



# City Research Online

## City St George's, University of London

**Citation:** Teichmann, J., Broom, M. & Alonso, E. (2014). The application of temporal difference learning in optimal diet models. *Journal of Theoretical Biology*, 340(7), pp. 11-16. doi: 10.1016/j.jtbi.2013.08.036

This is the unspecified version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

**Permanent repository link:** <https://city-ref-test.eprints-hosting.org/id/eprint/2787/>

**Link to published version:** <https://doi.org/10.1016/j.jtbi.2013.08.036>

**Copyright and Reuse:** Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

# The Application of Temporal Difference Learning in Optimal Diet Models

Jan Teichmann<sup>a,\*</sup>, Mark Broom<sup>a</sup>, Eduardo Alonso<sup>b</sup>

<sup>a</sup>*Department of Mathematical Science, City University London, Northampton Square,  
London EC1V 0HB*

<sup>b</sup>*Department of Computer Science, City University London, Northampton Square, London  
EC1V 0HB*

---

## Abstract

An experience-based aversive learning model of foraging behaviour in uncertain environments is presented. We use Q-learning as a model-free implementation of Temporal Difference learning motivated by growing evidence for neural correlates in natural reinforcement settings. The predator has the choice of including an aposematic prey in its diet or to forage on alternative food sources. We show how the predator's foraging behaviour and energy intake depends on toxicity of the defended prey and the presence of Batesian mimics. We introduce the precondition of exploration of the action space for successful aversion formation and show how it predicts foraging behaviour in the presence of conflicting rewards which is conditionally suboptimal in a fixed environment but allows better adaptation in changing environments.

*Keywords:* optimal diet, Batesian mimicry, predator-prey, taste sampling, Temporal Difference learning

---

## 1. Introduction

Predators have to secure a high energy intake in the face of changing and uncertain environments. Through the evolution of predator-prey interactions manifold mechanisms have emerged to avoid predation. So called secondary defences commonly involve the possession of toxins or deterrent substances which are not directly observable by predators. However, many defended species use conspicuous signals as warning flags in combination with their secondary defences (aposematism).

There is a wide body of theory which addresses the emergence and evolution of aposematism [1, 2, 3, 4, 5, 6]. However, the field of aposematism has a renewed interest in the role of the predator and details of the predator's aversive learning process. In particular, the role of aposematism in memory formation has been widely studied [7, 8, 9, 10, 11]. As the selective agent, aversive learning is an important aspect of predator avoidance. It has been shown that aversion of defended prey is rather a state dependent decision and predators can increase

---

\*Corresponding author

26 their attack rates on defended prey e.g. when particularly hungry [12, 13].  
27 There have been suggestions of an interaction of appetitive learning with aversive  
28 learning to explain the paradox of ingesting toxins in these situations [14].

29 An interesting perspective is to look at the predator and the consequences of  
30 aposematism in combination with aversive learning on the predator’s diet and  
31 energy intake. In particular, the role of mimics in the evolution of aposematism  
32 and their effect on foraging is not very well understood [15, 16, 8, 17]. A predator  
33 may utilise sampling to distinguish between the toxic model and the mimic  
34 [15, 18, 17].

35 The traditional way of analysing and predicting foraging behaviour is the  
36 application of optimal foraging theory (OFT) which maximises the predator’s  
37 net fitness per unit time [19, 20, 21]. However, OFT has well known limitations:  
38 OFT usually fails to correctly predict foraging behaviour on mobile prey in com-  
39 plex environments [21, 22, 23]. It can be argued that OFT was never intended  
40 for predictions in the case of mobile prey and that the optimisation per unit  
41 time omits the uncertainty of more complex environments. There are models  
42 which address optimal foraging under the constraints of risk and uncertainty  
43 and previously extended OFT with learning [24]. The two main approaches  
44 to optimal behaviour in dynamic decision making are dynamic programming  
45 (DP) and stochastic optimal control methods (e.g. Bayesian decision theory)  
46 [25, 26, 24, 27, 28]. Especially dynamic programming found wider application  
47 in behavioural ecology and has been used in models of dynamic decision mak-  
48 ing to identify optimal behaviour numerically [29]. These models have all in  
49 common that they are *model based*: they depend on a representation of the  
50 environment in the form of a model developed from expert knowledge and the  
51 learning objective is to find the parameters which optimise the representational  
52 model.

