Philosophia (2021) 49:473–493
https://doi.org/10.1007/s11406-020-00255-w

Constructing a Naturalistic Theory of Intentionality

J. H. van Hateren[1]

Received: 22 July 2019 / Revised: 24 February 2020 / Accepted: 10 August 2020 /
Published online: 20 August 2020

Abstract

A naturalistic theory of intentionality (in the sense of 'aboutness') is proposed that differs from previous evolutionary and tracking theories. Full-blown intentionality is constructed through a series of evolvable refinements. A first, minimal version of intentionality originates from a conjectured internal process that estimates an organism's own fitness and that continually modifies the organism. This process produces the directedness of intentionality. The internal estimator can be parsed into intentional components that point to components of the process that produces fitness. It is argued that such intentional components can point to mistaken or non-existing entities. Different Fregian senses of the same reference correspond to different components that have different roles in the estimator. Intentional components that point to intentional components in other organisms produce directedness towards semi-abstract entities. Finally, adding a general, population-wide means of communication enables intentional components that point to fully abstract entities. Intentionality thus naturalized has all of its expected properties: being directed; potentially making errors; possibly pointing to non-existent, abstract, or rigid entities; capable of pointing many-to-one and one-to-many; distinguishing sense and reference; having perspective and grain; and having determinate content. Several examples, such as 'swampman' and 'brain-in-a-vat', illustrate how the theory can be applied.

Keywords Intentionality · Naturalism · Evolution · Meaning · Reference

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11406-020-00255-w) contains supplementary material, which is available to authorized users.

✉ J. H. van Hateren
j.h.van.hateren@rug.nl

[1] Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, P.O. Box 407, 9700 AKGroningen, The Netherlands

1 Introduction

The terms 'intentionality' and 'intentional' are used, throughout this article, 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 article 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; 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 article is to offer such a possibility. It proposes and explains a 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). 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 parameter. Estimation is fundamentally different from causal, correlational, or informational tracking, because it is one-sided (see Section 3.1). 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 3 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 4–6 then use this minimum to build a construct that approaches the conventional, human kind of full-blown intentionality. However, the article only sketches the contours of the latter. Human language, a major means of human intentionality, is only addressed briefly.

Human intentionality is closely associated with consciousness and agency. Such phenomena can be tentatively explained with variants of the theory presented here (van Hateren 2015a, 2019). However, these variants are too complex to be summarized within the space constraints of this article. Nevertheless, they imply 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 van Hateren 2017, 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.

2 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 of 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 3). 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 4 to 6). Examples of how to apply the theory can be found for fairly simple cases (depending on the explanations up to Section 4) in the Supplementary Material (Online Resource 1), and for more complex cases in Section 6.2.

The sections below gradually develop the theory in detail. Section 3 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 (van Hateren 2015b; for a biological motivation and general explanation see van Hateren 2017).

3 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 parameter 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 physico-chemical 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 upon next.

(3) The main conjecture made in this article 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 for 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 (van Hateren 2015b) 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 parameter (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).

3.1 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 with 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.

4 The Parsing of Minimal Intentionality

Section 3 showed that organisms can evolve 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, 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 parameter of the weather (e.g., the temperature at a particular place). Estimating extended processes may involve estimating many parameters at once, as well as estimating the dynamics and coherence of components of the process. Estimating needs 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.

4.1 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 5 and 6.

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 2 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 the Supplementary Material (Online Resource 1). However, the focus of this article is not on minimal intentionality, but on full-blown intentionality. Several applications of the latter are presented in Section 6.2. The required extension to abstract entities is the topic of Sections 5 and 6.

5 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.

6 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 3 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. The latter 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 (stochastic) clustering process is optimal when R is driven by a redefined x. This x must then estimate the redefined f (i.e., inclusive fitness). In a similar way, also the resulting fitness-to-be then refers to inclusive fitness.

Interestingly, this analysis suggests that there is a further way to enhance clustering. 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. This 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 (van Hateren 2015c) show that this mechanism is indeed evolvable under the right conditions. Organisms that help organisms with similar form then outcompete organisms that help only kin. Similarity of form as such becomes heritable because of the establishment of attractor forms in the population. 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. 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 5) 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 5 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 $\langle X_A \rangle$ 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.

