Decision mechanisms from cognitive science for human-robot learning

Paul B. Reverdy

Abstract— **Robots that will be effective in performing tasks in cluttered real-world environments will have to solve problems analogous to those solved by humans in negotiating their environments. Cognitive science, therefore, can provide a fertile ground for finding decision mechanisms with the kinds of robustness required for automated robots. Furthermore, considering decision mechanisms for robots that are similar to those that can model human behavior facilitates the development of shared representations for both the human and the robot. Such shared representations provide natural interfaces from humans to robots and back again. In this work, we argue that the cognitive science paradigm of value-based decision making is a useful one for linking human decision making and robot control. We illustrate this argument through two examples, one in the context of decision making under uncertainty and the other using a value-based framework to perform real-time composition of high-level controllers.**

I. Introduction

Humans negotiate and manipulate their environments with a facility that far surpasses the abilities of even the best current robots. This facility is likely the fruit of a complex set of systems that allow humans to process noisy stimuli about the world, form internal representations of the patterns contained in these stimuli, and use the representations to make decisions that are expressed through motor actions.

This perception-action loop is analogous to the one used to design robots, although the physical (i.e., biological) hardware used by humans is different in kind to that used by robots. The biological hardware has significantly less computational power than that of robots, so human capabilities are likely underpinned by approximations that enable high performance at low computational cost. Understanding these approximations is of great use to robotics, and this paper argues that the framework of value-based decision making is a fruitful one to pursue this understanding.

II. Value-based decision making is optimization

Value-based decision making [1] is a paradigm for studying decision-making behavior. In the paradigm, subjects (who may be animal, human, or algorithmic) are given tasks where they are shown stimuli and have to react by performing certain actions, whereupon they receive rewards. The utility of this paradigm is due to the fact that the subject's optimal strategy can be written in terms of an optimization problem, namely, that of maximizing total rewards or reward rate. With such an optimal (often termed "normative") strategy defined, cognitive scientists can then empirically study deviations from optimal behavior and explain them in terms of heuristics and biases [2].

The properties of the value-based decision-making paradigm that make it useful for studying behavior also make it useful for robotics. It is natural to formulate robotics and control tasks as optimization problems; this is the basis of optimal control. Solving optimal control problems can be computationally costly, so understanding the heuristics and biases that permit biological systems to cheaply find approximate solutions is of great value to robotics. Conversely, when these biases create predictable patterns of suboptimal behavior, it is of interest to design computational aids to help humans improve their performance.

A generic value-based decision-making task has the following structure: a subject is presented with a set of possible actions $\mathcal{A}$ which may be finite or infinite. At each of a sequence of times $t \in \mathbb{N}$, the subject is presented with a stimulus s_t which may be visual, auditory, tactile, etc. The subject then selects an action $a_t \in \mathcal{A}$ and receives a reward $r_t \in \mathbb{R}$. Both the stimulus and the reward may be stochastic in the sense that they are corrupted by random noise. The subject's goal is to pick actions that maximize expected rewards:

$$\max_{\{a_t | \mathcal{F}_t\}_{t \in \mathbb{N}}} \sum_{t \in \mathbb{N}} \mathbb{E}\left[r_t\right], \tag{1}$$

where $\mathcal{F}_t$ represents the information available to the subject at time t. This paradigm is analogous to a Markov Decision Process (MDP) in obvious ways.

The normatively-optimal strategy to solve (1) can, in principle, be computed by exploiting the analogy to MDPs and employing an appropriate MDP solution algorithm, such as value iteration, Q-learning, etc. This provides a baseline for comparison with observed human behavior.

In solving control problems, one often employs the separation principle and designs a controller based on 1) a state-feedback control law, and 2) an estimator which produces estimates of the state that can be fed into the control law. Behavioral scientists build models that separate similarly. This introduces structure in two places: 1) heuristics, analogous to control laws, and 2) Bayesian statistics, which are used to build representations and estimators.