53 Contrary, a normative framework of rational decision making in a chang-  
54 ing and complex environment is reinforcement learning (RL). RL combines the  
55 computational task of maximising rewards and the algorithmic implementation  
56 of natural learning without an explicit supervisory control signal.

57 Neural correlates of behaving animals show that reinforcement signals in  
58 the brain represent the reward prediction error rather than a direct reward-  
59 reinforcement relation. Temporal difference (TD) learning reflects these insights  
60 by representing states and actions in terms of predictions about future rewards  
61 [30, 31]. Additionally, TD learning is *model-free*: the environment is repre-  
62 sented by moving targets rather than by a model and the learning objective  
63 is to iteratively update the targets towards its true values based on experience  
64 from interactions with the environment. TD learning has been widely used in  
65 artificial systems to choose appropriate actions in complex non-stationary envi-  
66 ronments. Furthermore, the computational theories are increasingly supported  
67 by experimental data describing the activity of dopaminergic neurons, medi-  
68 ate reward-processing and reward-dependent learning [32, 33, 34, 35]. In the  
69 greater picture of learning algorithms, TD learning resides between dynamic  
70 programming and Monte Carlo methods [36].

71 We will apply a TD learning algorithm in our model to gain insights on

72 how aversive learning influences foraging in uncertain environments and discuss  
73 similarities and differences to the optimisation approach of traditional OFT.  
74 In particular, we will compare TD learning with methodology from McNamara  
75 and Sherratt, and we will conclude that TD learning is a new approach to OFT  
76 which is better suited for modelling foraging in dynamic environments with  
77 learning.

## 78 2. Methodology

79 In our model the predator interacts with its environment to find an optimal  
80 foraging strategy to optimise its rewards. The predator’s environment offers a  
81 stable background of alternative food sources. Additionally, the predator has  
82 the choice to include a conspicuous looking type of prey into its diet. However,  
83 the conspicuous prey population may consist of an aposematic model species  
84 and a Batesian mimic species. We assume the environment to be uncertain  
85 with non-stationary parameters over a predator’s lifespan.

### 86 2.1. Temporal Difference learning

87 The predator is not able to distinguish models and mimics based on their ap-  
88 pearance and utilises experience to learn the optimal foraging behaviour. Based  
89 on the growing understanding of learning at the computational and neural level  
90 we use Temporal Difference (TD) learning to implement the predator’s aversive  
91 learning: in particular, we use Q-learning [37]. The learning process consists of  
92 a reward prediction termed the *action-value function* (1) of taking action  $a$  in  
93 state  $s$  at iteration  $k$ ,

$$Q(s, a) = E\{R_k \mid s_k = s, a_k = a\} . \quad (1)$$

94 The condition for the action-value function and Q-learning is for the Markov  
95 property to hold (2),

$$P\{s_{k+1} = s', r_{k+1} = r \mid s_k, a_k\} . \quad (2)$$

96 The reinforcement signal consists of the TD error of the reward prediction  
97 based on experienced rewards following an undertaken action  $a$ . Finally, the Q-  
98 learning update rule is utilised in order to minimise the prediction error [38, 36].

99 Each action taken has a state dependent subsequent reward signal termed  
100  $r_{k+1}$ . The predator not only takes immediate rewards into account but also  
101 the sum of discounted future rewards (3) with  $K$  being the end of an episode  
102 and  $\gamma$  being the discount factor. This combines an ubiquitous interest into

103 rewards with the uncertainty of future events, as follows:

$$\begin{aligned}
R_k &= \sum_{i=0}^K \gamma^i r_{k+i+1} \\
&= r_{k+1} + \sum_{i=1}^K \gamma^i r_{k+i+1} \\
&= r_{k+1} + \gamma \sum_{i=0}^T \gamma^i r_{k+i+2} \\
&= r_{k+1} + \gamma R_{k+1} \quad .
\end{aligned} \tag{3}$$

