a novel and strongly emergent causal factor that has thus been added to nature (by having evolved through natural selection). The same applies when estimation is inverted for the purpose of communication. The existence of strongly

emergent causes implies that the physical world is not completely causally closed (Chapter 14). Applying philosophical concepts like ‘grounding’ and ‘supervenience’ to the system that produces consciousness is problematic, because ontological randomness across a stretch of time does not correspond to physical ‘facts’ (see Section 14.2.4). Even ‘what is going on’ would be undefined in all its details (unless one tacitly and falsely presupposes determinism).

16.2.4 Duplicates 1 (Tye 2017, p. 21)

Problem. A philosophical zombie is taken to be a perfect material duplicate of a conscious being, except that it completely lacks phenomenal consciousness. Otherwise, it has identical behaviour and identical mental processes. Usually, it is not claimed that such zombies are physically possible (given the features of the world that is), but rather that they are imaginable or logically possible or metaphysically possible. If they are, then consciousness seems separable from its material substrate.

Solution. Consciousness is not separable from its material substrate, and philosophical zombies are therefore not possible—neither with ‘possible’ in the sense of feasible, nor in the sense of conceivable, nor in the sense of non-self-contradictory. Once there is a valid and accepted explanation of how consciousness arises, one is not free any more to use one’s imagination or logic or metaphysical assumptions in a way that conflicts with that explanation. That would amount to basing an argument on false or implausible premisses. Thus, arguments based on philosophical zombies are unlikely to be sound.

16.2.5 Duplicates 2 (Tye 2017, p.22)

Problem. One might simulate the brain in arbitrarily fine detail in another system, such as might be realized by one billion carefully instructed people. Intuitively, one would think that such a system (as a whole) would not be conscious, even if it would perform a perfect simulation.

Solution. Perfect simulation is not possible in this way. The main problem with this kind of simulation is that there can be no internal estimate of fitness (which is required for modulating random structural change in the system) because there is no fitness to estimate. One billion people do not survive and reproduce as a unitary entity (such as by multiplying at once to two or three billion, or by dying at once to zero). Moreover, there is no competition and cooperation with other such entities in a shared environment, nor a well-defined, unitary heredity of structure; therefore, there is no evolution by natural selection. Estimation and intentionality depend on sustained evolution by natural selection. Without real estimation and intentionality, there can be no strongly emergent and distinct entity that is felt as consciousness.

16.2.6 The inverted spectrum (Tye 2017, p.23)

Problem. Suppose that Tom has a very peculiar visual system (perhaps produced by a neurosurgical rewiring at birth), such that he experiences red where others experience green,

and vice versa. But nobody is aware of this difference, because otherwise Tom functions as anybody else. Thus, there is a phenomenal difference without a functional difference. More generally, one may suppose that such phenomenal inversion can occur even in microphysical duplicates.

Solution. Subjective experience is the sensed entity that is produced by inverting intentional components. Hence, the quality of the experience depends on the content of the corresponding intentional components. The content of an intentional component depends on the fitness component that it estimates, including the role that this component has in the process that produces fitness. The colour red has approximately the same fitness associations in a group of culturally and functionally similar people (think of typical red things: strawberries, sunsets, fires, roses, traffic lights, socialism, blood, and so on). Therefore, their intentional components concerning red are roughly similar, and thus is their phenomenal experience of red. This is no different for Tom and his peers. The assumption that Tom is possible (in any sense of the term) is false. Any rewiring at birth would still produce the same phenomenal experience if Tom indeed functions as anybody else.

16.2.7 Transparency (Tye 2017, p.24)

Problem. When attending to a visual experience, one becomes aware of what is seen (such as a particular object and its qualities), but not of the experience as such. Thus, phenomenal consciousness seems to be transparent. Why, then, is it felt?

Solution. It is felt because it equals a distinct, strongly emergent, transient, and spatially localized entity (which is identical to the strongly emergent causes produced by inverting intentional components). The content of this entity is the content of the corresponding intentional components (pointing to a particular object and its qualities). Thus, the entity has no additional content. Having no additional content may be interpreted, incorrectly, as transparency. The interpretation is incorrect, because entity and content are not separable.

16.2.8 Unity (Tye 2017, p.25)

Problem. There is a unity to conscious awareness. The different items that make up a specific conscious experience (e.g., the perceived objects, actions, and sensory impressions in a particular setting) are not experienced as fully separate. Rather, they are perceived as integrated in the whole. Similarly, conscious experiences stay integrated across time. How can that be?

Solution. Consciousness at any time consists of a large set of inverted intentional components. Their content equals the content of the corresponding intentional components. Intentional components estimate fitness components (aspects of an individual's fitness) in such a way that together they produce a unitary (scalar) estimate of the individual's fitness, x. This estimate has strongly emergent causal efficacy, which is, ultimately, the reason why consciousness is felt. The intentional components (as well as their inverted versions) are automatically integrated by x. This is not only true at any point in time, but also across time,

because X, the process that produces x, is maintained across time (even as it changes gradually).

16.2.9 Divided consciousness (Tye 2017, p.27)

Problem. In split-brain patients the corpus callosum is cut (for medical reasons), which drastically reduces the communication between left and right half of the cortex. When conflicting information is presented to the left and right half of a patient's brain, perception seems to occur locally, without being communicated to the other half. Thus, perception is divided. Does such a patient, then, have a split consciousness too?

Solution. Consciousness is conjectured to be produced by the second stage of a dual use of the corticothalamic feedback loop (van Hateren 2019). This second stage inverts intentional components that are presumably produced by a wider loop involving thalamus, cortex and basal ganglia, with important inputs from the upper brain stem. Together these establish x, the (distributed) estimate of an individual's evolutionary fitness. Specific parts of the left or right cortex are then participating in specific intentional components, corresponding, for example, to specific visual perceptions. However, the unity of consciousness itself does not fully depend on the unity of left and right cortex. It also depends on the left-right unity of thalamus, basal ganglia and upper brainstem (as these produce x too). The latter unity remains intact in split-brain patients. Hence, there is no reason to assume that these patients have a fully split consciousness. Moreover, they are still one individual with one fitness, thus they are likely to learn compensating strategies that repair the unity of their x, even if it were compromised initially. This may explain why split-brain patients still feel as one.

16.2.10 Animal consciousness (Tye 2017, p.28)

Problem. How can we decide which other creatures have consciousness?

Solution. For consciousness, creatures need to have evolutionary fitness and need to make an internal estimate of that fitness (which then stochastically drives structural changes in the creature's brain). Moreover, they need to invert components of this estimate, in preparation for internal or external communication. The capacity to communicate intentionality to conspecifics in a cooperative setting (thus typically increasing inclusive fitness) must be present at least, as a basis for more elaborate external or internal communication. Then inverting estimated fitness components produces strongly emergent causes, which constitute the distinct, strongly emergent entity that is felt as consciousness. In summary: in order to have consciousness, creatures must have evolutionary fitness, an internal fitness estimate driving a specific stochastic mechanism, inversions of this estimate's components, and cooperative communication of intentionality with at least one conspecific. These conditions are sufficient, and they are in principle amenable to empirical assessment through neurophysiological and behavioural research.

As a poor man's test of consciousness, one may try to engage a creature in a dialogue of (nonverbal) intentionality, thus establishing some sense of mutual rapport. When establishing a mutual empathic bond is easy (as with mammals and birds), this indicates the

presence of consciousness, and when this seems impossible (as with worms and even with social insects), this indicates its absence.

Note that the above considerations assume a specific mechanism for producing consciousness. Although it is highly specific and may well be the only one capable of producing consciousness, it cannot be ruled out—at this point in time—that alternative mechanisms exist.

