Do Reinforcement Learning Models Explain Neural Learning?

Svenja Stark
Fachbereich 20 - Informatik
TU Darmstadt
svenja.stark@stud.tu-darmstadt.de

Abstract

Because the functionality of our brains is still a blank sheet, this paper shall take a glimpse at what possibilities Reinforcement Learning has offered the for finding a few parts to fill it up in last years. It will also be illuminated partly how close or far away of reality these approaches are.

Therefore we first reflect the most successful idea, the temporal difference reward prediction hypothesis of phasic dopamine, a model for the classical conditioning. This is followed by the biologically related hypothesis of tonic dopamine and open fields around the dopamine functionality. At the end further open challenges will be mentioned, which could be described within the Reinforcement Learning framework, but have not been studied well yet.

1 Introduction

Scientists from different backgrounds such as biology and psychology found out a lot of data about the physical aspects of our brain and also the neurons, which it is built of. Even the ways how information is transported both inside the neuron and from one neuron to the next by transmitters are well explored. For different brain areas at least the functionality is suggested. But then it becomes soon really complicated when looking at which neurons do interact and whether they inhibit or excite, so still we have no idea how the brain really works, e.g. as we make decisions.
Since analyzing all neurons, their structure and their influences to each other would take a long time, taking a more abstract view by combining the related field of RL with psychology will be more fruitful and lead to earlier success. Who wants to care about the hardware architecture when you want to know what the program does?
For taking an abstract view, we have to develop a model fitting reality or normative framework. Now the resulting question is: Can we find a model which explains the brain?
What does it mean to explain the brain? As long as the model fits every known data, it may be assumed to be good which means we could trust its predictions for situations where no data has been collected yet. This first leads to see what kind of psychological experiments exist and which learning models have been made up to describe the result.

To answer this, we first have to find out: What do we do with our brain?

The basic of our behaviour is easily described: We want to maximize getting some reward as food, sex or Snickers and we want to minimize costs as illness or injuries. In short, we have goals and want to achieve them. For being successful, we can both learn to predict reward or punishment and select actions. This means we make optimal decisions which is a very hard task because of possible delayed reward or maybe more than one action which leads to it! So a model needs to include some algorithm which explains why some action has been taken and which other action will be taken. And that is exactly what the field of Reinforcement Learning is about.

2 Conditioning and Approached Models

2.1 Pavlovian Conditioning

A rather old behaviour study which treats the simplier case, prediction learning, is Pavlov's famous conditioning experiment:
While dogs usually start to drool if they get their food, most dogs do nothing if they hear a bell ring. The food is an affective, an unconditioned stimulus (US) and the bell ringing can be used as conditioned stimulus (CS), means the stimulus to which the dog shall be conditioned. Now in the experiment every time the bell is ringed when the dog gets its food. After some trials, the dog has combined the bell ringing with the food and starts drooling just after bell ringing, without any food around. According to that, the dog is now conditioned to the bell signal, it has learned to predict the reward "food" after the CS bell ringing.

As the Reinforcement Learning algorithms differ in being model-based or model-free, model-based means the acting agent builds an internal model of its environement which then is used for calculating optimal action selection, while the model-free approach just learns reward based values dependent on the state. As mentioned earlier, we do not know the whole dependencies, so it could be useful to pick the model-free RL approach.