104 The predator uses the experienced immediate reward  $r_{k+1}$  to minimise the  
105 prediction error by updating its state dependent action-value function using the  
106 *Q-learning* method. The algorithmic representation of the Q-learning update  
107 process is presented in (4) with  $\alpha$  being the learning rate following the derivation  
108 in (3), as follows:

$$Q'(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \underbrace{\left( \overbrace{r_{k+1} + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1})}^{\text{target}} - Q(s_k, a_k) \right)}_{\text{TD error}} \quad . \tag{4}$$

109 Q-learning is an iterative algorithm which uses the immediate experienced re-  
110 ward to form a target with  $Q'$  being the new estimate for  $Q$ . Thereby, Q-learning  
111 bases its update partially on a prevailing estimate  $Q(s_{k+1}, a_{k+1})$  which is known  
112 as bootstrapping. Q-learning is widely used to model Markov decision problems  
113 and under certain conditions, Q-learning has been proved to converge to opti-  
114 mality [39]. For a more detailed introduction of the Q-learning algorithm we  
115 refer to the supplementary material in AppendixA.

116 Finally, the predator uses the Gibbs soft-max policy which is the probability  
117 of taking action  $a$  in state  $s$  under stochastic policy  $\pi$  to translate its action-value  
118 predictions into foraging behaviour (5),

$$\begin{aligned}
\pi(s, a) &= P\{a_k = a \mid s_k = s\} \\
&= \frac{\exp(Q(s, a))}{\sum_a \exp(Q(s, a))} \quad .
\end{aligned} \tag{5}$$

## 119 2.2. The predator's interaction with conspicuous prey

120 We term the action of falling back on the alternative background food sources  
121 as  $a = 0$  and the action of attacking conspicuous prey as  $a = 1$ .

122 We assume the population of conspicuous prey consists of a fraction  $p$  of  
123 Batesian mimics and a fraction  $1 - p$  of defended models. The reward signal  
124 for the alternative stable background food source is  $r_{k+1} = \{1 \mid a = 0\}$ . The  
125 reward signal for ingesting a mimic individual is  $r_{k+1} = \{2 \mid a = 1, i = \text{mimic}\}$

and  $r_{k+1} = \{1 - t^2 | a = 1, i = \text{model}\}$  for ingesting a model individual with toxicity  $t$ . These reward signals do not have to represent necessarily fitness related entities and in our model we simply assume mimics to be rewarding [22].

We consider two different cases (Figure 1):

1. The predator has the ability to use taste-sampling to distinguish models from mimics assuming that the model's toxicity  $t$  operates as a clue to the predator. This foraging strategy is also called *go-slow behaviour* [40]. The probability of rejecting a model based on taste-sampling is given as follows:

$$d(t) = 1 - \frac{1}{1 + d_0 * t} . \quad (6)$$

2. The predator has no ability to distinguish mimics and models and the encounter is solely frequency dependent i.e.  $d_0 = 0$  in equation (6).

### 3. Results

In the case of the predator being unable to distinguish models from mimics ( $d_0 = 0$ ) the average reward signal is solely frequency dependent and given as

$$R = \begin{cases} 1 & \text{if } a = 0 \\ 2p + (1 - t^2)(1 - p) & \text{if } a = 1 \end{cases} . \quad (7)$$

If the predator utilises taste-sampling it can distinguish models from mimics based on the model's toxicity and will not ingest the toxic model with probability  $d(t)$  given in (6). After the predator rejects a conspicuous prey individual it will stay in the locality and forage for another conspicuous prey individual. The average reward signal incorporating taste sampling derives from the geometric series and is given as follows:

$$R = \begin{cases} 1 & \text{if } a = 0 \\ 2p \frac{1}{1 - (1-p)d(t)} + (1 - t^2)(1 - p) \frac{(1-d(t))}{1 - (1-p)d(t)} & \text{if } a = 1 \end{cases} . \quad (8)$$

