Running Head: FITNESS BEATS TRUTH

Fitness Beats Truth in the Evolution of Perception

Chetan Prakash

Department of Mathematics

California State University, San Bernardino

Kyle D. Stephens and Donald D. Hoffman

Department of Cognitive Sciences

University of California, Irvine

Manish Singh

Department of Psychology and Center for Cognitive Science

Rutgers University, New Brunswick

Chris Fields

Caunes Minervois, France

Manish Singh

Department of Psychology and Center for Cognitive Science

Rutgers University

New Brunswick, NJ 08901

Email: manish.singh@rutgers.edu

Abstract

Does natural selection favor veridical percepts—those that accurately (if not exhaustively) depict objective reality? Perceptual and cognitive scientists standardly claim that it does. Here we formalize this claim using the tools of evolutionary game theory and Bayesian decision theory. We state and prove the "Fitness-Beats-Truth (FBT) Theorem" which shows that the claim is false: If one starts with the assumption that perception involves inference to states of the objective world, then the FBT Theorem shows that a strategy that simply seeks to maximize expected-fitness payoff, with no attempt to estimate the "true" world state, does consistently better. More precisely, the FBT Theorem provides a quantitative measure of the extent to which the *fitness-only* strategy dominates the *truth* strategy, and of how this dominance increases with the size of the perceptual space. The FBT Theorem supports the *Interface Theory of Perception* (e.g. Hoffman, Singh & Prakash, 2015), which proposes that our perceptual systems have evolved to provide a species-specific interface to guide adaptive behavior, and not to provide a veridical representation of objective reality.

Keywords: Perception; Veridicality; Evolutionary Psychology; Bayesian Decision Theory; Fitness; Evolutionary Game Theory; Interface Theory of Perception

Introduction

It is standard in the perceptual and cognitive sciences to assume that more accurate percepts are fitter percepts and, therefore, that natural selection drives perception to be increasingly veridical, i.e. to reflect the objective world in an increasingly accurate manner. This assumption forms the justification for the prevalent view that *human* perception is, for the most part, veridical. For example, in his classic book *Vision*, Marr (1982) argued that:

> "We ... very definitely do compute explicit properties of the real visible surfaces out there, and one interesting aspect of the evolution of visual systems is the gradual movement toward the difficult task of representing progressively more objective aspects of the visual world". (p. 340)

Similarly, in his book *Vision Science,* Palmer (1999) states that:

> "Evolutionarily speaking, visual perception is useful only if it is reasonably accurate ... Indeed, vision is useful precisely because it is so accurate. By and large, *what you see is what you get.* When this is true, we have what is called veridical perception ... perception that is consistent with the actual state of affairs in the environment. This is almost always the case with vision."

In discussing perception within an evolutionary context, Geisler and Diehl (2003) similarly assume that:

> "In general, (perceptual) estimates that are nearer the truth have greater utility than those that are wide off the mark."

In their more recent book on human and machine vision, Pizlo et al. (2014) go so far as to say that:

> "…veridicality is an essential characteristic of perception and cognition. It is absolutely essential. *Perception and cognition without veridicality would be like physics without the conservation laws.*" (p. 227, emphasis theirs.)

These statements reflect three assumptions that are useful to distinguish. The first is that all organisms are embedded in and are continually interacting with an "objective" world whose properties can be specified entirely independently of the state or even the existence of any particular organism. The second—the assumption of *veridicality*—is that at least for humans (and presumably for other "higher" organisms), the apparent state of the world is homomorphic to the actual state of the world. The third, generally implicit, assumption is that (at least for humans) this veridical state of the world is the apparent state of the world that is consciously experienced. The actual state of the world is, in particular, assumed to objectively have the attributes assigned to it by human perceptual experience.

Our *perceived* world is three-dimensional, and is inhabited by objects of various shapes, colors, and motions. Perceptual and cognitive scientists thus typically assume that the *objective* world is three-dimensional, and is inhabited by objects of those very shapes, colors, and motions. In other words, they assume that the vocabulary of our perceptual representations is the correct vocabulary for describing the objective world and, moreover, that the specific attributes we perceive typically reflect the actual attributes of the

objective world. These assumptions are embodied within the standard Bayesian framework for visual perception, as we will see in the next section.

This standard assumption of veridical percepts goes hand-in-hand with the framework of *inverse optics* in vision science: It is standardly assumed that the goal of the visual system is to "undo" the effects of optical projection (or rendering) from 3D scenes to 2D images (e.g. Adelson & Pentland, 1996; Pizlo, 2001). This presumably allows vision to "recover" the 3D scene that is most likely to have produced any given image(s). As we will see in Section 2, the inverse optics conception also forms the basis of the standard Bayesian framework for vision. And within this Bayesian formulation, veridicality corresponds to the strategy of finding the 3D scene—the perceived interpretation—that has the highest probability of being the "correct" one, given any image(s).

Many vision scientists agree that perception is not always veridical, but most visual scientists believe that it is at least approximately veridical. While we might not visually experience the correct three-dimensional shape of some, or even most objects, for example, it is almost universally assumed that there are objects in the world that have well-defined, fully-objective and hence completely observer-independent three-dimensional shapes that could, at least in principle, be accurately experienced by an "ideal" observer.

Some proponents of embodied cognition deny the claim that perception is normally veridical (e.g. Chemero, 2009). We agree with them to the extent that they also emphasize guidance of adaptive behavior, rather than veridicality, as the key force in the

evolution of perceptual systems. We disagree, however, with other aspects of the embodied-cognition framework in so far as it minimizes the contribution of information processing and/or representations.

In what follows, veridicality will be represented by the "truth" strategy. We define an alternative "fitness-only" strategy that does not assume veridical perception. In order to compare the two strategies—*truth* vs. *fitness-only*—we place them in competition in an evolutionary resource-game context. Given some sensory inputs, the *truth* strategy attempts to estimate the most probable world interpretation for each input. (In some instances, this process may result in two or more equally probable interpretations, such as in the well-known Necker cube.) It then compares the fitness of these most probable world estimates, and picks the one with the highest fitness. The *fitness-only* strategy, on the other hand, makes no attempt to estimate the most likely world state corresponding to each sensory input. It simply computes the expected fitness corresponding to each input directly, via the posterior distribution (so that the fitness of any world state is weighted by its posterior probability). We prove that organisms whose resource-collecting behavior is governed by the *fitness-only* strategy dominate those utilizing the *truth* strategy. Indeed, our "fitness beats truth" theorem provides a quantitative measure of the extent to which the *fitness-only* strategy dominates the *truth* strategy, as well as how this dominance varies with the size of the perceptual space.

The "fitness beats truth" theorem therefore strongly calls into question the standard view that more accurate percepts are fitter, and hence that natural selection evolves perceptual systems to have more and more veridical percepts. This, in turn, calls into question the received textbook view in vision science that human vision is mostly

veridical, and even that its goal is to "invert optics"—i.e. to "undo" the effects of optical projection from 3D objects to 2D images. If our percepts do not correspond to reality, then 3D objects themselves are simply part of our own species-specific perceptual interface, and not part of objective reality. Capturing this new understanding in formal terms requires a novel framework—different than the standard Bayesian model for vision—which we call *Computational Evolutionary Perception*, and which we outline in the Discussion section (see also Hoffman & Singh, 2012; Singh & Hoffman, 2013; Hoffman, Singh, & Prakash, 2015).