16.2.11 Causal efficacy (Searle 2017, p.330)

Problem. One can initiate behaviour by a conscious decision. How is that possible if the brain is fully functioning through neural mechanisms?

Solution. It would indeed not be possible if neural mechanisms were deterministic or at least were ontologically reducible. But the neural mechanism that produces consciousness is neither. Consciousness consists of sets of inverted intentional components that can be used as internal input to produce intentional components, which are subsequently inverted and then used as internal input once more, and so on (van Hateren 2019). The totality of intentional components changes through time in this way, which is equivalent to a change of the structure of the X process that produces x. The latter drives random structural changes in the brain, including ones that affect behavioural dispositions. Behavioural dispositions that produce large x appear to be sticky (because large x produces a low rate of structural change), whereas behavioural dispositions that produce small x appear to be repellent (because small x produces a high rate of structural change). Which particular behavioural dispositions produce small or large x is determined not only by the input to the X process but also by its structure, and is, thus, partially controlled by how consciousness proceeds. Hence, consciousness affects which behavioural dispositions are present. Therefore, it affects the resulting behaviour—albeit by a slow, stochastic process. Instant behavioural decisions need to be prepared in advance, as stored dispositions that can be utilized nonconsciously or preconsciously (see also van Hateren 2015b).

16.2.12 Dancing qualia (Chalmers 2017, p.369)

Problem. Two functionally isomorphic systems must have the same sort of experiences. For example, a conscious biological organism may be gradually replaced, neuron by neuron and cell by cell, by silicon equivalents (this is utterly unrealistic[26], but let us suppose for the sake of argument that it could be done). If one claims that the final, silicon version has different consciousness, or no consciousness at all, then there might be, at some point along the transition, a significant shift in experience. Moving back and forth across this point would produce dancing qualia (qualities of experience). This seems counterintuitive, thus functional isomorphism must imply equal subjective experience.

[26] Neurons work and communicate, as all biological cells, at a molecular level; it is difficult to see how such specific processes could be replaced by processes with a different material basis without producing considerable consequences for fitness. What about the mass, energy requirements and volume of the replacement? Heat dissipation? Functional noise? Structural changeability? Reproduction? Repairability? Molecular defences against disease? And so on and so forth.

Solution. Replacing biology by silicon may not leave fitness intact, that is, the final silicon version may have lost the capacity to reproduce and the propensity to die. If that is so, the silicon version cannot make an internal fitness estimate (other than a fake one that would quickly fall short). Even if the silicon version had fitness (the propensity to survive and reproduce) it would not have inclusive fitness if the system were the only one of its kind. Then inverting intentional components would not be sustainable (for lack of inclusive fitness), and neither would be consciousness. Assuming that fitness is indeed lost, the thought experiment would not show a sharp transition between the presence and absence of consciousness. Rather, the silicon version would gradually loose more and more of its consciousness when it senses—implicitly or explicitly—that it is getting more and more alienated from its former conspecifics. Being indefinitely alone in the world, without any prospect of a meaningful future, is not consistent with sustaining consciousness.

16.2.13 Machine consciousness

Problem. Can machines become conscious?

Solution. Short answer: no, unless machines become alive first; but it is doubtful (or at least a definitional issue) whether one could still call such a living system—an organism—a machine. Long answer: consciousness arises in a system when intentionality is transformed such that it can be communicated (externally or internally) and thus can increase inclusive fitness. Intentionality is a strongly emergent phenomenon that depends on a fitness estimate that modulates random structural change of the system. This mechanism is only sustainable when the fitness estimate accurately estimates a real fitness, with real reproduction (because the exponential growth of reproduction is needed in order to compensate for the inefficiency of random structural change; this ultimately depends on evolution by natural selection). Fitness, random structural change, and evolution by natural selection are defining features of life. Therefore, a system needs to be alive in order to have intentionality and subsequently consciousness. Whether a living system could be called a machine is debatable. In any case, building such a system would be risky, because it would try to replicate without bounds, and would thus compete with humans and other biological life forms.

16.2.14 How could having consciousness produce evolutionary benefits?

Problem. Apparently, having consciousness is an evolved property in some species. If so, which evolutionary advantages would it confer on these organisms?

Solution. Consciousness is not a trait that can be separated from the mechanism that produces it as a strongly emergent entity. Thus, the evolutionary advantages of consciousness are equal to the evolutionary advantages of this mechanism. The mechanism is the transformation of intentional components into a form that can be communicated, at the very least to conspecifics that are inclined to cooperate. Communication of intentionality will then increase inclusive fitness, on average. Therefore, this transformation is evolvable, and the strong emergence of consciousness is the automatic consequence. Note that this does not make consciousness an epiphenomenon, because it is identical to the occurrence of the transformation. The accuracy of the transformation is a strongly emergent entity with

causal power, and is, thus, by no means an epiphenomenon. In summary, the evolutionary benefits of consciousness are identical to the evolutionary benefits of transforming intentional components for communication. The latter enhances inclusive fitness, on average.

16.2.15 Wouldn't a fully non-communicative species still benefit from experiencing pain?

Problem. If consciousness is, in its most basic form, communicative rather than perceptive, wouldn't this imply that a species that has no use for communicating intentionality has no consciousness? But wouldn't experiencing, such as experiencing pain, provide evolutionary benefits anyway?

Solution. If a species lacks the capacity to communicate intentionality, it has indeed no subjective experience. Stimuli or internal states that would indicate harm can then still lead to behaviour that alleviates the problem, but there would be no associated subjective experience. This is so, because such a species lacks a neuronal system that transforms intentional components for communication to others or for further internal processing. Experiencing is inseparably coupled to this transformation (as a strong emergent), and talk about the benefits of the experience as such makes no sense. Any benefits must arise from the prospective communication to others, because benefits depend on the parts of inclusive fitness that go beyond direct (individual) fitness.

16.2.16 Consciousness and quantum physics both seem weird. Is there a link?

Problem. Is there a link between consciousness and quantum weirdness, such as entanglement and wave function collapse upon observation?

Solution. Indirectly. The theory of consciousness and intentionality depends on randomness that is ontological (thus 'out there' and not just a consequence of insufficient knowledge). The source of such randomness is thermal in practice, specifically in the form of random fluctuations of the fairly small number of molecules that are typically involved in (neuro)physiological processes. Such molecular randomness may ultimately depend on quantum randomness, because nonlinear dynamical systems can amplify submicroscopic fluctuations to microscopic and macroscopic ones. Quantum randomness appears to be fundamental. But apart from ontic randomness, there does not seem to be a link between other forms of quantum weirdness and consciousness.

A potential issue here is that the correct interpretation (or foundation) of quantum physics is not yet clear, with some interpretations seemingly suggesting determinism. If the theory of consciousness discussed here acquires empirical support, then full determinism becomes less tenable. If, on the other hand, a fully deterministic physics acquires empirical support, then this particular theory of consciousness becomes less tenable.

16.2.17 Could mind and consciousness be uploaded to a computer?

Problem. If one assumes that mind and consciousness are produced by some specific kind of information processing in the brain, shouldn't it be possible to upload the relevant information to a computer, and then simulate or emulate consciousness?

Solution. No, this is not possible. Consciousness is not produced by a specific kind of information processing. Rather, it requires a physical body that participates in sustained evolution by natural selection and that incorporates a causally effective internal estimator of its evolutionary fitness (see also the discussion of the 'brain in a vat' thought experiment in Section 10.5.2). To the extent that neural processing can be described as information processing, this always concerns meaningful information. Such information is necessarily about something, and thus depends on intentionality. The assumption that information processing produces intentionality and consciousness, and can be used for explaining them, is viciously circular.