To obtain the optimal diet we find the correct, discounted action-value function by solving the TD learning problem

$$0 = R + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k) . \quad (9)$$

Figures 2 and 3 show the probability of an experienced predator attacking conspicuous prey based on the frequency of mimics ( $p$ ) and the model's toxicity ( $t$ ). We define aversiveness as  $\pi(a = 1) < 0.5$  with the threshold toxicity ( $t^*$ ) given in (10) for which conspicuous prey becomes aversive and  $R(a = 0, t^*) = R(a = 1, t^*)$  holds, as follows:

$$t^* = \begin{cases} \sqrt{-\frac{p}{p-1}} & \text{if } s_0 = 0 \\ -\frac{\sqrt{p^2 d_0^2 - 4p^2 + 4p + p d_0}}{2p-2} & \text{otherwise} \end{cases} . \quad (10)$$

154 We see that taste-sampling lowers the aversiveness of defended conspicuous prey  
155 when mimics are present.

156 Figures 4 and 5 show the average reward ( $R$ ) of an experienced predator.  
157 Mimics increase the average reward of the predator through increased foraging  
158 on non-averse conspicuous prey. Conversely, increasing toxicity of the models  
159 reduces the average reward for the predator until the increasing toxicity intake  
160 from mistakenly ingested models becomes aversive.

#### 161 4. Discussion

162 We apply Q-learning to the problem of optimal foraging behaviour of an  
163 experienced predator in an uncertain environment. Our motivation lays in the  
164 recognised importance of aversive learning in aposematism and the difficulties  
165 of the classical OFT approach to predict foraging behaviour on mobile prey [21].  
166 In the case of mobile prey additional factors of prey handling and uncertainty  
167 need to be considered, making the OFT model increasingly complex [17]. In-  
168 stead, reinforcement learning offers a normative framework of rational decision  
169 making in a changing and complex environment with growing evidence of neural  
170 correlates.

171 The TD learning based approach puts the emphasis on experience including  
172 discounted future rewards and requires exploration of the action space. This  
173 is fundamentally different to the OFT models of net fitness maximisation per  
174 unit time. It has been long argued that a learning animal cannot be foraging  
175 optimally and vice versa [41].

176 We hypothesise that a non-stationary environment introduces great uncer-  
177 tainty on the prey-population's parameters  $t$  and  $p$  which selects for learning in  
178 evolving predators to adapt quicker to their changing environment. Evidence  
179 for this claim has to come from an evolutionary model and is subject to future  
180 work. To coincide widely with the original OFT methodology, we assume that  
181 the learning process is sufficiently faster than the frequency of change of the  
182 environment to concentrate solely on the experienced predator and to exclude  
183 the iterative learning phase. Furthermore, we assume that the conspicuous prey  
184 inhabit a distinct locality. These assumptions allow us to solve the TD learn-  
185 ing problem directly (9) and we present the policy a predator adopts through  
186 Q-learning.

187 In the context of previous foraging models which incorporated learning, our  
188 learning methodology is model-free. Relevant models, among others, are from  
189 McNamara et al. [24] and Sherratt [13]. McNamara's learning rule describes a  
190 Monte Carlo method using past events to learn the maximum possible long-term  
191 rate as defined by the marginal value theorem [42]. It uses discounted experience  
192 from past interactions with the environment to optimize a current parameter  
193 estimation. The corresponding concept in TD learning is termed *eligibility trace*  
194 and is bridging TD learning with Monte Carlo methods. Eligibility traces can  
195 make TD learning more efficient but as we exclude the iterative learning phase  
196 it has no application in our model. Nevertheless, TD learning is conceptually

different as it's learning objective is based on bootstrapping future rewards rather than optimising the current estimate of a parameter from past events.

Sherratt's model [13] uses Bayesian learning based on dynamic programming. The learning objective is to infer the Bayesian posterior mean estimate of the fraction of defended prey in an unknown population from past experience. The model uses Beta distributions in the Bayesian inference to represent an assumed underlying binomial distribution of defence in a group of prey. The main assumption for the application of dynamic programming is the existence of a finite time horizon where the predator ceases attacking completely. Sherratt's model provides an optimal sampling strategy for novel prey populations with constant values for cost and benefit of an attack. However, the model can't provide optimal foraging policies in changing populations or when defence is not just binomial distributed.

