A Definition of Happiness
for Reinforcement Learning Agents[*]

Mayank Daswani[**] and Jan Leike[**]

Australian National University
{mayank.daswani|jan.leike}@anu.edu.au

Abstract. What is happiness for reinforcement learning agents? We seek a formal definition satisfying a list of desiderata. Our proposed definition of happiness is the *temporal difference error*, i.e. the difference between the value of the obtained reward and observation and the agent's expectation of this value. This definition satisfies most of our desiderata and is compatible with empirical research on humans. We state several implications and discuss examples.

Keywords: temporal difference error, reward prediction error, pleasure, well-being, optimism, machine ethics

1 Introduction

People are constantly in search of better ways to be happy. However, philosophers and psychologists have not yet agreed on a notion of human happiness. In this paper, we pursue the more general goal of defining happiness for intelligent agents. We focus on the reinforcement learning (RL) setting [11] because it is an intensively studied formal framework which makes it easier to make precise statements. Moreover, reinforcement learning has been used to model behaviour in both human and non-human animals [7].

Here, we decouple the discussion of happiness from the discussion of consciousness, experience, or qualia. We completely disregard whether happiness is actually consciously experienced or what this means. The problem of consciousness has to be solved separately; but its answer might matter insofar that it could tell us which agents' happiness we should care about.

Desiderata. We can simply ask a human how happy they are. But artificial reinforcement learning agents cannot yet speak. Therefore we use our human "common sense" intuitions about happiness to come up with a definition. We arrive at the following desired properties.

[**] Both authors contributed equally.

- *Scaling.* Happiness should be invariant under scaling of the rewards. Replacing every reward r_t by $cr_t + d$ for some $c, d \in \mathbb{R}$ with $c > 0$ (independent of t) does not change the reinforcement learning problem in any relevant way. Therefore we desire a happiness measure to be independent under rescaling of the rewards.
- *Subjectivity.* Happiness is a subjective property of the agent depending only on information available to the agent. For example, it cannot depend on the true environment.
- *Commensurability.* The happiness of different agents should be comparable. If at some time step an agent A has happiness x, and another agent B has happiness y, then it should be possible to tell whether A is happier than B by computing $x - y$. This could be relaxed by instead asking that A can calculate the happiness of B *according to A's subjective beliefs.*
- *Agreement.* The happiness function should match experimental data about human happiness.

It has to be emphasised that in humans, happiness cannot be equated with pleasure [8]. In the reinforcement learning setting, pleasure corresponds to the reward. Therefore happiness and reward have to be distinguished. We crudely summarise this as follows; for a more detailed discussion see Section 3.

$$\text{pleasure} \;=\; \text{reward} \;\neq\; \text{happiness}$$

The happiness measure that we propose is the following. An agent's happiness in a time step t is the difference between the value of the obtained reward and observation and the agent's expectation of this value at time step t. In the Markov setting, this is also known as the *temporal difference error* (TD error) [10]. However, we do not limit ourselves to the Markov setting in this paper. In parts of the mammalian brain, the neuromodulator dopamine has a strong connection to the TD error [7]. Note that while our definition of happiness is not equal to reward it remains highly correlated to the reward, especially if the expectation of the reward is close to 0.

Our definition of happiness coincides with the definition for *joy* given by Jacobs et al. [6], except that the latter is weighted by 1 minus the (objective) probability of taking the transition which violates subjectivity. Schmidhuber's work on 'intrinsic motivation' adds a related component to the reward in order to motivate the agent to explore in interesting directions [9].

Our definition of happiness can be split into two parts. (1) The difference between the instantaneous reward and its expectation, which we call *payout*, and (2) how the latest observation and reward changes the agent's estimate of future rewards, which we call *good news.* Moreover, we identify two sources of happiness: *luck*, favourable chance outcomes (e.g. rolling a six on a fair die), and *pessimism*, having low expectations of the environment (e.g. expecting a fair die to be biased against you). We show that agents that know the world perfectly have zero expected happiness.

In the rest of the paper, we use our definition as a starting point to investigate the following questions. Is an off-policy agent happier than an on-policy one? Do

monotonically increasing rewards necessarily imply a happy agent? How does value function initialisation affect the happiness of an agent? Can we construct an agent that maximises its own happiness?

