Natural Selection and Shape Perception

Manish Singh and Donald D. Hoffman

Abstract We present a formal framework that generalizes and subsumes the standard Bayesian framework for vision. While incorporating the fundamental role of probabilistic inference, our Computational Evolutionary Perception (CEP) framework also incorporates fitness in a fundamental way, and it allows us to consider different possible relationships between the objective world and perceptual representations (e.g., in evolving visual systems). In our framework, shape is not assumed to be a reconstruction of an objective world property. It is simply a representational format that has been tuned by natural selection to guide adaptive behavior. In brief, shape is an effective code for fitness. Because fitness depends crucially on the actions of an organism, shape representations are closely tied to actions. We model this connection formally using the Perception-Decision-Action (PDA) loop. Among other things, the PDA loop clarifies how, even though one cannot know the effects of ones actions in the objective world itself, one can nevertheless know the results of those effects back in our perceptions. This, in turn, explains how organisms can interact effectively with a fundamentally unknown objective world.

1 Introduction

Our perception of shape is, like all of our perceptions, a product of evolution by natural selection. This entails that our perception of shape is a satisficing solution to certain problems faced by our ancestors, e.g., the need to stalk prey, secure mates, elude predators, and predict outcomes of actions. Natural selection produces *satisficing* solutions, rather than *optimizing* solutions, because selection favors sur-

Manish Singh
Department of Psychology, Center for Cognitive Science, Rutgers University, New Brunswick, NJ.
e-mail: manish@ruccs.rutgers.edu

Donald D. Hoffman
Department of Cognitive Science, University of California, Irvine, CA. e-mail: ddhoff@uci.edu

vival of the *fitter*, not of the *fittest*: A gene need confer only a slight edge over the competition—a standard far lower than optimality—to proliferate in later generations.

It is standard in vision research to assume that more accurate perceptions are fitter perceptions, and that therefore natural selection tunes our perceptions to be veridical, i.e., to be accurate reflections of the objective world. For instance, Palmer argues that "Evolutionarily speaking, visual perception is useful only if it is reasonably accurate . . . This is almost always the case with vision" [28]. Geisler and Diehl argue that "In general, (perceptual) estimates that are nearer the truth have greater utility than those that are wide of the mark" [11].

If perception is indeed veridical, then the world of our visual experience shares the attributes of the objective world. Our visual world has three spatial dimensions, a temporal dimension, and contains 3D objects with shapes, colors, textures and motions. Vision researchers standardly assume that the objective world does also. In other words, they standardly assume that the language of our visual representations is the correct language for describing objective reality.

In this chapter we propose, contrary to standard assumptions, that natural selection does not in general favor veridical perceptions. The reason, in short, is that fitness is distinct from truth; it depends not only on the objective world, but also on the organism, its state, and the action class in question. A gazelle, for instance, offers lots of "fitness points" to a hungry cheetah seeking to eat, but none to a cheetah seeking to mate. Natural selection favors fitness, not truth. It is straightforward to produce evolutionary games in which true perceptions are driven to extinction by nonveridical perceptions that simply report fitness [25].

The consequences of this for shape perception are profound. If our perceptions of 3D shape are not veridical reconstructions of objective 3D shapes, then a new framework, entirely different from the standard, is required to properly understand shape perception. In this chapter we sketch such a formal framework that incorporates the role of evolution in a fundamental way, and in which perceived shape is an adaptive guide to behavior, not a reflection of objective reality. This framework is consistent with the *interface* theory of perception [15].

Because natural selection has tuned our perception of shape to be an adaptive guide to behavior, our perception of shape has evolved to be tightly coupled with our actions, a coupling that we formalize here with a commuting diagram that we call the "perception-decision-action" loop, or PDA loop. Thus the detailed properties of perceived shapes, such as their symmetries and parts, are not depictions of the true properties of shapes in an objective world, but simply guides to adaptive action.[1]

[1] We use "action" in the broadest sense of the word—to include not only visually-guided manipulation of objects ("dorsal stream"), but also visual categorizations ("ventral stream") that inform subsequent behavior, e.g., whether or not to eat a fruit that has some probability of being poisonous.

2 Bayesian Decision Theory

A common framework for modeling vision in general, and the "recovery" of 3D shape from 2D images in particular, is Bayesian decision theory (BDT) [12, 17, 18, 21, 23, 24]. BDT provides a probabilistic framework at the computational (or competence) level [26], at which visual problems are analyzed in terms input-output relations (e.g., the formal constraints needed to derive desired outputs from given inputs)—independently of performance considerations involving specific algorithms or their implementations.

Given the basic inductive problem that any image is consistent with many different 3D interpretations, the visual system can resolve this ambiguity only by bringing additional constraints (or biases) to bear—based on regularities observed in the terrestrial environment in which our species evolved—and comparing the relative probabilities of different scene interpretations. For example, in estimating 3D shape from shading, human vision appears to assume that light comes from above (e.g., [19, 24]). Similarly, theories of shape-from-contours often assume that the 3D shapes are symmetric, or maximally compact (e.g., [30]).