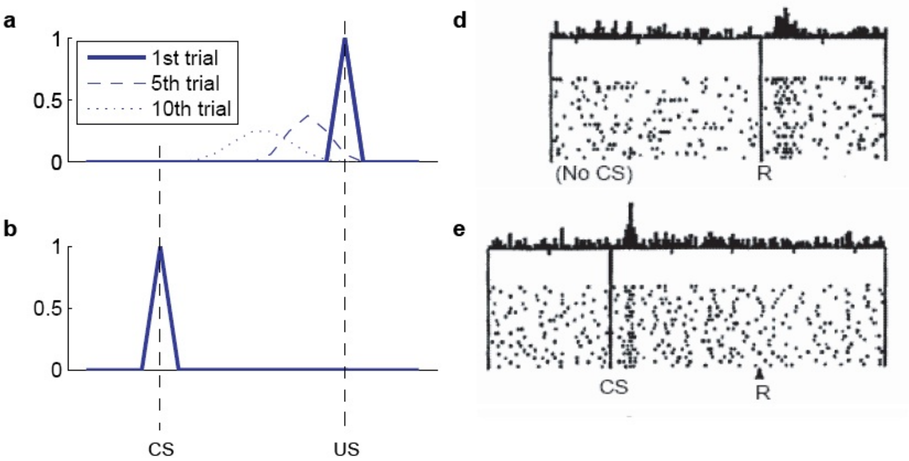
The first published RL model which should describe this prediction learning is the Rescorla-Wagner model of Pavlovian conditioning. It came up in 1972 and contained two (still) basic ideas: Learning occurs only when events violate expectations. And predictions of different stimuli sum up linear to the total prediction. This model could explain several psychology results and also was able to predict phenomena unknown at this time.

Even if it has achieved a huge success, it is not perfect: First, it differentiates between CS and US. This makes modeling of second order condition not possible: When CS A predicts another CS B which itself predicts some reward, the reward predictive value is shifted to A. Also higher order conditioning occurs in every-day life, so here the model is a drift away from reality.
The second limiting shortcome is the handling of a conditioning trial as a discrete temporal object. Thereby continuous events can not be parsed usefully and conditioning to temporal differences between the stimuli in a trial can not be regarded.

To improve this first attempt, Sutton and Barto presented the temporal difference learning rule in 1990, which extends the Rescorla-Wagner model and turned out to be more consistent with neural data. It included two new ideas: the observation of states or situations at a time t S_t and the idea of an estimated value of a state $V(S_t)$. As we do not know how the world works, the actual value is just what we predict as sum of all future rewards.
Hence it is indirectly included what we expect as reward at time t: the difference between $V(S_{t+1})$ and $V(S_t)$, for whom the former can be discounted with $\gamma \leq 1$, we will see more about it later. We can compare the result with the real observed reward r_t by subtracting the former from the latter. Now we have the error of our prediction, δ_t. For updating $V(S_t)$ we add this to the original value, decreased by the learning rate η which can avoid fast changements or to much adapting to extreme values. Iterating over the states can now improve our values.

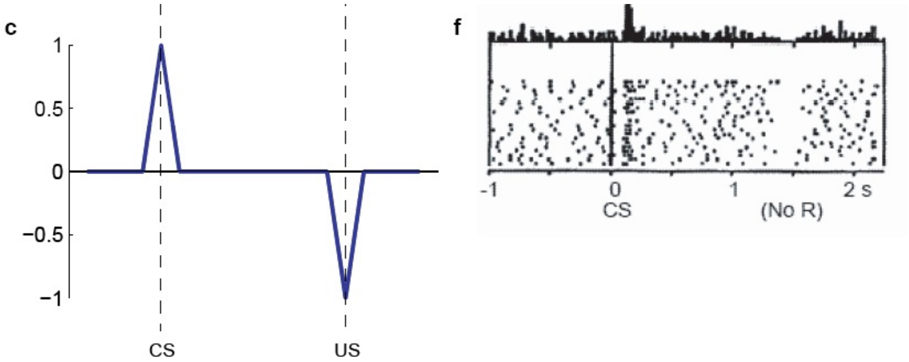
Already in 1978 an hypothesis that viewed dopamine as the brains reward signal was postulated by Wise, recognizing the connections between dopamine and learning. But as regarding the dog sobbering being a rather less quantitative and therefore not really prooving experiment, since the 90s modern medicine allows a deeper view inside an behaving animal's brain. This provides interesting measurements showing how the calculation of the prediction error is performed: Firing patterns of neurons can be recorded today so it is visible exactly which neuron is firing at which time. One of them is dopamine which is supposed to be involved in lots of fields, such as drug addiction, parkinson's disease or reward learning and working memory.