2   Reinforcement Learning

In reinforcement learning (RL) an *agent* interacts with an *environment* in cycles: at time step t the agent chooses an *action* $a_t \in \mathcal{A}$ and receives an *observation* $o_t \in \mathcal{O}$ and a real-valued *reward* $r_t \in \mathbb{R}$; the cycle then repeats for time step $t+1$ [11]. The list of interactions $a_1 o_1 r_1 a_2 o_2 r_2 \ldots$ is called a *history*. We use h_t to denote a history of length t, and we use the shorthand notation $h := h_{t-1}$ and $h' := h_{t-1} a_t o_t r_t$. The agent's goal is to choose actions to maximise cumulative rewards. To avoid infinite sums, we use a *discount factor* γ with $0 < \gamma < 1$ and maximise the discounted sum $\sum_{t=1}^{\infty} \gamma^t r_t$. A *policy* is a function π mapping every history to the action taken after seeing this history, and an *environment* μ is a stochastic mapping from histories to observation-reward-tuples.

A policy π together with an environment μ yields a probability distribution over histories. Given a random variable X over histories, we write the π-μ-expectation of X conditional on the history h as $\mathbb{E}_{\mu}^{\pi}[X \mid h]$.

The *(true) value function* V_{μ}^{π} of a policy π in environment μ maps a history h_t to the expected total future reward when interacting with environment μ and taking actions according to the policy π:

$$V_{\mu}^{\pi}(h_t) := \mathbb{E}_{\mu}^{\pi} \left[\sum_{k=t+1}^{\infty} \gamma^{k-t-1} r_k \mid h_t \right]. \tag{1}$$

It is important to emphasise that $\mathbb{E}_{\mu}^{\pi}$ denotes the *objective* expectation that can be calculated only by knowing the environment μ. The *optimal value function* V_{μ}^{*} is defined as the value function of the optimal policy, $V_{\mu}^{*}(h) := \sup_{\pi} V_{\mu}^{\pi}(h)$.

Typically, reinforcement learners do not know the environment and are trying to learn it. We model this by assuming that at every time step the agent has (explicitly or implicitly) an estimate $\hat{V}$ of the value function V_{μ}^{π}. Formally, a *value function estimator* maps a history h to a value function estimate $\hat{V}$. Finally, we define an *agent* to be a policy together with a value function estimator. If the history is clear from context, we refer to the output of the value function estimator as *the agent's estimated value*.

If μ only depends on the last observation and action, μ is called *Markov decision process (MDP)*. In this case, $\mu(o_t r_t \mid h_{t-1} a_t) = \mu(o_t r_t \mid o_{t-1} a_t)$ and the observations are called states ($s_t = o_t$). In MDPs we use the Q-value function, the value of a state-action pair, defined as $Q_{\mu}^{\pi}(s_t, a_t) := \mathbb{E}_{\mu}^{\pi} \left[\sum_{k=t+1}^{\infty} \gamma^{k-t-1} r_k \mid s_t a_t \right]$. Assuming that the environment is an MDP is very common in the RL literature, but here we will not make this assumption.

3   A Formal Definition of Happiness

The goal of a reinforcement learning agent is to maximise rewards, so it seems natural to suppose an agent is happier the more rewards it gets. But this does

not conform to our intuition: sometimes enjoying pleasures just fails to provide happiness, and reversely, enduring suffering does not necessarily entail unhappiness (see Example 3 and Example 7). In fact, it has been shown empirically that rewards and happiness cannot be equated [8] (p-value < 0.0001).

There is also a formal problem with defining happiness in terms of reward: we can add a constant $c \in \mathbb{R}$ to every reward. No matter how the agent-environment interaction plays out, the agent will have received additional cumulative rewards $C := \sum_{i=1}^{t} c$. However, this did not change the structure of the reinforcement learning problem in any way. Actions that were optimal before are still optimal and actions that are slightly suboptimal are still slightly suboptimal *to the same degree*. For the agent, no essential difference between the original reinforcement learning problem and the new problem can be detected: in a sense the two problems are *isomorphic*. If we were to define an agent's happiness as received reward, then an agent's happiness would vary wildly when we add a constant to the reward while the problem stays structurally exactly the same.

We propose the following definition of happiness.