Formally, given an image y_0, the visual system must compare the posterior probability $p(x|y_0)$ for different scene interpretations x. By Bayes' Theorem, this posterior probability is proportional to the the product of the likelihood of the scene $p(y_0|x)$ and its prior probability $p(x)$. The likelihood term captures the extent to which the scene interpretation x is consistent with—and hence can "explain"—the image y_0. In theories of shape-from-X, it is usually taken to be a projective mapping from 3D to 2D (orthographic or perspective), plus some model of noise. Because many different 3D interpretations are typically consistent with any given image, the likelihood cannot generally resolve the ambiguity by itself (i.e., the likelihood may be equally high for a large number of 3D interpretations). The other source of information—the prior probability—reflects the observer's internalized beliefs about fact that certain scenes, shapes, or states of the world are more likely than others—e.g., light tends to come from above, objects tend to be compact, there is a prevalence of symmetric objects, etc. [30, 19, 24].

The combined use of the prior and likelihood—via Bayes—yields a posterior distribution on the space of scene interpretations. It is common to use the maximum-a-posteriori (MAP) estimate as one's "best" interpretation. More generally, however, the choice of a "best" point estimate depends on the loss function one assumes—namely, the consequences of errors, or deviations from the "true" (but unknown) interpretation. If the loss function is essentially a Dirac-delta function (i.e., no loss for the correct answer, equal loss for every other answer) the value that minimizes expected loss is the mode of the posterior distribution, i.e., the MAP estimate. However, if the loss function is quadratic (i.e., squared-error), the value that minimizes expected loss is the mean of the posterior distribution. Hence different choices of loss functions lead to different strategies for picking a single "best" scene interpretation from the posterior distribution (e.g., [24]).

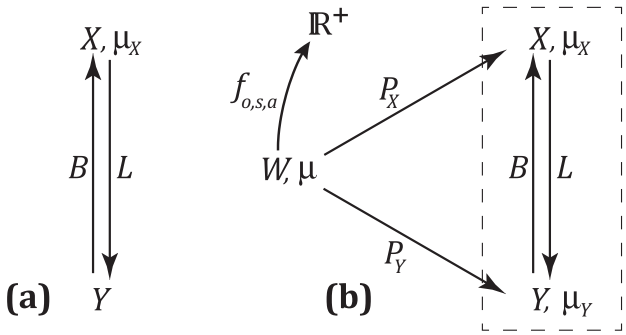

Fig. 1 (a) The standard Bayesian framework for vision. (b) The computational evolutionary perception (CEP) framework. In CEP, the objective world W lies outside of the probabilistic inferential apparatus for vision. There are perceptual channels P_X and P_Y to the two representational spaces X and Y, respectively. And there are specific fitness functions on W that assign, for a given organism o, its state s, and the type of action a in question, "fitness points" to each $w \in W$.

3 A general framework for perception and its evolution

Bayes' Theorem provides a provably optimal way of combining the two probabilistic sources of information embodied in the likelihood and prior [17]. Hence there is strong, principled justification for using Bayes, once a likelihood model and a prior have been specified on a particular space of possible interpretations. However, the Bayesian framework as it is standardly applied to vision involves important assumptions about the choice of interpretation space that we will argue are too restrictive.

Consider the standard Bayesian setup for vision shown in Figure 1a. X is the space of scene interpretations (say, 3D shapes), with prior probability distribution μ_X. Y is the space of 2D images. The likelihood mapping L is the projective map from 3D to 2D (possibly with noise). B is the Bayesian posterior map from Y to X. Technically, L and B are both Markovian kernels [31]. Thus, for each $x \in X$, the projective map L specifies a probability distribution on Y (in the noise-free case, this distribution is supported on a single point). And for each $y \in Y$, the Bayesian posterior B gives a probability distribution on the space X of 3D shapes.

Importantly, note that in this setup the space X plays two distinct roles: (i) it corresponds to the space of objective world states; and (ii) it corresponds to the space of possible perceptual interpretations from which the visual system must "choose." This dual role is entirely consistent with the *inverse optics* approach to vision—according to which the goal of vision is essentially to invert or "undo" the effects of optical projection (e.g., [1, 29]). It is also consistent with the historical roots of Bayesian methods, namely, as techniques for computing "inverse probability"—a prototypical case being to infer the relative probabilities of possible underlying causes $p(C|E)$ given some observed event E, when what one actually knows are the probabilities of obtaining various events $p(E|C)$ from particular causes C [22].

This dual role played by X makes it clear how BDT embodies the common assumption that human vision has evolved to see the truth. It is *not* the case, of course, that a BDT observer always makes veridical perceptual inferences. Indeed, it cannot. Because a BDT observer embodies specific assumptions about regularities in the world ("light tends to come from above," "objects tend to be mostly convex," etc.) it is always possible to place it within a context where its assumptions are violated. At a more fundamental level, however, BDT makes the basic assumption that the *language* of scene interpretations X is the correct language for describing objective reality. In other words, BDT assumes that the representational space X contains somewhere within it a true description of the objective world—even if the observer's estimate misses it in any given instance. It is in this more fundamental sense that BDT assumes that human vision has evolved to see the truth.