In 1997 Schultz, Dayan and Montague presented recordings of dopamine firing neurons, in which it is possible to recognize that after Pavlovian conditioning there are large peaks (means phasic firing) at the CS, but not at the US. This contradicts the 1978's hypothesis of dopamine being a reward signal (as then there should be firing at the US, too) and led to the temporal difference reward prediction error hypothesis of phasic dopamine, developed by the paper's authors. Apparently the contrasted computed TD-learning simulation by Niv fits quite well and shows that this hypothesis is a computationally precise theory.



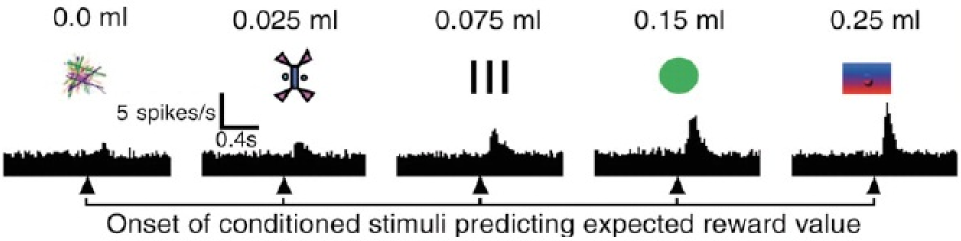
It is suggested that the dopaminergic neuron's afferents deliver information about the reward which is a significant event accordant to the current motivation, and the predictive value of the current state - exatly, what the TD learning algorithms would request. Hence, the nucleus has to calculate the TD error and the cell fires dopamin accordingly.

Further the firing patterns show even what happens if the reward is omitted, which was not observable before:



Here the negative prediction can not have a higher absolute magnitude than the baseline firing.

There are some more results from experiments which underpin this suggested functionality of dopamine, e.g. Tobler et al (2005). The diagram shows the firing rate being proportional to the amount of reward, which the TD error would calculate similarly.



In 2003 Fiorillo et al. showed that a higher probability of reward leads to higher transferring to signal, means there has been more firing at the CS, less at the reward. Also Roesch et al. additioned in 2007 a dopamine firing pattern which shows the interconnection between the dopamine magnitude at the CS and the predicted delay to the reward which we will see later in more detail. Hence, these also fit well in the TD error hypothesis.

2.2 Instrumental Conditioning - Action Selection

All what has been observed and modeled above has been just about prediction. The animals did not choose the action, as their response (e.g. the drooling) was rather a reflex. But only predicting what will happen does not lead to the reward necessarily. Performing an action may be helpful if the action was well chosen. So what about improving the action selection for maximizing the reward? This learning of taking the right action (to the right time or in the right state) is called instrumental learning and is more complex than prediction learning, as some actions may have long time consequences. This makes it difficult to find out which action caused the reward and whether this actually has been the reward.

A well known example for a simple instrumental conditioning experiment is Skinner's box in which a rat has learned to press a lever for earning some pellets. If we are critical, this is quite a bit simplified beacuse it may allow insights in learning of simple actions but not in the complex causing and interacting of different stimuli and possibilities.

But before we go deeper into that, a model for the easier case would also be informative. Of course, it is almost obvious that here the prediction learning could be useful: if we calculate the value of every state, we just have to go to the highest valued state. But here appears the problem: How to find out which action leads to this state? We need the knowledge of the probability for a transition into a state, given the action - what may not be available, as we do not know everything about the world around.

