(Draft, version 27 June 2016, available at: http://philpapers.org/rec/VANTES-7
or https://sites.google.com/site/jhvanhateren/home/self/self.pdf)

The evolved self has agency, purpose, and unity

J. H. van Hateren

Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, Groningen, The Netherlands

e-mail: j.h.van.hateren@rug.nl

Abstract

Recently developed extensions of evolutionary theory are used to explain the human self as an evolved, unitary, and purposeful phenomenon. A basic mechanism that can generate life's agency and goal-directedness is combined with mechanisms that can account for awareness by and of the self, and for the social characteristics of humans. The new theory is largely consistent with major existing theories of the self, in particular theories centred on self-esteem, self-determination theory, and terror management theory. It can therefore be regarded as a meta-theory that brings these theories, and related ones, within a common evolutionary framework. The theory suggests two primary dimensions of the self, the depth of awareness of the self and the social extent of the self.

Keywords: agency, meaning, self-esteem, unity of self, evolution

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 article 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 two recent additions to evolutionary theory (van Hateren, 2015a, c). The first addition explains the agency and intrinsic goal-directedness of living organisms. Agency is defined here as the ability to initiate meaningful behaviour. Throughout this article, the terms “goal-directedness” and “goal” are used in a general biological sense, not necessarily associated with human goals and human motivations. The second addition to evolutionary theory that is used below explains the strong human reliance on social and cultural processes. Both additions will be summarized below and subsequently used to develop an evolutionary understanding 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 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 approximation of an individual's 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. It therefore also acts as a mechanism that has a direct ultimate interpretation, as relevant for the timescale of evolution.

The article is organized in a somewhat unconventional way. Rather than starting with a historical overview of the field, and then zooming in on the new contributions of the present study, the order is reversed here. The theory is presented first, and comparisons with existing work are postponed until the "Discussion". The reason for this is that the theory originated from computational modelling that was performed in order to understand the agency and goal-directedness of living organisms in general (van Hateren, 2015a) and the social characteristics of humans in particular (van Hateren, 2015c). Only with hindsight, the relation to the social psychology of the self became clear. This indirect and ahistorical origin of the theory is reflected in how it is presented here.

The basis of the theory is summarized in the sections "Agency and Goal-Directedness" (on the evolved mechanism that can generate these in living organisms), "Subjective Awareness and Awareness of the Self" (on the conjectured origin of awareness and how that can fold back onto the self), and "Human Fitness Extends Beyond the Individual" (on the extended form of fitness in humans). Two of these sections consist of three parts, "General Explanation", "Consequences", and "Computational Summary". The latter part is more technical, and may be skipped by readers who only want to get the general idea. The core of the theory of the self is explained in the section "The Human Self". In the first sections of the "Discussion", this is further elaborated and 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.

AGENCY AND GOAL-DIRECTEDNESS

An important aspect of the self is that it is a prime source of agency, which is taken here as an individual's ability to initiate novel behaviour that is significant for that individual. All living organisms have some form of agency. From a fundamental, naturalistic perspective, agency has been difficult to understand, because it seems to suffer from an internal contradiction (van Hateren, 2015b). One would expect that significant behaviour should, typically, not be caused randomly, but should follow from certain criteria and rules. But such rules suggest a determinate mechanism, which, by its nature, could not initiate anything really novel. Initiating truly novel behaviour suggests a mechanism involving stochasticity (i.e., randomness, chance). But such a mechanism would produce behaviour that is random rather than significant. In this section I explain a mechanism that avoids this conundrum (van Hateren, 2015a). The explanation in this section is generally valid for any living organism, and not yet focussed on humans.

General Explanation

Current living organisms are descendants of organisms that were successful in terms of surviving and reproducing in the past. The chances of surviving and reproducing depend on many factors, both internal and external to each organism. Together these factors form a process that results in a likelihood of evolutionary success. Throughout this article, both process and likelihood are denoted by X_{true}, which stands for the evolutionary fitness of the organism. Thus X_{true} has a structure (because it is a process with many interacting inputs) as well as a value (the output of the process, a number that quantifies the likelihood of success). Because X_{true} is a process, it is more complex than simple empirical measures of evolutionary fitness (such as the actual number of offspring of an organism, or how much the organism contributes to the gene pool of the next generation). A second reason why

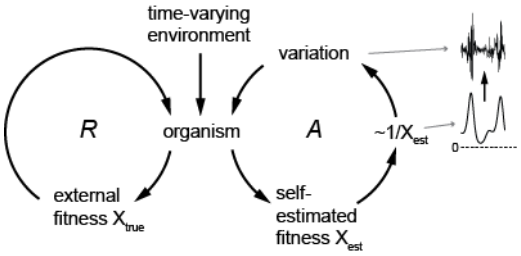

Fig. 1. Origin of agency and goal-directedness. A reproductive cycle *R* produces basic Darwinian evolution based on the fitness X_{true}, and an active cycle *A* generates agency and intrinsic goal-directedness. The latter cycle continually updates an organism's structure and behaviour, with the amount of change being modulated by an internal estimate of fitness, X_{est}.

X_{true} is complex is that its structure and value change continually during an organism's life. It depends, for example, on a time-varying environment that includes other organisms. The value of X_{true} becomes low when circumstances are poor, such as during a famine, but it can subsequently recover. It becomes zero when the organism dies. A final reason why X_{true} is complex is that it can be defined in a broader sense than just concerning survival and procreation (by including social and cultural factors, as detailed in sections below).

Standard evolutionary theory explains evolutionary change by using the value of X_{true} as acting in a reproductive cycle *R* (left part of Fig. 1). Hereditary changes that prove beneficial have a good chance to be transferred to next generations. However, there is an additional, secondary pathway through which fitness can influence evolutionary success. This pathway was conjectured by van Hateren (2015a) as a necessary component for explaining agency and goal-directedness. It contains a process within the organism, X_{est}, with an output value that approximates ("estimates") the output value of X_{true}. Also X_{est} is a complex process, but now entirely produced within the organism. It depends on many factors, both internal and external to the organism, such as obtained through its senses and memory. The output value of the process is taken to be represented only implicitly, distributed throughout the process (analogously to how parameter values may be diffusely represented in a neural network). It affects the organism in a distributed way as well.