16.3 Conclusion

All problems discussed above have a clear solution if the proposed theory of consciousness turns out to be correct.

Epilogue

Having reached this point, it may be a good moment to look back at what has been presented in this book. Where has it brought us, what needs to be done further, and what are the perspectives?

To the author, the main surprise of this line of research has been the realization that mind presumably presupposes life, and that both have non-standard causal properties. This is not what I expected when I started this project. My basic view then was that only the brain was special, possibly in terms of an ingenious kind of complexity, since it produces the enigma of consciousness. In contrast, the phenomenon of life seemed basically solved in principle, as a material process that uses a genetic code and a huge range of molecular control circuits. It would still take the scientific community a long time and a lot of hard work to figure it out in detail, but it would not produce anything other than a model of a highly complex chemical machine. Now it seems that this view needs revision. Both life and mind presumably give rise to a strongly emergent form of causation that utilizes noise rather than that it is fully dependent on fundamental laws. This was the first surprise.

The second surprise was the nature of consciousness, which I presumed to be perceptual—as would befit my background in visual neuroscience. But the clearest way to produce something of which the presence may be felt, as the feeling of consciousness, requires a special form of communication that transfers components of intentionality. Then the fundamental nature of consciousness is communicative, rather than perceptual. Perceptual consciousness is then derived, through development and learning, from the more basic, communicative kind. With hindsight, this makes sense, but it is not where I—and presumably most of those studying the neural basis of consciousness—would be inclined to start.

As further results of the overall theoretical effort, I found that there are literally existing forms of agency, goal-directedness, value and meaning. This was again not anticipated, since genuine goals and values are absent from the natural world as studied by natural scientists. Though this absence remains true of the abiotic parts of the world, it is presumably not true of life and mind. It was a surprise to find this possibility even for forms of life that do not have a mind.

The major caveat here is that the theory requires empirical substantiation (see van Hateren 2015e, 2019, and Chapter 13 for some suggestions for testing it). As it stands, it is a working hypothesis that is unproven. It may be wrong. But its explanatory power is quite extraordinary, across many different fields. I suspect that the situation here is analogous to when the concept of atoms arose in the second half of the 19th century. The assumption that all matter consists of atoms could explain a large number of physical phenomena, but there was no direct proof of atoms. Their existence remained in doubt by many physicists for a considerable time. Only in the first half of the 20th century, combinations of theory and observations (such as of Brownian motion and scattering experiments) made the idea of atoms well established. But it took until the late 20th century before atoms could be visualized directly.

If there is a viable theory of agency, intentionality and consciousness, one may wonder whether this might be applied in technology. Unfortunately, this seems far from easy. The

main issue is that the key part of the required mechanism, a self-modulated stochastic variation of structural change, is a very inefficient process. Random changes only rarely work well, even if the randomness were constrained (e.g., to certain sections of form-space). The only way to make the mechanism work in a sustainable way is by compensating this inefficiency with a strongly expansive nonlinearity, such as the exponential growth in numbers produced by self-replication. In other words, what is needed is either self-replication combined with self-modulated structural change—that is, life and evolution by natural selection—or an equivalent. It is not clear if such an equivalent could be created. If not, then technology can only produce genuine forms of agency, intentionality and consciousness if it implements self-replication and self-modulated structural change. Then, such systems would be literally alive. But even if technically feasible, creating such new forms of life would be highly dangerous: they would tend to replicate beyond bounds and compete with existing life. They should do so, because they must follow their own intrinsic goals and interests. There is no reason why the latter would stably align with those of humans.

The conclusion here is that the prospects for agency, intentionality and consciousness seem unfavourable in human-serving technology. Without intentionality and consciousness, mimicked cognitive tasks—such as machine translation of text from one language into another—will keep producing disturbing errors, occasionally, because of a fundamental lack of understanding about the world. Moreover, I suspect that having intentionality is a minimal requirement for having Artificial General Intelligence. The prospects for the latter are then not so good either.

Even if there may be no direct technological perspectives for the theory, there are many applications in other fields. If corroborated by experiments, the theory can guide thinking about evolution, life, biological function and biological meaning. When applied to the human mind and consciousness, it can explain parts of human psychology that are otherwise difficult to understand (for a first attempt see Chapter 13). It may help to explain pathologies of consciousness and of sense of purpose. Moreover, it readily explains why human thinking and behaviour often seem irrational when they become mixed up with perceived values and goals that conflict with reality (see Chapter 15 for how reality, the understanding of reality, and the valuation of reality are interconnected). Understanding the basic mechanisms of such phenomena may lead to effective and transparent compensatory strategies.

Appendix A:
Summaries of computational simulations

Several of the mechanisms that were discussed qualitatively in this book were investigated quantitatively by computational (van Hateren 2015a, c) and theoretical (van Hateren 2015f) means. The kind of analysis that was done and several major results are summarized here. Details and additional results can be found in the original publications. Reading this section is not necessary for understanding the material in this book, so it could be skipped.

The mechanisms of Figs. 2 and 3, as well as variants, are investigated computationally in van Hateren (2015a). The G-loop can act on various timescales, in particular those of evolution, behaviour, and neural processing. This summary focusses on agency, where only changes within and restricted to an organism's lifetime are relevant. Then the x of Fig. 3 modulates the rate of structural and behavioural change on that timescale only, without hereditary transfer. Changes may occur directly, but also more sophisticated variants are studied, such as when x does not immediately drive such changes, but only after the possible effects of changes are simulated first within the organism.

Variants are simulated using simplified model systems, with organisms that have a limited lifespan and that acquire hereditary or behavioural changes along a single dimension. Along the same dimension, the environment varies in time, unpredictably across a wide range of timescales. The fitness f quantifies the expected reproductive rate of each organism. It is the outcome of a function F that quantifies by how much the momentary environment differs from the optimal one for the momentary combination of behavioural disposition and heredity (with the latter fixed for a given organism). For readers consulting van Hateren 2015a, a note on notation may be helpful. In that study, fitness is denoted by a function *f*, with a function form (i.e., the way in which it processes its inputs) and a function value (i.e., the single-valued outcome of the function). In this book, the function form is denoted by F and the function value by f. In several other early studies, *f* is written as f_{true}. In van Hateren 2015a, a fitness estimate is denoted by a function $\hat{f}$, which has a form and value that is denoted by X and x, respectively, in this book. In other studies, $\hat{f}$ is written as f_{est}.

In the simulations, two populations share an environment with limited resources. Thus, organisms must compete for resources in order to be able to reproduce. The two populations consist of organisms that differ in a specific way between the populations. For example, one population consists of organisms with a behavioural variability that is fixed to an optimal (i.e., most competitive) level. The other population then consists of organisms with a G-loop and an x (as an estimate of f) that modulates behavioural variability. The two population sizes start out equal, but fluctuate because the environment varies over time (across a wide range of timescales) and each organism varies randomly. Simulations are repeated with different realizations of how the environment varies over time and different realizations of the random variations of each organism. Invariably, the population that consists of organisms lacking an x-driven G-loop becomes extinct. Such organisms are less capable of adapting to environmental change than organisms with an x-driven G-loop. The simulations show that fitness f is effectively increased by having this particular mechanism, or various

variants of it. These computational results have been corroborated by mathematical analysis (van Hateren, 2015f).