While the TD learning rule above already provides states, we now also need actions in our model. This means, we now have to learn values for e.g. state-action pairs. This value contains the predicted sum of all future rewards following the taken action. An example would be Q-Learning, where the value of the state-action pair is learned. The updating rule looks quite familiar:

$$Q(S_t, a_t)_{new} = Q(S_t, a_t)_{old} + \eta\delta_t$$

For updating the value, an adapted TD error δ_t is used which still compares the current value with the sum of the reward for taking the action plus the discounted best future reward.

There are recent reasearch results (Morris et al., 2006 recordings in non-human primates, and Roesch et al., 2007, reacordings in rats) which suggest that the brain uses state-action values similar to Q-learning or a variation of it, SARSA.

2.3 Measurements In Human Brains

Instead of the original dog, mostly rodents' and monkeys' dopamine patterns have been studied. But what about the human brain, how can the hypothesis above be prooven or at least maintained? It is not possible or rather not allowed to put electrodes in human brains, that is why there are no such direct firing rate patterns. As alternative, functional magnetic resonance imaging (fMRI) can be used for getting some pictures of activity in the living human brain. This method measures the blood oxygen level dependent (BOLD), which is suggested to be high if there is some brain activity. The results have a low signal-to-noise-ratio, which requires averaging of many trials to erase it. Also they are not very precise for finding any prooves: Whether the observed activity means dopamine activity can not be guaranteed. But at least the results seem to maintain the hypothesis and do not contradict it, as the recordings show correlations to the prediction error in the striatum, dopamin's major target.

Newer experiments in 2006 (Pessiglione et al.) have been made by using dopamine manipulators additionally. They showed that dopamine enhancers and blockers seem to influence the BOLD measurement results as well as learning and action selection.

2.4 Completed?

So what about the question, does the Reinforcement Learning explain the brain as mentioned above? At least it is a good abstract description yet. Using that model makes it possible to describe parts, but not everything. Until today, most studies used just simple scenarios, means including only one explicit CS and one reward. Here, the rather abstract Reinforcement Learning rule explains the reactions. Niv himself recommends testing the theory now in more complex tasks varying the standard by using more (conflicting) CS, more than one action needed for earning the reward or more than one possible reward so there can be chosen between.This could draw the made experiments closer to reality with hard decisions between lots of influencing signals or dependencies on actions.

The experiments already dealing with these extensions provided some rather different than conforming results about which algorithms are used for action selection. Since they analyzed different animal species, brain locations and task learning, they may not be comparable or they could have observed different task managing in the brain. So as Niv mentioned, some more studies are required.

A related field, where the current models may not suffice to describe the procedures, is the prediction of aversive events. These are not correlated to dopamine while the absence of an appetitive event, i.e. the reward, is signaled via dopamine. So the loss of a reward and punishment are not treated similarly as it could be supposed to be.

2.5 Vigor

The examined prediction error theory of dopamine describes only the phasic firing of dopamine which can change in milliseconds. As the firing patterns above may have already shown that dopamine neurons do not only fire if there is a calculated prediction error, there is also a tonic mode in background changing rather slowly (assumed in minutes). A second aspect leading to dopamine in the extrasynaptic fluid is that the fired dopamine rests some time and does not disappear directly. So firing can increase the concentration of dopamine.
This ever-present baseline of dopamine, called tonic dopamine level, seems to be involved in another part of action selection: our whole life takes place in continuous time, so our decisions and actions are also continuous, which was not respected in the already presented models. But on that account we do not only choose the correct action, we also choose the vigor or speed with which an action is performed, or at which rate. The vigor of course is also dependent from motivation as an hungry animal will run faster to its meal, means the food has become more valuable.

Some additional thought: even if the omitted background firing of dopamine looks like the TD error and may be calculated respectively, the effect (for learning) of the missing dopamine should need some time until it can be recognized and have some influence. May the learning through the loss of the reward has another velocity as learning through goal achievement.