Depending on the value of X_{est}, the organism is conjectured to vary its structure and behaviour. Theoretically, a low value should produce high variability, whereas a high value should produce low variability (symbolized by $\sim 1/X_{est}$ in Fig. 1). The reason is that when the value of X_{est} is high, also the actual fitness X_{true} is likely to be high (because evolutionary pressure promotes an X_{est} that approximates X_{true} with reasonable accuracy). Then the organism is doing all right, and there is little reason to change. However, when the value of X_{est} is low, this indicates that the evolutionary success of the individual is at risk. Change is then necessary. Both the process X_{est} and the resulting changes are continually updated during the lifetime of the organism (the continuously running *A* cycle in Fig. 1). Eventually, the organism is likely to find behaviour with a high value of X_{est}, at least on average. This subsequently produces low variability. Low variability means that the organism changes only slowly. This state will continue until X_{est} happens to become low, often as a result of changes in the organism's environment. Then variability increases once more, and the random search for behaviour with higher X_{est} starts again. Model simulations (see the "Computational Summary" below) show that this mechanism is indeed evolvable. It increases the X_{true} of organisms equipped with it, on average (that is, probabilistically expected, but not guaranteed).

The mechanism produces behaviour that is neither completely random not completely determinate. It is not completely random, because the size of the variations are driven by the value of X_{est}. It is also not completely determinate, because the variations are random in detail (the rightmost diagrams in Fig. 1 illustrate this point). Moreover, the *A* cycle continues to mix these determinate and random factors. The result is agency, because new, unforeseen behaviour is initiated that is nevertheless driven by a rule-based process (X_{est}). A second result is an intrinsic goal-directedness within the organism, in the form of striving towards high X_{est}.

The value of X_{est} drives the magnitude of random changes in the organism's structure and behaviour. In practice, the total change will consist of such random changes on top of determinate

change. Determinate change will happen when the fitness consequences of a particular change are implicitly known beforehand. Fitness consequences may be partly known based, for example, on genetic information formed by previous natural selection, or on neural information formed by previous learning. Although determinate changes themselves are not generated by agency, they may be implicitly used in the *A* cycle as components of the process producing agency. This leads to more complex forms of agency than the basic one depicted in Fig. 1, as discussed in van Hateren (2015b). The latter also argues (pp. 996-997) that agency as explained here is quite different from enactivist proposals (e.g., Di Paolo, 2005; Thompson, 2007; see also van Hateren, 2013, p. 498), essentially because it does not depend on the assumption that self-maintenance is valuable. Moreover, the goal-directedness associated with agency is quite different, and more fundamental, than the standard goal-directedness one may perceive in cybernetic systems (e.g., in control systems using feedback).

Consequences

The mechanism producing agency and goal-directedness described above has a number of consequences. Most importantly, it solves the problem that agency and goal-directedness pose to conventional Darwinian theory. In that theory there is no place for meaningfulness and significance. Even if there is something resembling meaning and agency, it must be understood as serving X_{true} directly. But X_{true} has nothing to do with meaningfulness, because it is just the consequence of a blind physico-chemical process. It is not goal-directed. However, X_{est} and the *A* cycle introduce a limited form of goal-directedness, albeit intrinsic to individual organisms and not extended to the overall evolutionary process. The latter is still undirected, but the agency of individual organisms is not completely undirected any more. Organisms have meaning and significance to themselves, because of X_{est}, and act accordingly. The structure of X_{est} incorporates all the factors that the organism implicitly takes into account for approximating its own X_{true}.

There is a fundamental reason why high X_{est} has to be the intrinsic overall goal of any living organism. It is the only overall goal that is stable on an evolutionary timescale. Because of its stochastic structure, the *A* cycle of Fig. 1 would produce goal-directedness also when X_{est} were replaced by another, arbitrary goal. But organisms with such an arbitrary goal would be outcompeted by organisms with high fitness as overall goal, that is, with high X_{est} as an approximation of high X_{true}. The better X_{est} approximates X_{true}, and the higher X_{est}, the more evolutionary success is to be expected. Thus there is evolutionary pressure to improve X_{est}, given the means available to the organism.

Although high X_{est} is the overall goal, it consists of a large set of sub-goals, in practice. These sub-goals are all expected to contribute to the overall goal. Such contributions should typically contribute to X_{true} as well, but they are not guaranteed to do so, because of the approximate nature of X_{est}. 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 explained above implies that the structure formed by all sub-goals together corresponds to the structure of X_{est}, including its dynamics. Sub-goals can only be fully understood, then, from their role in constituting the structure of X_{est}.

Computational Summary

The evolvability of mechanisms such as in Fig. 1 was investigated in van Hateren (2015a). The *A* cycle can act on various timescales—evolutionary across generations, behaviourally, and in neural processing. However, for agency, only changes within the organism's lifetime are relevant, thus the X_{est} of Fig. 1 then only modulates the rate of structural and behavioural change on that timescale, without hereditary transfer. Change may occur directly, but also more sophisticated variants were studied, such as when X_{est} does not immediately drive such changes, but only after the possible effects of changes are simulated first within the organism.

Variants were simulated using simplified model systems, with (mortal) organisms mutating and changing behaviour along a single dimension. Along the same dimension, the environment varies in time, unpredictably across a wide range of timescales. X_{true} quantifies the expected reproductive rate of each organism. It is a function of how much the momentary environment differs from the combination of (momentary) behavioural disposition and the organism's heredity (fixed for a given organism).

In simulations, two populations share an environment with limited resources. Thus, organisms must compete in order to reproduce. One population may consist, for example, of organisms with variability that is fixed and not modulated by X_{est}. The other population may consist of organisms with an *A* cycle and an X_{est} approximating X_{true}. Population sizes start out equal, but fluctuate because the environment varies over time. Simulations using different realizations of the environmental time course invariably show that the population with organisms lacking X_{est} becomes extinct. Such organisms are less capable of adapting to environmental change than organisms with X_{est}. The simulations show that having X_{est} modulate variability increases X_{true}. These computational results have recently been corroborated by mathematical analysis (van Hateren, 2015d).

SUBJECTIVE AWARENESS AND AWARENESS OF THE SELF

In a sense, the process X_{est} as discussed above can be viewed as a proto-self. The structure of X_{est} defines, for any living organism, what the organism implicitly considers important for its own survival and reproduction. The identity of the organism can be equated to the form of X_{est}, that is, to which internal and external factors the organism takes into account for X_{est}, and how. It is important to stress that only X_{est}—and not X_{true}—can produce a self, because X_{est} is the source of agency (through the *A* cycle of Fig. 1). 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 do have agency and goal-directedness according to the theory presented here. But they clearly, or at least almost certainly, lack subjective awareness and a self in any psychological sense.