Definition 1 (Happiness). *The* happiness *of a reinforcement learning agent with estimated value $\hat{V}$ at time step t with history ha_t while receiving observation o_t and reward r_t is*

$$\odot(ha_t o_t r_t, \hat{V}) := r_t + \gamma \hat{V}(ha_t o_t r_t) - \hat{V}(h). \tag{2}$$

If $\odot(h', \hat{V})$ is positive, we say the agent is happy*, and if $\odot(h', \hat{V})$ is negative, we say the agent is* unhappy*.*

It is important to emphasise that $\hat{V}$ represents the agent's *subjective* estimate of the value function. If the agent is good at learning, this might converge to something close to the true value function V_μ^π. In an MDP (2) is also known as the *temporal difference error* [10]. This number is used used to update the value function, and thus plays an integral part in learning.

If there exists a probability distribution ρ on histories such that the value function estimate $\hat{V}$ is given by the expected future discounted rewards according to the probability distribution ρ,

$$\hat{V}(h) = \mathbb{E}_\rho^\pi \left[\sum_{k=t+1}^{\infty} \gamma^{k-t-1} r_k \mid h \right], \tag{3}$$

then we call $E := \mathbb{E}_\rho^\pi$ the agent's *subjective expectation*. Note that we can always find such a probability distribution, but this notion only really makes sense for *model-based agents* (agents that learn a model of their environment). Using the agent's subjective expectation, we can rewrite Definition 1 as follows.

Proposition 2 (Happiness as Subjective Expectation). *Let E denote an agent's subjective expectation. Then*

$$\odot(h', \hat{V}) = r_t - E[r_t \mid h] + \gamma \left(\hat{V}(h') - E[\hat{V}(haor) \mid h] \right). \tag{4}$$

Proposition 2 states that happiness is given by the difference of how good the agent thought it was doing and what it learns about how well it actually does. We distinguish the following two components in (4):

- *Payout:* the difference of the obtained reward r_t and the agent's expectation of that reward $E[r_t \mid h]$.
- *Good News:* the change in opinion of the expected future rewards after receiving the new information $o_t r_t$.

$$\odot(h', \hat{V}) = \underbrace{r_t - E[r_t \mid h]}_{\text{payout}} + \gamma \big(\underbrace{\hat{V}(h') - E[\hat{V}(haor) \mid h]}_{\text{good news}} \big)$$

Example 3. Mary is travelling on an air plane. She knows that air planes crash very rarely, and so is completely at ease. Unfortunately she is flying on a budget airline, so she has to pay for her food and drink. A flight attendant comes to her seat and gives her a free beverage. Just as she starts drinking it, the intercom informs everyone that the engines have failed. Mary feels some happiness from the free drink (*payout*), but her expected future reward is much lower than in the state before learning the *bad news*. Thus overall, Mary is unhappy.

For each of the two components, payout and good news, we distinguish the following two sources of happiness.

- *Pessimism:*[1] the agent expects the environment to contain less rewards than it actually does.
- *Luck:* the outcome of r_t is unusually high due to randomness.

$$r_t - E[r_t \mid h] = \underbrace{r_t - \mathbb{E}_\mu^\pi[r_t \mid h]}_{\text{luck}} + \underbrace{\mathbb{E}_\mu^\pi[r_t \mid h] - E[r_t \mid h]}_{\text{pessimism}}$$

$$\hat{V}(h') - E[\hat{V}(haor) \mid h] = \underbrace{\hat{V}(h') - \mathbb{E}_\mu^\pi[\hat{V}(haor) \mid h]}_{\text{luck}}$$
$$+ \underbrace{\mathbb{E}_\mu^\pi[\hat{V}(haor) \mid h] - E[\hat{V}(haor) \mid h]}_{\text{pessimism}}$$

Example 4. Suppose Mary fears flying and expected the plane to crash (*pessimism*). On hearing that the engines failed (*bad luck*), Mary does not experience very much change in her future expected reward. Thus she is happy that she (at least) got a free drink.

The following proposition states that once an agent has learned the environment, its expected happiness is zero. In this case, underestimation cannot contribute to happiness and thus the only source of happiness is luck, which cancels out in expectation.

---

[1] *Optimism* is a standard term in the RL literature to denote the opposite phenomenon. However, this notion is somewhat in discord with optimism in humans.

Proposition 5 (Happiness of Informed Agents). *An agent that knows the world has an expected happiness of zero: for every policy π and every history h,*

$$\mathbb{E}_\mu^\pi[\odot(h', V_\mu^\pi) \mid h] = 0.$$

Analogously, if the environment is deterministic, then luck cannot be a source of happiness. In this case, happiness reduces to how much the agent underestimates the environment. By Proposition 5, having learned a deterministic environment perfectly, the agent's happiness is equal to zero.

