האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

REINFORCEMENT LEARNING AND HUMAN BEHAVIOR

By

HANAN SHTEINGART, and YONATAN LOEWENSTEIN

Discussion Paper # 656 January 2014

מרכז לחקר הרציונליות

CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

Reinforcement Learning and Human Behavior

Hanan Shteingart[1] and Yonatan Loewenstein[1, 2]

[1]Edmond and Lily Safra Center for Brain Sciences and [2]Dept. of Neurobiology, The Alexander Silberman Institute of Life Sciences, Dept. of Cognitive Science and Center for the Study of Rationality, The Hebrew University, Jerusalem 91904, Israel.

Corresponding author: Yonatan Loewenstein. Edmond and Lily Safra Center for Brain Sciences, the Hebrew University of Jerusalem 91904, Israel.

Email: yonatan@huji.ac.il. Phone: +972-2-6585996, Fax: +972-2-6585997.

Abstract

The dominant computational approach to model operant learning and its underlying neural activity is model-free reinforcement learning (RL). However, there is accumulating behavioral and neuronal-related evidence that human (and animal) operant learning is far more multifaceted. Theoretical advances in RL, such as hierarchical and model-based RL extend the explanatory power of RL to account for some of these findings. Nevertheless, some other aspects of human behavior remain inexplicable even in the simplest tasks. Here we review developments and remaining challenges in relating RL models to human operant learning. In particular, we emphasize that learning a model of the world is an essential step prior or in parallel to learning the policy in RL and discuss alternative models that directly learn a policy without an explicit world model in terms of state-action pairs.

Model-free RL

The computational problem in many operant learning tasks can be formulated in a framework known as *Markov Decision Processes* (MDP) [1]. In MDPs, the world can be in one of several *states*, which determine the consequences of the agent's *actions* with respect to the future *rewards* and world states. A *policy* defines the agent behavior at a given situation. In MDP, a policy is a mapping from the states of the environment to actions to be taken when in those states [1]. Finding the optimal *policy* is difficult because actions may have both immediate and long-term consequences. However, this problem can be simplified by estimating *values*, the expected cumulative (discounted) rewards associated with these states and actions and using these values to choose the actions (for detailed characterization of the mapping from values to actions in humans, see [2]).

Model-free RL, as its name suggests, is a family of RL algorithms devised to learn the values of the states without learning the full specification of the MDP. In a class of model-free algorithms, known as temporal-difference learning, the learning of the values is based on the *reward-prediction error* (RPE), the discrepancy between the expected reward before and after an action is taken (taking into account also the ensuing obtained reward). The hypothesis that the brain utilizes model-free RL for operant learning holds considerable sway in the fields of neuroeconomics. This hypothesis is supported by experiments demonstrating that in primates, the phasic activity of mid-brain dopaminergic neurons is correlated with the RPE [3,4]. In mice, this correlation was also shown to be causal: optogenetic activation of dopaminergic neurons is sufficient to drive operant learning, supporting the hypothesis that the dopaminergic neurons encode the RPE , which is used for operant learning [5]. Other putative brain regions for this computation are the striatum, whose activity is correlated with values of the states and / or actions [6,7] and the nucleus accumbens and pallidum, which are involved in the selection of the actions [8]. In addition to its neural correlates, model-free RL has been used to account for the trial-by-trial dynamics (e.g., [2]) and for several robust aggregate features of human behavior such as risk aversion [9],

recency [10] and primacy [2]. Moreover, model-free RL has been proven useful in the field of computational psychiatry as a way of diagnosing and characterizing different pathologies [11–14].

However, there is also evidence that the correspondence between dopaminergic neurons and the RPE is more complex and diverse than was previously thought [15]. First, dopaminergic neurons increase their firing rate in response to both surprisingly positive and negative reinforcements [16,17]. Second, dopaminergic activity is correlated with other variables of the task, such as uncertainty [18]. Third, the RPE is not exclusively represented by dopamine, as additional neuromodulators, in particular serotonin, are also correlated with the RPE [19]. Finally, some findings suggest that reinforcement and punishment signals are not local but rather ubiquitous in the human brain [20]. These results challenge the dominance of the anatomically-modular model-free RL as a model for operant learning.