Consideration of vision in other species, especially those with simpler visual systems, suggests that this implicit identification of the representational space X with the objective world is too simplistic. As we will see, it is also too restrictive if one wants a formal framework that is general enough to encompass the evolution of visual systems.

In discussing simpler visual systems, such as those of the fly and the frog, Marr [26] noted that they "...serve adequately and with speed and precision the needs of their owners, but they are not very complicated; very little objective information about the world is obtained. The information is all very subjective..."; and that "...it is extremely unlikely that the fly has any explicit representation of the visual world around him—no true conception of a surface, for example, but just a few triggers and some specifically fly-centered parameters..." (p. 34). Thus Marr seemed to acknowledge that visual systems that do not compute objective properties of the world can serve the needs of their owners well enough for them to survive, even thrive, in their respective niches. This should not be surprising; after all, what matters in evolution is fitness, not truth, and even visual systems that compute only simple, purely "subjective," properties can confer sufficient fitness. Despite this, Marr held that the properties computed by *human* vision—such as object shape—are objective properties of the world that exist independently of any observer. There is no reason to believe, however, that the representational spaces that evolved in the species *Homo sapiens* must correspond to objective reality. The evolution of *Homo sapiens* is guided no less by fitness than the evolution of any other species. And fitness is clearly distinct from objective truth because it depends not only on the objective world, but also on the *organism* (fly vs. elephant), its *state* (hungry vs. satiated), and the *type of action* under consideration (eating vs. mating). Therefore one's formal framework must be broad enough to include the possibility that *human* visual representations also do not capture objective truth.

Thus, rather than simply assuming, or postulating, that the space of interpretations X is identical to (or in one-to-one correspondence with) the objective world—let's call it W—one's formal framework must consider different possible relationships between X and W. We make no assumptions about W, except that it is meaningful to talk about probabilities in W, governed by some (unknown) probability measure μ on an event space $\mathscr{W}$. We define a *perceptual strategy* as a measurable function $P : W \to X$. One can think of P as a channel between W and X, that allows information to flow from the objective world to the organism. In the general case, P is a Markovian kernel which specifies, for each $w \in W$, a probability distribution on X.[2] One can then consider four classes of perceptual strategies corresponding to different relationships between X and W (see [16, 25]): (i) the *naïve realist* strategy assumes that $X = W$ and that P preserves all structures on W; (ii) the *strong critical realist* strategy assumes only that $X \subset W$ but requires that P projects all structures of W onto X; (iii) the *weak critical realist* strategy allows that $X \not\subset W$ but requires that P projects all structures of W onto X; and (iv) an *interface strategy* allows that $X \not\subset W$ and does not require that P projects all structures of W onto X. The interface strategy *need not see the truth* in the more fundamental sense that the very language of the space X may be the wrong language to capture the structure of the objective world W.

Most vision researchers today are weak critical realists. They recognize—contrary to the claim of naive realism and strong critical realism—that perceptual representations are distinct from objective reality, but assume that perceptual representations are isomorphic, or at least homomorphic, to objective reality. We call these two versions "isomorphic realism" and "homomorphic realism."

We generalize BDT to a framework we call Computational Evolutionary Perception (CEP; [16]). In CEP, the objective world W lies outside of the Bayesian inferential apparatus (see Figure 1b). X and Y are simply two representational spaces—neither corresponds to the objective world W (nor are they assumed to be isomorphic to W). For example, Y may be a lower-level representation (say, a 2D representation of image structure) that evolved earlier, whereas X may be a higher-level representation, involving some 3D structure, that evolved later. There are perceptual channels P_X and P_Y from the world W to X and Y, respectively. As noted above, in the general case, P_X and P_Y are also Markovian kernels. Thus, for each $w \in W$, P_X specifies a probability measure on X, and P_Y specifies a probability measure on Y. In particular, the measure μ on W yields, via P_X, a pushdown measure μ_X on X, and similarly via P_Y, a measure μ_Y on Y.[3] In the diagram in Figure 1b, therefore, all four mappings shown (L, B, P_X and P_Y) are Markovian kernels. It is therefore meaningful to take their compositions, which are also Markovian kernels (such as the composition $P_X L : W \to Y$).[4] An important constraint in the CEP framework is that the diagram in Figure 1b must commute. As a result, for example, $P_Y = P_X L$. This is a coherence constraint on perceptual representations that allows observers to predict the percep-

[2] Hence, formally, P is a mapping $P : W \times \mathscr{X} \to [0,1]$, where $\mathscr{X}$ is the event space on X. One can view P as a linear operator that maps probability measures on W to probability measures on X. In the discrete case, it would be represented by a stochastic matrix whose rows add up to 1. For more on Markovian kernels see [3, 31].

[3] Thus, whereas in BDT μ_X is taken to be the world prior, in CEP μ_X is the pushdown, via the perceptual channel P_X, of the prior μ on the objective world.