6.1 Concluding Evaluation

Intermediate evaluations above have already discussed and explained most desiderata of the list in Section 2, 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 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 4.1), 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.

6.2 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 van Hateren 2017). 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).

7 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 article, because of space constraints. 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 6.2) 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 6.2), 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 6.2, 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.

References

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences of the USA, 108*, 10918–10925.

Brandom, R. B. (2008). *Between saying and doing: Towards an analytic pragmatism*. Oxford: Oxford University Press.

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

Claidière, N., Scott-Phillips, T. C., & Sperber, D. (2014). How Darwinian is cultural evolution? *Philosophical Transactions of the Royal Society B, 369*, 20130368.

Dennett, D. C. (1989). *The intentional stance*. Cambridge: Bradford Books.

Dennett, D. C. (2009). Intentional systems theory. In B. P. McLaughlin, A. Beckermann, & S. Walter (Eds.), *The Oxford handbook of philosophy of mind* (pp. 339–350). Oxford: Oxford University Press.

Dretske, F. I. (1981). *Knowledge and the flow of information*. Cambridge: The MIT Press.
Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience, 9*, 292–303.
Fodor, J. A. (1990). *A theory of content and other essays*. Cambridge: Bradford/The MIT Press.
Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik, N.F., 100*, 25–50.
Hamilton, W. D. (1964). The genetical evolution of social behaviour I & II. *Journal of Theoretical Biology, 7*, 1–52.
Horgan, T., & Graham, G. (2012). Phenomenal intentionality and content determinacy. In R. Schantz (Ed.), *Prospects for meaning* (pp. 321–344). Berlin: De Gruyter.
Hutto, D. D., & Satne, G. (2015). The natural origins of content. *Philosophia, 43*, 521–526.
Jacob, P. (2014). Intentionality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (winter 2014 edition). https://plato.stanford.edu/archives/win2014/entries/intentionality/
Kriegel, U. (2013). The phenomenal intentionality research program. In U. Kriegel (Ed.), Phenomenal intentionality. Oxford Scholarship Online, DOI: https://doi.org/10.1093/acprof:oso/9780199764297.001.0001.
Kriegel, U. (2016). Brentano's mature theory of intentionality. *Journal for the History of Analytical Philosophy, 4*, 1–15. https://doi.org/10.15173/jhap.v4i2.2428.
Kripke, S. (1980). *Naming and necessity*. Cambridge: Harvard University Press.
McLaughlin, P. (2001). *What functions explain: Functional explanation and self-reproducing systems*. Cambridge: Cambridge University Press.
Mendelovici, A., & Bourget, D. (2014). Naturalizing intentionality: Tracking theories versus phenomenal intentionality theories. *Philosophy Compass, 9*, 325–337.
Millikan, R. G. (1984). *Language, thought, and other biological categories*. Cambridge: Bradford/The MIT Press.
Neander, K. (2017). *A mark of the mental: In defense of informational Teleosemantics*. Cambridge: The MIT Press.
Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science, 7*, 131–193.
Quine, W. V. (1960). *Word and object*. Cambridge: The MIT Press.
Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge: Cambridge University Press.
Searle, J. R. (1987). Indeterminacy, empiricism, and the first person. *The Journal of Philosophy, 84*, 123–146.
Shea, N. (2013). Naturalising representational content. *Philosophical Compass, 8*, 496–509.
Strawson, G. (2008). Real intentionality 3: Why intentionality entails consciousness. In G. Strawson (Ed.), *Real materialism and other essays* (pp. 281–305). Oxford: Oxford University Press.
van Hateren, J. H. (2015a). The origin of agency, consciousness, and free will. *Phenomenology and the Cognitive Sciences, 14*, 979–1000.
van Hateren, J. H. (2015b). Active causation and the origin of meaning. *Biological Cybernetics, 109*, 33–46.
van Hateren, J. H. (2015c). Extensive fitness and human cooperation. *Theory in Biosciences, 134*, 127–142.
van Hateren, J. H. (2017). A unifying theory of biological function. *Biological Theory, 12*, 112–126.
van Hateren, J. H. (2019). A theory of consciousness: Computation, algorithm, and neurobiological realization. *Biological Cybernetics, 113*, 357–372.