Nevertheless, subjective awareness is presumably much older than the human species. How and when it originated is quite uncertain, but the theory explained above suggests one particularly likely point of origin (van Hateren, 2015b). This follows from the fact that the stochastic *A* cycle that contains X_{est} already produces novel, emergent phenomena—agency and goal-directedness. Such phenomena are absent from the non-living material world. Because X_{est} embodies intrinsic goals in an organism with agency, such goals implicitly represent values to the organism. If the goals were not valuable, the organism would not pursue them, because it has some behavioural freedom—it is not an automaton.

For most species, the goals and their values are not made explicit, but are merely contained within the organism's structure and dynamics. But this changes once important aspects of X_{est} are transferred when organisms engage in reciprocal communication, that is, in (usually nonverbal) dialogue. Then internal goals and values have to be made explicit. They must be transformed into regular physical signals—such as touch, posture, gestures, and sounds—that can be interpreted by both partners in the dialogue. This externalization of goals and values is another emergent phenomenon, also absent from the non-living material world. The conjecture is that this leads to subjective experience whenever the very form of X_{est} is changed as a result of the externalization (and subsequent internalization) necessary for dialogue (van Hateren, 2015a). Merely engaging in a dialogue is assumed to be already sufficient to produce such a change, both during sending and receiving meaningful messages. The terms “meaningful” and “meaning” are used here not in the linguistic sense, but in the broad sense of indicating importance and significance for the individual. Significance and meaning depend on the fact that X_{est} is involved.

Changing the form of X_{est} goes beyond regular learning and simple forms of communication, such as between social insects. Such learning and communication would only involve changing some parameter settings of a given general form of X_{est}, within a fixed and predictable range of possibilities. Instead, changing X_{est} as meant here requires a substantial change in the structure of X_{est}. This change concerns which specific aspects are incorporated in X_{est} and how specifically. The conjecture is that an X_{est}-changing form of dialogue 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 within mother-infant and pair bonds. Such bonds presumably induce a significant reconfiguration of the overall goals and sensed meaning, thus a significant reorganization of X_{est} in the individuals involved (van Hateren, 2015b).

In its simplest form, awareness then occurs when two subjects communicate in such a dialogue. When this changes the form of X_{est}, this produces a subjective sense of awareness, roughly

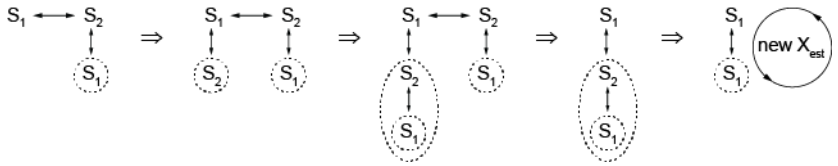

Fig. 2. Development of the self as a perceived object. Subject S_1 communicating with subject S_2 gradually develops a continually updated model of the self (final diagram). Dialogues are symbolized by double-headed arrows, models by dashed lines.

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 2 sketches how more complex forms of awareness may arise during development. This is analogous to what was previously suggested by Mead (1934) and it is related to work in developmental psychology (Tomasello, 1993; Trevarthen and Aitken, 2001). It is shown in Fig. 2 specifically for the self, but similar schemas can be made for awareness of others, of objects (van Hateren, 2015b), and of groups of others. The schema is not meant as a detailed theoretical proposal, but merely as a minimalist tool for explaining the general 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 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 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. 2). 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 herself. In a final stage (last diagram), the modelled S_2 is not needed any more. Then S_1 can directly engage with her own simulated self and be aware of her own self. The final stage thus represents a primary form of awareness of the self.

Because the model can contribute to the X_{est} 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_{est}, 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 updating of X_{est} is conjectured to be accompanied by subjective experience, as discussed above.

The above explanation applies to any kind of internal model, including nonverbal ones with only non-symbolic simulation. In human development, the schema of Fig. 2 is presumably executed several times, probably in overlapping, continuous, and more complex ways. This then results in increasingly sophisticated internal models of the self (Reddy, 2003), as a form of Theory of Mind. In particular, symbolic (language-based) and social (communal) layers are gradually added (Tomasello, 1993; Tomasello and Carpenter, 2007).

HUMAN FITNESS EXTENDS BEYOND THE INDIVIDUAL

The above theories of agency and awareness may be adequate for understanding such phenomena in nearly all species where they occur. X_{est} then estimates X_{true} in the form of the standard evolutionary fitness, that is, inclusive fitness (explained below). However, in particular for humans, the standard fitness requires elaboration in order to include social factors not targeted at inclusive fitness. The way this is done here is different from previous accounts in the literature, because it extends fitness itself rather than adding factors that amplify inclusive fitness.

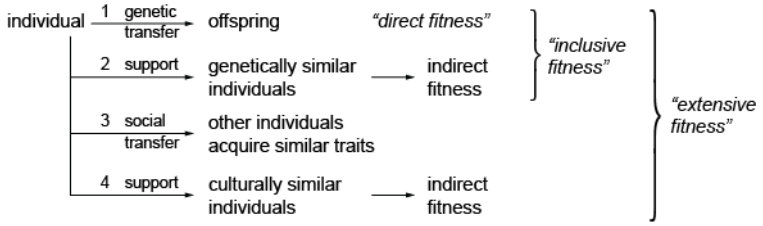

Fig. 3. Various forms of fitness. Direct fitness (pathway 1) is the expected rate of producing offspring. Inclusive fitness combines direct fitness with indirect fitness (pathway 2) produced by helping genetically related individuals. Extensive fitness combines inclusive fitness with fitness produced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

General Explanation

In its most general sense, evolutionary fitness at the organismal level can be defined as the rate by which an organism produces organisms that are like itself, that is, organisms with similar properties. The extent of the similarity and the means by which it is produced are not specified, and there are indeed various possibilities. The most straightforward way is through what is known as direct fitness (Fig. 3, pathway 1), by directly producing offspring. When reproduction is asexual, offspring is nearly identical to the parent. Similarity is lower with sexual reproduction, but it still involves direct fitness.

However, sexual reproduction and a social lifestyle imply that direct fitness is only part of total fitness (Hamilton, 1964). An individual might support individuals that are genetically similar, such as close relatives. That individual thereby increases the fitness of those being helped. The excess fitness should then partly count as fitness attributable to the supporting individual. By supporting, that individual indirectly increases the production of organisms like itself. For spreading genetic similarity (i.e., correlated genes) it does not matter if it is produced directly through own offspring, or indirectly through offspring of related individuals. Obviously, the degree of relatedness should be taken into account as a weighting factor when compounding the resulting fitness. When relatedness is only little above the average relatedness expected in the relevant population, the contribution to fitness should be small as well. Nevertheless, when the group of related individuals is large, the contribution of this so-called indirect fitness can be large as well. The combined effect of direct and indirect fitness (Fig. 3, pathways 1 and 2) is known as inclusive fitness (Hamilton, 1964).