[4] Kernel composition is defined as follows: Let M be a kernel from $(X, \mathscr{X})$ to $(Y, \mathscr{Y})$, and N be a kernel from $(Y, \mathscr{Y})$ to $(Z, \mathscr{Z})$. Then the composition kernel MN from $(X, \mathscr{X})$ to $(Z, \mathscr{Z})$ is defined, $\forall x \in X$ and $A \in \mathscr{Z}$, by $MN(x,A) = \int_Y M(x,dy)N(y,A)$. This is simply a generalization to the continuous case of the familiar multiplication of (stochastic) matrices. For details, see [31].

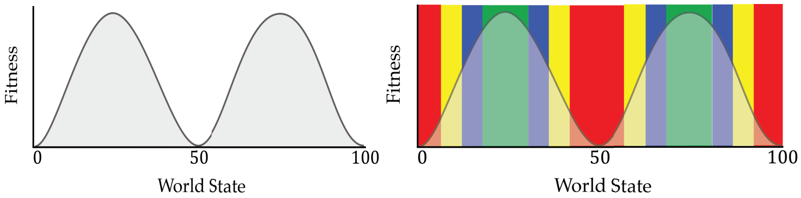

Fig. 2 (a) A specific fitness function defined on a world containing a resource that varies in quantity from 0 to 100. Resource quantities around 25 and 75 confer the greatest fitness, whereas resource values around 0, 50, and 100 confer the least fitness. (b) The construction of a message set with 4 messages, based on a simple clustering of fitness values into four categories: “very high” (green), “somewhat high” (blue), “somewhat low” (yellow) and “very low” (red).

tual consequences of their actions, despite the fact that they are ignorant about the objective world itself (see also Section 4).

What shapes the evolution of perception is, of course, fitness. We therefore expect that natural selection tunes perceptual channels (and their corresponding representational spaces) to the only signal that matters for evolution, namely, fitness. In order to bring fitness into our formalism, we view organisms as gathering “fitness points” as they interact with the world. As we noted, fitness depends not only on the objective world, but also on the organism, its current state, and the type of action in question. Thus we define a *global fitness function* $f : W \times O \times S \times A \rightarrow \mathbb{R}^+$, where O is the set of organisms, S of their possible states, and A of possible action classes. Once we fix a particular organism $o \in O$, state $s \in S$, and action class $a \in A$, the *specific fitness function* $f_{o,s,a} : W \rightarrow \mathbb{R}^+$ assigns fitness points to each possible $w \in W$ (say, of a starving lion eating a gazelle).

Given a specific fitness function $f_{o,s,a}$, evolution shapes a source message about fitness and a channel to communicate that message, that results in hill-climbing toward greater expected-fitness payout to the organism. This means that a perceptual channel P_X from W to X may be expressed as the composition of two Markovian kernels: a message construction kernel P_{C_X} from W to a set of messages M, and a transfer kernel P_{T_X} from M to X. The message *construction* kernel P_{C_X} is needed because the messages to be transmitted depend not only on the world W, but also on the fitness values associated with elements of W (for a particular organism o, its state s, and action class a). Hence, given the same W, but a different specific fitness function $f_{o,s,a}$, the set of messages to be transmitted may be different. Consider an example of a simple world with multiple territories, each of which contains a resource whose quantity varies from 0 to 100. Thus each value from 0 to 100 may be considered to be a different world state. Now consider the specific fitness function $f_{o,s,a}$ shown in Figure 2a. As shown, resource quantities around 25 and 75 confer the greatest fitness, whereas resource values around 0, 50, and 100 confer the least fitness. Assume that the representational space X contains 4 elements, say, $X = \{A, B, C, D\}$. Then an efficient way to construct a message set might be to have four messages, obtained by clustering the fitness values into four categories: “very high” (green), “somewhat high” (blue), “somewhat low” (yellow) and “very low” (red) (see Fig Figure 2b). The received messages are then highly informative about fitness, and would allow the organism to choose between territories in a manner that will result in high expected-fitness payout (e.g., given a choice between a "green" territory vs. a "blue" one).[5] (Note that this occurs despite the fact that the received messages carry little information about the actual number of resources.) We use the term *Darwinian Observer* to refer to a perceptual channel P_X that has been shaped by natural selection as a satisficing solution for a specific fitness function.

The above analysis assumed that the representational space X was fixed, and the perceptual channel P_X was being tuned to increase expected-fitness payout. Another way, however, to increase expected-fitness payout is to evolve the representational space itself: $X_1 \rightarrow X_2 \rightarrow \ldots$. Presumably, there would be selection pressure to evolve a more complex representational space (e.g., a representation that captures some 3D structure) when the expected-fitness payout with the current space is insufficient to survive or compete, and going to the more complex representational space would allow a substantial increase in expected-fitness payout.