Model-based RL

When training is intense, task-independent reward devaluation, e.g. through satiety, has only a little immediate effect on behavior. This habitual learning is consistent with model-free RL because in this framework, the value of an action is updated only when it is executed. By contrast, when training is moderate, the response to reward devaluation is immediate and substantial [21]. This and other behaviors (e.g., planning) is consistent with an alternative RL approach, known as *model-based* RL, in which a model of the world, i.e., the parameters that specify the MDP is learned prior to choosing a policy. The effect of reward devaluation after moderate training can be explained by model-based RL because a change in a world parameter (e.g., the value of the reward as a result of satiety) can be used to update (off-line) the values of other states and actions.

If the parameters of the MDP are known, one can compute the values of all states and actions, for example by means of dynamic programming or Monte-Carlo simulation. Alternatively, one could choose an action by expanding a look-ahead decision tree on-line [1]. However, a full expansion of a look-ahead tree is computationally difficult because the number of branches increases exponentially with the height

of the tree, so that pruning of the tree is a necessary approximation. Indeed, a recent study has suggested that humans prune the decision-tree, by trimming branches associate with large losses [14].

Whether or not model-based and model-free learning are implemented by two anatomically distinct systems is a subject of debate. In support of anatomical modularity are findings that the medial striatum is more engaged during planning whereas the lateral striatum is more engaged during choices in extensively trained tasks [22]. In addition, the *state prediction error*, which signals the discrepancy between the current model and the observed state transitions is correlated with activity in the intraparietal sulcus and lateral prefrontal cortex, spatially separated from the main correlate of the RPE in the ventral striatum [23]. Findings that tend to negate anatomical modularity include reports of signatures of both model-based and model-free learning in the ventral striatum [24].

The curse of dimensionality and the blessing of hierarchical RL

There are also theoretical reasons why the RL models described above cannot fully account for operant learning in natural environments. First, the computational problem of finding the values is bedeviled by the "curse of dimensionality": the number of states is exponential with the number of variable, which define a state [1]. Second, when the state of the world is only partially known, (i.e., the environment is a partially observable MDP (POMDP), applying model-free algorithms such as Q-learning may converge to a solution that is far from optimality or may fail to converge altogether [25]. One approach to addressing these problems is to break down the learning task into a hierarchy of simpler learning problems, a framework known as *Hierarchical Reinforcement Learning* (HRL) [26]. Neuroimaging studies have indeed found neural responses that are consistent with sub-goal-related RPE, as is predicted by HRL [27].

Challenges in relating human behavior to RL algorithms

Despite the many successes of the different RL algorithms in explaining some of the observed human operant learning behaviors, others are still difficult to account for. For example, humans tend to

alternate rather than repeat an action after receiving a positively surprising payoff. This behavior is observed both in simple repeated two-alternative force choice tasks with probabilistic rewards (also known as the 2-armed bandit task, Fig. 1A) and in the stock market [28]. Moreover, a recent study found that the behavior of half of the participants in a 4-alternaive version of the bandit task, known as the Iowa gambling task, is better explained by the simple ad-hoc heuristic "win-stay, lose-shift" (WSLS) than by RL models [29]. Another challenge to the current RL models is the tremendous heterogeneity in reports on human operant learning, even in simple bandit tasks, measured in different laboratories in slightly different conditions. For example, as was described above, the WSLS observed in 4-arm bandit [29] is inconsistent with the alternation after positively surprising payoffs discussed above [28]. Additionally, *probability matching*, the tendency to choose an action in proportion to the probability of reward associated with that action, has been a subject of debate over half-a-century. On one hand, there are numerous reports supporting this law of behavior both in the laboratory and when humans gamble substantial amounts of money on the outcome of real-life situations [30]. On the other hand, there is abundant literature arguing that people deviate from probability matching in favor of choosing the more rewarding action (maximization) [31]. Finally, there is substantial heterogeneity not only between subjects and laboratories but also within subjects over time. A recent study has demonstrated substantial day-to-day fluctuations in learning behavior of monkeys in the two-armed bandit task and has shown that these fluctuations are correlated with day-to-day fluctuations in the neural activity in the putamen [32].

