# Fitness Beats Truth in the Evolution of Perception

Chetan Prakash[a*], Kyle D. Stephens[b], Donald D. Hoffman[b],

Manish Singh[c], Chris Fields[d]

[a] *Department of Mathematics, California State University, San Bernardino, CA 92407;*

[b] *Department of Cognitive Sciences, University of California, Irvine*

[c] *Department of Psychology and Center for Cognitive Science, Rutgers University, New Brunswick,*

[d] *243 West Spain Street, Sonoma, CA 95476*

## Abstract

Does natural selection favor veridical perceptions—those which accurately, though perhaps not exhaustively, depict objective reality? Prominent vision scientists and evolutionary theorists claim that it does. Here we formalize this claim using the tools of evolutionary game theory and Bayesian decision theory. We then present and prove a "Fitness-Beats-Truth (FBT) Theorem" which shows that the claim is false. We find that increasing the complexity of objective reality, or perceptual systems, or the temporal dynamics of fitness functions, increases the selection pressures against veridical perceptions. We illustrate the FBT Theorem with a specific example in which veridical perception minimizes expected fitness payoffs. We conclude that the FBT Theorem supports the "interface theory of perception," which proposes that our senses have evolved to hide objective reality and guide adaptive behavior. It also supports the assertion of some proponents of embodied

[*] Corresponding author. cprakash@csusb.edu; +1 (909) 537-5390

cognition that "representing the animal-independent world is not what action-oriented representations are supposed to do; they are supposed to guide action" (Chemero, 2009).

**Keywords:** Replicator Dynamics; Interface Theory of Perception; Evolutionary Game Theory; Universal Darwinism; Sensation

## 1. Introduction

It is standard in the perceptual and cognitive sciences to assume that more accurate perceptions are fitter perceptions 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.)

If human perception is in fact veridical, it follows that the objective world shares the attributes of our 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 so inhabited. 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, which we consider in the next section.

Some proponents of embodied cognition reject the claim that perception is normally veridical. For instance, Chemero (2009) argues that "… perceptual systems evolved to guide behavior. Neither humans nor beetles have action-oriented representations that represent

the animal-independent world exactly correctly. Indeed, representing the animal-independent world is not what action-oriented representations are supposed to do; they are supposed to guide action. So the set of human affordances, that is, action-oriented representeds, is just as tightly geared to human needs and sensorimotor capacities as those of other types of animal. This leaves us with a multiplicity of conflicting sensorimotor systems, each of which is appropriate for guiding the adaptive behavior of animals whose systems they are." The FBT Theorem, which we present below, supports Chemero's claim. It is supported, in turn, by specific examples of non-veridical perceptions, such as those discussed by Loomis (2004) and Koenderink et. Al. (2010).

## 2. The standard Bayesian framework for visual perception

The standard approach to visual perception treats it as a problem of inverse optics: The "objective world"—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; Feldman, 2013; 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). 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 scenes 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, in a well-defined sense, it yields 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. By this 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. That is, the vocabulary of perceptual experiences is assumed to 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.

## 3. Evolution and Fitness

The Bayesian framework, summarized above, focuses on estimating the world state that has the highest probability of being the "true" one, given some sensory inputs. This estimation

involves no notion of evolutionary fitness.[2] In order to bring evolution and fitness into the picture, we think of organisms as gathering fitness points as they interact with their environment. Thus each element $w$ of the world $W$ has associated with it a fitness value. In general, however, the fitness value depends not only on the world, but also on the organism $o$ in question (e.g., lion vs. rabbit), its state $s$ (e.g., hungry vs. satiated), and the action class $a$ in question (e.g., feeding vs. mating). Given such a fitness landscape, natural selection favors perceptions and choices that yield more fitness points.

We may thus define a *global fitness function* as a (non-negative) real-valued function $f(w, o, s, a)$ of these four variables. However, once we fix an organism, its state and a given action class, i.e., once we fix $o$, $s$ and $a$, a *specific fitness function* is 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 turn 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. That is, $W$ is here interpreted as depicting different quantities of some relevant resource.