The evolvability of extensive fitness was investigated with models containing the bare minimum for producing extensive fitness (van Hateren 2015c). These models utilize behavioural plasticity, but heredity is only genetic. That is, they do not contain explicit social or cultural transmission, and also no explicit psychological mechanisms. The most basic model has only direct fitness (Fig. 4, pathway 1). Fitness is then modelled as a simple reproductive rate of each individual. For inclusive fitness, pathway 2 (Fig. 4) is added in the form of a fitness multiplier. This factor increases the fitness of an individual if they help others of similar hereditary type (i.e., with similar genes, such as present in kin). This is called h-helping. Helping and being helped is more likely when the group that matches an individual's heredity is large (as each individual has then more opportunities for helping), hence the fitness multiplier increases with group size. Simulations use two populations, consisting either of individuals without h-helping (only direct fitness) or of individuals with h-helping (inclusive fitness, pathways 1 and 2 together). As expected, simulations with different realizations of the environmental time course invariably show that the population of individuals without h-helping is driven to extinction.

As an alternative to h-helping, a fitness multiplier was used that increases an individual's fitness if they are involved in helping based on phenotypic similarity (called p-helping). Phenotypes depend on both heredity and behaviour. Heredity can only change across generations, and is fixed for a particular individual. Behaviour can change dynamically within an individual's lifetime. Thus, individuals belonging to a phenotypically similar group need not have similar heredity. The resulting p-helping directly implements a simple form of pathway 4 (Fig. 4). It also produces pathway 3, indirectly, because of the fitness multiplier. When an individual has acquired a certain phenotype, they contribute to the size of the corresponding phenotypic group. Thus, they increase the fitness of all group members, because larger groups produce more helping. Therefore, the group effectively attracts other individuals as they vary their phenotype behaviourally. Their x quantifies this attractiveness, as depending on phenotype and environmental state. In effect, then, the individual induces others to get a similar phenotype. As stated above, this model is the minimum needed to produce this effect. It could be amplified by adding explicit social and psychological mechanisms.

Individuals with p-helping but no h-helping (i.e., using pathways 1, 3 and 4) outperform individuals with h-helping but no p-helping (i.e., using pathways 1 and 2). Simulations invariably show that populations with h-helping are driven to extinction if they share resources with populations with p-helping. At first sight, this is a surprising result, because p-helping seems inferior to h-helping for keeping beneficial genes in the gene pool. However, evolution has two sides: one is that good heredity is retained, but the other is that organisms interact successfully with their environment. It is the phenotype, not the genotype, that confronts the environment. It can be shown theoretically (Appendix of van Hateren 2015c) that h-helping and p-helping counterbalance their relative strengths and weaknesses. They should perform equally well if all else were equal. But all else is not equal, because of the fitness multipliers that implement benefits for h-groups or p-groups. Phenotypes can adapt more quickly than genotypes to a changing environment. Therefore, groups of individuals with similar phenotypes can become larger than groups of individuals with similar heredity. Then p-helping can outperform h-helping, on average, since the

fitness multiplier increases with group size. Obviously, there are many potential complications here, because p-helping is cognitively more demanding than h-helping and more vulnerable to cheating (as is discussed in the literature on altruism, e.g., Rand and Nowak 2013).

Appendix B:
Examples of minimal intentionality

Applying the theory of intentionality of Chapter 10 (up to Section 10.3) to a particular case involves several steps. First, one needs to identify the relevant F-component, that is, the subprocess of F that is involved. Second, one needs to assess whether this F-component has a corresponding X-component. Such assessment is ideally an empirical one, by experimentally investigating X. In the absence of such empirical data, evolutionary arguments and common sense can often produce a plausible assessment. If one concludes that there is no corresponding X-component, then there is no intentionality, nor content (other than in the eye of the beholder). If one concludes that a corresponding X-component is likely, then one can make an educated guess of its structure from the assumption that the fitness estimate x (the outcome of X) has evolved to be a reasonable estimate of the actual fitness f (the outcome of F). The inferred structure of the X-component then establishes what it refers to and what role it has in X, that is, it specifies its content. Note that intentionality and content can only be ascribed to an X-component as a whole; trying to ascribe clear-cut micro-content to the micro-constituents of an X-component runs into the same (usually unsolvable) problems as trying to ascribe a clear-cut role to a single neuron in a large neural network.

We will now apply the above procedure to several examples of minimal intentionality that have been discussed in the literature. One such example is bacterial magnetotaxis (Millikan 1984). The discussed magnetotactic bacteria contain tiny magnets that help them to align to the geomagnetic field, and to swim downward, away from (detrimental) oxygen-rich water. Bacteria in the northern and southern hemispheres incorporate the magnets in opposite orientations with respect to swimming direction, which complies with the fact that the geomagnetic polarity in these hemispheres is reversed with respect to the sea floor. This system raises two questions. First, can one say that the orientation of the magnet represents the external world, potentially erroneously (e.g., in a bacterium from the northern hemisphere that has been transposed to the southern hemisphere)? And second, what would it then represent: the polarity of the geomagnetic field, or the direction of lower oxygen?

Applying the theory to this case goes as follows. The relevant F-component here consists of two parts contributing to F: first, the ambient oxygen level and how that affects the bacterium, and, second, the mechanism that keeps the bacterium out of oxygen-rich water. The latter mechanism consists of the local geomagnetic field, the bacterial system producing alignment, and the bacterial system that lets the bacterium move accordingly. It is likely that the bacterium utilizes some estimate of the ambient oxygen level for its X, because this level is typically highly important for its fitness and may vary. Thus, minimal intentionality is not absent here. The corresponding X-component needs to estimate how well the bacterium is kept out of oxygen-rich water. An obvious way to do that is to monitor the current oxygen level, for example through a physiological variable that is strongly influenced by (detrimental) oxygen-rich conditions. Such a measure would be readily evolvable, also because there may be more than one reason (apart from depth) why local oxygen levels would vary. When X, which includes this X-component, produces a small x,

then the bacterium needs to increase the variability of its form. For example, it may increase the variability of its behavioural dispositions to move in any direction, which may increase the chances that it eventually reaches water with lower oxygen levels. Alternatively, it may increase the variability of its metabolic pathways, which can increase the chances that it finds a way to cope with higher oxygen levels (assuming that it is already using any determinate mechanism that has been evolved specifically for coping with this situation).

One might think that X could have evolved a system that specifically monitors the veracity of the bacterial geomagnetic subsystem, instead of or in addition to an oxygen monitor. However, that seems implausible. The veracity of the geomagnetic subsystem is unlikely to vary during the lifespan of an individual bacterium, given the robustness of the geomagnetic field and the robustness of the response of magnets to magnetic fields. Thus, there is probably no evolutionary benefit from having X monitor the performance of this system. When a bacterial species migrates gradually from northern to southern hemisphere, the magnetic subsystem may become realigned, but only through evolutionary change. However, this does not involve minimal intentionality in the sense discussed here.

In conclusion, the geomagnetic subsystem does not, by itself, represent anything, because it belongs to an F-component and not to an X-component. It is not an origin of minimal intentionality. Nevertheless, it may be regarded as an intentionality aid, that is, as a factor that shapes the intentionality of an X-component (the literature sometimes ascribes 'derived intentionality' to such intentionality aids, in contrast to 'original intentionality', which is here ascribed exclusively to X). This is so because the geomagnetic subsystem is part of an F-component that is targeted by an X-component that estimates how well the bacterium is kept out of oxygen-rich water. The latter is, then, the content (in a minimal sense) of that X-component, that is, what it represents (in a weak, minimal sense). If the geomagnetic subsystem is not working as implicitly expected (such as in a hemispherically misplaced bacterium), this affects how accurately the X-component can contribute to estimating fitness. The X-component implicitly expects (through X and x) that the geomagnetic subsystem is working to keep the bacterium out of oxygen-rich water. In other words, the geomagnetic subsystem can produce errors from the point of view of X. The two questions above are thus answered as follows: the orientation of the magnet can indeed represent, to X, external reality in an erroneous way, and, to X, it represents the direction of lower oxygen, not the polarity of the geomagnetic field. However, this is only a derived form of representation, as it depends on having a role in the process that is estimated by X, rather than on being part of the estimator itself.