The CEP framework is thus more general than the BDT framework for vision. First, while incorporating the fundamental role of probabilistic inference, it allows us to consider different possible relationships between the space of interpretations X and the objective world W (rather than simply assuming that $X = W$, or that X is isomorphic to W). Second, it explicitly incorporates the role of fitness into the formal framework, in a way that does not simply reduce fitness to the gain/loss function of BDT. And third, by using Markovian kernels to map the relationship between W and X, it allows us to articulate precisely different ways in which perceptual evolution can proceed (e.g., by tuning a perceptual channel to a fixed representational space, or evolving the representational space itself).

4 Shape as a code for fitness

4.1 Implications for shape perception

With our general framework in place, the implications for shape perception now follow straightforwardly. First, our framework makes it clear that we really have no basis for assuming—as is standardly done—that shape is an objective property of the world. For example, it is fairly standard among shape researchers to speak of "shape recovery" when referring to the computation of 3D shape from different 2D cues. This nomenclature reflects the identification of the representational space X with the objective world W that is assumed in the *inverse optics* approach to vision (and, as noted above, is commonly made in Bayesian approaches to vision). When one sees the 3D shape of an object, the undulations in its surface, etc., one

[5] In this example, a simple clustering based on fitness values was sufficient. More generally, however, multi-dimensional scaling may be required. Indeed, MDS-type solutions may also provide an explanation of how dimensional structure can arise in perceptual representations.

sees, according to the inverse optics approach, geometric properties that correspond to objective properties of the world[6]—properties that exist independently of any observer. However, as we noted above, this is too simplistic. It is certainly much more than can be claimed based on available facts. There is surely an objective world W, but there is no basis for saying that *shape* is a property of that world. Rather, shape is simply a representational format used by our visual systems to guide interactions with the objective world. It is part of the representational space X, not W. It should be clear from this that our position is strictly weaker—not stronger—than the standard *inverse optics* or *shape recovery* approach. Whereas the standard approach assumes, or postulates, that $X = W$ or that X is isomorphic to W, we are open to different possible relations between X and W.

Second, our framework entails that *shape*, as a representational format, most likely evolved because it made possible the development of a perceptual channel with high expected-fitness payout. Thus the property we call *shape* is essentially an effective coding scheme that has been tuned by natural selection: it conveys to an organism—in a compact and efficient format—the various ways in which the organism could interact with objects in the world to gain more "fitness points." Therefore when we perceive the 3D shape of an object—the undulations on its surface, its symmetries, its part structure—all of these are different aspects of a representational format that natural selection has fashioned, one which compactly summarizes the different possible actions that we could take, and that allows us to predict the perceptual consequences of those actions (e.g., how the perception of a 3D object would change were we to rotate it slightly to left, pick it up in a certain way, etc.), and what the fitness consequences would be (e.g., would we successfully eat that apple or evade that tiger).

This last point raises a natural question: How is it possible for us to interact successfully with the objective world if we are fundamentally ignorant of it, and can assume no simple correspondence between our perceptions and that objective world? This is where the third implication of our framework comes in, namely, that action (broadly construed) plays a central role in the evolution of shape perception. In brief, it is perfectly possible to interact successfully with a fundamentally unknown objective world because (i) there is a regularity in the perceptual mapping; (ii) there is regularity in the consequences of our actions in the objective world; and (iii) these mappings are linked in a coherent manner. This is a fundamental point for our framework and, to develop it fully, we need to introduce some more formalism, namely that of the *perception-decision-action* (or PDA) loop. Before we do this in the next subsection, however, we provide an example that should help fix intuitions.

Consider the desktop interface of a PC. A file's icon on the desktop might be green, rectangular and in the middle of the screen. Does this entail that the file

[6] The inverse optics approach allows for misperceptions—e.g., that observers tend to perceive an object from a certain viewpoint as being less elongated in depth than physical measurements of the object tell us it is. But the inverse optics approach nevertheless assumes that *one* of the shapes in X is the "correct" one in the objective world W. In other words, at a more fundamental level, the inverse optics approach assumes that the very property we call *shape* is an intrinsic property of the objective world W itself.

itself is green, rectangular and in the middle of the computer? Of course not. The shape, position and color of the icon are merely conventions that allow the user to interact with the computer despite being ignorant of the complex details of its diodes, resistors, software, voltages and magnetic fields. The desktop interface is useful not because it reveals the truth about the computer, but because it hides the complex truth, and instead provides simple symbols that guide useful interactions with the computer. In like manner, natural selection has shaped our perceptions to be an interface that hides the true nature of the objective world, and guides adaptive behavior [14, 15, 20]. Spacetime is the desktop, and objects with their shapes, colors, textures and motions are icons in the desktop. Spacetime and objects are not the objective truth, and do not resemble the truth. Instead, they are a species-specific adaptation shaped by natural selection to guide adaptive behaviors and to allow us to survive long enough to reproduce. Perception has been shaped by the imperative to produce offspring, not to see truth.