The standard Bayesian framework for visual perception

The standard approach to visual perception treats it as a problem of inverse optics: The "objective world"— generally taken to be 3D scenes consisting of objects, surfaces, and light sources—projects 2D images onto the retinas. Given a retinal image, the visual system's goal is to infer the 3D scene that is most likely to have projected it (e.g. Adelson & Pentland, 1996; Knill & Richards, 1996; Mamassian, Landy, & Maloney, 2002; Shepard, 1994; Yuille & Bülthoff, 1996). Since a 2D image does not uniquely specify a 3D scene, the only way to infer a 3D scene is to bring additional assumptions or "biases" to bear on the problem, based on prior experience—whether phylogenetic or ontogenetic (Feldman, 2013; Geisler et al. 2001). For example, in inferring 3D shape from image shading, the visual system appears to make the assumption that the light source is more likely to be overhead (e.g. Kleffner & Ramachandran, 1992). Similarly, in inferring 3D

shape from 2D contours, it appears to use the assumption that 3D objects are maximally compact and symmetric (e.g. Li et al., 2013).

Formally, given an image x_0, the visual system aims to find the "best" (generally taken to mean "most probable") scene interpretation in the world. In probabilistic terms, it must compare the posterior probability $\mathbb{P}(w|x_0)$ of various scene interpretations w, given the image x_0. By Bayes' Rule, the posterior probability is given by:

$$\mathbb{P}(w|x_0) = \frac{\mathbb{P}(x_0|w)\cdot\mathbb{P}(w)}{\mathbb{P}(x_0)}$$

Since the denominator term $\mathbb{P}(x_0)$ does not depend on w, it plays no essential role in comparing the relative posterior probabilities of different scene interpretations *w*. The posterior probability is thus proportional to the product of two terms: The first is the likelihood $\mathbb{P}(x_0|w)$ of any candidate scene interpretation *w*; this is the probability that the candidate scene *w* could have projected (or generated) the given image x_0. Because any 2D image is typically consistent with many different 3D scenes, the likelihood will often be equally high for a number of candidate scenes. The second term is the prior probability $\mathbb{P}(w)$ of a scene interpretation; this is the probability that the system implicitly assigns to different candidate scenes, even prior to observing any image. For example, the visual system may implicitly assign higher prior probabilities to scenes where the light source is overhead, or to scenes that contain compact objects with certain symmetries. Thus, when multiple scenes have equally high likelihoods (i.e. are equally consistent with the image), the prior can serve as a disambiguating factor.

Application of Bayes' Rule yields a probability distribution on the space of candidate scenes—the posterior distribution. A standard way to pick a single "best" interpretation from this distribution is to choose the world scene that has the maximal posterior probability—one that, statistically speaking, has the highest probability of being the "correct" one, given the image x_0. This is the maximum-a-posteriori or MAP estimate. More generally, the strategy one adopts for picking the "best" answer from the posterior distribution depends on the choice of a loss (or gain) function, which describes the consequences of making "errors," i.e. picking an interpretation that deviates from the "true" (but unknown) world state by varying extents. The MAP strategy follows under a Dirac-delta loss function—no loss for the "correct" answer (or "nearly correct" within some tolerance), and equal loss for everything else. Other loss functions (such as the squared-error loss) yield other choice strategies (such as the mean of the posterior distribution; see e.g. Mamassian et al., 2002). But we focus on the MAP estimate here because it yields, in a well-defined sense, the highest probability of picking the "true" scene interpretation within this framework.

This standard Bayesian approach embodies the "veridicality" or "truth" approach to visual perception. We do not mean, of course, that the Bayesian observer *always* gets the "correct" interpretation. Given the inductive nature of the problem, that would be a mathematical impossibility. It is nevertheless true that:

(i) The space of hypotheses or interpretations from which the Bayesian observer chooses is assumed to correspond to the objective world (i.e. to the space of possible objective-world states). That is, the vocabulary of perceptual

experiences is assumed to be the right vocabulary for describing objective reality.

(ii) Given this setup, the MAP strategy maximizes—statistically speaking—the probability of picking the "true" world state.

The framework described above constitutes the standard Bayesian framework for vision—i.e. the way in which Bayes' Theorem is standardly applied in modern vision science to model various problems in visual perception. This is not, however, the only way to apply Bayes' Theorem to problems in vision, and indeed in the Discussion section we will provide and alternative framework that incorporates not only Bayesian inference, but evolutionary fitness as well.

Evolution and Fitness

The Bayesian framework, with the standard interpretation summarized above, focuses on estimating the world state that has the highest probability of being the "true" one, given some sensory input. This Bayesian estimation involves no explicit notion of evolutionary fitness (although by defining prior probabilities over states *w* of the world, it implicitly builds in the assumption that truer percepts are more fit). As we noted above, approaches based on Bayesian Decision Theory (BDT) do involve a loss (or gain / utility) function. It is important to note, however, that this is quite distinct from a fitness function (defined below). The loss function of BDT describes the consequences of making "errors," i.e. picking an interpretation that deviates from the "true" world state by varying extents. It is therefore a function of two variables: (i) the observer's estimate / interpretation, and

(ii) the "true" state of the world. By contrast, the evolutionary fitness function involves no dependence on the observer's estimate (whereas it does depend on the observer, its state, and the action class in question; see below).

In evolutionary theory, *fitness* is a measure of the probability of transferring genes, and therefore characteristics, into the next generation (Maynard Smith, 1989). The effects on fitness of different decisions or behaviors by an organism or population can be represented by a global fitness function $f(w, o, s, a)$ that depends, in general, on the state w of the world W in which behavior takes place, the organism o executing the behavior (e.g., a lion vs. a rabbit), the organism's state s (e.g., hungry vs. satiated), and the action a that is executed (e.g., feeding vs. mating). Fitness functions vary widely between organisms; indeed the diversity of extant organisms indicates that the correlation between fitness functions for distinct organisms can be arbitrarily small. For any particular organism, the complexity of the fitness function can be expected to increase rapidly as the number of its possible states and actions increases; even the fitness function for a bacterium is extraordinarily complex.

To examine the behavior of $f(w, o, s, a)$ in a game-theoretic context, we can think of organisms of different kinds as competing to gather "fitness points" as they interact within the shared "environment" W (Maynard Smith, 1982). In such a competitive game, natural selection favors percepts and choices that yield more fitness points. For simplicity, we consider evolutionary games between organisms of the same type o, in the same state s, and with only a single available action a. In this case, we can model a *specific fitness function* as simply a (non-negative) real-valued function $f: W \rightarrow [0, \infty)$ defined on the world W.

In order to compare the fitness of different perceptual and/or choice strategies, one pits them against one another in an evolutionary *resource game* (for simulations exemplifying the results of this paper, see, e.g., Mark, Marion, & Hoffman, 2010; Marion, 2013; and Mark, 2013). In a typical game, two organisms employing different strategies compete for available territories, each with a certain number of resources. The first player observes the available territories, chooses what it estimates to be its optimal one, and receives the fitness payoff for that territory. The second player then chooses *its* optimal territory from the remaining available ones. The two organisms thus take turns in picking territories, seeking to maximize their fitness payoffs.

In this case, the quantity of resources in any given territory is the relevant world attribute *w*. That is, *W* is here interpreted as depicting different quantities of some relevant resource. We can then consider a perceptual map $P : W \rightarrow X$, where X is the set of possible perceptual states, together with an ordering on it: *P* picks out the "best" element of *X* in a sense relevant to the perceptual strategy. One may, for instance, imagine a simple organism whose perceptual system has only a small number of distinct perceptual states. Its perceptual map would then be some way of mapping various quantities of the resource to the small set of available perceptual states. As an example, Figure 1 shows two possible perceptual mappings, i.e. two ways of mapping the quantity of resources (here, ranging from 0 through 100) to four available sensory categories (here depicted here by the four colors R, Y, G, B).