The structures of heuristics and priors are representations of task-relevant information that can be shared between humans and robots. By understanding the structures of heuristics that guide human decision making and the priors that inform these heuristics, we can develop algorithms

The author is with the Department of Aerospace and Mechanical Engineering, University of Arizona, Tucson, AZ 85721 USA. preverdy@email.arizona.edu. This work has been supported in part by startup funds provided by the University of Arizona, as well as grant number FA8650-15-D-1845 subcontract 669737-6 from the Air Force Research Laboratory.

that naturally interface between humans and machines. For example, a robotic system could learn from a human with high performance by fitting a model to the human's observed behavior and then using those model parameters in an algorithm with similar structure.

III. Prototypical tasks from cognitive science

Several prototypical tasks from cognitive science are particularly relevant to our goal of using value-based decision making in robotics. In this section we introduce three of them, and in the following two sections we show how we have begun to connect them to robotics.

A. *The explore-exploit problem*

The explore-exploit problem captures the central tradeoff at the heart of decision making under uncertainty: based on my current information, do I select an action whose rewards appear uncertain (explore), or do I select an action which appears highly rewarding (exploit)? If I explore too much, I will gain precise knowledge of the rewards associated with each action, but I will forego rewards. On the other hand, if I exploit too much, I may get stuck in a local maximum and fail to find highly rewarding actions.

Achieving high performance when making decisions under uncertainty therefore requires carefully balancing exploration and exploitation. Such decision-making scenarios occur frequently in daily life, so it is of interest to study how humans balance exploration and exploitation. The explore-exploit problem has received significant attention in the cognitive science community [3], [4], as well as in the statistics and machine learning communities [5].

B. *Perceptual decision making*

In its most straightforward form, the problem of perceptual decision making is the problem of filtering a perceptual stimulus and classifying it into one of several categories. In this sense, it is the behavioral analog to the classification problem in statistical signal processing.

Often, the stimulus comes as a signal to be integrated over time, and in the simplest case the subject must decide among two competing hypothetical classes to which the signal may belong. In this case, the optimal strategy, the sequential probability ratio test (SPRT), was investigated by Wald [6], and the analogous task is called a two-alternative forced choice task [7]. Such tasks have been studied in human as well as animal subjects using visual and auditory stimuli. The drift-diffusion model (DDM), a form of SPRT, has been shown to provide a unifying account for a wide variety of behavioral experimental results [7]. The DDM can incorporate a variety of behavioral biases, for example those arising from prior experience.

Standard signal processing theory can solve the stimulus classification problem using a variety of hypothesis testing techniques. These techniques yield classifiers which integrate the stimulus over time and output the corresponding category. Signal processing theory facilitates constructing performance guarantees for these classifiers, such as bounding the probability of making errors. A robot can then use the output category to decide on an action to take.

In certain applications, however, the time required to integrate sufficient evidence to make a confident decision may be of the same order as the time available to execute the corresponding action. In such an application, it may be beneficial to provide a less-filtered signal (such as the probability ratio at the heart of the SPRT) to the decision-making algorithm, since this gives real-time information about the relative likelihood of the different classifications.

C. *Affordance competition and action selection*

Situations where evidence must continuously be integrated and used to select among alternatives are at the heart of the so-called affordance competition hypothesis. An affordance is an opportunity for action defined by the environment around an animal or robot. The affordance competition hypothesis states that, in animals, the processes of action selection and movement planning operate simultaneously and in an integrated manner [8].

The cognitive science literature provides a body of evidence for the affordance competition hypothesis [8], [9]. Often this evidence takes the form of brain structures which interweave responsibility for perception and motion planning and execution. In the context of robotics, an analogous sensorimotor system would similarly integrate processing of perceptual stimuli and motion planning. In Section V below I present the details of such a sensorimotor system, called *motivation dynamics*, which I am actively developing.

IV. Multi-armed bandits and UCL

In a series of recent papers [10], [11], [12], I and several colleagues connected the generic explore-exploit problem to a particular problem of interest to the robotics community, namely, spatial search. We did this by linking the spatial search problem to the multi-armed bandit problem [5], [13]. The standard multi-armed bandit problem is a sequential value-based decision-making task of the form (1) where $\mathcal{A}$ is a finite set of N actions and the reward r_t at time t is sampled from a stationary probability distribution associated with the action a_t.