Heterogeneity in world model

The lack of uniformity regarding behavior even in simple tasks could be due to heterogeneity in the prior expectations of the participants. From the experimentalist point of view, the two-armed bandit task, for example, is simple: the world is characterized by a single state and two actions (Figure 1A). However, from the participant point of view there is, theoretically, an infinite repertoire of possible world models characterized by different definitions of sets of states and actions. This could be true even when precise instructions are given due, for example, lack of trust, inattention or forgetfulness. With respect to

the actions set, the participant may assume that there is only a single available action, the pressing of any button, regardless of its properties (Figure 1B). Alternatively, differences in the timing of the button press, the finger used, etcetera, could all define different actions. Such precise definition of action, which is irrelevant to the task, may end with non-optimal behavior. With respect to the states set, the participant may assume that there are several states that depend on the history of actions and / or rewards. For example, the participant may assume that the state is defined by the last action (Fig. 1C), the last action and the last reward, or a function of the long history of actions [33]. Finally, the participant may assume a strategic game setting such as a matching-pennies game (Figure 1D). These and other possible assumptions (Figure 1E) may lead to very different predictions on behavior [34]. In support of this possibility, experimental manipulations such as instructions, which are irrelevant to the reward schedule, but may change the prior belief about the number of states can have a considerable effect on human behavior [35]. Finally, humans and animals have been shown to develop idiosyncratic and stereotyped superstitious behaviors even in simple laboratory settings [36]. If participants fail to recognize the true structure of a learning problem in simple laboratory settings, they may also fail to identify the relevant states and actions when learning from rewards in natural environments. For example, professional basketball players have been shown to overgeneralize when learning from their experience [37].

Learning the world model

Many models of operant learning often take as given that the learner has already recognized the available sets of states and actions (Fig. 2A). Hence, when attempting to account for human behavior they fail to consider the necessary preliminary step of identifying them (correctly or incorrectly). In machine learning, classification is often preceded by an unsupervised dimension-reduction for feature extraction [38,39]. Similarly, it has been suggested that operant learning is a two-step process (Figure 2B): in the first step, the state and action sets are learned from the history (possibly using priors on the world), where in the second step RL algorithms are utilized to find the optimal policy given these sets [40]. An interesting alternative is that the relevant state-action sets and the policy are learned in parallel (Figure

2C). For example, a new approach in RL, known as *feature* RL, the state set and the values of the states are learned simultaneously from the history of observations, actions and rewards. One crucial property of feature RL is that it neither requires nor learns a model of the complete observation space, but rather learns a model that is based on the reward-relevant observations [41].

Learning without States

Operant learning can also be accomplished without an explicit representation of states and actions, by directly tuning a parametric policy (Fig. 2D). A plausible implementation of such direct policy learning algorithms is using stochastic policy-gradient methods [42–44]. The idea behind these methods is that the gradient of the average reward (with respect to policy parameter) can be estimated on-line by perturbing a neural network model with noise and considering the effect of these perturbations on the stream of payoffs delivered to the learning agent. Changes in the policy in the direction of this estimated gradient are bound, under certain assumptions, to improve performance. However, local minima may prevent the learning dynamics from converging to the optimal solution.

Direct policy methods have been proposed to explain birdsong learning [45] and have received some experimental support [46,47]. In humans, a model for gradient learning in spiking neurons [48,49] has been shown to be consistent with the dynamics of human learning in two-player games [50]. Under certain conditions, gradient-like learning can be implemented using covariance-based synaptic plasticity. Interestingly, operant matching (not to be confused with probability matching) naturally emerges in this framework [51,52]. A model based on attractor dynamics and covariance-based synaptic plasticity has been shown to quantitatively account for free operant learning in rats [53]. However, the experimental evidence for gradient-based learning, implemented at the level of single synapses, awaits future experiments.

Concluding remarks

RL is the dominant theoretical framework to operant learning in humans and animals. RL models were partially successful in quantitative modeling of learning behavior and provided important insights into the putative role of different brain structures in operant learning. Yet, substantial theoretical as well as experimental challenges remain, indicating that these models may be substantially oversimplified. In particular, how state-space representations are learned in operant learning remain important challenges for future research.