Inclusive fitness focusses on genetic transfer of traits, in particular on how much an organism contributes to the gene pool of the next generation. In addition, similar traits in organisms may also be produced through social learning and cultural transmission (e.g., Boyd et al., 2011). Such mechanisms are present in many species, and can be explained from their positive effects on inclusive fitness. Thus, they depend only on pathways 1 and 2, not on a separate form of fitness. However, specifically for humans, an actual extension of fitness beyond inclusive fitness was recently proposed by van Hateren (2015c; see the "Computational Summary" below). It is still a form of biological fitness, depending on behavioural plasticity, and it is part of both X_{true} and X_{est}. It belongs to individual organisms, that is, it concerns organismal fitness, not the fitness of cultural traits in a population of organisms, nor the replicative fitness of cultural memes themselves. The extension can arise when organisms are capable of helping their con-specifics not based on genetic relatedness (as in the indirect form of inclusive fitness), but based on phenotypic similarity. The phenotype of an organism is the total of its properties, as interacting with environment and other organisms.

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. The primary conditions for this mechanism to work as a separate form of fitness—reliably recognizing intentional behaviour and flexibly copying it to and from others—are presumably only fully developed in humans (as discussed in van Hateren, 2015c).

Crucially, the mechanism can only work if there is an intrinsic X_{est} and an *A* cycle, because it requires that fitness is evaluated continually. 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 (see van Hateren, 2015c, p. 136). The mechanism proposed here depends on behavioural plasticity, and is therefore different from genetic greenbeards (Gardner and West, 2010). Moreover, it should also not be confused with group selection. Although group membership confers benefits on individuals, evolution in this model still happens at the level of individual organisms, not at the level of groups.

The assumed extended form of fitness in humans is called extensive fitness. It is illustrated by the four pathways of Fig. 3 combined. It still includes the two fitness components of inclusive fitness (pathways 1 and 2), but, in addition, it has two forms of transfer that are not genetic, but social. The first, direct form (pathway 3) is analogous to direct fitness (pathway 1). Through social transfer an individual may, in effect, reproduce some of his traits in other individuals, independent of whether they are related or not. This transfer enlarges the individual's phenotypic group. It thereby increases the probability of helping and thus confers benefits on the individual, on average. Examples of such transfer are active teaching, facilitating or allowing others to learn from or copy one's behaviour, and in general acting as a role model. Often this form of fitness will align with other forms of fitness, such as with direct fitness when raising children. It also aligns with the indirect part of inclusive fitness when an individual helps to raise the children of relatives or teaches related individuals.

The second, indirect form of fitness that involves social transfer (pathway 4) is analogous to the indirect part of inclusive fitness (pathway 2). Now individuals are supported that are already similar in their culturally shaped traits, independent of whether these individuals are genetically related or not. When this support increases the extensive fitness of those individuals, part of the increase should, by analogy with pathway 2, be attributed to the extensive fitness of the original individual. When similar individuals can be clearly attributed to a specific group, this leads to a supported in-group. The overall effect is that the original individual will enhance her own fitness. Fitness is increased in the general sense of increasing the rate by which individuals arise that are similar, i.e., that have similar traits. As before, this form of fitness will often align with other forms of fitness.

Consequences

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 thus 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 the core 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 (X_{est} approximating pathways 1 and 2 in Fig. 3), but on self-estimated extensive fitness (X_{est} approximating all four pathways). The *A* cycle of Fig. 1 functions as before, thus the process X_{est} is still the source of agency, goal-directness, and meaning. These factors then automatically acquire social, non-genetic aspects through X_{est}. Importantly, the extended form of X_{est} is necessary for understanding the origin and nature of the human self (see below).

Helping based on phenotypic similarity makes the structure of X_{est} and X_{true} 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 very important. The internal structure of X_{est} 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_{est}. For example, genetic alliances may

conflict with phenotypic alliances, different group alliances may conflict, 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_{est} of specific individuals, the properties of X_{est} can also produce tension and conflict between individuals and between groups. In general, pathways 1 and 3 imply competition between individuals. There is mate selection, raising children partly depends on shared resources, and 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 limited cultural absorbance. The balance between prosocial and anti-social behaviour then depends on the details of how an individual engages pathways 1 to 4, on the specific in-groups to which the individual belongs, and on how these in-groups overlap or have conflicting interests.

Computational Summary

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. 3, pathway 1). Fitness is then modelled as a simple reproductive rate of each individual. For inclusive fitness, pathway 2 (Fig. 3) is added in the form of a fitness multiplier. This factor increases the fitness of an individual if he helps others of similar hereditary type (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, 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 she is 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. 3). It also produces pathway 3, indirectly, because of the fitness multiplier. When an individual has acquired a certain phenotype, it contributes to the size of the corresponding phenotypic group. It thus increases 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_{est} 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 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 (van Hateren 2015c) that h-helping and p-helping counterbalance their relative strengths and weaknesses. They should perform equally well if all else is equal. But all else is not equal, because of the fitness multipliers that implement benefits for h- 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, because the fitness multiplier increases with group size. Obviously, there are many potential complications here, because p-helping is more vulnerable to cheating and cognitively

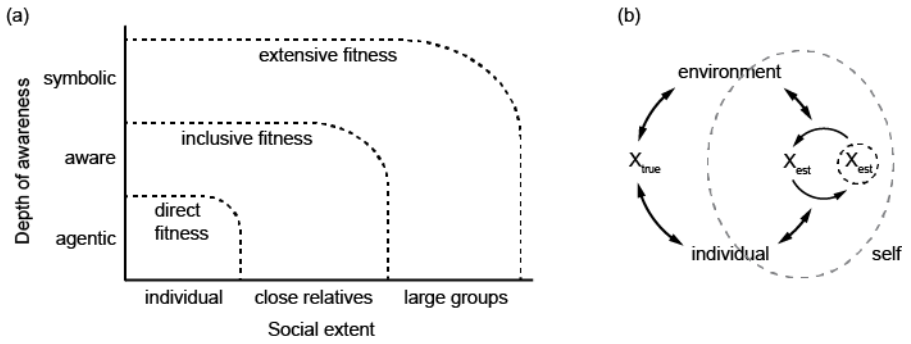

Fig. 4. Evolutionary Extensive Self Theory (EEST). (a) The self has evolved along two dimensions, depth of awareness and social extent. (b) The self is aware of itself in a continual cycle involving X_{est} and its model. It incorporates parts of individual and environment, and it depends on and modifies X_{true}.

more demanding than h-helping (as discussed in the literature on altruism, e.g., Rand and Nowak, 2013).