In addition, there is a fitness function on *W*, $f : W \rightarrow [0, \infty)$, which assigns a non-negative fitness value to each resource quantity. One can imagine fitness functions that are monotonic (e.g. fitness may increase linearly or logarithmically with the number of

resources), or highly non-monotonic (e.g. fitness may peak for a certain number of resources, and decrease in either direction). Monotonic fitness functions can be expected to be rare; it is possible to have "too much" of a typical resource. The vast majority of fitness functions will be non-monotonic (such as the one shown in Figure 2): too little water and one dies of thirst, too much water and one drowns. Similar arguments apply to the level of salt, or to the proportion of oxygen and indeed most other resources. Given the ubiquitous need for organisms to maintain homoeostasis, and the invariably limited energetic resources available to do so, one expects most fitness functions to be non-monotonic. In what follows, we will consider fitness functions generically, among which monotonic functions constitute an extremely small subset.[1]

One simple consequence of non-monotonic fitness functions is that fitness and "truth" are not in general correlated. It is sometimes argued that a strategy based on fitness works simply because it allows one to approximate the truth. In the absence of any generic correlations between fitness and truth, however, this argument carries little weight. It is in fact not meaningful to view fitness functions as "approximating" the truth. Recall also that, while fitness clearly depends on the world ("truth"), it also depends on the organism, its state, and the action class in question. Thus, considering a different organism, or the same organism in a different state, or in the context of a different action class, will result in very different fitness values—even as the world remains unchanged.

Comparing perceptual strategies: "Truth" vs. "Fitness-only"

In the context of these evolutionary games, in which perceptual strategies compete for resource acquisition, we assume that the organism's behavior depends on three fixed elements: the specific fitness function (in a particular state and for a particular action class), its prior, and its perceptual map from world states (i.e. resource-containing territories) to sensory states (see Figure 3). On any given trial, the organism observes a number of available territories through its sensory states, say $x_1, x_2, \ldots, x_n$. Its goal is to pick one of these territories, seeking to maximize its fitness payoff. One can now consider two possible resource strategies:

The "Truth" strategy: For each of the *n* sensory states, the organism estimates the world state or territory - the Bayesian MAP estimate - that has the highest probability of being the "true" one, given that sensory state. It then compares the fitness values for these n "true" world states. Finally, it makes its choice of territory based on the sensory state x_i that yields the highest fitness. Its choice is thus mediated through the MAP estimate of the world state: it cannot choose a territory that does not qualify as "true." The "Truth" strategy *ignores* any fitness information about possible states of the world other than the one selected as being the "true" state.

The "Fitness-only" strategy: In this strategy, the organism makes no attempt to estimate the "true" world state corresponding to each sensory state. Rather it directly computes the expected fitness payoff that would result from each possible choice of x_i. For a given sensory state x_i, there is a posterior probability distribution (given, as with the Truth strategy, by Bayes' formula) on the possible world states, as well as a fitness value

corresponding to each world state. The organism weights these fitness values by the posterior probability distribution, in order to compute the expected fitness that would result from the choice x_i. And it picks the one with the highest expected fitness.

Results from Evolutionary Game Theory

In an evolutionary game between the two strategies, say *A* and *B,* the *payoff matrix* is as follows:

	$against\ A$	$against\ B$
$A\ plays$	a	b
$B\ plays$	c	d

Here *a*, *b*, *c*, and *d* denote the various payoffs to the row player when playing against the column player. E.g., *b* is the payoff *to A* when playing *B*. We will refer to three main theorems from evolutionary game theory relevant to our analysis, as follows.

We first consider games with infinite populations. These are investigated by means of a deterministic differential equation, called the *replicator equation*, where time is the independent variable and the relative population sizes x_A, x_B are the dependent variables, with $x_A + x_B = 1$ (Taylor and Jonker, 1978, Hofbauer and Sigmund, 1990, Nowak 2006). In this context, there are four generic behaviors in the long run:

Theorem 1. (Nowak 2006) *In a game with an infinite population of two types, A and B, of players, either*

(i) A dominates B (in the sense that a non-zero proportion of A players will eventually take over the whole population), *if* $a \geq c$ *and* $b \geq d$, with at least one of the inequalities being strict;

(ii) B dominates A, if $a \leq c$ *and* $b \leq d$, with at least one of the inequalities being strict;

(iii) A and B coexist, if $a \leq c$ *and* $b \geq d$ (with at least one of the inequalities being strict)*, at a stable equilibrium given by* $x_A^* = \frac{b-d}{b+c-a-d}$ *(and* $x_B^* = 1 - x_A^*$*);*

(iv) The system is bistable, if $a \geq c$ *and* $b \leq d$ (with at least one of the inequalities being strict) *and will tend towards either all A or all B from an unstable equilibrium at the same value of* x_A^* *as above.*

A fifth, non-generic possibility is that $a = c$ *and* $b = d$, in which case we have that *A* and *B* are neutral variants of one another: any mixture of them is stable.

Games with a *finite* population size *N* can be analyzed via a *stochastic*, as against deterministic, approach. The dynamics are described by a birth-death process, called the Moran process (Moran 1958). The results are more nuanced than in the infinite population sized case: there are now *eight* possible equilibrium behaviors, and they are population dependent, not just payoff dependent.

Let ρ_{AB} denote the *fixation probability* of a single *A* individual in a population of *N*-1 *B* individuals replacing (i.e., taking over completely) that population. Similarly, let Let ρ_{BA} denote the *fixation probability* of a single *B* individual in a population of *N*-1 of *A* individuals replacing (i.e., taking over completely) that population. In the absence of any selection, we have the situation of *neutral* drift, where the probability of either of

these events is just $\frac{1}{N}$. We say that *selection favors A replacing B* if $\rho_{AB} > \frac{1}{N}$ and that *selection favors B replacing A* if $\rho_{BA} > \frac{1}{N}$.

By analyzing the probabilities of a single individual of each type interacting with an individual of either type, or of dying off, we can use the payoff matrix above to compute the fitness F_i, when there are *i* entities of type *A*, and the fitness G_i of (the *N-i* individuals) of type *B*. If we set $h_i = F_i - G_i$ $(i = 1, \ldots, N)$, we can see that $h_1 > 0$ implies that *selection favors A invading B,* while $h_{N-1} > 0$ implies that *selection favors B invading A*. There are now sixteen possibilities, depending upon whether selection favors *A* replacing *B* or not; *B* replacing *A* or not; whether selection favors *A* invading *B* or not; and whether selection favors *B* invading *A* or not. Of these, eight are ruled out by a theorem of Taylor, Fudenberg, Sasaki and Nowak (2004). A full description is provided in that paper, along with a number of theorems detailing the possibilities in terms of the payoff values and population size. Their Theorem 6, interpreted below as our Theorem 2, is most relevant to our analysis of evolutionary resource games: it gives conditions under which selection is *independent* of population size and is reproduced below. Interestingly, for finite populations the relationship between payoffs *b* and *c* becomes relevant:

Theorem 2. *In a game with a finite population of two types of players, A and B, if* $b > c, a > c$ *and* $b > d$*, we have for all N,* $h_i > 0\ \forall i$ *and* $\rho_{AB} > \frac{1}{N} > \rho_{BA}$*: selection favors A.*

Finally, we also consider, within large finite populations, the limit of *weak selection*. In order to model the strength of selection, a new parameter *w* is introduced. This parameter, lying between 0 and 1, is a measure of the strength of selection: we write

the fitness of *A* now as $f_i = 1 - w + wF_i$ and the fitness of *B* now as $g_i = 1 - w + wG_i$. When $w = 0$, there is no selection: the fitnesses are equal and we have neutral drift. When $w = 1$, we have selection at full strength. An analysis of the dynamics of the Moran process under weak selection (i.e., in the limit as $w \to 0$), reveals (following Nowak 2006, equation 7.11) that:

Theorem 3. *In a game with a finite population of two types of players, A and B, and with weak selection,* $(a - c) + 2(b - d) > \frac{2(a-c)-(b-d)}{N}$ *implies that* $\rho_{AB} > \frac{1}{N}$. *Thus, if* $a > c$ *and* $b > d$*, for large enough N, selection favors A.*[2]

Evolutionary Resource Games

For our situation of two resource strategies, we may define the payoff matrix as follows:

a: to *Fitness-Only* when playing against *Fitness-Only*	*b*: to *Fitness-Only* when playing against *Truth*
c: to *Truth* when playing against *Fitness-Only*	*d*: to *Truth* when playing against *Truth*

In a game with a very large (effectively infinite) number of players, the *Fitness-Only* resource strategy dominates the *Truth* strategy (in the sense that *Fitness-Only* will eventually drive *Truth* to extinction) if the payoffs to *Fitness-Only* as first player always exceed those of *Truth* as first player, regardless of who the second player is, i.e. if $a \geq c$

and $b \geq d$ and at least one of these is a strict inequality. If neither of these inequalities is strict, then at the least *Fitness-Only* will never be dominated by *Truth*.