---

[2] As noted above, Bayesian approaches often involve a loss (or gain) function. However, this is quite distinct from a fitness function, as defined below. Specifically, loss functions are functions of two variables $l(x, x^*)$, where $x^*$ is the "true" world state, and $x$ is a hypothetical estimate arrived at by the observer. A fitness function is, however, not a function of the observer's *estimate* $x$.

We can then consider a perceptual map $P : W \to X$, where X is the set of possible sensory 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 sensory states. Its perceptual map would then be some way of mapping various quantities of the resource to the small set of available sensory 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).

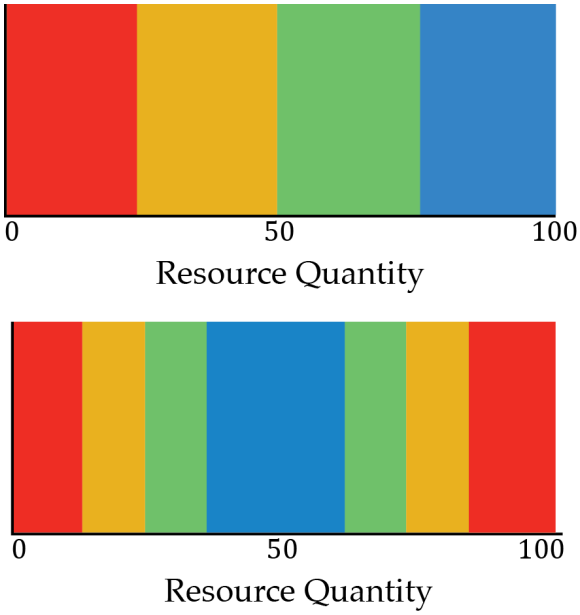



*Figure 1.* A simple example showing two different perceptual mappings $P : W \to X$ from world states, *W* = [*1, 100*] to sensory states *X* = {*R, Y, G, B*}.

In addition, there is a fitness function on *W*, $f : W \to [0, \infty)$, 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). Non-monotonic fitness functions (such as the one shown in Figure 2) are in fact quite common: 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 any number of other resources. Indeed, given the ubiquitous need for organisms to main homoeostasis, one expects non-monotonic fitness functions to be prevalent. (Moreover, 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.)

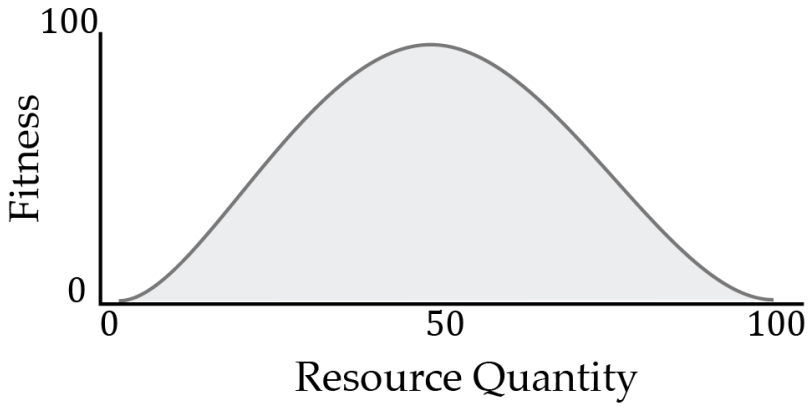


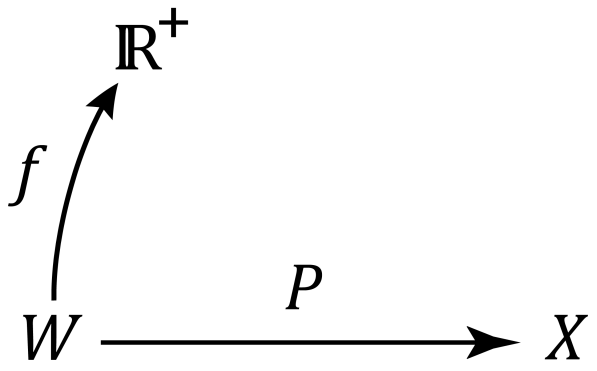

*Figure 2.* An example of a non-monotonic fitness function $f: W \to [0, \infty)$. Fitness is maximal for an intermediate value of the resource quantity and decreases in either direction. Given the ubiquitous need for organisms to main homoeostasis, one expects that such fitness functions are quite common.