The normatively-optimal solution to the multi-armed bandit problem is computationally intractable except for several special cases. However, there is well-known upper bound to performance [13] and a large number of authors in the statistics literature have found heuristic-based algorithms that effectively match the bound. In [10], we considered a so-called *spatial multi-armed bandit problem*, where the actions $\mathcal{A}$ are embedded in an ambient space. In this way, the spatial multi-armed bandit problem models stochastic function optimization. The spatial embedding provides an opportunity for subjects to integrate structural knowledge in the form of spatial dependencies among arms, for example arising from a spatial correlation length scale.

In experiments reported in [10], we found that a significant number of human subjects were able to achieve performance in spatial multi-armed bandit tasks better than that implied by the bound of [13]. We attributed this improved performance to subjects' understanding of the spatial structure of the problem and showed how to build a heuristic-based algorithm called UCL that could model human decision making in the task. Notably, we showed that UCL could produce behavior that qualitatively matched the major categories of human behavior exhibited by our subjects by changing a small number of parameters. We also showed in [10] that UCL could achieve optimal performance with appropriate parameter tunings. Therefore, UCL could act both as a model of human behavior and a decision-making algorithm for an automated system. By learning good parameter tunings from humans with high performance in a given scenario, an automated system using UCL could achieve higher performance than it could with parameter tunings designed to perform well in a generic scenario.

In [11], we considered the problem of learning the UCL parameters from human behavioral data in detail. We developed a parameter estimator for the UCL parameters and proved performance guarantees for the estimator. By employing the estimator on experimental data previously reported in [10], we showed that there were statistically-significant differences between subjects with high performance in different task scenarios. This empirical result provides a proof-of-concept of the idea of using high-performing individuals to train automated systems by estimating their UCL algorithm parameters.

In [12], we considered the multi-armed bandit problem with a satisficing objective, as opposed to the standard maximizing one. Satisficing is a behavioral concept by which the decision-making agent seeks performance above a certain threshold rather than seeking the absolute best-possible performance. It implicitly accounts for the fact that seeking high performance is costly, and that seeking performance above a certain level may not be worth the resulting costs. This tradeoff is intuitive to anyone who has tuned parameters for an algorithm.

We are currently pursuing a variety of applications of the multi-armed bandit framework to robotics. In work currently in revision [14] we are applying satisficing UCL to the problem of tuning parametric gaits in a quadrupedal robot. In [15] we considered applying a distributed version of UCL to a multi-robot foraging task. In future work we will extend these applications to include training the UCL algorithms using parameter tunings from high-performing human supervisors.

V. Motivation dynamics

In ongoing work, I am pursuing a framework for value-based sensorimotor systems that implements a form of affordance competition and naturally interfaces with low-level signal processing models of the type used to study perceptual decision making. I call this framework *motivation dynamics*.

The motivation dynamics framework, introduced in [16], can be thought of as a convex relaxation of hybrid dynamical systems in the following sense. Like a hybrid dynamical system, a motivation dynamics system with continuous state x has a finite set of low-level controllers $F_a(x), a \in \mathcal{A} = \{1, \dots, N\}$ called *modes*. A hybrid dynamical system follows one mode a_t at time t, where a is the state variable of a discrete finite automaton. The continuous dynamics are then $\dot{x} = F_{a_t}(x)$, and various guard functions control the transitions of the automaton. Instead of the mode variable $a \in \mathcal{A}$, the motivation dynamics system maintains a *motivation* state $m \in \Delta^N = \{x \in \mathbb{R}^{N+1} | x_i \geq 0, \sum_i x_i = 1\}$, where Δ^N is the N-simplex. The vertices of Δ^N are analogous to the modes $a \in \mathcal{A}$, while other elements of Δ^N consist of convex combinations of the vertices. The dynamics of x under motivation dynamics are given by $\dot{x} = \sum_{i=1}^N m_i F_i(x)$, which can be thought of a convex relaxation of the continuous dynamics of the hybrid dynamical system.