A second example in the literature on minimal intentionality is the bug-detecting system of frogs and toads (Millikan 1984; Neander 2017). It is assumed here that this is at most a form of minimal intentionality, and not the type of intentionality that requires a mind of some kind (assuming that frogs do not have minds, which is actually an open question at this point in time). Bug-detection in these animals involves a visual subsystem that responds to nearby, small objects that cross the visual field within a defined range of speeds. The animal responds by attempting to catch the object with its tongue. Usually, the object is an edible insect, but the animal responds to other small objects as well, for example objects that are tossed by an experimenter. Again, there are two questions. Does the system involve minimal intentionality, and, if so, what does it represent: 'edible bugs' or 'moving objects'?

The F-component in this case is the system that keeps the animal nourished by letting it catch and eat edible insects. This system consists of environmental factors (such as potential

moving prey and visual circumstances) and organismal factors (such as parts of the visual and motor systems). The fitness consequences of this F-component are large and may vary, which means that it is likely that a corresponding X-component exists. This X-component then represents how well the bug-catching and bug-eating system works: if most catches fail, or if most prey is inedible or poisonous (as evaluated by the same X-component), this should lead to a low value of x. Subsequently, the animal should change the variability of its behavioural dispositions (e.g., change posture more often, change hunting spots more often, and so on). On average, higher variability increases the chances of higher fitness (again assuming that the repertoire of evolved and learned strategies for directed improvement has already been utilized).

Part of the F-component are 'bug-detecting' neurons, which respond to suitable visual motion. Such neurons do not have minimal intentionality by themselves, because they are part of an F-component and not of an X-component. The performance of such neurons is presumably quite robust, and could be regulated by conventional cybernetic control (for example, by recalibrating a neuron if it has drifted away from its proper operating range). Therefore, it is unlikely that their performance is explicitly monitored by an X-component. Nevertheless, the X-component identified above estimates the F-component to which these neurons belong, and implicitly expects them to operate in a certain way. They operate correctly when they respond to the right kind of visual motion, irrespective of what causes it. They only operate erroneously, from the point of view of X, when they respond in a way that is different from what X expects (e.g., when they fail to respond at all to suitable movement). But the X-component itself is presumably more discriminating: it checks if the caught objects are indeed edible. In other words, the minimal intentionality of this X-component is directed towards 'edible bugs' and not towards 'moving objects'. However, the part of the X-component that evaluates the edibility of a caught object is separate from the neural system that detects visual motion. Thus 'bug-detecting' neurons represent, to X, just 'moving objects', not 'edible bugs'. Note that this kind of representation is again not an original, but a derived one, as it depends on the original intentionality of an X-component.

References

Ardourel V, Guay A (2018) Why is the transference theory of causation insufficient? The challenge of the Aharonov-Bohm effect. Stud Hist Philos M P 63:12–23

Audi R (2011) Epistemology: a contemporary introduction to the theory of knowledge (3rd edition). Routledge, New York

Baars BJ (1988) A cognitive theory of consciousness. Cambridge University Press, Cambridge

Balázsi G, van Oudenaarden A, Collins JJ (2011) Cellular decision making and biological noise: from microbes to mammals. Cell 144:910–925

Barbieri M (2008) Biosemiotics: a new understanding of life. Naturwissenschaften 95:577–599

Bateson G (1979) Mind and nature: a necessary unit. E. P. Dutton, New York

Baumeister RF (2010) The self. In: Baumeister RF, Finkel EJ (eds) Advanced social psychology. Oxford University Press, New York, pp 139–175

Baumeister RF, Leary MR (1995) The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychol Bull 117:497–529

Bechtel W, Abrahamsen A (2005) Explanation: a mechanistic alternative. Stud Hist Philos Biol Biomed Sci 36:421–441

Becker M, Vignoles VL, Owe E, et al. (2014) Cultural bases for self-evaluation: Seeing oneself positively in different cultural contexts. Pers Soc Psychol Bull 40:657–675

Bedau MA (1997) Weak emergence. In: Tomberlin J (ed) Philosophical perspectives: mind, causation, and world, vol. 11. Blackwell, Malden MA, pp 375–399

Bedau MA (2007) What is life? In: Sarkar S, Plutynski A (eds) A companion to the philosophy of biology. Blackwell, New York, pp 455–471

Bedau MA (2008) Is weak emergence just in the mind? Minds Mach 18:443–459

Bedau MA, Humphreys P (2008) Emergence: contemporary readings in philosophy and science. MIT Press, Cambridge

Bell G (2008) Selection: the mechanism of evolution (2nd ed). Oxford University Press, Oxford

Bell G (2010) Fluctuating selection: the perpetual renewal of adaptation in variable environments. Phil Trans R Soc B 365:87–97

Benner SA (2010) Defining life. Astrobiology 10:1021–1030

Berry MV (1988) The electron at the end of the universe. In: Wolpert L, Richards A (eds) A passion for science. Oxford University Press, Oxford, pp 39–51

Bigelow J, Pargetter R (1987) Functions. J Philos 84:181–196

Bird A (2016) Overpowering: how the powers ontology has overreached itself. Mind 125:341–383

Block N (1995) On a confusion about a function of consciousness. Behav Brain Sci 18:227–287

Boorse C (1976) Wright on functions. Philos Review 85:70–86

Boorse C (1977) Health as a theoretical concept. Philos Sci 44:542–573

Boyd R, Richerson PJ, Henrich J (2011) The cultural niche: why social learning is essential for human adaptation. Proc Natl Acad Sci USA 108: 10918–10925

Brandom RB (2008) Between saying and doing: towards an analytic pragmatism. Oxford University Press, Oxford

Brembs B (2011) Towards a scientific concept of free will as a biological trait: spontaneous actions and decision-making in invertebrates. Proc R Soc B 278:930–939

Buller DJ (1998) Etiological theories of function: a geographical survey. Biol Philos 13:505–527

Burgin M (2010) Theory of information: fundamentality, diversity and unification. World Scientific, Singapore

Call J, Tomasello M (2008) Does the chimpanzee have a theory of mind? 30 years later. Trends Cogn Sci 12:187–192

Carver CS, Scheier MF (1982) Control-theory: a useful conceptual framework for personality-social, clinical, and health psychology. Psychol Bull 92:111–135

Carver CS, Scheier MF (2002) Control processes and self-organization as complementary principles underlying behaviour. Pers Soc Psychol Rev 6:304–315

Chalmers DJ (2002) On sense and intension. Noûs 36(suppl. s16):135–182

Chalmers D (2006) Strong and weak emergence. In: Clayton P, Davies P (eds) The re-emergence of emergence: the emergentist hypothesis from science to religion. Oxford University Press, Oxford, pp 244–254

Chalmers D (2017) Naturalistic dualism. In: Schneider S, Velmans M (eds) The Blackwell companion to consciousness (2nd edition). Wiley-Blackwell, Chichester, pp 363–373

Chandler D (2007) Semiotics: the basics. Routledge, Oxon

Claidière N, Scott-Phillips TC, Sperber D (2014) How Darwinian is cultural evolution? Phil Trans R Soc Lond B 369:20130368

Cleland CE, Chyba CF (2002) Defining 'life'. Orig Life Evol Biosph 32:387–393

Cleland CE (2012) Life without definitions. Synthese 185:125–144

Craver CF (2007) Explaining the brain: mechanisms and the mosaic unity of neuroscience. Oxford University Press, Oxford