Acknowledgements

We would like to thank Ido Erev for many fruitful discussions and David Hansel, Gianluigi Mongillo, Tal Neiman and Ran Darshan for carefully reading the manuscript.

This work was supported by the Israel Science Foundation (Grant No. 868/08), grant from the ministry of science and technology, Israel and the ministry of foreign and European affairs and the ministry of higher education and research France and the Gatsby Charitable Foundation.

References

1. Sutton RS, Barto AG: *Introduction to Reinforcement Learning*. MIT Press; 1998.

- 2. Shteingart H, Neiman T, Loewenstein Y: **The role of first impression in operant learning.** *J. Exp. Psychol. Gen.* 2013, **142**:476–88.

3. Schultz W: **Updating dopamine reward signals.** *Curr. Opin. Neurobiol.* 2013, **23**:229–38.

4. Montague PR, Hyman SE, Cohen JD: **Computational roles for dopamine in behavioural control.** *Nature* 2004, **431**:760–7.

- 5. Kim KM, Baratta M V, Yang A, Lee D, Boyden ES, Fiorillo CD: **Optogenetic mimicry of the transient activation of dopamine neurons by natural reward is sufficient for operant reinforcement.** *PLoS One* 2012, **7**:e33612.

6. Samejima K, Ueda Y, Doya K, Kimura M: **Representation of action-specific reward values in the striatum.** *Science* 2005, **310**:1337–40.

7. O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ: **Dissociable roles of ventral and dorsal striatum in instrumental conditioning.** *Science* 2004, **304**:452–4.

8. Nicola SM: **The nucleus accumbens as part of a basal ganglia action selection circuit.** *Psychopharmacology (Berl).* 2007, **191**:521–50.

9. Denrell J: **Adaptive learning and risk taking.** *Psychol. Rev.* 2007, **114**:177–187.

10. Hertwig R, Barron G, Weber EU, Erev I: **Decisions from experience and the effect of rare events in risky choice.** *Psychol. Sci.* 2004, **15**:534–9.

11. Yechiam E, Busemeyer JR, Stout JC, Bechara A: **Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits.** *Psychol. Sci.* 2005, **16**:973–8.

12. Maia T V, Frank MJ: **From reinforcement learning models to psychiatric and neurological disorders.** *Nat. Neurosci.* 2011, **14**:154–62.

13. Montague PR, Dolan RJ, Friston KJ, Dayan P: **Computational psychiatry.** *Trends Cogn. Sci.* 2012, **16**:72–80.

- 14. Huys QJM, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP: **Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees.** *PLoS Comput. Biol.* 2012, **8**:e1002410.

15. Lammel S, Lim BK, Malenka RC: **Reward and aversion in a heterogeneous midbrain dopamine system.** *Neuropharmacology* 2013, doi:10.1016/j.neuropharm.2013.03.019.

16. Iordanova MD: **Dopamine transmission in the amygdala modulates surprise in an aversive blocking paradigm.** *Behav. Neurosci.* 2010, **124**:780–8.

17. Joshua M, Adler A, Mitelman R, Vaadia E, Bergman H: **Midbrain dopaminergic neurons and striatal cholinergic interneurons encode the difference between reward and aversive events at different epochs of probabilistic classical conditioning trials.** *J. Neurosci.* 2008, **28**:11673–84.

18. Fiorillo CD, Tobler PN, Schultz W: **Discrete coding of reward probability and uncertainty by dopamine neurons.** *Science* 2003, **299**:1898–902.

19. Seymour B, Daw ND, Roiser JP, Dayan P, Dolan R: **Serotonin selectively modulates reward value in human decision-making.** *J. Neurosci.* 2012, **32**:5833–42.

20. Vickery TJ, Chun MM, Lee D: **Ubiquity and specificity of reinforcement signals throughout the human brain.** *Neuron* 2011, **72**:166–77.

21. Tricomi E, Balleine BW, O'Doherty JP: **A specific role for posterior dorsolateral striatum in human habit learning.** *Eur. J. Neurosci.* 2009, **29**:2225–32.