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

In the context of these evolutionary games, in which perceptual strategies compete for resource acquisition, we take as fixed and known to the organism: the specific fitness function, its prior (in a particular state and for a particular action class) and its perceptual map (see Figure 3). On any given trial, the organism observes a number of available territories through its sensory states, say $x_1$, $x_2$,..., $x_n$. 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 those estimated 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 MAP estimate of the world 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.

212 


213 *Figure 3.* The framework within which we define the two resource strategies. We assume a
214 fixed perceptual map $P : W \to X$ as well as a fixed fitness function $f : W \to [0, \infty)$. Given a
215 choice of available territories sensed through the sensory states, say $x_1, x_2, \ldots, x_n$, the
216 organism's goal is to pick one of these, seeking to maximize its fitness payoff.

217

218 **5. Theorems from Evolutionary Game Theory**

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

221

| | *against A* | *against B* |
|---|---|---|
| *A plays* | *a* | *b* |
| *B plays* | *c* | *d* |

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

225 We first consider games with infinite populations. These are investigated by means of a
226 deterministic differential equation, called the *replicator equation*, where time is the
227 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 $\rho_{AB}$ 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 $\rho_{BA}$ 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****.*[3]

## 6. 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***.

[3] 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* $b > d$.

Our main claim in this paper is that the ***Truth*** strategy—attempting to infer to the "true" state of the world that is most likely correspond 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.

### 6.1 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 food, while $w_2$ and $w_3$ correspond to moderately healthful 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.

| | **Likelihood: $x_1$ given $w_j$** $\mathbb{P}(x_1 \mid w_j)$ | **Likelihood: $x_2$ given $w_j$** $\mathbb{P}(x_2 \mid w_j)$ | **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.

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 utility and the ***Fitness-Only*** observer's choice maximizes expected utility, 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).

## 7. 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 mathematical definitions and assumptions regarding these spaces and maps.

It will suffice for a basic understanding of the development in what follows, to think of $W$ as a finite set (as in the example in 6.1).[4] In general, we take the ***world*** $W$ to be a compact regular Borel space whose collection of measurable events is a $\sigma$-algebra, denoted $\mathcal{B}$.[5] We assume that $< W, \mathcal{B} >$ comes equipped with an ***a priori probability measure*** $\mu$ on

[4] in which case all the integral signs below can be replaced by summations.

[5] 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.

$\mathcal{B}$. We will consider only those probability measures $\mu$ that are absolutely continuous with respect to the Borel measure on $\mathcal{B}$. That is, if we write 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 $\mu$, 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 biological and perceptual situations.

We 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 $\sigma$-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 $X$.

### 7.1 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 *a 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 $\mu$, the prior on $W$, has a density $g$ with respect to the Borel measure 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[6] 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.

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

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.

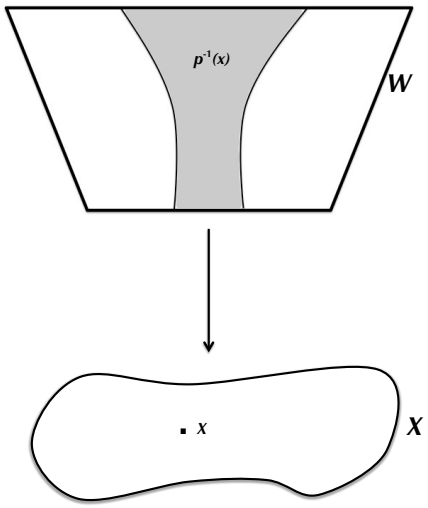



*Figure 4.* The expected fitness of $x$ is the average, using the **posterior** probability, over the fiber $p^{-1}(x)$.

### 7.2 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

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

### 7.3 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.

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

## 8. Results

### 8.1 The "Fitness Beats Truth" Theorem

The following theorem applies to infinite populations, or to large finite populations including those 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:

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

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: $1 - \frac{2}{n-1} = \frac{n-3}{n-1}$.