Niv himself has proposed a model of optimal responding rates in which the vigor coming along with the choice of action is represented by a latency for performing the action. Regarding vigor being costly, the goal is now to obtain the highest possible net rate of rewards minus costs per unit time. For this field being not really well explored, there are only hints about how a differenet dopamine level influences the vigor.
An experiment of Salmone & Correa (2002) showed that killing dopaminergic neurons, which project to the ventral striatum, reduces the rate of instrumental responding rather than showing influence to the learning process. Therefore Niv et al. suggested the tonic levels of dopamine (in the striatum) represent the net rate. Some other facts maintaining this theory are the faster responding under medicaments or drugs such as amphetamine, which increase the dopamine level, as well as the lethargy of people suffering from dopamine depletion, caused by Parkinson.

3 Open Fields and Theoretical Approaches

3.1 Rather Concrete and Correlated to Dopamine

The temporal discounting of far away rewards γ happening is rather certain and shown with several experiments, e.g. by Roesch et al. in 2007:

Calculating the discount is the actual problem: Exponential discounting has attractive theoretical properties, but hyperbolic discounting seems to be more common in real brains. As this can be seen as optimizing long run rewards, it also correlates with free operant choice task. That leads again to the rate of reward which is assumed to be coded by the tonic dopamine level as mentioned

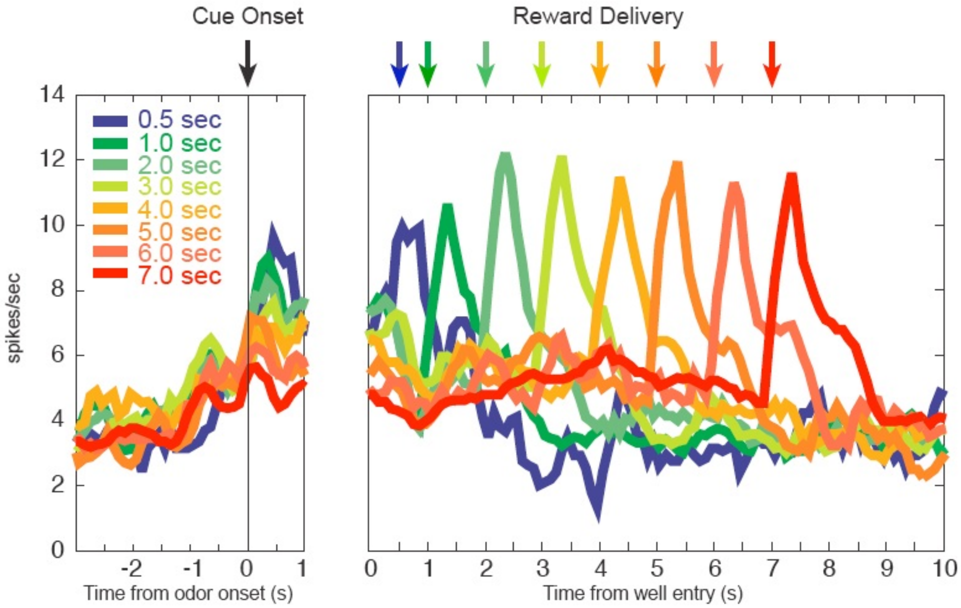


Figure 1: Dopamine firing (for not completely expected reward) in rats shows the discount factor: Rats which are trained to delayed rewards (delivered after 7 seconds) had a lower peak at the CS than rats trained on sooner reward delivery.

above. But experiments with diets changing the serotonin level showed that it seems to be not only dopamine which influences the discounting. Probands made more risky choices under serotonin diet than they did without. So here the interaction will be more complicated than actual models are able to simulate.

Another challenge correlating to prediction is about extinction which means that the predicted reward ceases to appear. Yet there is no satisfactory computational model which describes the behaviour properly. TD models may treat the extinction as unlearning of predictive values, but this lacks in reality because spontaneous recovery is possible indeed. Also, a modelling problem is conditioned inhibition, which means a CS predicts that the reward will not occur, e.g. a flash besides the bell means that even if the bell has ringed, there will be no food. Even if the flash is presented alone, which the TD model would pair with zero reward, in reality the inhibitory value does not extinguish - but in the TD model it does. So no model has been found yet, either.