22. Wunderlich K, Dayan P, Dolan RJ: **Mapping value based planning and extensively trained choice in the human brain.** *Nat. Neurosci.* 2012, **15**:786–91.

23. Gläscher J, Daw N, Dayan P, O'Doherty JP: **States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning.** *Neuron* 2010, **66**:585–95.

- 24. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ: **Model-based influences on humans' choices and striatal prediction errors.** *Neuron* 2011, **69**:1204–15.

25. Jaakkola T, Singh SP, Jordan MI: **Reinforcement learning algorithm for partially observable Markov decision problems**. In *NIPS 1994*. 1994:1143.

26. Barto AG, Mahadevan S: **Recent Advances in Hierarchical Reinforcement Learning**. *Discret. Event Dyn. Syst.* 2003, **13**:341–379.

- 27. Diuk C, Tsai K, Wallis J, Botvinick M, Niv Y: **Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia.** *J. Neurosci.* 2013, **33**:5797–805.

28. Nevo I, Erev I: **On surprise, change, and the effect of recent outcomes.** *Front. Psychol.* 2012, **3**:24.

- 29. Worthy DA, Hawthorne MJ, Otto AR: **Heterogeneity of strategy use in the Iowa gambling task: a comparison of win-stay/lose-shift and reinforcement learning models.** *Psychon. Bull. Rev.* 2013, **20**:364–71.

30. McCrea SM, Hirt ER: **Match madness: probability matching in prediction of the NCAA basketball tournament**. *J. Appl. Soc. Psychol.* 2009, **39**:2809–2839.

31. Vulkan N: **An economist's perspective on probability matching**. *J. Econ. Surv.* 2000, **14**:101–118.

32. Laquitaine S, Piron C, Abellanas D, Loewenstein Y, Boraud T: **Complex Population Response of Dorsal Putamen Neurons Predicts the Ability to Learn**. *PLoS One* 2013, **8**:e80683.

33. Loewenstein Y, Prelec D, Seung HS: **Operant matching as a Nash equilibrium of an intertemporal game.** *Neural Comput.* 2009, **21**:2755–73.

34. Green CS, Benson C, Kersten D, Schrater P: **Alterations in choice behavior by manipulations of world model.** *Proc. Natl. Acad. Sci. U. S. A.* 2010, **107**:16401–6.

35. Morse EB, Willard N. Runquist: **Probability-matching with an unscheduled random sequence**. *Am. J. Psychol.* 1960, **73**:603–607.

36. Ono K: **Superstitious behavior in humans**. *J. Exp. Anal. Behav.* 1987, **47**:261–271.

37. Neiman T, Loewenstein Y: **Reinforcement learning in professional basketball players**. *Nat. Commun.* 2011, **2**:569.

38. Hinton GE, Salakhutdinov RR: **Reducing the dimensionality of data with neural networks.** *Science* 2006, **313**:504–7.

39. Hinton G: **Where Do Features Come From?** *Cogn. Sci.* 2013, doi:10.1111/cogs.12049.

40. Legenstein R, Wilbert N, Wiskott L: **Reinforcement learning on slow features of high-dimensional input streams.** *PLoS Comput. Biol.* 2010, **6**.

41. Nguyen P, Sunehag P, Hutter M: **Context tree maximizing reinforcement learning**. In *Proc. of the 26th AAAI Conference on Artificial Intelligence.* 2012.

42. Williams R: **Simple statistical gradient-following algorithms for connectionist reinforcement learning**. *Mach. Learn.* 1992, **8**:229–256.

43. Sutton RS, Mcallester D, Singh S, Mansour Y, Avenue P, Park F, Satinder S: **Policy gradient methods for reinforcement learning with function approximation**. *Adv. Neural Inf. Process. Syst. 12* 1999, **12**:1057–1063.

44. Bartlett PL, Baxter J, Jonathan Baxter PLB: **Infinite-horizon policy-gradient estimation**. 2001, doi:10.1613/jair.806.

45. Fiete IR, Fee MS, Seung HS: **Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances.** *J. Neurophysiol.* 2007, **98**:2038–2057.