Our main claim in this paper is that the *Truth* strategy—attempting to infer the "true" state of the world that most likely corresponds to a given sensory state—confers no evolutionary advantage to an organism. In the next section, we state and prove a theorem—the "Fitness Beats Truth" theorem—which states that *Fitness-Only* will never be dominated by *Truth*. Indeed, the *Truth* strategy will generally result in a lower expected-fitness payoff than the *Fitness-Only* strategy, and is thus likely to go extinct in any evolutionary competition against the *Fitness-Only* strategy. (The statement of the FBT theorem articulates the precise way in which this is true.) We begin, first, with a numerical example that exemplifies this.

Numerical Example of Fitness Beating Truth

We give a simple example to pave the way for the ideas to follow. Suppose there are three states of the world, $W = \{w_1, w_2, w_3\}$ and two possible sensory stimulations, $X = \{x_1, x_2\}$. Each world state can give rise to a sensory stimulation according to the information contained in Table 1. The first two columns give the likelihood values, $\mathbb{P}(x|w)$, for each sensory stimulation, given a particular world state; for instance, $\mathbb{P}(x_1|w_2) = 3/4$. The third column gives the prior probabilities of the world states. The fourth column shows the fitness associated with each world state. If we think of the world states as three different kinds of food that an organism might eat, then these values correspond to the fitness benefit an organism would get by eating one of the foods. With this analogy, w_1 corresponds to an extremely healthful (but relatively rare) food, while

w_2 and w_3 correspond to moderately healthful (and more common) foods, with w_2 being more healthful than w_3 (see Table 1). This setup is the backdrop for a simple game where observers are presented with two sensory stimulations and forced to choose between them.

Using Bayes' theorem we have calculated (see Appendix) that for x_1 the *Truth* (i.e. the maximum-a-posteriori) estimate is w_2, and that for x_2 this estimate is w_3. Thus, if a *Truth* observer is offered a choice between two foods to eat, one that gives it stimulation x_1 and one that gives it stimulation x_2, it will perceive that it has been offered a choice between the foods w_2 and w_3. Assuming that it has been shaped by natural selection to choose, when possible, the food with greater fitness, it will always prefer w_2. So, when offered a choice between x_1 and x_2, the *Truth* observer will always choose x_1, with an expected utility of 5.

Now suppose a *Fitness-only* observer is given the same choice. The *Fitness-only* observer is not at all concerned with which "veridical" food these signals most likely correspond to, but has been shaped by natural selection to only care about which stimulus yields a higher expected fitness. We have calculated (see Appendix) that the expected fitness of sensory stimulation x_1 is 5 and the expected utility of stimulation x_2 is 6.6. Thus, when offered a choice between x_1 and x_2, the *Fitness-only* observer will always, maximizing expected fitness, choose x_2.

The implications of these results are clear. Consider a population of *Truth* observers competing for resources against a population of *Fitness-only* observers, both occupying the niche described by Table 1. Since, in this case, the *Truth* observer's choice minimizes expected fitness and the *Fitness-only* observer's choice maximizes expected

fitness, the *Fitness-only* population will be expected to drive the population of *Truth* observers to extinction. Seeing truth can minimize fitness; thereby leading to extinction. This conclusion is apart from considerations of the extra *energy* required to keep track of truth (see Mark, Marion and Hoffman 2010, for discussion on energy resources).

In psychological terms, the advantage of Fitness-only over Truth in the niche described by Table 1 is due to perceptual ambiguity: the percept x_2 is ambiguous between the excellent resource w_1 and the moderately-good resource w_3. Ambiguous percepts are, however, common; percepts can indeed be ambiguous between very good and very bad outcomes, as stock pickers, sushi connoisseurs and practitioners of serial romance know all too well. The *Truth* player executes a correct MAP estimate, but ignores the possibility of ambiguity along the dimension that actually matters, i.e. fitness. A high value of fitness associated with x_2 is not plausible given the low value of the prior $\mathbf{P}(w_1)$. By choosing the MAP estimate instead of performing a full expected fitness calculation, the *Truth* player "jumps to conclusions" along the most important dimension. From this perspective, employing the Truth strategy is a fallacy of practical reasoning, analogous in its effects to ignoring priors when estimating posterior probabilities (Kahneman, 2011).

Mathematical Background for the Main Theorem

We assume that there is a fixed preliminary map, *p*, which associates to each world state-$w \in W$ a sensory state $x \in X$. And we assume a fitness map on W (recall Figure 3). This places the *Truth* strategy and the *Fitness-only* strategy on a common

footing where they can be set in direct competition against each other within the context of an evolutionary resource game.

We begin with some definitions and assumptions regarding these spaces and maps.

It will suffice for a basic understanding of what follows to think of W as a finite set (as in the example in 6.1).[3] In general, we take the *world* W to be a compact regular Borel space whose collection of measurable events is a σ-algebra, denoted $\mathcal{B}$.[4] We assume that $< W, \mathcal{B} >$ comes equipped with an *a priori probability measure* μ on $\mathcal{B}$. We will consider only those probability measures μ that are absolutely continuous with respect to the Borel measure on $\mathcal{B}$. That is, if we write $\mathrm{d}w$ for the uniform, or Borel, probability measure on W, then the a priori measure satisfies $\mu(dw) = g(w)\,\mathrm{d}w$. Here $g: W \to \mathbb{R}_+$ is some non-negative measurable function, called the *density* of μ, satisfying $\int g(w)\,\mathrm{d}w = 1$. We will take any such density to be continuous, so that it always achieves its maximum on the compact set W. This constitutes the structure of the world: a structure that applies to most of the studied biological and perceptual situations.

We may assume that a given species interacts with its world, employing a perceptual mapping that "observes" the world via a measurable map $p: W \to X$. We refer to this as a *pure perceptual map* because it involves no dispersion: each world state can yield only a single sensory state x. We assume that the set of perceptual states X is a finite set, with the standard discrete σ-algebra $\mathcal{X}$, i.e., its power set (so that *all* subsets of X are measurable). In the general case, the perceptual map may have dispersion (or noise), and is mathematically expressed as a Markovian kernel $p: W \times \mathcal{X} \to [0,1]$. That is, for every element w in W, the kernel p assigns a probability distribution on X (hence it assigns a

probability value to each measurable subset of *X*). Because *X* is finite and all of its subsets are measurable, here the kernel may be viewed simply as assigning, for every element *w* in *W*, a probability value to each element of *X*.