We conclude that TD learning is a new approach to optimal foraging in dynamic environments where cost-benefit values of attacking prey do not necessarily follow simple distributions. TD learning uses a model free objective which makes it an ideal method for learning in complex and dynamic environments where parameters are subject to constant change.

Our model confirms expected results such as that mimics in general lower the aversiveness of the conspicuous prey population and undermine aposematism. Nevertheless, highly toxic models can sustain aversion even for high frequencies of mimics especially in predators not utilising taste sampling. However, it requires exploration for a predator to gain insights about its environment and to form aversive memory. Therefore, even an aversive prey population experiences some level of predation.

Our model predicts that a taste-sampling predator increases its attack rate on mixed conspicuous prey populations in the case of moderately defended models and rewarding mimics. The taste-sampling predator gains increased rewards from moderately defended models as it allows for better discrimination of models and mimics. This is a contrary finding to [17] in which mimics benefit from moderately defended models. This difference is founded on the representation of toxins as recovery time in the OFT maximisation approach and the missing occasional ingestion of models to maintain aversion for highly toxic models.

An interesting paradox is the foraging behaviour on aversive prey which reduces the reward for the predator further before recovering through increasingly falling back on alternative background food sources. (The adopted attack policy for certain parameters results in an average reward  $R$  which lays in the shaded area in Figures 4 and 5, and is suboptimal.) This is a result of the conflicting reward signals of mimics and models and the necessity of exploration of the action space in the face of uncertainty for successful aversion formation. Additionally, an increasing frequency of mimics slows the switching to alternative food sources through further extended uncertainty. Similar results have been observed in counter conditioning and operant conflict situations [43, 44, 45, 46]. Our model predicts a fixed amount of average toxicity which a predator tolerates motivated either by the higher reward signal of ingested mimics or as a consequence of uncertainty. This foraging behaviour on aversive prey for a spe-

243 cific parameter space is conditionally suboptimal in a stationary environment  
244 (even if only during an individuals lifetime) but we note that a) it reflects what  
245 real animals do, and b) it is a good policy precisely because environments are  
246 inherently uncertain.

247 Summarising, our main conclusions are as follows:

- 248 • TD learning is a suitable approach to optimal foraging in changing envi-  
249 ronments.
- 250 • Even aversive prey experience some level of predation as part of the preda-  
251 tor's aversive memory formation.
- 252 • Taste-sampling lowers the effective aversiveness of conspicuous prey if  
253 mimics are present.
- 254 • Intermediate toxicity of aposematic models increases the predator's for-  
255 aging on conspicuous prey through increased discrimination from taste-  
256 sampling and higher average rewards when mimics are rewarding.
- 257 • The conflicting reward signals from mimics and models cause uncertainty  
258 and conditionally suboptimal foraging behaviour on aversive prey.
- 259 • The uncertainty is linked to a fixed amount of average toxicity intake  
260 which predators tolerate in order to forage on rewarding mimics before  
261 switching to mediocre background food sources.
- 262 • Taste-sampling extends the range of parameters where suboptimal foraging  
263 occurs.