□

### 8.1 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.

## 9. Discussion

As we noted in the *Introduction*, it is standard in the literature to assume that more accurate perceptions are fitter perceptions 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 perception is, 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 world corresponding to a given a sensory state, confers no evolutionary benefit whatsoever. Rather 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 "Fitness Beats Truth" Theorem). Indeed, this "Fitness-only" strategy does not estimate any single world state; it simply

averages over all possible world states to compute the expected-fitness payoff corresponding to any given sensory state (this is analogous to a model-averaging strategy in model selection). And yet, as the theorem shows, in an evolutionary competition, this strategy is likely to drive the "truth" strategy to extinction.

At first glance, this expected-fitness strategy, based on averaging over all possible world states, may seem implausible: After all, in our own perceptual experience, we perceive things to be one particular way; we certainly don't experience a superposition or "smear" resulting from averaging over various ways that the world *could* be. While this is undoubtedly true, one should note that this is a fact about *perceptual experience*, and provides no support whatsoever for a strategy that involves estimating the "true" state of the world. In what follows, we sketch out a more complete answer to the seeming implausibility of averaging, based on our *Interface Theory of Perception* (Hoffman, Singh, & Prakash, 2015).

For the purpose of the current analysis, it was essential to 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 world, $W$ (recall Figure 3). This allowed us to place the two strategies on the same footing, so they could compete directly against each other. However, this result strongly supports our belief 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 happens in a space of perceptual representations, and not in an objective world.

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 visual perception conflates the interpretation space (or the space of perceptual hypotheses from which the visual system much choose) with the objective world. This is a mistake; it is essentially the assumption that the language of our perceptual representation is the correct language for describing objective reality—rather than simply a species-specific interface that has been shaped by natural selection. In our ITP framework, the probabilistic inference that results in perceptual experience takes place in a space of perceptual representations, say, $X_1$, that may have no isomorphic or even homomorphic relation whatsoever to *W*. The extended framework of this *Computational Evolutionary Perception* is sketched in Figure 5 (see Hoffman & Singh, 2012; Hoffman, Singh, & Prakash, 2015; Singh & Hoffman, 2013).

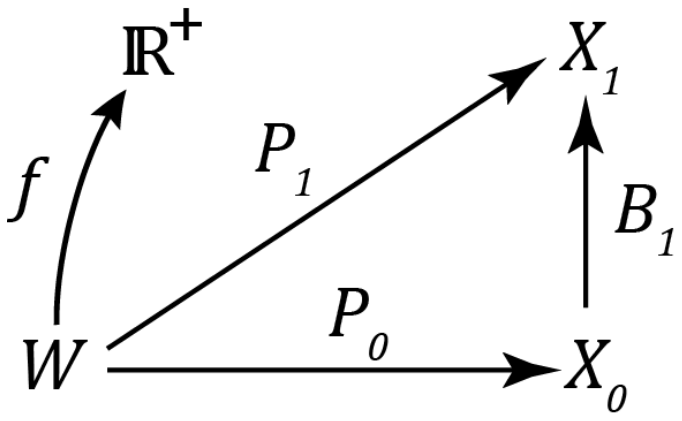



*Figure 5.* The framework of *Computational Evolutionary Perception* in which perceptual inferences take place in a space of representations $X_1$ that is not isomorphic or homomorphic to *W*. The more complex representational format of $X_1$ evolves because it permits a higher-capacity channel $P_1 : W \rightarrow X_1$ 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).

Thus, the reason we generally perceive a single interpretation is because the probabilistic inference in the perceptual space $X_1$ generally results in a unique 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). In other words, a more complex or higher-dimensional representational format (e.g. involving 3D representations in $X_1$, in place of 2D representations in $X_0$) evolves because it permits a higher-capacity channel $P_1 : W \rightarrow X_1$ 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. For example, 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.[7]