General Perceptual Mappings and Bayesian Inference

We use the letter $\mathbb{P}$ to indicate any relevant probability. Bayesian inference consists in a computation of the conditional probability measure $\mathbb{P}(\mathrm{d}w \mid x)$ on the world, given a particular perception x in *X*. The *likelihood* function is the probability $\mathbb{P}(x \mid w)$ that a particular world state w could have given rise to the observed sensory state x. Then the conditional probability distribution $\mathbb{P}(dw \mid x)$ is the *posterior* probability distribution in a (partially) continuous version of Bayes formula:

$$\mathbb{P}(\mathrm{d}w \mid x) = \frac{\mathbb{P}(x \mid w)\ \ \mathbb{P}(\mathrm{d}w)}{\mathbb{P}(x)}.$$

Since μ, the prior on *W*, has a density g with respect to the Borel measure $\mathrm{d}w$, we can recast this formula in terms of g: indeed, $\mathbb{P}(\mathrm{d}w \mid x)$ also has a *conditional density,* $g(w \mid x)$, with respect to the Borel measure[5] and we obtain

$$g(w \mid x) = \frac{\mathbb{P}(x \mid w)\ \ g(w)}{\int \mathbb{P}(x \mid w')\ g(w')}.$$

We now define a *maximum a posteriori estimate* for x in *X* to be any w_x at which this conditional density is maximized: $g(w_x \mid x) = \max\{\, g(w \mid x) \mid w \in W\}$. At least one such maximum will exist, since g is bounded and piecewise continuous; however, there could be multiple such estimates for each x.

For a given sensory state x, the only world states that could have given rise to it lie in the *fiber over* x, i.e., the set $p^{-1}\{x\} \subset W$. So, for a given x, the mapping $w \to \mathbb{P}(x \mid w)$ takes the value 1 on the fiber, and is zero everywhere else. This mapping may thus be viewed as the *indicator function* of this fiber. We denote this indicator function by $1_{p^{-1}\{x\}}(w)$.

For a pure mapping the *conditional density* is just

$$g(w \mid x) = \frac{g(w) \cdot 1_{p^{-1}\{x\}}(w)}{\mu(p^{-1}\{x\})},$$

where $\mu(p^{-1}\{x\})$ is the *a priori* measure of the fiber.

In this special case of a pure mapping that has given rise to the perception x, we can diagram the fiber over x on which this average fitness is computed. This is the shaded region in figure 3 below.

Expected Fitness

Given a *fitness function* $f: W \to [0, \infty)$ that assigns a non-negative fitness value to each world state, the *expected fitness* of a perception x is

$$F(x) = \int f(w)\, \mathbb{P}(\mathrm{d}w \mid x) \; = \int f(w) g(w \mid x)\, \mathrm{d}w\,.$$

Two Perceptual Strategies

We may build our two perceptual strategies P_T, P_F, called "*Truth*" and "*Fitness-Only*" respectively, as compositions of a "sensory" map $p: W \to X$ that recognizes territories and "ordering" maps $d_T, d_F: X \to X$, where $P_T = d_T \circ p$ and $P_F = d_F \circ p$. That is, the map d_T re-names the elements of *X* by re-ordering them, so that the best one, in terms of its *Bayesian MAP* estimate, is now the first, x_1, the second best is x_2 etc. The map d_F, on the other hand, re-orders the elements of *X* so that the best one, in terms of its *expected fitness* estimate, is x_1, the second best is x_2 etc. The organism picks x_1 if it can, x_2 otherwise.

The "Fitness Beats Truth" Theorem

We can now state our main theorem, which applies in various contexts of evolutionary games: with infinite populations, finite populations with full selection, and (sufficiently) large finite populations with weak selection.

Theorem 4: *Over all possible fitness functions and a priori measures, the probability that the Fitness-only perceptual strategy strictly dominates the Truth strategy is at least* $(|X| - 3)/(|X| - 1)$*, where* $|X|$ *is the size of the perceptual space. As this size increases, this probability becomes arbitrarily close to 1: in the limit, Fitness-only will generically strictly dominate Truth, so driving the latter to extinction.*

Proof: For any given x, the Bayesian MAP estimate is a world point w_x (it is the w_x such that $g(w_x|x) = \max\{ g(w|x) \mid w \in W\}$). This point has fitness $f(w_x)$; let x_M be that x

for which the corresponding $f(w_x)$ is maximized. Then this x_M is, if available, is chosen by *Truth* and $F(x_M)$, its expected fitness, is the payoff to *Truth*.

On the other hand, the fitness payoff to the *Fitness-only* strategy is, by definition the maximum expected fitness $F(x_I)$ over *all* fibers, so clearly, $F(x_M) \leq F(x_I)$.

As defined earlier, our evolutionary game has as payoffs, *a*: to *Fitness-only* when playing against *Fitness-only; b*: to *Fitness-only* when playing against *Truth*; *c*: to *Truth* when playing against *Fitness-only; d*: to *Truth* when playing against *Truth*.

We need to estimate the probability that $a \geq c$ and $b \geq d$. We assume that if both strategies are the same, then each has an even chance of picking its best territory first. Thus if, in any given play of the game, two competing strategies both take a particular territory as their most favored one, then each strategy has an even chance of picking that territory and then the other strategy picks its next-best choice of territory.

If *Fitness-only* meets *Fitness-only,* then each has an even chance of choosing its best territory, say x_I; the second to choose then chooses its second best territory, say x_I'. Since each player has an equal chance of being first, we have

$$a = [F(x_I) + F(x_I')]/2.$$

If *Truth* meets *Fitness-only,* its choice will be x_M, as long as this value differs from x_I. In this instance, we have $a > c$. If, however, $x_M = x_I$, half the time *Truth* will choose x_M and the other half x_M', where x_M' is the second best of the optimal territories for *Truth*. Hence

$$c = \begin{cases} F(x_M), & \text{if different best territories} \\ \dfrac{F(x_I) + F(x'_M)}{2}, & \text{if same best territories} \end{cases}$$

and since $F(x_M) \leq F(x_I)$ and $F(x'_M) \leq F(x'_I)$ we get $a \geq c$.

What happens when *Fitness-only* meets *Truth*? If *Fitness-only* goes first, the payoff will be $b = F(x_I)$. The same is true if *Truth* goes first and the two best territories are different. If, however, the two best territories are the same, then the payoff to *Fitness-only* is its second-best outcome:

$$b = \begin{cases} F(x_I), & \text{if different best territories} \\ F(x'_I), & \text{if same best territories} \end{cases}$$

Finally, when *Truth* meets *Truth,* we have that

$$d = \frac{[F(x_M) + F(x'_M)]}{2}.$$

So it is clear that $b \geq d$, as long as the two best territories are different. If they are the same, this may or may not be true: it depends on the relative size of the average d and $F(x'_I)$ (which, in this instance, also lies in between $F(x'_M)$ and $F(x_I) = F(x_M)$).

Now, *a priori*, there is no canonical relation between the functions *f* and *g*, both of which can be pretty much arbitrary (in fact, *f* need not even be continuous anywhere, and could have big jumps as well as bands of similar value separated from each other in *W*). Also, generically the maximum for each strategy will be unique and also the expected fitnesses for the different territories will all be distinct.

Thus, generically, $F(x_M)$ and $F(x'_M)$ will be different from and indeed strictly less than $F(x_I)$ (and also $F(x'_M) < F(x'_I)$). The only impediment to the domination of *Fitness-only* can come from the situation where the best territories for both strategies are the same. Let X have size $|X| = n$. There are n ways the two strategies can output the same territory, out of the $n!/[2!\,(n-2)!]$ ways of pairing territories. Thus, across all possibilities for f and g, the probability that randomly chosen fitness and *a priori* measures would result in choosing the *same* territory for both strategies, i.e., that $F(x_M) = F(x_I)$, will happen with a probability of

$$\frac{n}{n!/_{2!\,(n-2)!}} = \frac{2}{n-1}$$

Finally, the probability of the two fibers being different is the complement: $1 - \frac{2}{n-1} = \frac{n-3}{n-1}$. □

Dynamic Fitness Functions

A possible objection to the applicability of this theorem is that it seems to assume a *static* fitness function, whereas realistic scenarios may involve *changing*, or even rapidly changing, fitness functions. However, a close scrutiny of the proof of the theorem reveals that at any moment, the fitness function *at that time* being the same for both strategies, the relative payoffs remain in the same *generic* relation as at any other moment. Hence the theorem also applies to dynamically changing fitness functions.