THE HUMAN SELF

We are now finally in a position to formulate the key thesis of this article. This thesis is that the human self is defined by the structure of the process X_{est}, because it is the source of human agency, subjective awareness, and sense of meaning and purpose. As discussed above, the process X_{est} is an evolved, layered phenomenon, which is summarized in Fig. 4a. 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 merely providing the basis for agency. The second dimension, 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 correlation may be lower, for example when the symbolic self is more strongly developed than the social self.

Although X_{est} is taken here to be the core of the self, there are two complications that need to be discussed. First, the individual can only be aware of the self, as object of awareness, to the extent that X_{est} is represented by a model (X_{est} enclosed by a dashed circle in Fig. 4b). However, this model is continually evaluated and modified through X_{est}, and itself continually evaluates and modifies X_{est} in return. This is shown in Fig. 4b as a cycle represented by the arrows connecting X_{est} and its modelled version. This is similar to the cycle at the far right in Fig. 2. Parts of the modelled self will thereby gradually become incorporated into X_{est} proper. The self is thus changing dynamically. The borders between the self as source of subjective awareness (X_{est}) and the self as object of awareness (X_{est} within a dashed circle) are therefore fluid.

The second complication is that the self can incorporate, as input factors, not only parts of the social environment, but also parts of the physical environment and parts of the individual's body. This is symbolized by the dashed line marked "self" in Fig. 4b. The self as perceived object is a model of one's estimated fitness. This estimate includes aspects of the physical and social environment that are judged to be vital for one's goals. For example, individuals who have lived on a particular piece of land for generations may see that land as vital for their X_{est}. The extended fitness X_{est} means not just fitness in terms of survival and biological reproduction, but also fitness in terms of socially

transmittable values. The land will then be part of X_{est} and therefore of the self. Similarly, parts of the social and cultural environment can contribute to the self. For example, political and religious ideas may become an important part of X_{est}.

Also parts of the individual itself will often be incorporated into the self. 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 essential for survival (essential for X_{true}), it is usually not part of the aware and symbolic parts of one's X_{est}. It could become part of the self when there is a special reason, such as a diagnosed liver disease. But usually, the liver is functioning autonomously. It is not directly participating in human agency and goal-directedness that belong to the aware self. However, other constituents of an individual may typically be important parts of X_{est}. For example, one's bodily features and general appearance are usually taken as important parts of the self.

As argued above, X_{est} has a particularly wide scope in humans. However, as in other species, it is still embodied in a physiological process restricted to the individual. This process gets its wide scope from the factors it takes into account for producing its outcome. These factors include abstract, social factors, and factors related to environmental or cultural processes that may be distant in space and time. Nevertheless, they are all processed within the individual with the information available to X_{est}.

Interpretation of X_{est} in Psychological Terms

X_{est} is assumed to be an internal, physiological process within the organism, and it is therefore interesting to see to what extent it can be interpreted in psychological terms. The parts of X_{est} that are situated in the lower-left corner of Fig. 4a presumably affect human behaviour through nonconscious processes. Such behaviours are still not automatic, because X_{est} inherently produces agency (through the *A* cycle of Fig. 1). In psychological terms, such nonconscious agency can be viewed as behaviour influenced by drives or needs. The dashed lines in Fig. 4a 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_{est} in a way that makes X_{est} align well with X_{true}. The result is agency that increases X_{true}, on average. For example, basic nutritional and sexual drives belong in this category. However, X_{est} only approximates X_{true}, and agency is not determinate. Thus, there is no guarantee that drives and needs work out well in specific cases.

The aware parts of X_{est} (Fig. 4a) lead to agency associated with consciousness, that is, volition. Motives influencing behaviour contribute to these parts of X_{est}. Motives are typically flexible and partly formed through learning. A motive adjusts the form of X_{est} in such a way that the value of X_{est} 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_{est} 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 also increase the actual X_{true} 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 forming beneficial motives. On the other hand, too loose constraints will increase the risk that motives take an adverse form.

The symbolic parts of X_{est} lead to agency that is strongly dependent on social, communicated factors, 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_{est}. Such goals and policies are the result of inter-individual communication (horizontally in Fig. 4a) combined with intra-individual drives, motives, and goals (vertically in Fig. 4a).

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_{est} and its model (Fig. 4b, inner cycle to the right). The model (X_{est} within dashed lines) represents knowledge, through its structure

and memory, which is used and modified by the agency of X_{est} and the *A* cycle (Fig. 1). The interpersonal self corresponds to the interpersonal aspects of extensive fitness (particularly pathways 3 and 4 of Fig. 3), where X_{est} and social environment interact. Finally, the self as agent corresponds to agency generated by X_{est} and *A* cycle.

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_{est} produces of its model (Fig. 2). The cognitive self corresponds to the modelled X_{est} as interacting with X_{est} (Fig. 4b), in particular when the interaction involves symbolic awareness (Fig. 4a). The executive self corresponds to how X_{est} and its model function as sources of agency.

Several other well-known approaches to the self can be similarly related to particular aspects of the current theory (abbreviated below as EEST, Evolutionary Extensive Self Theory). I will discuss here in some detail how EEST relates to self-esteem, self-determination theory, terror management theory, and evolutionary psychology. Subsequently, I discuss how explanations based on X_{est} differ from those based on X_{true}, and how EEST understands the self with respect to unity, continuity, and stability. As stated in the first paragraph of this article, EEST 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 EEST is based. Nevertheless, empirical testing should be possible, as discussed in the final section below.

Self-esteem

Explicit self-esteem presumably corresponds to the subject's perception of that part of X_{est} that estimates the subject's role in producing extensive fitness, and 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_{est} are produced with a direct, active role for the subject. For example, external factors, such as disease and war, can strongly affect X_{est}. They thereby affect general psychological well-being, but not self-esteem, at least not directly. However, in reasonably favourable circumstances, X_{est} is likely to be strongly determined by agency and how the individual functions within the social environment. Because the process X_{est} has a complex, heterogeneous structure, self-esteem is heterogeneous as well (Heppner and Kernis, 2011). If the dynamics of X_{est} 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_{est} (Fig. 4a), implicit self-esteem (Heppner and Kernis, 2011) corresponds to nonconscious, agentic parts of X_{est}.