## Acknowledgments

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

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

## References


1. 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: Cambridge University Press).
2. Chemero, A. (2009). Radical Embodied Cognitive Science. Cambridge, MA: MIT Press.
3. Dennett, D. C. (1995). *Darwin's Dangerous Idea*. New York: Simon & Schuster.
4. Feldman J. (2013) Tuning your priors to the world. *Topics in Cognitive Science*, 5, 13-34.
5. Geisler, W. S. and Diehl, R. L. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. Cognitive Science 27, 379-402.
6. Hofbauer, J., and Sigmund, K. (1998). Evolutionary games and population dynamics. Cambridge: Cambridge University Press.
7. Nowak, Martin A. (2006). *Evolutionary Dynamics*. Harvard University Press.
8. 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.
9. Hoffman, D. D., & Prakash, C. (2014). Objects of consciousness. Frontiers of Psychology, 5:577. DOI: 10.3389/fpsyg.2014.00577.
10. Hoffman, D. D., & Singh, M. (2012). Computational evolutionary perception. Perception, 41, 1073–1091.
11. 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.
12. Hoffman, D. D., Singh, M., & Prakash, C. (2015). Interface Theory of Perception. Psychonomic Bulletin & Review. DOI 10.3758/s13423-015-0890-8
13. Kleffner, D., & Ramachandran, V. (1992). On the perception of shape from shading. *Perception and Psychophysics, 52*, 18–36.
14. Knill, D. and Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press, Cambridge, UK.
15. Koenderink, J. J., van Doorn, A., de Ridder, H., Oomes, S. (2010). Visual rays are parallel. Perception, 39, 1163-1171.
16. 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.
17. 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.
18. Koenderink, J. J. (2014). The all seeing eye? Perception, 43, 1–6.
19. 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.
20. 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.
21. Mamassian P, Landy M, Maloney L T, 2002 “Bayesian modeling of visual perception”, in Probabilistic Models of the Brain: Perception and Neural Function, Ed. R Rao, B Olshausen, M Lewicki (Cambridge, MA: MIT Press) pp 13 – 36.

22. Marion, B. The Impact of Utility on the Evolution of Perceptions. Dissertation, University of California, Irvine, 2013.
23. Mark, J., Marion, B. and Hoffman, D. D. (2010). Natural selection and veridical perception. Journal of theoretical Biology 266, 504-515.
24. Marr, D. (1982). *Vision*. San Francisco: Freeman.
25. 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.
26. Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Belknap Harvard University Press, Cambridge, MA.
27. Palmer, S. (1999). Vision Science: Photons to Phenomenology. MIT Press, Cambridge, MA.
28. Pizlo, Z., Li, Y., Sawada, T., & Steinman, R.M. (2014). *Making a machine that sees like us.* New York: Oxford University Press.
29. Shepard, R. (1994). Perceptual-cognitive universals as reflections of the world. Psychonomic Bulletin & Review, 1(1), 2-28.
30. 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.
31. Taylor, C, Fudenberg, D., Sasaki, A. & Nowak, M. Evolutionary Game Dynamics in Finite Populations. *Bulletin of Mathematical Biology* (2004) **66**, 1621–1644
32. Taylor, P. D., and L. B. Jonker. 1978. "Evolutionary stable strategies and game dynamics." Math. Biosci. 40: 145– 156.
33. von Uexküll, J. (1934). *A stroll through the worlds of animals and men: A picture book of invisible worlds*. Reprinted in Instinctive behavior: Development of a modern concept, C.H. Schiller (1957). New York: Hallmark Press.
34. 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

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

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

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

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:

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

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

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

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 $x_1$.

## Highlights

- We make rigorous mathematical definitions of two perceptual strategies employable by a given species, for a given action class and within a given environment: ***Truth***, based on Bayesian estimation of assumed objective properties of the world, and ***Fitness***, tuned to an arbitrary fitness function;
- Under the assumption of universal Darwinism (Dennett, 1995) we subject the two strategies to an evolutionary game analysis;
- We conclude that the ***Fitness*** will generally drive ***Truth*** to extinction, for generic fitness functions and priors;
- The likelihood of ***Fitness*** dominating ***Truth*** exceeds 1/2 as soon as the sensorium has more than five elements, and rises monotonically to 1 as the size of the sensorium grows towards infinity;
- This theorem holds in the presence of changing fitness functions and for large finite populations.