As already mentioned above, while phasic dopamine firing is suggested to represent appetitive prediction errors, it does not seem to be involved in aversive predictions. A reason could be that the latter appears being subject to complex inputs which can lead to more different reactions than the former. Depending on the distance or the own condition, the choice can hover between withdrawal or freezing and fighting. These reactions are also vigorous, which involves the tonic dopamine level. Furthermore, an aversive prediction can lead to the active avoidance which then can be seen as an achievement of safety, which would be an appetitive prediction. Actually, fMRI showed the same area being activated as when reward was received.

Those challenges above may have indicated already: it is not dopamine alone which is involved in learning. Even more, 2005 and 2007 Berridge showed that despite the absence of dopamine, animals were able to learn correct action selection. This supports the idea of humans having more than one system making decisions.
At the beginning of an action learning experiment, the behaviour of the test animal may be rather flexible and can be changed by task variation or the animals' own motivation, while after some training, the behavior is more inflexible and not easy to influence anymore. Despite over-training is not the only reason leading to these hard-wired action selection, this behaviour is called ”habitual responding”, the former ”goal-directed”. So the systems can be identified by the fact that behaviour is able to have different characteristics.

Why not always choose the best one? The answer is rather easy: the best system depends to the situation as they have different advantages and disadvantages. Daw et al. suggested in 2005 that habitual action selection is based on cached values which have been learned through prediction errors. This sounds familiar, doesn't it? As we already have mentioned, there are model-based and model-free learning methods - and until now we looked at the model-free approach mostly. They just handle with values, which are updated. This means that some difference in our environment will change that value rather slowly and for good values we need a lot of training. But in return it is really fast in picking up the right action. According to that the model-based approach allows updating its model directly and has therefore always the (current) best estimated world model. Of course here the best action has to be calculated each time by using the model which makes the action selection rather slowly and costly. This suits the goal-directed behaviour quite well, which is suggested to calcualte the values by using forward search or forward simulation of consequences of actions on the world model.

Now that we know the options, how to find out which system is the best one for a given situation? As we have seen, "goal-directed" deciding is much more costly than simple "habitual responding". Hence, it should only be used if the model-free system has not had enough experience to be informed. And this is really close to what happens in our brain: We choose the model-based system early in training as mentioned above or for selecting between several actions with several outcomes. For contrary situations the model-free system is preferred, when there would be some long action chain to evaluate in the model or also for the other extreme, simple scenarios which demands just one action evaluation. Of course after some long training the "habitual responding" will be chosen, too.

Here future reasearch is needed doubly. Not only there is no implementation of deciding between the two systems, but also the flexible model-based system is not well explored yet, as may have been clear by just having dealt with the other system.

3.2 More Abstract

Leaving now the comparatively well explored field of dopamine, we will take a glance at some more comprehensive fields.

Most approaches just concentrate on optimizing exploitation, i.e. finding the best combination of values and parameters so the calculation fits best. Those approaches just use past experience for future trials. More ambitious agents will also optimize their exploration, as this updates their past experiences. This behaviour can also be caused by situations of uncertainty or novelty.
In 2002 Kakade and Dayan proposed a 'novelty bonus' for the latter situations: It is also possible to learn that something new is good, as new stimuli may predict some future reward. This leads to more dopamine boosting at new stimuli, that also could be shown in humans by using fMRI. Hence more exploration follows the new stimuli. The former situations appear to influence learning, too: Experiments using fMRI showed that people adapt their learning (and their forgetting) rates to their environment. So higher volatility leads to a higher rate, which is an interesting aspect for our future since we have created a world still accelerating its speed.