Discussion

As we noted in the *Introduction*, it is standard in the literature to assume that more accurate percepts are fitter percepts and that, therefore, natural selection drives perception to increasing veridicality—i.e. to correspond increasingly to the "true" state of the objective world. This assumption informs the prevalent view that human percepts are, for the most part, veridical.

Our main message in this paper has been that, contrary to this prevalent view, attempting to estimate the "true" state of the objective world corresponding to a given sensory input confers no evolutionary benefit whatsoever. Specifically: If one assumes that perception involves inference to states of the objective world, then the FBT Theorem shows that a strategy that simply seeks to maximize expected-fitness payoff, with no attempt to estimate the "true" world state, does consistently better (in the precise sense articulated in the statement of the FBT Theorem). In an evolutionary competition, this "fitness" strategy would drive the "truth" strategy to extinction.

In our view, the very idea of attempting to estimate the "true" state of the world is wrong-headed. Perceptual scientists generally take "objects," "surfaces," "light sources," etc. to be part of the objective world that perceptual systems are trying to "recover." But these entities are all still part of our own perceptual interface (Hoffman, Singh & Prakash, 2015), though perhaps enhanced by precise measurement procedures—which themselves, of course, take place within the interface. For the purpose of the current analysis, it was important that we place the two strategies to be compared—"Truth" and "Fitness-only"—within a common framework involving Bayesian inference from the space of sensory

states, X, to the objective world, W (recall Figure 3). This allowed us to place the two strategies on the same footing, so that they could be placed in direct competition against each other. However, the basic FBT result strongly supports the view that the very idea of perception as probabilistic inference *to states of the objective world* is misguided. Perception is indeed fruitfully modeled as probabilistic inference, but the inference is over a space of perceptual representations, not over a space of objective world states.

These ideas are part of larger theory, the *Interface Theory of Perception*, that we have described in detail elsewhere (Hoffman, 2009; Hoffman & Prakash, 2014; Hoffman & Singh, 2012; Hoffman, Singh, & Prakash, 2015; see also Koenderink, 2011; 2013; 2014; von Uexküll, 1934). For the purposes of the current discussion, the key point is that the *standard Bayesian framework for vision* conflates the interpretation space (or the space of perceptual hypotheses from which the visual system much choose) with the objective world (or, to be more precise, with the space of possible objective-world states). This is a mistake; it is essentially the assumption that the language of our perceptual representations is the correct language for describing objective reality—rather than simply a species-specific interface that has been shaped by natural selection. In our framework, the probabilistic inference that results in perceptual experience takes place in a space of perceptual representations, say, X_1, that may have no homomorphic relation whatsoever to W. This extended framework of *Computational Evolutionary Perception* (which incorporates fitness as well) is sketched in Figure 5 (see Hoffman & Singh, 2012; Hoffman, Singh, & Prakash, 2015; Singh & Hoffman, 2013).

Thus, when we see an object as having a certain 3D shape, it is because the probabilistic inference in the relevant perceptual space X_1 resulted in that 3D

interpretation. But the perceptual space X_1 is not the objective world, nor is it homomorphic to it. It is simply a representational format that has been crafted by natural selection in order to support more effective interactions with the environment (in the sense of resulting in higher expected-fitness payoff, and of better predicting the results of our actions back in our perceptual space). In other words, a more complex or higher-dimensional representational format (such as one involving 3D representations in X_1, in place of just 2D representations in X_0) evolves because it permits a higher-capacity channel $P_1 : W \to X_1$ for expected fitness (see Figure 5). But this does not in any way entail that this representational format somehow more closely "resembles" the objective world. Evolution can fashion perceptual systems that are, in this sense, ignorant of the objective world because natural selection depends only on fitness and not on seeing the "truth."

These considerations strongly undermine the standard assumptions that seeing more veridically enhances fitness, and that therefore one can expect that human perception is largely veridical. As human observers, we are prone to imputing structure to the objective world that is properly part of our own perceptual experience. Our *perceived* world is three-dimensional and populated with objects of various shapes, colors, and motions, and so we tend to conclude that the *objective world* is as well. But if, as the *Fitness-beats-Truth Theorem* shows, evolutionary pressures do not push perception in the direction of being increasingly reflective of objective reality, then such imputations have no logical basis whatsoever.[6]

On the narrower question of whether perceptual systems simply pick a single interpretation (a point estimate) based on the posterior distribution, or store and use the full posterior distribution, a version of this question remains applicable even once we drop the idea of making perceptual inferences back in the objective world (and the concomitant idea of maximizing "truth"). It is clear that using the full posterior distribution allows for greater power and flexibility, e.g. in tailoring the posterior distribution to different contexts and task demands that involve different utility functions. Indeed, empirical evidence suggests that human observers represent at least the mean and variance of posterior distributions, and use this information in a near-optimal manner in making perceptual and sensorimotor decisions (e.g. Trommershäuser, Maloney, & Landy, 2003; Graf, Warren, & Maloney, 2005; Koerding & Wolpert, 2006). According to this approach, contexts such as the conscious visual perception of an object (where we typically see a single interpretation, rather than a distribution or "smear" of possible percepts) result from the application of specific utility functions that collapses the full posterior distribution to a single "best" interpretation (e.g. Maloney, 2002; Maloney & Mamassian, 2009). Formally treating such cases within the context of ITP would require incorporating aspects of Bayesian Decision Theory into our Computational Evolutionary Perception framework, something we plan to do in future work.

Acknowledgments

We thank, Federico Faggin for illuminating discussions. This work has been partially funded by the Federico and Elvia Faggin Foundation.

References

Adelson E. H., & Pentland A. (1996). The perception of shading and reflectance. In: D Knill and W Richards (Eds.), *Perception as Bayesian Inference*, pp. 409-423. New York, NY: Cambridge University Press.

Chemero, A. (2009). *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press.

Feldman J. (2013) Tuning your priors to the world. *Topics in Cognitive Science*, 5, 13-34.

Geisler, W. S. and Diehl, R. L. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. Cognitive Science, 27, 379-402.

Graf, E. W., Warren, P. A. & Maloney, L. T**.** (2005). Explicit estimation of visual uncertainty in human motion processing. *Vision Research*, *45,* 3050-3059.

Hofbauer, J., and Sigmund, K. (1998). *Evolutionary games and population dynamics.* Cambridge: Cambridge University Press.

Hoffman, D. D. (2009). The interface theory of perception. In S. Dickinson, M. Tarr, A. Leonardis, B. Schiele, (Eds.) *Object categorization: computer and human vision perspectives.* New York: Cambridge University Press, pp. 148–165.

Hoffman, D. D., & Prakash, C. (2014). Objects of consciousness. Frontiers of Psychology, 5:577. DOI: 10.3389/fpsyg.2014.00577.

Hoffman, D. D., & Singh, M. (2012). Computational evolutionary perception. Perception, 41, 1073–1091.

Hoffman, D. D., Singh, M., & Mark, J. T. (2013). Does natural selection favor true perceptions? Proceedings of the SPIE 8651, Human Vision and Electronic Imaging XVIII, 865104. DOI: 10.1117/12.2011609.