4.2 The role of action in the evolution of shape perception

In this section we incorporate action and decision into our formalism, and draw out implications for shape perception. Natural selection necessarily couples perception and action because fitness, to which perception is tuned, depends crucially on the actions of the observer. Different classes of action are, in general, coupled with different expected fitnesses. The fitness points gleaned from an apple for the action of eating is greater than for the action of mating. Since natural selection tunes perceptual channels to convey information about fitness, one expects tight coupling between perceptual channels and the actions they inform.

When an observer receives a perceptual experience $x \in X$, it must decide what action to take. We will denote the set of available actions by a set G, where we think of G as including a group that acts on X. Recall that if a group G acts on X, then for every $g \in G$ the mapping $x \mapsto gx$ is a bijective map from X to X. Common examples

$$\begin{tikzcd}
& G \arrow[dl, "A"'] & \\
W \arrow[rr, "P"'] & & X \arrow[ul, "D"']
\end{tikzcd}$$

Fig. 3 The Perception-Decision-Action (PDA) loop. W denotes the objective world, X a space of perceptual representations of an organism, and G the related set of actions the organism can take. P is a perception kernel, D a decision kernel, and A an action kernel. All kernels are Markovian.

are the actions of translation and rotation on Euclidean spaces. We also allow there to be actions in G other than group actions.

Thus, given a perceptual experience $x \in X$ the observer must decide which action $g \in G$ to take. The natural formalism to describe such a decision is again a Markovian kernel, D, from $(X, \mathscr{X})$ to $(G, \mathscr{G})$. We call D the decision kernel.

Once an action g is chosen, the observer must then act on the objective world W. We model this action by a Markovian kernel A from $(G, \mathscr{G})$ to $(W, \mathscr{W})$, which we call the action kernel. Given this formalism, we can think of action as sending a message from the observer to the objective world.

Thus we have three kernels: P, D, and A. P maps from W to X; D maps from X to G; A maps from G back to W (see Figure 3). So together they form a loop, which we call the PDA loop. We have a PDA loop for each perceptual representation space X. So, in the CEP example discussed in Section 3, there is a PDA loop for the 2D image space Y and another PDA loop for the 3D space X.

However, just as we assume that the observer does not know the objective world W, and therefore does not know the perception kernel P, so also the observer does not know the action kernel A. Informally, this means that when we act, we don't really know what effects we are having in the objective world W itself; however we do know the results of those effects back in our perceptual experiences X. Formally, even though the observer cannot know the kernels P and A, it can know the kernel AP from $(G, \mathscr{G})$ to $(X, \mathscr{X})$, which is formed by the kernel composition of A and P. It can also know the kernel DAP from $(X, \mathscr{X})$ to $(X, \mathscr{X})$ (i.e., from X back to itself). This allows the observer to learn how to interact with W, even while being ignorant of W. The observer can try different actions $g \in G$ and note their consequences for perceptual experiences in X. If the consequences are unexpected, the observer can update its decision kernel D to correct this.

This applies to actions with objects and shapes. If, for instance, the observer acts in a way that leads it to perceive that its body moves through space via an element of the Galilean group, or that its hand is grasping an object and rotating it, then, given its perceptions of the relative position of an object, and the symmetries and parts of that object, it can predict what the consequences of its action should be for changes in the relative position and perceived shape of that object.

This also applies to object categorization. Such categorization allows the observer to predict the fitness consequences of various current and future interactions with the object (such as eating it). We are thus using the word "action" broadly to include not only "dorsal stream" visually-guided motor behavior, but also "ventral stream" perception and categorization that inform future behavior.

Let's return to the desktop metaphor discussed above. A new generation of desktops now employ 3D interfaces. In such a desktop, if the icon of a file has a particular 3D shape, say the shape of a book, and the desktop contains a 3D bookshelf with a book-shaped gap, then the user can be guided by the shape and position of the 3D icon to grasp it and place it in the bookshelf. In one sense, this is unremarkable. But the key concept here is that the file itself in the computer has no 3D shape, and in particular is not shaped like a book. Moreover, the directory system in the computer has no 3D shape, and in particular is not shaped like a bookshelf. These 3D shapes are mere conveniences for guiding effective interactions of the user, not insights into the true nature of files and directories—and certainly not of the myriads of voltages and magnetic fields in the computer.

4.3 Perceptual organization of shape

Apart from computing 3D shape from 2D image cues, another fundamental aspect of shape perception is the perceptual organization of shape. A great deal of psychophysical work indicates that human vision organizes complex shapes hierarchically in terms of parts and their spatial relationships (e.g., [5, 7, 13, 32]). This "structural" approach to shape separates the representation of individual parts from that of their spatial relationships—thereby allowing a shape to be identified as comprising the same parts, but in somewhat different spatial relations (e.g., a sleeping cat vs. a standing cat). It is also closely related to the axis or skeleton-based approach, which provides a compact "stick-figure" representation of a complex shape that captures its structural aspects (e.g., its branching structure) [4]. A recent probabilistic approach to the computation of shape skeletons yields a one-to-one correspondence between parts and skeletal branches—indicating that parts and skeletons are indeed complementary aspects of the perceptual organization of shape [8].