In place of the discrete finite automaton that selects modes in a hybrid system, motivation dynamics framework [16] uses a bio-inspired dynamical system from [17] which implements a value-based decision-making model. Specifically, the motivation dynamics framework associates a value state $v_i > 0$ to each mode $i \in \mathcal{A}$ (v_i may have its own dynamics and depend on the environment or external stimuli) and the motivation dynamics $\dot{m} = f_m(m, v)$ is such that the motivation state m will tend towards a point that puts most weight on the highest-value mode. By tightly coupling valuation, action (i.e., mode) selection, and physical dynamics, the motivation dynamics framework implements a form of affordance competition as discussed in [8].

A. Natural interface for perceptual decision models

The value state $v \in \mathbb{R}^N_+$ in the motivation dynamics provides a natural interface between the motivation dynamics framework and low-level perceptual decision models like the DDM discussed in Section III-B above. For example, the likelihood or log-likelihood of a given category of stimulus can be used as the value input associated with the mode that should be triggered when that stimulus is detected. Such a connection was suggested in [18], which studied the dynamics $\dot{m} = f_m(m, v)$ in their original biological context. In recent work [19], we show that using log-likelihoods as value states permits a robot to smoothly select correct actions in response to noisy stimuli.

B. Connections to LTL controller synthesis

A more standard approach for building a robot that could carry out actions in response to noisy stimuli as demonstrated in [19] would be to construct a statistical classifier which processes the raw stimulus and outputs a discrete classification of the stimulus. Then a logic-based framework such as the Linear-Temporal Logic (LTL) synthesis approaches of [20], [21] could be used to synthesize a hybrid system that would select the appropriate sequence of actions in response to the classified stimuli.

In contrast to the discretized approach required by the LTL synthesis framework, the motivation dynamics framework retains a continuous representation of all the relevant signals, including physical dynamics, action selection mode (i.e., motivation state m), and mode valuations, such as the likelihood values for various stimuli considered in [19]. This facilitates the interfacing between motivation dynamics and more complex models of human perceptual decision-making behavior, which could encode insights from human domain experts into the low-level stimulus processing.

Motivation dynamics is not intended as a substitute for LTL methods but rather as a complement to them acting at a lower level, closer to the sensorimotor physical hardware. Nevertheless, motivation dynamics is already able to encode some elements of LTL in a continuous dynamical system. For example, we provided a formal guarantee in [16] that motivation dynamics can encode a limit cycle that corresponds to a robot repeatedly visiting two desired locations, a behavior which is referred to in the LTL literature as persistent surveillance. In [19], we showed how to implement the motivation dynamics limit cycle on a physical robot and that the resulting persistent surveillance behavior was robust to a variety of environmental perturbations.

We are actively pursuing research on the motivation dynamics framework in a variety of directions. One direction seeks to encode more logical elements of the LTL framework in terms of motivation dynamical systems. Another seeks to integrate the limit cycle behavior shown in [16] and [19] with the stimulus response behavior demonstrated in [19]. Ultimately, we seek to develop more complicated action valuation schemes that can take into account a variety of stimuli and contextual information from the environment. One natural way to develop such schemes would be to learn them from human behavior. The value-based nature of the motivation dynamics decision mechanism would facilitate learning from human behavior using decision-making models such as the UCL algorithm discussed above or a more generic model such as a utility function, which can encode human preferences over possible outcomes [22].

VI. Conclusions

To develop more capable robots, roboticists should more carefully investigate the connections between human decision making and robot control. These connections can facilitate developing more effective methods for robots to learn from humans.

I argued that the behavioral science paradigm of value-based decision making is a fruitful framework for this project. I considered three prototypical tasks from cognitive science: the explore-exploit problem, the perceptual decision making problem, and the concept of affordance competition and the action selection problem.

I provided an overview of our work connecting human decision making and robot control through these three prototypical tasks, discussing a number of algorithms and frameworks we have developed in the process.

I hope that our argument for taking the value-based decision-making viewpoint is convincing and that other researchers are encouraged to pursue work along similar lines. I have attempted to point out opportunities for future work and welcome collaboration on this exciting project.