4   Matching the Desiderata

Here we discuss in which sense our definition of happiness satisfies the desiderata from Section 1.

Scaling. If we transform the rewards to $r_t' = cr_t + d$ with $c > 0$, $d \in \mathbb{R}$ for each time step t without changing the value function, the value of $\odot$ will be completely different. However, a sensible learning algorithm should be able to adapt to the new reinforcement learning problem with the scaled rewards without too much problem. At that point, the value function gets scaled as well, $V_{new}(h) = cV(h) + d/(1 - \gamma)$. In this case we get

$$\begin{aligned}
\odot(ha_to_tr', V_{new}) &= r_t' + \gamma V_{new}(ha_to_tr_t') - V_{new}(h) \\
&= cr_t + d + \gamma cV(ha_to_tr_t') + \gamma \frac{d}{1 - \gamma} - cV(h) - \frac{d}{1 - \gamma} \\
&= c\big(r_t + \gamma V(ha_to_tr_t') - V(h)\big),
\end{aligned}$$

hence happiness gets scaled by a positive factor and thus its sign remains the same, which would not hold if we defined happiness just in terms of rewards.

Subjectivity. The definition (4) of $\odot$ depends only on the current reward and the agent's current estimation of the value function, both of which are available to the agent.

Commensurability. The scaling property as described above means that the exact value of the happiness is not useful in comparing two agents, but the sign of the total happiness can at least tell us whether a given agent is happy or unhappy. Arguably, failing this desideratum is not surprising; in utility theory the utilities/rewards of different agents are typically not commensurable either.

However, given two agents A and B, A can still calculate the A-subjective happiness of a history experienced by B as $\odot(haor_B, \hat{V}^A)$. This corresponds to the human intuition of "putting yourself in someone else's shoes". If both agents are acting in the same environment, the resulting numbers should be commensurable, since the calculation is done using the same value function. It is entirely possible that A believes B to be happier, i.e. $\odot(haor_B, \hat{V}^A) > \odot(haor_A, \hat{V}^A)$, but also that B believes A to be happier $\odot(haor_A, \hat{V}^B) > \odot(haor_B, \hat{V}^B)$, because they have different expectations of the environment.

Agreement. Rutledge et al. measure subjective well-being on a smartphone-based experiment with 18,420 participants [8]. In the experiment, a subject goes through 30 trials in each of which they can choose between a sure reward and a gamble that is resolved within a short delay. Every two to three trials the subjects are asked to indicate their momentary happiness.

Our model based on Proposition 2 with a very simple learning algorithm and no loss aversion correlates fairly well with reported happiness (mean $r = 0.56$, median $r^2 = 0.41$, median $R^2 = 0.27$) while fitting individual discount factors, comparative to Rutledge et al.'s model (mean $r = 0.60$, median $r^2 = 0.47$, median $R^2 = 0.36$) and a happiness=cumulative reward model (mean $r = 0.59$, median $r^2 = 0.46$, median $R^2 = 0.35$). This analysis is inconclusive, but unsurprisingly so: the expected reward is close to 0 and thus our happiness model correlates well with rewards.

The *hedonic treadmill* [2] refers to the idea that humans return to a *baseline level of happiness* after significant negative or positive events. Studies have looked at lottery winners and accident victims [3], and people dealing with paralysis, marriage, divorce, having children and other life changes [5]. In most cases these studies have observed a return to baseline happiness after some period of time has passed; people learn to make correct reward predictions again. Hence their expected happiness returns to zero (Proposition 5). Our definition unfortunately does not explain why people have different baseline levels of happiness (or hedonic set points), but these may be perhaps explained by biological means (different humans have different levels of neuromodulators, neurotransmitters, hormones, etc.) which may move their baseline happiness. Alternatively, people might simply learn to associate different levels of happiness with "feeling happy" according to their environment.

5   Discussion and Examples

5.1   Off-policy Agents

In reinforcement learning, we are mostly interested in learning the value function of the optimal policy. A common difference between RL algorithms is whether they learn *off-policy* or *on-policy*. An *on-policy* agent evaluates the value of the policy it is currently following. For example, the policy that the agent is made to follow could be an ε-greedy policy, where the agent picks $\arg\max_a Q^\pi(h, a)$ a fraction $(1 - \varepsilon)$ of the time, and a random action otherwise. If ε is decreased to zero over time, then the agent's learned policy tends to the optimal policy in MDPs. Alternatively, an agent can learn *off-policy*, that is it can learn about one policy (say, the optimal one) while following a different *behaviour policy*.