They key point, for current purposes, is that the perceptual organization of shape in terms of parts and axes has no natural interpretation in terms of inverse optics. There is no objective "ground truth" regarding whether an object "really" has one part or two, or whether an axis that continues from one portion of a shape to another is "really" the same or a different axial branch (e.g., consider a U-shape vs. a V-shape, and a morphing sequence between them). The organization of shape in terms of segmented parts, or in terms of axes, is something that the visual system *imposes* on perceptual objects—it is not an objective property of the world. This does not mean that a Bayesian analysis of the problem is not possible. However, the likelihood or the "forward" mapping in that case has a different interpretation; it is not a projective or rendering map, but the visual system's own *generative model* concerning how objects are formed [8]. This is easily accommodated within the current framework, since for us the space of interpretations X is distinct from the world W. Hence, in this case, the space X would consist of all possible interpretations of a shape as a hierarchical organization using segmented parts (e.g., different partitions of a shape, and different tree structures capturing possible part hierarchies). In the context of perceptual organization of shape, it is therefore especially clear that elements of X have no simple correspondence to the objective world W.

A natural question is: Why have shape representations based on parts and axes evolved, if they have no simple correspondence to the objective world W? The answer, as expected, has to do with fitness. Organisms that can predict, upon seeing an object at one time, what that object might look like on other occasions, are likely to interact with it much more successfully—and thus have greater fitness—than those that cannot. And a shape representation based on parts and axes goes a long way in conferring this ability: Upon seeing an animal in one particular articulated pose (configuration of limbs), for example, it is much easier to predict other possible (unseen) articulated poses if one's shape representation is part-based than, say, if one's representation consists simply of an unstructured template of the shape as a whole. In sum, a framework that allows X and W to be distinct, and incorporates the role of fitness, makes it much easier to understand the perceptual organization of shape.

5 Discussion

We sketched a formal framework—Computational Evolutionary Perception—that subsumes and generalizes the standard Bayesian framework for vision. While incorporating the role of probabilistic inference, CEP also incorporates fitness in a fundamental way, and it allows us to consider different possible relationships between the objective world and perceptual representational spaces. In our framework, shape is not an objective property of the world. It is simply a representational format employed by our visual systems to guide adaptive interactions with the world. This representational format evolved because it allows a high-capacity channel for fitness. In other words, *shape is an effective code for expected fitness that has been tuned by natural selection*. Because fitness depends crucially on the actions of an organism, shape representations in our framework are closely tied to actions. Thus when we perceive the 3D shape of an object—the undulations of its surface, its local and global symmetries, its part and skeletal structure—these are various aspects of a code that compactly summarizes the possible actions that one could take (including future actions based on current categorization), and to predict the fitness consequences of those actions. To model this formally, we introduced the perception-decision-action (PDA) loop. Among other things, the PDA loop clarifies how, even though one cannot know the effects of one's actions in the objective world itself, one can nevertheless know (because of the coherent coupling between perception and action) the results of those effects back in our perceptual experience. This explains how organisms can interact effectively with a fundamentally unknown objective world. Finally, CEP and the PDA loop provide a new framework for understanding the perceptual organization of shape using parts and skeletons—something that is difficult to accommodate within a standard inverse-optics approach to shape.

Acknowledgments

For helpful discussions, we thank Jacob Feldman, Pete Foley, Brian Marion, Justin Mark, Darren Peshek, and Kyle Stevens. MS was supported by NIH EY021494 (joint with Jacob Feldman); DH was supported by a grant from Procter & Gamble.