A still very unexplored field is hierarchical task planning. We perform it nearly all time since most people do not plan in tiny steps like 'I want to put down my feets on the floor, then step up, take the right foot forward, balance my weight on it, turn the left...'. Instead they would just plan like 'I want to make coffee now!'. Therefore we are able to group small tasks to larger ones, which allows easier forward planning. Also we can divide a task in smaller subtasks or modules, which then have to be performed in a specified order.
We must have learned those modules before - but not imperatively in that content (that's why they are modules). Which brings up a second unexplored field: the ability to generalize. We can abstract an action module from its content and use this solution in another one. This provides large benefits as it limits search space for new situations and also the time we would need for exploration of all similar situations. How to learn and combine modules? How to transfer from one task to another? This still has to be explored.

4 Finally, Is The Reinforcement Learning Framework Successfull?

Although we have now seen that there is still a lot to do, we can not deny that the Reinforcement Learning brought up some great basic insights by its highly simplified description of reality. We even found a part that can roughly be described with a RL model and some others which could be described (by some adapted) models. While most of them still handle just primitive tasks, this is a good motivation for future reasearch which could lead to success, means finding or maybe specializing and combining models.

And even if it failed, the difference between assumed and real functionality can teach a lot about how we learn and what are our preferences and so help to improve learning algorithms - for our brain still having the best known performance. So maybe Reinforcement Learning Models describes some part, the real basics, of learning in brains but we do know that current models do not describe it completely.

5 What Would Interest Me?

While writing this paper, some questions arised in my head and here I just want to present a few of them which were the most well-figured ones.

First, for studies about depression, depressive laboratory animals are required and must be set up. Therefore rats or mice are set in a world of randomness: The rat can 'do what it wants', it will be punished or rewarded randomly. So at least the rat learns that it is not able to influence its fate by acting in a specific way and stops acting. Later, even if it could avoid punishment or save its life, the rat will still not select any action, e.g. if it is thrown into water, it will not start swimming although it can swim. So it interests me if a model describing vigor would also describe this behaviour no matter what the rat does, it can't increase its reward, it only has the costs for trying. So the best policy is doing nothing?

Also, I could not find information about imitation learning in the paper. But for animals and humans imitation learning is an essential part in the whole learning process. I think this can not (only) be modeled by building on the Rescorla-Wagner assumption, that learning only happens if anything violates our expectations. Because we can learn something just by watching someone else performing a task, we are actually not in the state where we would calculate the assumed prediction error but we are able to improve our behaviour anyway! Additionally humans do not even have to see a teacher they can learn by just being told how to do so.

What I also see being far away from reality is the initialisation of algorithms. E.g. Q-values are often set to zero, but I doubt that this counts for humans and animals. I think there is a lot of basic set while we are unable to act or maybe even think: while pregnancy. There are several studies about what influences us in our prenatal phasis (and also lot of unprooved superstition) and of course it is still not well explored how our mother's actions and our connection forms our initialization and maybe even what Niv calls the hard-wired responses or reflexes, which later will be hard to change.

Furthermore, and this is really philosophical and so not very scientific anymore: How does our body structure influences learning and action selection? Often, we are not able to take the best action, because we can't perform it. But in exchange we use a lot of cleverness to find another way reaching the same state (even if this way contains building airplanes or something else). So maybe our body forced us to be better calculaters? Or are we just that weak (compared to other animals) because our brain allowed us to be?

References

[1] Yael Niv: Reinforcement Learning in the Brain.
The Journal of Mathematical Psychology, 2009 / Vol. 53, No. 3, 139-154.

[2] Peter Dayan & Yael Niv: Reinforcement Learning: The Good, The Bad and The Ugly.
Current Opinion in Neurobiology, 2008 / Vol. 18, No. 2, 185-196.

[3] Yael Niv's presentation based on the paper mentioned in [1]
http://www.princeton.edu/ yael/ICMLTutorialNoPrint.pdf