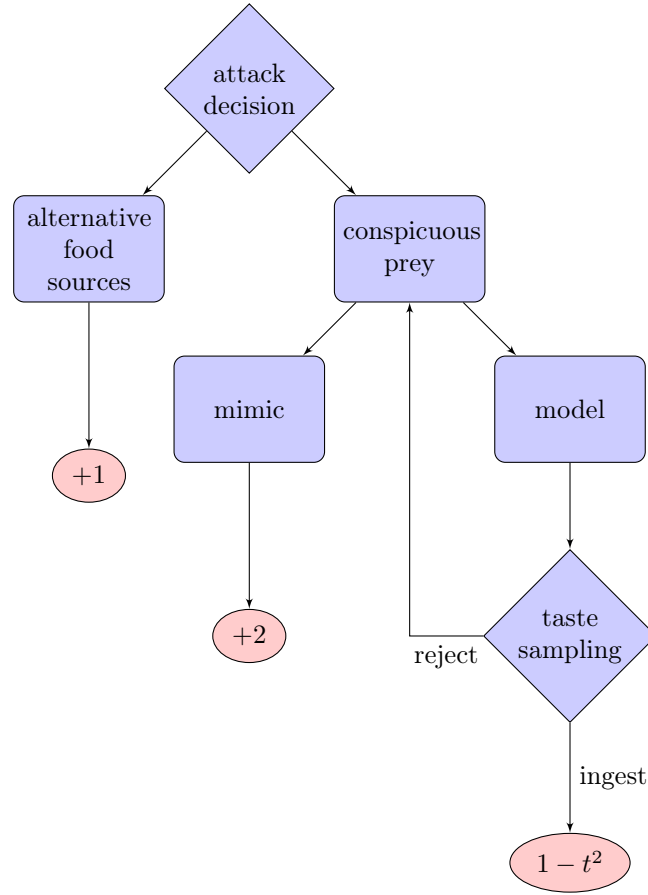


Figure 1: The predator's interaction with its environment and possible reward signals. The predator has the ability to recognise toxic models by taste-sampling.  $t$  stands for the toxicity of defended models.

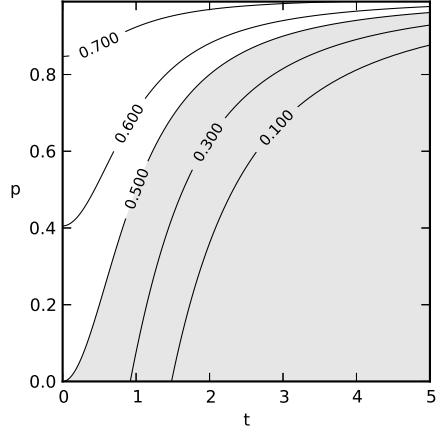


Figure 2: Predator attack probability ( $\pi$ ) of conspicuous prey without taste-sampling ( $d_0 = 0$ ) and discount rate  $\gamma = 0.5$  following soft-max policy (5).  $t$  stands for the toxicity of models and  $p$  for the fraction of mimics. The shaded area indicates aversive toxicity.

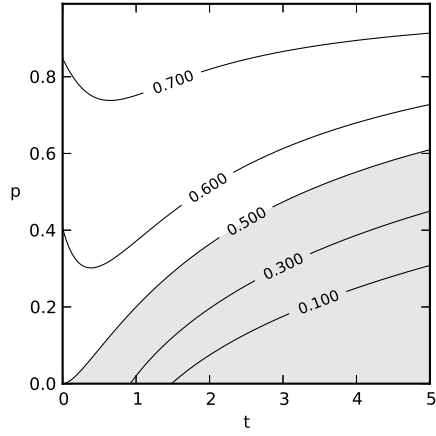


Figure 3: Predator attack probability ( $\pi$ ) of conspicuous prey utilising taste-sampling ( $d_0 = 3$ ) (6) and discount rate  $\gamma = 0.5$  following Gibbs soft-max policy (5).  $t$  stands for the toxicity of models and  $p$  for the fraction of mimics. The shaded area indicates aversive toxicity.

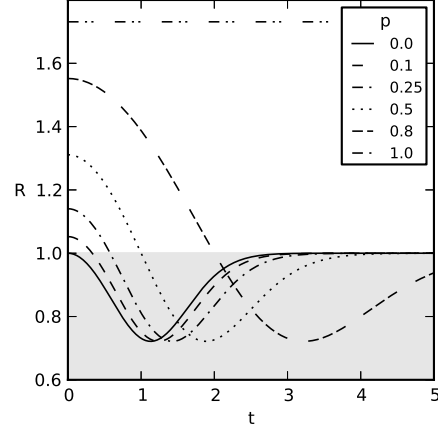


Figure 4: The predator’s average reward ( $R$ ) from interacting with its environment without taste-sampling ( $d_0 = 0$ ) and discount rate  $\gamma = 0.5$ .  $t$  stands for the toxicity of models and  $p$  for representative fractions of mimics. The shaded area indicates suboptimal rewards due to foraging on aversive prey.