46. Andalman AS, Fee MS: **A basal ganglia-forebrain circuit in the songbird biases motor output to avoid vocal errors.** *Proc. Natl. Acad. Sci. U. S. A.* 2009, **106**:12518–12523.

47. Tumer EC, Brainard MS: **Performance variability enables adaptive plasticity of "crystallized" adult birdsong.** *Nature* 2007, **450**:1240–1244.

48. Urbanczik R, Senn W: **Reinforcement learning in populations of spiking neurons.** *Nat. Neurosci.* 2009, **12**:250–2.

49. Friedrich J, Urbanczik R, Senn W: **Spatio-temporal credit assignment in neuronal population learning.** *PLoS Comput. Biol.* 2011, **7**:e1002092.

50. Friedrich J, Senn W: **Spike-based decision learning of Nash equilibria in two-player games.** *PLoS Comput. Biol.* 2012, **8**:e1002691.

51. Loewenstein Y, Seung HS: **Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity.** *Proc. Natl. Acad. Sci. U. S. A.* 2006, **103**:15224–9.

52. Loewenstein Y: **Synaptic theory of replicator-like melioration.** *Front. Comput. Neurosci.* 2010, **4**:17.

53. Neiman T, Loewenstein Y: **Covariance-based synaptic plasticity in an attractor network model accounts for fast adaptation in free operant learning.** *J. Neurosci.* 2013, **33**:1521–34.