The behaviour policy (π_b) determines how the agent acts while it is learning the optimal policy. Once an off-policy learning agent has learned the optimal value function V_μ^*, then it is not happy if it still acts according to some other (possibly suboptimal) policy.

Proposition 6 (Happiness of Off-Policy Learning). *Let π be some policy and μ be some environment. Then for any history h*

$$\mathbb{E}_\mu^\pi[\Theta(h', V_\mu^*) \mid h] \le 0.$$

Q-learning is an example of an off-policy algorithm in the MDP setting. If Q-learning converges, and the agent is still following the sub-optimal behaviour policy then Proposition 6 tells us that the agent will be unhappy. Moreover, this means that SARSA (an on-policy RL algorithm) will be happier than Q-learning on average and in expectation.

5.2   Increasing and Decreasing Rewards

Intuitively, it seems that if things are constantly getting better, this should increase happiness. However, this is not generally the case: even an agent that obtains monotonically increasing rewards can be unhappy if it thinks that these rewards mean even higher negative rewards in the future.

Example 7. Alice has signed up for a questionable drug trial which examines the effects of a potentially harmful drug. This drug causes temporary pleasure to the user every time it is used, and increased usage results in increased pleasure. However, the drug reduces quality of life in the long term. Alice has been informed of the potential side-effects of the drug. She can be either part of a placebo group or the group given the drug. Every morning Alice is given an injection of an unknown liquid. She finds herself feeling temporary but intense feelings of pleasure. This is evidence that she is in the non-placebo group, and thus has a potentially reduced quality of life in the long term. Even though she experiences pleasure (increasing rewards) it is evidence of very *bad news* and thus she is unhappy.

Analogously, decreasing rewards do not generally imply unhappiness. For example, the pains of hard labour can mean happiness if one expects to harvest the fruits of this labour in the future.

5.3   Value Function Initialisation

Example 8 (Increasing Pessimism Does Not Increase Happiness). Consider the deterministic MDP example in Figure 1. Assume that the agent has an initial value function $\hat{Q}_0(s_0, \alpha) = 0$, $\hat{Q}_0(s_0, \beta) = -\varepsilon$, $\hat{Q}_0(s_1, \alpha) = \varepsilon$ and $\hat{Q}_0(s_1, \beta) = 0$. If no forced exploration is carried out by the agent, it has no incentive to visit s_1. The happiness achieved by such an agent for some time step t is $\Theta(s_0 \alpha s_0 0, \hat{V}_0) = 0$ where $\hat{V}_0(s_0) := \hat{Q}_0(s_0, \alpha) = 0$. However, suppose the agent is (more optimistically) initialised with $\hat{Q}_0(s_0, \alpha) = 0, \hat{Q}_0(s_0, \beta) = \varepsilon$. In this case, the agent would take action β and arrive in state s_1. This transition would have happiness $\Theta(s_0 \beta s_1 - 1, \hat{V}_0) = -1 + \gamma \hat{Q}_0(s_1, \alpha) - \hat{Q}_0(s_0, \beta) = -1 - 0.5\varepsilon$. However, the next transition is $s_1 \alpha s_1 2$ which has happiness $\Theta(s_1 \alpha s_1 2, \hat{V}_0) =$
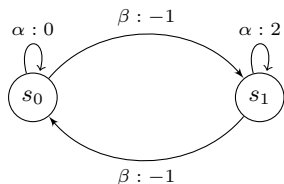
Fig. 1: MDP of Example 8 with transitions labelled with actions α or β and rewards. We use the discount factor $\gamma = 0.5$. The agent starts in s_0. Define $\pi_0(s_0) := \alpha$, then $V^{\pi_0}(s_0) = 0$. The optimal policy is $\pi^*(s_0) = \beta$, so $V^{\pi^*}(s_0) = 1$ and $V^{\pi^*}(s_1) = 4$.
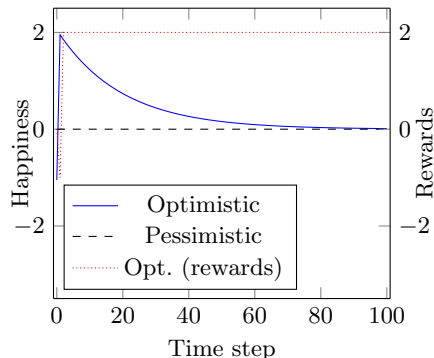
Fig. 2: A plot of happiness for Example 8. We use the learning rate $\alpha = 0.1$. The pessimistic agent has zero happiness (and rewards), whereas the optimistic agent is initially unhappy, but once it transitions to state s_1 becomes happy. The plot also shows the rewards of the optimistic agent.