Crocker J, Park LE (2004) The costly pursuit of self-esteem. Psychol Bull 130:392-414

Cummins R (1975) Functional analysis. J Philos 72:741–765

Cummins R (2002) Neo-Teleology. In: Ariew A, Cummins R, Perlman M (eds) Functions: new essays in the philosophy of psychology and biology. Oxford University Press, New York, pp 164–174

Damasio A (1999) The feeling of what happens: body and emotion in the making of consciousness. Hartcourt Brace & Company, New York

Darwin C (1859) On the origin of species by means of natural selection. John Murray, London

Davies PS (2009) Subjects of the world: Darwin's rhetoric and the study of agency in nature. University of Chicago Press, Chicago

Deacon TW (1997) The symbolic species: the co-evolution of language and the brain. Norton, New York

Deci EL, Ryan RM (2000) The "what" and "why" of goal pursuits: Human needs and the self-determination of behaviour. Psychol Inq 11:227–268

Dehaene S (2014) Consciousness and the brain: deciphering how the brain codes our thoughts. Viking Press, New York

Dehaene S, Sergent C, Changeux J-P (2003) A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proc Natl Acad Sci USA 100:8520–8525

Dennett DC (1984) Elbow room: the varieties of free will worth wanting. Clarendon Press, Oxford

Dennett DC (1989) The intentional stance. Bradford Books, Cambridge MA

Dennett DC (1991a) Real patterns. J Philos 88:27–51

Dennett DC (1991b) Consciousness explained. Little, Brown and Company, Boston

Dennett D (1992) The self as a centre of narrative gravity. In: Kessel FS, Cole PM, Johnson DL (eds) Self and consciousness: multiple perspectives. Erlbaum Associates, Hillsdale, pp 103–115

Dennett DC (2009) Intentional systems theory. In: McLaughlin BP, Beckermann A, Walter S (eds), The Oxford handbook of philosophy of mind. Oxford University Press, Oxford, pp 339–350

Di Paolo EA (2005) Autopoiesis, adaptivity, teleology, agency. Phenomenol Cogn Sci 4:429–452

Dretske FI (1981) Knowledge and the flow of information. The MIT Press, Cambridge MA

Edelman GM, Tononi G (2000). A universe of consciousness: how matter becomes imagination. Basic Books, New York

Eigen M (1971) Selforganization of matter and evolution of biological macromolecules. Naturwissenschaften 58:465–523

El-Hani CN, Queiroz J, Emmeche C (2006) A semiotic analysis of the genetic information system. Semiotica 160:1–68

Faisal AA, Selen LPJ, Wolpert DM (2008) Noise in the nervous system. Nature Rev Neurosci 9:292–303

Feinberg TE, Mallatt J (2016) The *nature* of primary consciousness. A new synthesis. Conscious Cogn 43:113–127

Fekete T, Edelman S (2011) Towards a computational theory of experience. Conscious Cogn 20:807–827

Fodor JA (1990) A theory of content and other essays. Bradford/The MIT Press, Cambridge MA

Frege G (1892) Über Sinn und Bedeutung. Z Philos Philos Krit, N.F. 100:25–50

Galhardo RS, Hastings PJ, Rosenberg SM (2007) Mutation as a stress response and the regulation of evolvability. Crit Rev Biochem Mol Biol 42:399–435

Garson J (2012) Function, selection, and construction in the brain. Synthese 189:451–481

Gibb S, Hendry RF, Lancaster T (eds) (2019) The Routledge handbook of emergence. Routledge, London

Gillett C (2016) Reduction and emergence in science and philosophy. Cambridge University Press, Cambridge

Glennan SS (1996) Mechanisms and the nature of causation. Erkenntnis 44:49–71

Glennan S (2017) The new mechanical philosophy. Oxford University Press, Oxford

Glimcher PW (2011) Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc Natl Acad Sci USA 108:15647–15654

Gordus A, Pokala N, Levy S, Flavell SW, Bargmann CI (2015) Feedback from network states generates variability in a probabilistic olfactory circuit. Cell 161:1–14

Gray JM, Hill JJ, Bargmann CI (2005) A circuit for navigation in *Caenorhabditis elegans*. Proc Natl Acad Sci USA 102:3184–3191

Graziano MSA, Kastner S (2011) Human consciousness and its relationship to social neuroscience: a novel hypothesis. Cogn Neurosci 2:98–113

Grice HP (1957) Meaning. Philos Rev 66:377–388

Griffiths PE (1993) Functional analysis and proper functions. Brit J Philos Sci 44:409–422

Guay A, Sartenaer O (2016) A new look at emergence. Or when *after* is different. Eur J Philos Sci 6:297–322

Hamilton WD (1964) The genetical evolution of social behaviour. I & II. J Theor Biol 7:1–52

Heft H (2013) An ecological approach to psychology. Rev Gen Psychol 17:162–167

Heppner WL, Kernis MH (2011) High self-esteem: Multiple forms and their outcomes. In: Schwartz SJ, Luyckx K, Vignoles VL (eds) Handbook identity theory and research. Springer, New York, pp 329–355

Hills TT (2006) Animal foraging and the evolution of goal-directed cognition. Cogn Sci 30:3–41

Hoffmeyer J (2008a) The semiotic niche. J Mediterr Ecol 9:5–30

Hoffmeyer J (2008b) Semiotic scaffolding of living systems. In: Barbieri M (ed) Introduction to biosemiotics. Springer, Dordrecht, pp 149–166

Hoffmeyer J (2012) The natural history of intentionality: a biosemiotic approach. In: Schilhab T, Stjernfelt F, Deacon T (eds) The symbolic species evolved. Springer, Dordrecht, pp 97–116

Horgan T, Graham G (2012) Phenomenal intentionality and content determinacy. In: Schantz R (ed) Prospects for meaning. De Gruyter, Berlin, pp 321–344

Hulswit M (2002) From cause to causation: a Peircean perspective. Kluwer, Dordrecht

Humphreys P (2016) Emergence: a philosophical account. Oxford University Press, Oxford

Hüttemann A, Papineau D (2005) Physicalism decomposed. Analysis 65:33–39.

Hutto DD, Satne G (2015) The natural origins of content. Philosophia 43:521–526

Illari P, Russo F (2014) Causality: philosophical theory meets scientific practice. Oxford University Press, Oxford

Illari PM, Williamson J (2012) What is a mechanism? Thinking about mechanisms *across* the sciences. Eur J Philos Sci 2:119–135

Jacob P (2014) Intentionality. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (winter 2014 edition). https://plato.stanford.edu/archives/win2014/entries/intentionality/

Kauffman SA (2000) Investigations. Oxford University Press, New York

Kauffman SA (2003) Molecular autonomous agents. Phil Trans R Soc Lond A 361:1089–1099

Kim J (2006) Emergence: core ideas and issues. Synthese 151:547–559

Kim J (2010) Essays in the metaphysics of mind. Oxford University Press, Oxford

Kiviet DJ, Nghe P, Walker N, Boulineau S, Sunderlikova V, Tans SJ (2014) Stochasticity of metabolism and growth at the single-cell level. Nature 514:376–379

Koch C, Massimini M, Boly M, Tononi G (2016) Neural correlates of consciousness: progress and problems. Nature Rev Neurosci 17:307–321

Kriegel U (2013) The phenomenal intentionality research program. In: Kriegel U (ed) Phenomenal intentionality. Oxford Scholarship Online, https://doi.org/10.1093/acprof:oso/9780199764297.001.0001

Kriegel U (2016) Brentano's mature theory of intentionality. J Hist Anal Philos 4:1–15 https://doi.org/10.15173/jhap.v4i2.2428 (open access)