Hoffman, D. D., Singh, M., & Prakash, C. (2015). Interface Theory of Perception. Psychonomic Bulletin & Review. DOI 10.3758/s13423-015-0890-8

Kleffner, D., & Ramachandran, V. (1992). On the perception of shape from shading. *Perception and Psychophysics, 52*, 18–36.

Knill, D. and Richards, W. (1996). *Perception as Bayesian inference.* Cambridge University Press, Cambridge, UK.

Koenderink, J. J., van Doorn, A., de Ridder, H., Oomes, S. (2010). Visual rays are parallel. *Perception, 39,* 1163-1171.

Koenderink, J. J. (2011). Vision as a user interface. Human Vision and Electronic Imaging XVI, Interface theory of perception 55 SPIE Vol. 7865. doi: 10.1117/12.881671.

Koenderink, J. J. (2013). World, environment, umwelt, and inner-world: a biological perspective on visual awareness, Human Vision and Electronic Imaging XVIII, SPIE Vol. 8651. doi:10.1117/12.2011874.

Koenderink, J. J. (2014). The all seeing eye? *Perception, 43,* 1–6.

Koerding, K.P. and Wolpert, D. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences, 10,* 319-326.

Li, Y., Sawada, T., Shi, Y., Steinman, R. M., & Pizlo, Z. (2013). Symmetry is the sine qua non of shape. In S. Dickinson & Z. Pizlo (Eds.), *Shape perception in human and computer vision* (pp. 21–40). London: Springer.

Loomis, J. M. (2014). Three theories for reconciling the linearity of egocentric distance perception with distortion of shape on the ground plane. Psychology & Neuroscience, 7, 3, 245-251.

Maloney, L. T. (2002), Statistical decision theory and biological vision. In Heyer, D. & Mausfeld, R. [Eds], *Perception and the Physical World: Psychological and Philosophical Issues in Perception*. New York: Wiley, pp. 145-189.

Maloney, L. T. & Mamassian, P. (2009), Bayesian decision theory as a model of visual perception: Testing Bayesian transfer. Visual Neuroscience, 26, 147-155.

Mamassian P, Landy M, Maloney L T (2002). Bayesian modeling of visual perception. In: *Probabilistic Models of the Brain: Perception and Neural Function*, Eds. R Rao, B Olshausen, M Lewicki (Cambridge, MA: MIT Press) pp 13 – 36.

Marion, B. (2013). The Impact of Utility on the Evolution of Perceptions. *Dissertation, University of California, Irvine,* 2013.

Mark, J., Marion, B. and Hoffman, D. D. (2010). Natural selection and veridical perception. Journal of theoretical Biology 266, 504-515.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Moran, P. A. P. 1958. Random processes in genetics. P. Camb. Philos. Soc. 54: 60– 71. 1962. *The statistical processes of evolutionary theory*. Oxford: Clarendon Press.

Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life.* Belknap Harvard University Press, Cambridge, MA.

Palmer, S. (1999). *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge, MA.

Pizlo, Z., Li, Y., Sawada, T., & Steinman, R.M. (2014). *Making a machine that sees like us.* New York: Oxford University Press.

Shepard, R. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review, 1(1),* 2-28.

Singh, M., & Hoffman, D. D. (2013). Natural selection and shape perception: shape as an effective code for fitness. In S. Dickinson and Z. Pizlo (Eds.) *Shape perception in human and computer vision: an interdisciplinary perspective.* New York: Springer, pp. 171–185.

Taylor, C, Fudenberg, D., Sasaki, A. & Nowak, M. Evolutionary Game Dynamics in Finite Populations. *Bulletin of Mathematical Biology* (2004) **66**, 1621–1644

Taylor, P. D., and L. B. Jonker. 1978. "Evolutionary stable strategies and game dynamics." *Math. Biosci. 40:* 145– 156.

Trommershäuser, J., Maloney, L. T. & Landy, M. S. (2003), Statistical decision theory and rapid, goal-directed movements. *Journal of the Optical Society A*, 1419-1433.

von Uexküll, J. (2010). A Foray into the Worlds of Animals and Humans: with a Theory of Meaning. Translated by J. D. O'Neil. University of Minnesota Press. (Original work published in 1934.)

Yuille, A., & Bülthoff, H. (1996). Bayesian decision theory and psychophysics. In: *Perception as Bayesian inference*, D. Knill and W. Richards, Eds., New York: Cambridge University Press.

Appendix: Calculations for the numerical example in Table 1

In this appendix we perform the Bayesian and expected-fitness calculations using the data given in Table 1.

To compute the *Truth* estimates, we first need the probability of each stimulation $\mathbb{P}(x_1)$ and $\mathbb{P}(x_2)$. These can be computed by marginalizing over the priors in the world as follows:

$$\mathbb{P}(x_1) = p(x_1|w_1)\mu(w_1) + \mathrm{p}(x_1|w_2)\mu(w_2) + \mathrm{p}(x_1|w_3)\mu(w_3) = \frac{1}{4}.\frac{1}{7} + \frac{3}{4}.\frac{3}{7} + \frac{1}{4}.\frac{3}{7} = \frac{13}{28}$$

$$\mathbb{P}(x_2) = p(x_2|w_1)\mu(w_1) + \mathrm{p}(x_2|w_2)\mu(w_2) + \mathrm{p}(x_2|w_3)\mu(w_3) = \frac{3}{4}.\frac{1}{7} + \frac{1}{4}.\frac{3}{7} +$$

$\frac{3}{4}.\frac{3}{7} = \frac{15}{28}$

By Bayes' Theorem, the posterior probabilities of the world states, given x_1, are

$$p(w_1|x_1) = p\big(x_1|w_1\big)\big).\frac{\mu(w_1)}{\mathbb{P}(x_1)} = \frac{1}{4}.\frac{1}{7}/\frac{13}{28} = \frac{1}{13}$$

$$p(w_2|x_1) = p\big(x_1|w_2\big)\big).\frac{\mu(w_2)}{\mathbb{P}(x_1)} = \frac{3}{4}.\frac{3}{7}/\frac{13}{28} = \frac{9}{13}$$

$$p(w_3|x_1) = p\big(x_1|w_3\big)\big).\frac{\mu(w_3)}{\mathbb{P}(x_1)} = \frac{1}{4}.\frac{3}{7}/\frac{13}{28} = \frac{3}{13}$$

Thus the maximum *a posteriori*, or *Truth* estimate for stimulus x_1 is w_2.

Posterior probabilities of the world states, given s_2, are:

$$p(w_1|x_2) = p\big(x_2|w_1\big)\big).\frac{\mu(w_1)}{\mathbb{P}(x_2)} = \frac{3}{4}.\frac{1}{7}/\frac{15}{28} = \frac{1}{5}$$

$$p(w_2|x_2) = p\big(x_2|w_2\big)\big).\frac{\mu(2)}{\mathbb{P}(x_2)} = \frac{1}{4}.\frac{3}{7}/\frac{15}{28} = \frac{1}{5}$$

$$p(w_3|x_2) = p\big(x_2|w_3\big)\big).\frac{\mu(w_3)}{\mathbb{P}(x_2)} = \frac{3}{4}.\frac{3}{7}/\frac{15}{28} = \frac{3}{5}$$

Thus the maximum *a posteriori*, or *Truth* estimate for stimulus x_2 is w_3.

Finally, the expected-fitness values of the different sensory stimulations x_1 and x_2 are, respectively:

$$F(x_1) = p(w_1|x_1)f(w_1) + p(w_2|x_1)f(w_2) + p(w_3|x_1)f(w_3)$$

$$= \frac{1}{13}.20 + \frac{9}{13}.4 + \frac{3}{13}.3 = 5;$$

$F(x_2) = p(w_1|x_2)f(w_1) + p(w_2|x_2)f(w_2) + p(w_3|x_2)f(w_3) = \frac{1}{5}.20 + \frac{1}{5}.4 + \frac{3}{5}.3 =$ 6.6.