Figure 1. Repertoire of possible world models

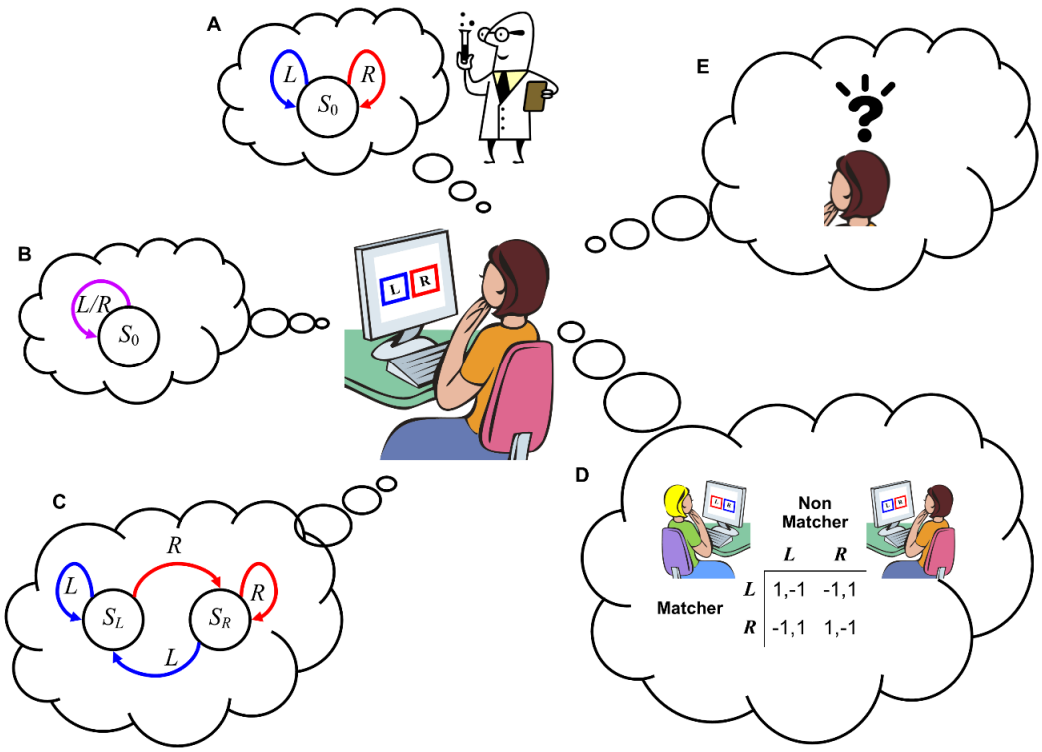

Figure 1. In this example, a participant is tested in the two-armed bandit task. (A) From the experimentalist's point of view (scientist caricature), the world is characterized by a single state (S_0) and two actions: left (blue, *L*) or right (red, *R*) button press. However, from the participant's point of view there is an infinite repertoire of possible world models characterized by different sets of states and actions. (B) With respect to the action sets, she may assume that there is only a single available action, pressing any button, regardless of its location (purple, *L*/*R*). (C) With respect to the state sets, the participant may assume that the state is defined by her last action (S_L and S_R, for previous *L* and *R* action, respectively). (D) Moreover, the participant may assume she is playing a penny-matching game with another human. (E) These and other possible assumptions may lead to very different predictions in the framework of RL.

Figure 2. Alternative models of operant learning

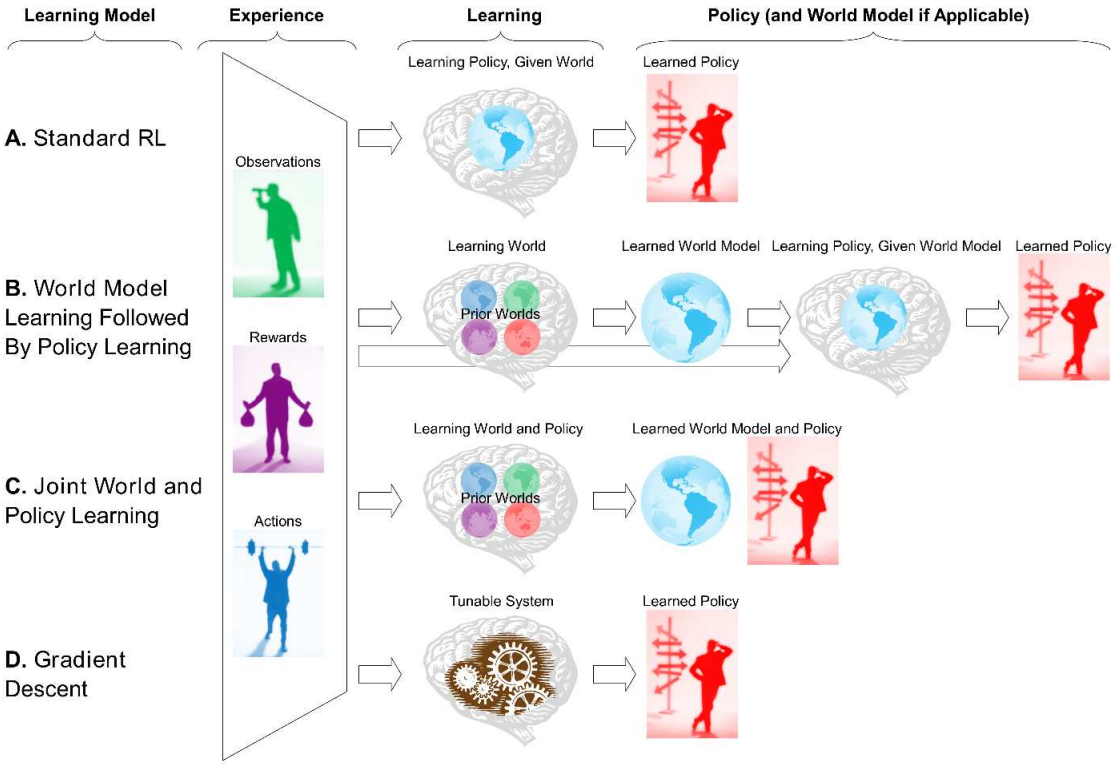

Figure 2. In operant learning, experience (left trapezoid), composed of present and past observations actions and rewards is used to learn a policy. (A) Standard RL models typically assume that the learner (brain gray icon) has access to the relevant states and actions set (represented by a bluish world icon) prior to the learning the policy. Alternative suggestions are that the state and action sets are learned from experience and from prior expectations (different world icons) before (B) or in parallel (C) to the learning of policy. (D) Alternatively, the agent may directly learn without an explicit representation of states and actions, but rather by tuning a parametric policy (cog wheels icon), e.g., using stochastic gradient methods on this policy's parameters.