$2 + \gamma \hat{Q}_0(s_1, \alpha) - \hat{Q}_0(s_1, \alpha) = 2 - 0.5\varepsilon$. If $\hat{Q}_0$ is not updated by some learning mechanism the agent will continue to accrue this positive happiness for all future time steps. If the agent does learn, it will still be some time steps before $\hat{Q}$ converges to Q^* and the positive happiness becomes zero (see Figure 2). It is arguable whether this agent which suffered one time step of unhappiness but potentially many time steps of happiness is overall a happier agent, but it is some evidence that absolute pessimism does not necessarily lead to the happiest agents.

5.4 Maximising Happiness

How can an agent increase their own happiness? The first source of happiness, luck, depends entirely on the outcome of a random event that the agent has no control over. However, the agent could modify its learning algorithm to be systematically pessimistic about the environment. For example, when fixing the value function estimation below $r_{\min}/(1-\gamma)$ for all histories, happiness is positive at every time step. But this agent would not actually take any sensible actions. Just as optimism is commonly used to artificially increase exploration, pessimism discourages exploration which leads to poor performance. As demonstrated in Example 8, a pessimistic agent may be less happy than a more optimistic one.

Additionally, an agent that explicitly tries to maximise its own happiness is no longer a reinforcement learner. So instead of asking how an agent can increase its own happiness, we should fix a reinforcement learning algorithm and ask for the environment that would make this algorithm happy.

6   Conclusion

An artificial superintelligence might contain subroutines that are capable of suffering, a phenomenon that Bostrom calls *mind crime* [1, Ch. 8]. More generally, Tomasik argues that even current reinforcement learning agents could have moral weight [12]. If this is the case, then a general theory of happiness for reinforcement learners is essential; it would enable us to derive ethical standards in the treatment of algorithms. Our theory is very preliminary and should be thought of as a small step in this direction. Many questions are left unanswered, and we hope to see more research on the suffering of AI agents in the future.

References

1. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press (2014)
2. Brickman, P., Campbell, D.T.: Hedonic relativism and planning the good society. Adaptation-Level Theory pp. 287–305 (1971)
3. Brickman, P., Coates, D., Janoff-Bulman, R.: Lottery winners and accident victims: Is happiness relative? Journal of Personality and Social Psychology 36, 917 (1978)
4. Daswani, M., Leike, J.: A definition of happiness for reinforcement learning agents. Technical report, Australian National University (2015), `http://arxiv.org/abs/1505.04497`
5. Diener, E., Lucas, R.E., Scollon, C.N.: Beyond the hedonic treadmill: Revising the adaptation theory of well-being. American Psychologist 61, 305 (2006)
6. Jacobs, E., Broekens, J., Jonker, C.: Joy, distress, hope, and fear in reinforcement learning. Conference on Autonomous Agents and Multiagent Systems pp. 1615–1616 (2014)
7. Niv, Y.: Reinforcement learning in the brain. Journal of Mathematical Psychology 53, 139–154 (2009)
8. Rutledge, R.B., Skandali, N., Dayan, P., , Dolan, R.J.: A computational and neural model of momentary subjective well-being. In: Proceedings of the National Academy of Sciences (2014)
9. Schmidhuber, J.: Formal theory of creativity, fun, and intrinsic motivation (1990–2010). In: IEEE Transactions on Autonomous Mental Development. vol. 2, pp. 230–247 (2010)
10. Sutton, R., Barto, A.: Time-derivative models of Pavlovian reinforcement. In: Learning and Computational Neuroscience: Foundations of Adaptive Networks, pp. 497–537. MIT Press (1990)
11. Sutton, R.S., Barto, A.G.: *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA (1998)
12. Tomasik, B.: Do artificial reinforcement-learning agents matter morally? Technical report, Foundational Research Institute (2014), `http://arxiv.org/abs/1410.8233`