Thus x_2 has a larger expected fitness than x_1.

Footnotes

[1] From a purely mathematical point of view, the set of monotonic fitness functions is an extremely small subset of the set of all functions on a given domain. That is to say, there are "many more" non-monotonic functions than monotonic ones; hence a random sampling of fitness functions is much more likely to yield a non-monotonic one.

[2] The value of N at which this happens depends upon the payoff matrix, but can be arbitrarily large over the set of all payoff matrices satisfying $a > c$ *and* $b > d$.

[3] In this case, all the integral signs can be replaced by summations.

[4] An example is a closed rectangle in some k-dimensional Euclidean space, such as the unit interval $[0, 1]$ in one dimension, or the unit square in two.

[5] That is, $\mathbb{P}(\mathrm{d}w \mid x) = g(w|x)\mathrm{d}w$.

[6] See also the *Invention of Space-Time Theorem* in Hoffman, Singh, & Prakash (2015).

Tables

	Likelihood of w_j *given* x_1 $\mathbb{P}(x_1 \mid w_j)$	*Likelihood of* w_j *given* x_2 $\mathbb{P}(x_2 \mid w_j)$	*Prior* $\mathbb{P}(w_j)$	*Fitness* $f(w_j)$
w_1	1/4	3/4	1/7	20
w_2	3/4	1/4	3/7	4
w_3	1/4	3/4	3/7	3

Table 1: Likelihood functions, priors and fitness for our simple example where the *Truth* observer *minimizes* expected fitness, while *Fitness-only* observer *maximizes* it.

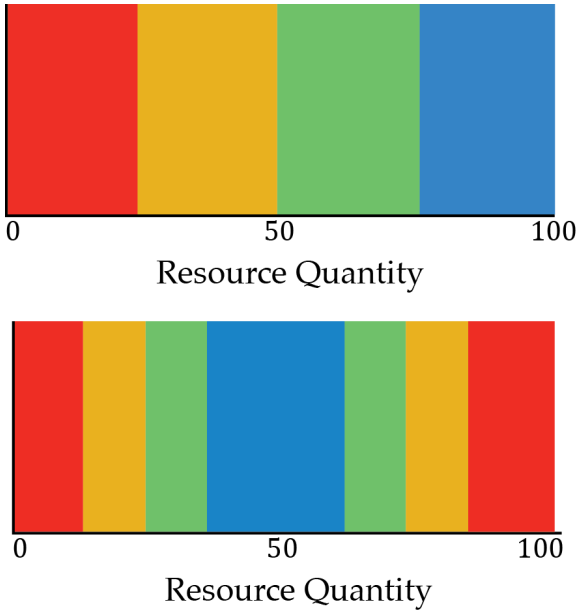

Figure 1. A simple example showing two different perceptual mappings $P : W \rightarrow X$ from world states, $W = [1, 100]$ to sensory states $X = \{R, Y, G, B\}$.

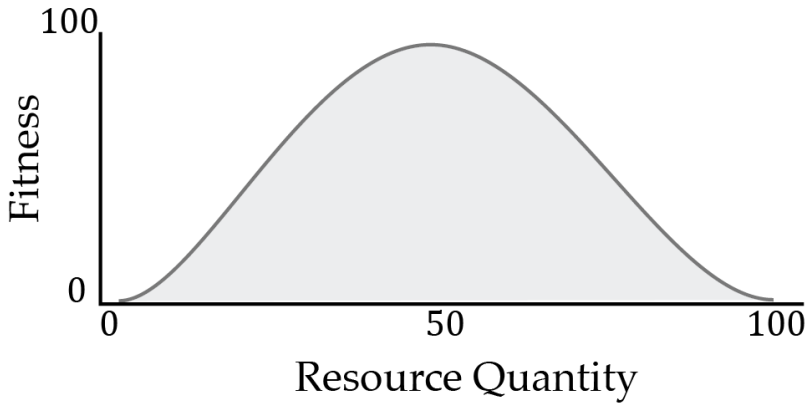

Figure 2. An example of a non-monotonic fitness function $f: W \to [0, \infty)$. Fitness is maximal for an intermediate value of the resource quantity and decreases in either direction. Given the ubiquitous need for organisms to main homoeostasis, one expects that such fitness functions are quite common.

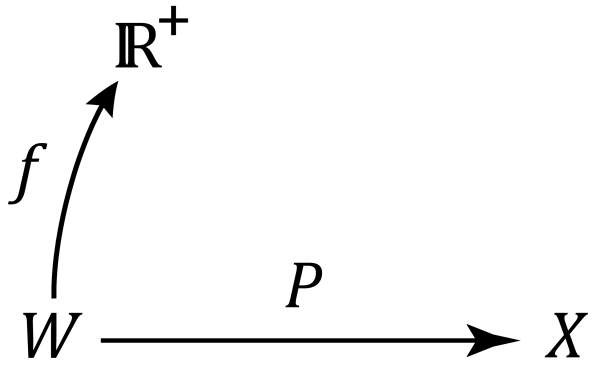

Figure 3. The framework within which we define the two resource strategies. We assume a fixed perceptual map $P : W \to X$ as well as a fixed fitness function $f : W \to [0, \infty)$. Given a choice of available territories sensed through the sensory states, say x_1, $x_2, \ldots, x_n$, the organism's goal is to pick one of these, seeking to maximize its fitness payoff. Note that the "Fitness only" strategy employs Bayesian estimation while rejecting the interpretative assumptions usually associated with it.

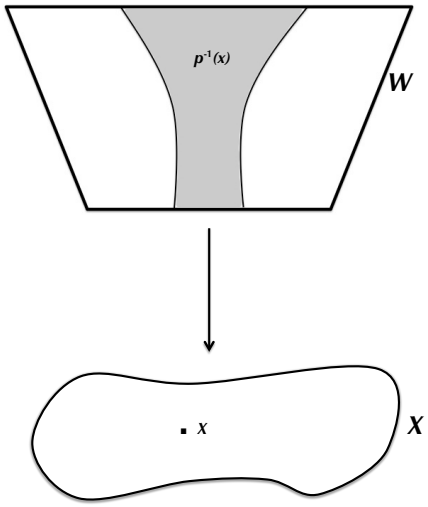

Figure 4. The expected fitness of x is the average, using the *posterior probability*, over the fiber $p^{-1}(x)$.

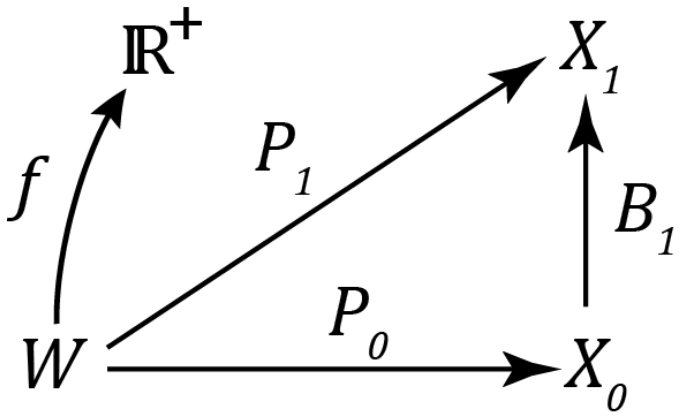

Figure 5. The framework of *Computational Evolutionary Perception* in which perceptual inferences take place in a space of representations X_1 that is not isomorphic or homomorphic to W. The more complex representational format of X_1 evolves because it permits a higher-capacity channel $P_1 : W \rightarrow X_1$ for expected fitness, thereby allowing the organism to choose and act more effectively in the environment (i.e. in ways that result in higher expected-fitness payoffs).