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. 3). 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 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_{est}. 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 her current agentic role as favourable, and therefore have high self-esteem. Nevertheless, X_{est} always integrates individual and social factors (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_{est}, because most fitness pathways (Fig. 3) have a social component. Socially held values are therefore vitally important for X_{true} (and thus for X_{est}). 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 negligant (e.g., Crocker and Park, 2004). From the perspective of EEST, it is important to note that high X_{est} does not necessarily reflect high X_{true}. X_{est} could be a poor, distorted approximation of X_{true}. X_{est} is about an unknown and uncertain X_{true}, including how that might change as a result of the agency of the individual and others. Individuals might therefore form an X_{est}, and goals associated with its structure, that are, objectively, not in the best interest of themselves and those involved in the prosocial parts of their X_{est}. 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 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_{est} in EEST. 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 produced by being aware of one's mortality. Moreover, it is regarded as a cultural construction (Pyszczynski et al., 2004, pp. 436-437). For EEST, striving for high self-esteem may be culturally shaped, but it has firm biological roots. High self-esteem is correlated with high X_{est}, high X_{est} is correlated with high X_{true}, and high X_{true} is necessary for sustaining life.

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 EEST. The need for autonomy is implied by the agency of the self combined with the goal of obtaining high X_{est} (and thus high X_{true} as well, probabilistically). In the *A* cycle of Fig. 1, it is the freedom to act—as called agency here and autonomy in SDT—that enables enhancing X_{true} and X_{est}. 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. 3 and 4a). In order to obtain high X_{true} and X_{est}, people need to relate to others. Prosociality in the form of pathways 2 and 4 (Fig. 3) is an intrinsic part of X_{true} and X_{est}. Finally, the need for competence can be understood from the fact that behaviour driven by X_{est} will typically only result in actual high X_{true} when it is carried out competently. Similarly, X_{est} is only likely to be successful if it is coherent. Because X_{true} is inherently coherent (i.e., unitary and continuous, see below), an incoherent X_{est} is likely to approximate X_{true} inadequately. It would then be maladaptive. Incompetently performed behaviour based on a coherent X_{est} would be maladaptive as well.

More generally, EEST 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 active *A* cycle of Fig. 1.

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 driving 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 his own death. Increasing mortality salience will generally induce him to defend his cultural worldviews more strongly. It also induces him 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 EEST. However, TMT assumes different primary causes of self and meaning than EEST. In EEST, the main factor driving the self is the need to get or keep a high X_{est}. Because the value of X_{est} is an internal estimate of the rate by which an individual induces others to become similar, X_{est} as well as X_{true} go to zero when the individual dies. Avoiding death is therefore an absolute condition for maintaining a positive X_{est}. 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_{est} strongly peak. Such a peak can occur through one or more of the pathways depicted in Fig. 3, for example when protecting offspring 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_{est} in the remaining lifespan.

Nevertheless, increasing mortality salience is clearly a particularly powerful way to let individuals feel that their X_{est} may be too low. A way to compensate for a possibly low X_{est} is to invest more in some of the pathways of Fig. 3, such as by giving more weight to the views of the in-group or in general by investing more in self-esteem. A belief in immortality, such as life after death, can be an effective way to increase X_{est} as well, even if such a belief does not correspond to the reality of X_{true}. The reason is that X_{est} is merely an estimate of X_{true}. It needs to be reasonably close—but not perfect or optimal—if it is to be adaptive.

Relationship to Evolutionary Psychology

EEST relies on evolutionary arguments and stresses the importance of evolutionary fitness for human psychology, in particular the internalized, estimated form 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, EEST 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 recognized. 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) stress that individuals are not fitness-maximizers in a teleological (goal-directed) sense, but rather adaptation-executors.

EEST 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 *A* cycle of Fig. 1 and its elaborations (see also van Hateren, 2015b). The *A* cycle does not produce behaviour that consists of executed, ready-to-go adaptations, but creates novel behaviour. Such novel behaviour will usually partly rely on existing behavioural adaptations, but these are used then as mere components. Novel behaviour is still evolutionarily constrained by the requirements that X_{est} approximates X_{true} and that both are sufficiently high. However, as mentioned above, such constraints must be rather abstract, high-level ones. They should protect the advantages of using a highly flexible X_{est}, because that can potentially increase X_{true}, as in a self-fulfilling prophecy. But the constraints should also protect X_{est} from becoming too different from X_{true}, or from producing maladaptive forms of X_{true}. 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 to increase fitness, albeit only in a statistical way. Making fitness high is a genuine, innate goal, thus the mechanism is in fact teleological, in a weak sense. The teleology is weak, because it only exists within organisms, and does not depend on an external teleology.

A major addition to conventional EP is 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_{est} as an estimate provides considerable freedom to individuals—and indirectly to society—to approximate X_{true} in different ways. Different forms of X_{est} subsequently affect X_{true} in a continual cycle (Fig. 4b). EEST 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.

Differences between Explanations based on X_{est} or X_{true}

The present theory explains the self and its motives as based on X_{est} rather than X_{true}. Evolutionary explanations of the self have been given before (e.g., Sedikides and Skowronski, 1997), based only on the conventional X_{true}. Because X_{est} has evolved to approximate X_{true}, explanations based on either X_{est} or X_{true} are often fairly close. If a particular behaviour increases X_{true}, it is likely to increase X_{est}, and vice versa. However, the form of X_{true} depends in important ways on the presence of X_{est}. Without X_{est}, the social pathways 3 and 4 (Fig. 3) would not exist as forms of fitness (van Hateren, 2015c). If only conventional X_{true} existed, all fitness effects of human sociality must be explained through their effects on inclusive fitness (pathways 1 and 2). Explaining 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 similar, actually or potentially. Such judgment is mediated by X_{est}, which can flexibly define similarity to varying degrees of inclusiveness (e.g., based on clan, region, nationality, or just being human).

A primary problem for explanations based purely on X_{true} 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_{est} offers more freedom. New behaviour, generated partly through trial-and-error, is evaluated, in real time, through X_{est}. Behaviour is then changed to varying degrees, depending on that evaluation. Although the structure of X_{est} 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 *A* cycle and its elaborations). In effect, it performs a fast form of evolution, albeit through a fitness proxy rather than through the actual fitness.