References

[1] A. Rangel, C. Camerer, and P. R. Montague, "A framework for studying the neurobiology of value-based decision making," *Nature Reviews Neuroscience*, vol. 9, no. 7, pp. 545–556, Jun 2008. [Online]. Available: http://dx.doi.org/10.1038/nrn2357

[2] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[3] J. D. Cohen, S. M. McClure, and J. Y. Angela, "Should I stay or should I go? how the human brain manages the trade-off between exploitation and exploration," *Philosph. Trans. of the Roy. Soc. B: Biological Sci.*, vol. 362, no. 1481, pp. 933–942, 2007.

[4] R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, and J. D. Cohen, "Humans use directed and random exploration to solve the explore-exploit dilemma," *Journal of Experimental Psychology: General*, vol. 143, no. 6, pp. 2074–2081, 2014.

[5] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the Amer. Math. Soc.*, vol. 58, pp. 527–535, 1952.

[6] A. Wald, "Sequential tests of statistical hypotheses," *Ann. of Math. Stat.*, vol. 16, no. 2, pp. 117–186, 1945.

[7] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen, "The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks." *Psychological review*, vol. 113, no. 4, p. 700, 2006.

[8] P. Cisek, "Cortical mechanisms of action selection: the affordance competition hypothesis," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 362, no. 1485, pp. 1585–1599, 2007.

[9] P. Cisek and J. F. Kalaska, "Neural mechanisms for interacting with a world full of action choices," *Annual review of neuroscience*, vol. 33, pp. 269–298, 2010.

[10] P. B. Reverdy, V. Srivastava, and N. E. Leonard, "Modeling human decision making in generalized Gaussian multiarmed bandits," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 544–571, 2014.

[11] P. Reverdy and N. E. Leonard, "Parameter estimation in softmax decision-making models with linear objective functions," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 1, pp. 54–67, 2016.

[12] P. Reverdy, V. Srivastava, and N. E. Leonard, "Satisficing in multi-armed bandit problems," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3788–3803, 2017.

[13] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[14] C. Zawacki, P. Reverdy, and D. E. Koditschek, "Gait optimization on a quadrupedal RHex using multiarmed bandits," *Preprint*, 2018. [Online]. Available: https://camzawacki.com/static/pdfs/bandit-paper.pdf

[15] P. Landgren, P. Reverdy, V. Srivastava, and N. E. Leonard, "Multi-robot foraging using the graphical multiarmed bandit framework," in *Workshop on Informative Path Planning and Adaptive Sampling, IEEE ICRA*, 2018.

[16] P. B. Reverdy and D. E. Koditschek, "A dynamical system for prioritizing and coordinating motivations," *SIAM Journal on Applied Dynamical Systems*, vol. 17, no. 2, pp. 1683–1715, 2018.

[17] T. D. Seeley, P. K. Visscher, T. Schlegel, P. M. Hogan, N. R. Franks, and J. A. Marshall, "Stop signals provide cross inhibition in collective decision-making by honeybee swarms," *Science*, vol. 335, no. 6064, pp. 108–111, 2012.

[18] D. Pais, P. M. Hogan, T. Schlegel, N. R. Franks, N. E. Leonard, and J. A. Marshall, "A mechanism for value-sensitive decision-making," *PloS one*, vol. 8, no. 9, p. e73216, 2013.

[19] P. B. Reverdy, V. Vasilopoulos, O. Arslan, and D. E. Koditschek, "Motivation dynamics for autonomous composition of navigation tasks," in *In preparation*, 2019.

[20] H. Kress-Gazit, G. E. Fainekos, and G. J. Pappas, "Temporal-logic-based reactive mission and motion planning," *IEEE Transactions on Robotics*, vol. 25, no. 6, pp. 1370–1381, 2009.

[21] J. Liu, N. Ozay, U. Topcu, and R. M. Murray, "Synthesis of reactive switching protocols from temporal logic specifications," *IEEE Transactions on Automatic Control*, vol. 58, no. 7, pp. 1771–1785, 2013.

[22] R. L. Keeney and H. Raiffa, *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge University Press, 1993.