Kripke S (1980) Naming and necessity. Harvard University Press, Cambridge MA

Kull K (2009) Biosemiotics: to know, what life knows. Cybernetics and Human Knowing 16:81–88

Kull K, Deacon T, Emmeche C, Hoffmeyer J, Stjernfelt F (2011) Theses on biosemiotics: prolegomena to a theoretical biology. In: Emmeche C, Kull K (eds) Towards a semiotic biology. Imperial College Press, London, pp 25–41

Laland KN, Sterelny K, Odling-Smee J, Hoppitt W, Uller T (2011) Cause and effect in biology revisited: is Mayr's proximate-ultimate dichotomy still useful? Science 334:1512–1516

Laland KN, Uller T, Feldman M, Sterelny K, Müller GB, Moczek A, et al. (2014) Does evolutionary theory need a rethink? Nature 514:161–164

Lamme VAF, Roelfsema PR (2000) The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci 23:571–579

Landau MJ, Solomon S, Pyszczynski T, Greenberg J (2007) On the compatibility of terror management theory and perspectives on human evolution. Evol Psychol 5:476–519

Laskar J, Gastineau M (2009) Existence of collisional trajectories of Mercury, Mars and Venus with the Earth. Nature 459:817–819

Lau H, Rosenthal D (2011) Empirical support for higher-order theories of conscious awareness. Trends Cogn Sci 15:365–373

Leary MR (1999) Making sense of self-esteem. Curr Dir Psychol Sci 8:32–35

Leary MR, Tangney JP (2012) The self as an organizing construct in the behavioural and social sciences. In: Leary MR, Tangney JP (eds) Handbook of self and identity (2nd ed). The Guilford Press, New York, pp 1–18

Levinson SC (2006) On the human 'interaction engine'. In: Enfield NJ, Levinson SC (eds) Roots of human sociality: culture, cognition and interaction. Berg, Oxford, pp 39–69

Levinson SC, Holler J (2014) The origin of human multi-modal communication. Philos Trans R Soc B 369:20130302

Lickliter R, Honeycutt H (2013) A developmental evolutionary framework for psychology. Rev Gen Psychol 17:184–189

Lineweaver CH, Egan CA (2008) Life, gravity and the second law of thermodynamics. Phys Life Rev 5:225–242

Machamer P, Darden L, Craver CF (2000) Thinking about mechanisms. Philos Sci 67:1–25

Macnab RM, Koshland DE Jr (1972) The gradient-sensing mechanism in bacterial chemotaxis. Proc Natl Acad Sci USA 69:2509–2512

Maner JK, Kenrick DT (2010) Evolutionary social psychology. In: Baumeister RF, Finkel EJ (eds) Advanced social psychology. Oxford University Press, New York, pp 613–653

Masel J, Trotter MV (2010) Robustness and evolvability. Trends Genet 26:406–414

Maynard Smith J (2000) The concept of information in biology. Philos Sci 67:177–194

Mayr E (1961) Cause and effect in biology. Science 134:1501–1506

McAdams DP (2001) The psychology of life stories. Rev Gen Psychol 5:100-122

McLaughlin P (2001) What functions explain: functional explanation and self-reproducing systems. Cambridge University Press, Cambridge

Mead GH (1934) Mind, self, and society. University of Chicago Press, Chicago

Mendelovici A, Bourget D (2014) Naturalizing intentionality: tracking theories versus phenomenal intentionality theories. Philos Compass 9:325–337

Michel D (2013) Life is a self-organizing machine driven by the informational cycle of Brillouin. Orig Life Evol Biosph 43:137–150

Millikan RG (1984) Language, thought, and other biological categories. Bradford/The

MIT Press, Cambridge Mass

Millikan RG (1989) In defense of proper functions. Philos Sci 56:288–302

Moreno A, Mossio M (2015) Biological autonomy: a philosophical and theoretical enquiry. Springer, Dordrecht

Mossio M, Saborido C, Moreno A (2009) An organizational account of biological functions. Brit J Philos Sci 60:813–841

Mumford S, Anjum RL (2011) Getting causes from powers. Oxford University Press, Oxford

Nanay B (2010) A modal theory of function. J Philos 107:412–431

Neander K (1991) The teleological notion of 'function'. Australas J Philos 69:454–468

Neander K (1996) Swampman meets swampcow. Mind Lang 11:118–129

Neander K (2017) A mark of the mental: in defense of informational teleosemantics. The MIT Press, Cambridge MA

Nurse P (2008) Life, logic and information. Nature 454:424–426

Nussey DH, Wilson AJ, Brommer JE (2007) The evolutionary ecology of individual phenotypic plasticity in wild populations. J Evol Biol 20:831–844

O'Connor T (2020) Emergent properties. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (fall 2020 edition). https://plato.stanford.edu/archives/fall2020/entries/properties-emergent/

Oizumi M, Albantakis L, Tononi G (2014) From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. PLOS Comput Biol 10:e1003588

Oyserman D, Elmore K, Smith G (2012) Self, self-concept, and identity. In: Leary MR, Tangney JP (eds) Handbook of self and identity (2nd ed). The Guilford Press, New York, pp 69–104

Papineau D (2008) Must a physicalist be a microphysicalist? In: Hohwy J, Kallustrup J (eds) Being reduced: new essays on reduction, explanation, and causation. Oxford University Press, Oxford, pp 126–148

Papineau D (2009) The causal closure of the physical and naturalism. In: Beckermann A, McLaughlin BP, Walter S (eds) The Oxford handbook of philosophy of mind. Oxford University Press, Oxford, pp 53–65

Pattee HH (2008) The necessity of biosemiotics: matter-symbol complementarity. In: Barbieri M (ed) Introduction to biosemiotics. Springer, Dordrecht, pp 115–132

Pearl J (2009) Causality: models, reasoning, and inference (2nd edition). Cambridge University Press, Cambridge

Peirce CS (2010) The logic of signs. In: Favareau D (ed) Essential readings in biosemiotics. Springer, Dordrecht, pp 115–148

Perlman M (2004) The modern philosophical resurrection of teleology. Monist 87:3–51

Perlman M (2009) Changing the mission of theories of teleology: DOs and DON'Ts for thinking about function. In: Krohs U, Kroes P (eds) Functions in biological and artificial worlds. The MIT Press, Cambridge Mass, pp 17–36

Pigliucci M, Müller GB (2010) Evolution – The extended synthesis. MIT Press, Cambridge Mass

Prinz JJ (2012) The conscious brain: how attention engenders experience. Oxford University Press, Oxford

Pross A (2004) Causation and the origin of life. Metabolism or replication first? Orig Life Evol Biosph 34:307–321

Putnam H (1975) The meaning of 'meaning'. Minn Stud Philos Sci 7:131–193

Pyszczynski T, Greenberg J, Solomon S, Arndt J (2004) Why do people need self-esteem? A theoretical and empirical review. Psychol Bull 130:435–468

Pyszczynski T, Greenberg J, Arndt J (2012) Freedom versus fear revisited. An integrative analysis of the dynamics of the defense and growth of self. In: Leary MR, Tangney JP (eds) Handbook of self and identity (2nd ed). The Guilford Press, New York, pp 378–404

Quine WV (1960) Word and object. The MIT Press, Cambridge

Ram Y, Hadany L (2012) The evolution of stress-induced hypermutation in asexual populations. Evolution 66:2315–2328

Rand DG, Nowak MA (2013) Human cooperation. Trends Cogn Sci 17:413–425

Reddy V (2003) On being the object of attention: implications for self–other consciousness. Trends Cogn Sci 7:397–402