Explanations based on X_{est} are particularly insightful when individual, social, or cultural behaviour is clearly maladaptive when judged by conventional X_{true}. Conventional X_{true} 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—and implicitly judged to be adaptive—by the individuals and social groups displaying them. The reason is that X_{est} may differ from X_{true}, or at least the X_{true} as inferred by independent observers. In some obvious cases, the latter X_{true} may be known to be the more accurate one, the one that indeed applies or will be realized eventually. Discrepancies between X_{est} and such accurate X_{true} can explain, for example, individual and group behaviour arising from mental delusions and delusional ideologies. Then one can conclude that the X_{est} of the individual or group is objectively wrong, that is, different from present and forthcoming X_{true}. However, in other cases, X_{true}, and how it will develop into the future, may be rather uncertain. Then the X_{est} of individuals or groups with an unconventional X_{est} may in fact turn out to be adaptive, eventually.

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 used here (e.g., part of the processes in Fig. 2). However, thus concluding that the self has no solid reality would be wrong when viewed from the perspective of EEST. X_{est} is taken here as a real physiological process with significant causal consequences that are crucial for life. Ultimately, it may decide, via its influence on X_{true}, between flourishing and becoming extinct. X_{est} may be partially constructed, but it is a constrained construction, because an X_{est} that becomes too different from X_{true} is maladaptive.

The unity of the self can be understood from the unity of fitness. Fitness is used in this article as the fitness of individual organisms, not as a trait fitness. Therefore, fitness is unitary, because individuals survive and reproduce as wholes. Thus, there is only one X_{true}, and therefore there should be only one X_{est}. In certain pathologies, the unity of X_{est} is poorly maintained, but that is a sign that something is wrong. It is then likely to produce low fitness because it implies a mismatch between X_{est} and X_{true}. Such a mismatch would not be sustainable on an evolutionary timescale. 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., home, work, 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 X_{true}. Also X_{true} 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 EEST, the self is expected to be continuous because X_{true} is continuous. Continuity does not imply gradualness, because abrupt changes are possible. For example, circumstances may suddenly change, or the individual may go through a 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 for X_{true}, but only by implication for X_{est} 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 EEST, such as self-esteem, efficacy (combining competence and agency), and belonging (similar to relatedness and communion). Distinctiveness is necessary in order to be competitive through pathways 1 and 3 (Fig. 3), 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 X_{true}. Also adverse behaviour could 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 EEST, because the structure of X_{est} represents the individual's goals and overall purpose. If one senses meaning in one's life, this essentially means that X_{est} indicates that X_{true} is high, or is likely to become high. Felt meaninglessness indicates that X_{est} needs work, along any of its pathways.

The Stable and Instable 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_{est} may just follow corresponding abrupt changes in X_{true}. The latter could be produced by abrupt environmental change or by abrupt internally generated change. On average, X_{est} should then remain dynamically aligned with X_{true}. Alternatively, changes in X_{est} may be produced by contingencies that do not similarly change X_{true}. This would correspond to variability of the self that generates a mismatch between X_{est} and X_{true}. Such a mismatch would affect fitness negatively, if sustained.

However, some mismatch between X_{est} and X_{true} is expected, as part of the regular, stochastic functioning of the *A* cycle of Fig. 1. Agency and goal-directedness require variability. Such variability modifies the behavioural dispositions of the individual, and thereby the subsequent X_{est} and X_{true}. In particular when the values of X_{est} and X_{true} are low, large variability and fast changes are to be expected. Then the structure of X_{est}, and thereby the self, may change quickly, and may thus induce a quick change in X_{true} 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 X_{true} 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_{est} and X_{true}.

If the instability continues without finding a favourable X_{est} and X_{true}, it may become maladaptive. The healthy self is thus expected to show at most transient instabilities, as a normal consequence of an adapting X_{est}. If X_{est} changes only slowly over time, this can indicate a situation where the values of both X_{est} and X_{true} 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_{est} can also indicate a situation where X_{est} is high, but X_{true} 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_{est} is low, but X_{true} 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_{est}, behaviour may be varied more than would be optimal. As a final example, X_{est} and X_{true} 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 X_{true} low, or it may cause a decrease of X_{true} even if X_{true} starts out as moderately high. Low X_{true} is then a self-fulfilling prophesy based on an X_{est} that is inaccurately low because of the depression.

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 EEST is an implicit self-estimate of fitness. But fitness is inherently difficult to measure. Straightforward statistical measurement of X, either X_{true} or X_{est}, would require similar individuals and similar circumstances that are difficult to realize for humans. Alternatively, a theoretical model of X might be developed that could be compared with experimental outcomes with variable individual properties and circumstances. But a theoretical model of 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 EEST is not simple is that there is an asymmetry when comparing it with existing theories of the self. As discussed above, EEST incorporates several of the key components of such theories. Therefore, empirical support for these theories will often also support EEST. For testing the current theory specifically, one needs predictions that distinguish it from other theories. One general prediction is that individuals with low X_{est} should show more behavioural variability than individuals with high X_{est}, at least under stable conditions. More specific predictions follow from Figs. 3 and 4, respectively, as detailed below.

An enhanced X_{est} implies an enhanced self-worth, as perceived by the self but partly based on how others are believed to value oneself. Fig. 3 implies that there are four major pathways to increase 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) can then compensate for a (properly scaled) decreased loyalty to the in-group (pathway 4), keeping state self-esteem approximately constant. Similarly, pathways 2 and 4 may be manipulated into opposite directions.

A specific prediction following from Fig. 4b 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_{est} is dynamic. 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_{est} is layered. The aware and symbolic layers of Fig. 4a are presumably produced by repeated internalizations as in Fig. 2. 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.

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 the structure of this process changes 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 ones concern, first, directly producing offspring and, second, supporting individuals with similar genes. The social ones 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. Remarkably, 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.

REFERENCES

Baumeister, R. F., Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*, 497-529.

Baumeister, R. F. (2010). The self. In R. F. Baumeister, E. J. Finkel (Eds.), *Advanced social psychology* (pp. 139-175). New York: Oxford University Press.

Becker, M., Vignoles, V. L., Owe, E., et al. (2014). Cultural bases for self-evaluation: Seeing oneself positively in different cultural contexts. *Personality and Social Psychology Bulletin, 40*, 657-675.

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

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

Carver, C. S., Scheier, M. F. (1982). Control-theory: A useful conceptual framework for personality-social, clinical, and health psychology. *Psychological Bulletin, 92*, 111-135.

Carver, C. S., Scheier, M. F. (2002). Control processes and self-organization as complementary principles underlying behaviour. *Personality and Social Psychology Review, 6*, 304-315.