References

1. Adelson E. H., Pentland A.: The perception of shading and reflectance. In: D Knill and W Richards (Eds.). Perception as Bayesian Inference, pp. 409–423. Cambridge University Press, Cambridge, UK (1996)
2. Appleby, D., Ericsson, A., Fuchs, C.: Properties of QBist state spaces. Foundations of Physics **41**, 564–579 (2011)
3. Bauer, H.: Probability Theory. de Gruyter, Berlin (1996)
4. Blum, H.: Biological shape and visual science: Part I. Journal of Theoretical Biology **38**, 205–287 (1973)
5. Cohen, E. H., Singh, M.: Perceived orientation of complex shape reflects graded part decomposition. Journal of Vision **6**, 805–821 (2006)
6. Cover, T., Thomas, J.: Elements of Information Theory. Wiley, New York (2006)
7. De Winter, J., Wagemans, J.: Segmentation of object outlines into parts: A large-scale integrative study. Cognition **99**, 275–325 (2006)
8. Feldman, J., Singh, M.: Bayesian estimation of the shape skeleton. Proceedings of the National Academy of Sciences **103(47)**, 18014–18019 (2006)
9. Fuchs, C.: QBism, the perimeter of quantum Bayesianism. arXiv:1003.5209v1 (2010)
10. Fuchs, C., Schack, R.: Quantum-Bayesian coherence. arXiv:0906.2187v1 (2010)
11. Geisler, W., Diehl, R.: A Bayesian approach to the evolution of perceptual and cognitive systems. Cognitive Science **27**, 379–402 (2003)
12. Geisler, W., Kersten, D.: Illusions, perception and Bayes. Nature Neuroscience **5**, 508–510 (2002)
13. Hayworth, K., Biederman, I.: Neural evidence for intermediate representations in object recognition, Vision Research **46**, 4024–4031 (2006)
14. Hoffman, D.: Visual Intelligence: How We Create What We See. Norton, New York (1998)
15. Hoffman, D: The interface theory of perception. In Dickinson, S., Tarr, M. Leonardis, A., Schiele, B. (eds.) Object Categorization: Computer and Human Vision Perspectives, pp. 148–165. Cambridge University Press, Cambridge, UK (2009)
16. Hoffman, D., Singh, M.: Computational evolutionary perception. Perception **41**, 1073–1091 (2012)
17. Jaynes, E.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, UK (2003)
18. Kersten, D., Mamassian, P., Yuille, A.: Object perception as Bayesian inference. Annual Review of Psychology **555**, 271–304 (2004)
19. Kleffner, D., Ramachandran, V.: On the perception of shape from shading. Perception and Psychophysics **52**, 18–36 (1992)
20. Koenderink, J.: Vision and information. In Albertazzi, L., Tonder, G., Vishnawath, D. (eds.) Perception Beyond Inference: The Information Content of Visual Processes. Cambridge University Press, Cambridge, UK (2011)
21. Knill, D., Richards, W.: Perception As Bayesian Inference. Cambridge University Press, Cambridge, UK (1996)
22. Laplace P. S.: Memoir on the Probability of the Causes of Events. Statistical Science **1**, 364–378 (1986) (English translation by S. M. Stigler. Original work published in 1774)
23. Maloney, L., Zhang, H.: Decision-theoretic models of perception and action. Vision Research **50**, 2362–2374
24. Mamassian, P., Landy, M., Maloney, L.: Bayesian modeling of visual perception. In: Rao, R., Olshausen, B., Lewicki, M. (eds.) Probabilistic Models of the Brain: Perception and Neural Function, pp. 13–36. MIT Press, Cambridge, MA (2002)
25. Mark, J., Marion, B., Hoffman, D.: Natural selection and veridical perception. Journal of Theoretical Biology **266**, 504–515 (2010)
26. Marr, D.: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Freeman, San Francisco (1982)

27. Mausfeld, R.: The physicalist trap in perception theory. In: Heyer, D., Mausfeld, R. (eds.) Perception and the Physical World: Psychological and Philosophical Issues in Perception, pp. 75–112. Wiley, New York (2002)
28. Palmer, S.: Vision Science. MIT Press, Cambridge, MA (1999)
29. Pizlo, Z.: Perception viewed as an inverse problem. Vision Research **41**, 3145–3161 (2001)
30. Pizlo, Z., Sawada, T., Li, Y., Kropatsch, W., Steinman, R.M.: New Approach to the Perception of 3D Shape Based on Veridicality, Complexity, Symmetry and Volume. Vision Research **50**, 1–11 (2010)
31. Revuz, D.: Markov Chains. North-Holland, Amsterdam (1984)
32. Singh, M., Hoffman, D.: Part-based representations of visual shape and implications for visual cognition. In T. Shipley & P. Kellman (Eds.), From fragments to objects: Segmentation and grouping in vision, pp. 401–459. Elsevier, New York (2001)

Appendix: Relation to Quantum Bayesianism

One possible objection to the framework proposed in this chapter might be: “It is naive for vision scientists to propose that our perceptions are not veridical, and that therefore the objective world need not be spatiotemporal and need not contain 3D objects with shapes. Surely physicists know otherwise, and would dismiss such a proposal out of hand.”

Although some physicists might dismiss such a proposal, there are others who, in trying to best interpret the formalism of quantum theory, have been led to a view about quantum states that comports well with our proposal. These physicists, who call their approach “quantum Bayesianism,” or QBism for short, claim that quantum states are not objective representations of the external world, but rather are compendia of beliefs about possible outcomes of measurements [9, 10, 2]. As Fuchs [9] puts it, “. . . there is no sense in which the quantum state itself represents (pictures, copies, corresponds to, correlates with) a part or a whole of the external world, much less a world that *just is*” and “. . . a quantum state is a *state of belief* about what will come about as a consequence of . . . actions upon the system.” So, for instance, according to QBism a state function of a quantum system, represented say in the basis of the position operator, has a particular shape in space that can be used to predict the consequences of actions on that system.

This is entirely consistent with the view we propose about our perceptual experiences in general, and our experiences of shape in particular. There is no sense in which the objects in our perceptual experiences picture, copy, correspond to, or correlate with a part or a whole of the external world. Instead such objects and their shapes, and perceived space-time itself, are states of belief about what will come about as a consequence of our actions (which could include measurement). The reason is that natural selection, which has tuned our perceptions, rewards fitness and nothing else. Therefore our perceptions have been tuned to inform us of the fitness consequences of our possible actions, not to copy or picture the objective world.