Ruiz-Mirazo K, Peretó J, Moreno A (2004) A universal definition of life: autonomy and open-ended evolution. Orig Life Evol Biosph 34:323–346

Ryan RM, Deci EL (2000) Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 55:68–78

Scalas LF, Morin AJS, Marsh HW, Nagengast B (2014) Importance models of the physical self: improved methodology supports a normative-cultural importance model but not the individual importance model. Eur J Soc Psychol 44:154–174

Schlomer GL, Del Giudice M, Ellis BJ (2011) Parent-offspring conflict theory: an evolutionary framework for understanding conflict within human families. Psychol Rev 118:496–521

Searle JR (1983) Intentionality: an essay in the philosophy of mind. Cambridge University Press, Cambridge

Searle JR (1987) Indeterminacy, empiricism, and the first person. J Philos 84:123–146

Searle J (2013) Theory of mind and Darwin's legacy. Proc Natl Acad Sci USA 110 (suppl 2):10343–10348

Searle J (2017) Biological naturalism. In: Schneider S, Velmans M (eds) The Blackwell companion to consciousness (2nd edition). Wiley-Blackwell, Chichester, pp 327–336

Sedikides C, Skowronski JJ (1997) The symbolic self in evolutionary context. Pers Soc Psychol Rev 1:80–102

Sharov AA (2014) Evolutionary constraints or opportunities? BioSystems 123:9–18

Shea N (2013) Naturalising representational content. Philos Compass 8:496–509

Smeekes A, Verkuyten M (2014) Perceived group continuity, collective self-continuity, and in-group identification. Self Identity 13:663–680

Solomon S, Greenberg J, Pyszczynski T (2004) The cultural animal: twenty years of terror management theory and research. In: Greenberg J, Koole SL, Pyszczynski T (eds) Handbook of experimental experiential psychology. The Guilford Press, New York, pp 13–34

Sperber D, Wilson D (1995) Relevance: communication and cognition (2nd edition). Wiley-Blackwell, Oxford

Strawson G (2008) Real materialism and other essays. Clarendon Press, Oxford

Swann WB Jr, Bosson JK (2010) Self and identity. In: Fiske ST, Gilbert DT, Lindzey G (eds) Handbook of social psychology (5th ed). John Wiley & Son, Hoboken NJ, pp 589–628

Szathmáry E, Maynard Smith J (1995) The major evolutionary transitions. Nature 374:227–232

Thompson E (2007) Mind in life: biology, phenomenology, and the sciences of mind. Belknap Press, Cambridge Mass.

Tomasello M (1993) On the interpersonal origins of self-concept. In: Neisser U (ed) The perceived self: ecological and interpersonal sources of self-knowledge. Cambridge University Press, New York, pp 174–184

Tomasello M, Carpenter M (2007) Shared intentionality. Devel Sci 10:121–125

Tomasello M, Carpenter M, Call J, Behne T, Moll H (2005) Understanding and sharing intentions: the origins of cultural cognition. Behav Brain Sci 28:675–735

Tønnesen M (2015) The biosemiotic glossary project: agent, agency. Biosemiotics 8:125–143

Tooby J, Cosmides L (1992) The psychological foundations of culture. In: Barkow JH, Cosmides L, Tooby J (eds) The adapted mind: evolutionary psychology and the generation of culture. Oxford University Press, New York, pp 19–136

Trevarthen C, Aitken KJ (2001) Infant intersubjectivity: research, theory, and clinical applications. J Child Psychol Psychiatry 42:3–48

Tsokolov SA (2009) Why is the definition of life so elusive? Epistemological considerations. Astrobiology 9:401–412

Tye M (2017) Philosophical problems of consciousness. In: Schneider S, Velmans M (eds) The Blackwell companion to consciousness (2nd edition). Wiley-Blackwell, Chichester, pp 17–31

Van Gulick R (2001) Reduction, emergence and other recent options on the mind/body problem: A philosophic overview. J Conscious Stud 8:1–34

van Hateren JH (2013) A new criterion for demarcating life from non-life. Orig Life Evol Biosph 43:491-500 https://doi.org/10.1007/s11084-013-9352-3 (no open access) or preprint

van Hateren JH (2015a) Active causation and the origin of meaning. Biol Cybern 109:33–46 https://doi.org/10.1007/s00422-014-0622-6 (no open access) or arXiv:1310.2063

van Hateren JH (2015b) The origin of agency, consciousness, and free will. Phenom Cogn Sci 14:979–1000 https://doi.org/10.1007/s11097-014-9396-5 (no open access) or preprint

van Hateren JH (2015c) Extensive fitness and human cooperation. Theory Biosci 134:127–142 https://doi.org/10.1007/s12064-015-0214-6 (open access)

van Hateren JH (2015d) The natural emergence of (bio)semiosic phenomena. Biosemiotics 8:403–419 https://doi.org/10.1007/s12304-015-9241-4 (open access)

van Hateren JH (2015e) Intrinsic estimates of fitness affect the causal structure of evolutionary change. Biol Philos 30:729–746 https://doi.org/10.1007/s10539-014-9463-x (no open access) or preprint

van Hateren JH (2015f) Causal non-locality can arise from constrained replication. EPL-Europhys Lett 112:20004 https://doi.org/10.1209/0295-5075/112/20004 (no open access) or arXiv:1506.00787

van Hateren JH (2015g) What does Maxwell's demon want from life? When information becomes functional and physical. arXiv:1407.8314

van Hateren JH (2017) A unifying theory of biological function. Biol Theory 12:112–126 https://doi.org/10.1007/s13752-017-0261-y (open access)

van Hateren JH (2019) A theory of consciousness: Computation, algorithm, and neurobiological realization. Biol Cybern 113:357–372 https://doi.org/10.1007/s00422-019-00803-y (open access)

van Hateren JH (2021a) Constructing a naturalistic theory of intentionality. Philosophia 49:473–493. https://doi.org/10.1007/s11406-020-00255-w (open access)

van Hateren JH (2021b) A mechanism that realizes strong emergence. Synthese 199:12463–12483. https://doi.org/10.1007/s11229-021-03340-z (open access)

van Hateren JH (2022) Minimal agency. In: Ferrero L (ed) The Routledge handbook of the philosophy of agency. Routledge, London and New York, pp 91–100

Varela FG, Maturana HR, Uribe R (1974) Autopoiesis: the organization of living systems, its characterization and a model. BioSystems 5:187–196

Vignoles VL, Regalia C, Manzi C, Golledge J, Scabini E (2006) Beyond self-esteem: influence of multiple motives on identity construction. J Pers Soc Psychol 90:308–333

Vignoles VL (2011) Identity motives. In: Schwartz SJ, Luyckx K, Vignoles VL (eds) Handbook identity theory and research. Springer, New York, pp 403–432

von Uexküll J (1982) The theory of meaning. Semiotica 42:25–82

Vygotsky L (1986) Thought and language. Translated from the original 1934 Russian edition. MIT Press, Cambridge Mass

Walker SI, Davies PCW (2012) The algorithmic origins of life. J R Soc Interface 10:20120869

Walsh DM, Ariew A (1996) A taxonomy of functions. Can J Philos 26:493–514

Wilson J (2015) Metaphysical emergence: Weak and strong. In: Bigaj T, Wüthrich C (eds) Metaphysics in contemporary physics. Brill, Leiden, pp 345–402

Woodward J (2013) Causation and manipulability. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (summer 2019 edition). https://plato.stanford.edu/archives/sum2019/entries/causation-mani/

Wouters A (2005) The functional debate in philosophy. Acta Biotheor 53:123–151

Wright L (1973) Functions. Philos Review 82:139–168

Zahavi D (2005) Subjectivity and selfhood. MIT Press, Cambridge, Mass