Crocker, J., Park, L. E. (2004). The costly pursuit of self-esteem. *Psychological Bulletin, 130,* 392-414.

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

Deci, E. L., Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behaviour. *Psychological Inquiry, 11,* 227-268.

Dennett, D. (1992). The self as a centre of narrative gravity. In F. S. Kessel, P. M. Cole, D. L. Johnson (Eds.), *Self and consciousness: multiple perspectives* (pp. 103-115). Hillsdale: Erlbaum Associates.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. Phenomenology and the Cognitive Sciences, 4, 429–452.

Gardner, A., West, S. A. (2010). Greenbeards. *Evolution, 64,* 25-38.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. I & II. *Journal of Theoretical Biology, 7,* 1-52.

Heft, H. (2013). An ecological approach to psychology. *Review of General Psychology, 17,* 162-167.

Heppner, W. L., Kernis, M. H. (2011). High self-esteem: Multiple forms and their outcomes. In S. J. Schwartz, K. Luyckx, V. L. Vignoles (Eds.), *Handbook identity theory and research* (pp. 329-355). New York: Springer.

Landau, M. J., Solomon, S., Pyszczynski, T., Greenberg, J. (2007). On the compatibility of terror management theory and perspectives on human evolution. *Evolutionary Psychology, 5,* 476-519.

Leary, M. R. (1999). Making sense of self-esteem. *Current Directions in Psychological Science, 8,* 32-35.

Leary, M. R., Tangney, J. P. (2012). The self as an organizing construct in the behavioural and social sciences. In M. R. Leary, J. P. Tangney (Eds.), *Handbook of self and identity* (2nd ed., pp. 1-18). New York: The Guilford Press.

Lickliter, R., Honeycutt, H. (2013). A developmental evolutionary framework for psychology. *Review of General Psychology, 17,* 184-189.

Maner, J. K., Kenrick, D. T. (2010). Evolutionary social psychology. In R. F. Baumeister, E. J. Finkel (Eds.), *Advanced social psychology* (pp. 613-653). New York: Oxford University Press.

Masel, J., Trotter, M. V. (2010). Robustness and evolvability. *Trends in Genetics, 26,* 406-414.

Mayr, E. (1961). Cause and effect in biology. *Science, 134,* 1501-1506.

McAdams, D. P. (2001). The psychology of life stories. *Review of General Psychology, 5,* 100-122.

Mead, G. H. (1934). *Mind, self, and society*. Chicago: University of Chicago Press.

Oyserman, D., Elmore, K., Smith, G. (2012). Self, self-concept, and identity. In M. R. Leary, J. P. Tangney (Eds.), *Handbook of self and identity* (2nd ed., pp. 69-104). New York: The Guilford Press.

Pyszczynski, T., Greenberg, J., Solomon, S., Arndt, J. (2004). Why do people need self-esteem? A theoretical and empirical review. *Psychological Bulletin, 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 M. R. Leary, J. P. Tangney (Eds.), *Handbook of self and identity* (2nd ed., pp. 378-404). New York: The Guilford Press.

Rand, D. G., Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences, 17*, 413-425.

Reddy, V. (2003). On being the object of attention: implications for self-other consciousness. *Trends in Cognitive Sciences, 7*, 397-402.

Ryan, R. M., Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*, 68-78.

Scalas, L. F., Morin, A. J. S., Marsh, H. W., Nagengast, B. (2014). Importance models of the physical self: Improved methodology supports a normative-cultural importance model but not the individual importance model. *European Journal of Social Psychology, 44*, 154–174.

Schlomer, G. L., Del Giudice, M., Ellis, B. J. (2011). Parent-offspring conflict theory: An evolutionary framework for understanding conflict within human families. *Psychological Review, 118*, 496-521.

Sedikides, C., Skowronski, J. J. (1997). The symbolic self in evolutionary context. *Personality and Social Psychology Review, 1*, 80-102.

Smeekes, A., Verkuyten, M. (2014). Perceived group continuity, collective self-continuity, and in-group identification. *Self and Identity, 13*, 663-680.

Solomon, S., Greenberg, J., Pyszczynski, T. (2004). The cultural animal: Twenty years of terror management theory and research. In J. Greenberg, S. L., Koole, T. Pyszczynski (Eds.), *Handbook of experimental experiential psychology* (pp. 13-34). New York: The Guilford Press.

Swann, W. B., Jr., Bosson, J. K. (2010). Self and identity. In S. T. Fiske, D. T. Gilbert, G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., pp. 589-628). Hoboken, NJ: John Wiley & Son, Inc.

Thompson, E. (2007). Mind in life: Biology, phenomenology, and the sciences of mind. Cambridge Mass.: Belknap Press.

Tomasello, M. (1993). On the interpersonal origins of self-concept. In U. Neisser (Ed.), *The perceived self: Ecological and interpersonal sources of self-knowledge* (pp. 174-184). New York: Cambridge University Press.

Tomasello, M., Carpenter, M. (2007). Shared intentionality. *Developmental Science, 10*, 121-125.

Tooby, J., Cosmides, L. (1992). The psychological foundations of culture. In J. H. Barkow, L. Cosmides, J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19-136). New York: Oxford University Press.

Trevarthen, C., Aitken, K. J. (2001). Infant intersubjectivity: Research, theory, and clinical applications. *Journal of Child Psychology and Psychiatry, 42*, 3-48.

van Hateren, J. H. (2013). A new criterion for demarcating life from non-life. *Origins of Life and Evolution of Biospheres, 43*, 491–500.

van Hateren, J. H. (2015a). Active causation and the origin of meaning. *Biological Cybernetics, 109*, 33-46.

van Hateren, J. H. (2015b). The origin of agency, consciousness, and free will. *Phenomenology and the Cognitive Sciences, 14*, 979-1000.

van Hateren, J. H. (2015c). Extensive fitness and human cooperation. *Theory in Biosciences, 134*, 127-142.

van Hateren, J. H. (2015d). Causal non-locality can arise from constrained replication. *EPL - Europhysics Letters, 112*, 20004

Vignoles, V. L., Regalia, C., Manzi, C., Golledge, J., Scabini, E. (2006). Beyond self-esteem: Influence of multiple motives on identity construction. *Journal of Personality and Social Psychology, 90*, 308-333.

Vignoles, V. L. (2011). Identity motives. In S. J. Schwartz, K. Luyckx, V. L. Vignoles (Eds.), *Handbook identity theory and research* (pp. 403-432). New York: Springer.

Zahavi, D. (2005). *Subjectivity and selfhood*. Cambridge, Mass.: MIT Press.