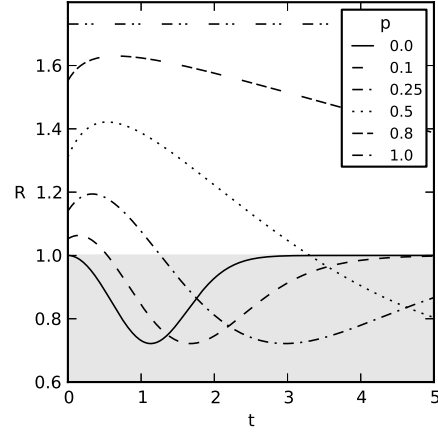


Figure 5: The predator’s average reward from interacting with its environment utilising taste-sampling ( $d_0 = 3$ ) and discount rate  $\gamma = 0.5$ .  $t$  stands for the toxicity of models and  $p$  for representative fractions of mimics. The shaded area indicates suboptimal rewards due to foraging on aversive prey.

- 264 [1] G. Ruxton, T. Sherratt, and M. Speed, *Avoiding Attack: The Evolutionary*  
265 *Ecology of Crypsis, Warning Signals and Mimicry*. Oxford University Press,  
266 2004.
- 267 [2] S. Yachi and M. Higashi, “The evolution of warning signals,” *Nature*,  
268 vol. 394, no. 6696, pp. 882–884, 1998.
- 269 [3] M. Broom, M. Speed, and G. Ruxton, “Evolutionarily stable defence and  
270 signalling of that defence,” *Journal of Theoretical Biology*, vol. 242, pp. 32–  
271 34, 2006.
- 272 [4] O. Leimar, M. Enquist, and B. Sillen-Tullberg, “Evolutionary stability of  
273 aposematic coloration and prey unprofitability: A theoretical analysis,”  
274 *American Society of Naturalists*, vol. 128, pp. 469–490, 1986.
- 275 [5] T. J. Lee, M. P. Speed, and P. A. Stephens, “Honest signaling and the uses  
276 of prey coloration,” *American Society of Naturalists*, vol. 178, pp. E1–E9,  
277 2011.
- 278 [6] N. M. Marples, D. J. Kelly, and R. J. Thomas, “Perspective: The evolution  
279 of warning coloration is not paradoxical,” *Evolution*, vol. 59, no. 5, pp. 933–  
280 940, 2005.
- 281 [7] M. P. Speed, “Warning signals, receiver psychology and predator memory,”  
282 *Animal Behaviour*, vol. 60, no. 3, pp. 269 – 278, 2000.
- 283 [8] K. Svádová, A. Exnerová, P. Štys, E. Landová, J. Valenta, A. Fučíková,  
284 and R. Socha, “Role of different colours of aposematic insects in learning,  
285 memory and generalization of naïve bird predators,” *Animal Behaviour*,  
286 vol. 77, no. 2, pp. 327 – 336, 2009.
- 287 [9] J. Skelhorn and C. Rowe, “Prey palatability influences predator learning  
288 and memory,” *Animal Behaviour*, vol. 71, no. 5, pp. 1111 – 1118, 2006.
- 289 [10] A. N. Johnston and T. H. Burne, “Aposematic colouration enhances mem-  
290 ory formation in domestic chicks trained in a weak passive avoidance learn-  
291 ing paradigm,” *Brain Research Bulletin*, vol. 76, no. 3, pp. 313 – 316, 2008.
- 292 [11] M. Speed and G. Ruxton, “Aposematism: what should our starting point  
293 be?,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 272,  
294 no. 1561, pp. 431–438, 2005.
- 295 [12] C. Barnett, M. Bateson, and C. Rowe, “State-dependent decision making:  
296 educated predators strategically trade off the costs and benefits of con-  
297 suming aposematic prey,” *Behavioral Ecology*, vol. 18, no. 4, pp. 645–651,  
298 2007.
- 299 [13] T. N. Sherratt, “State-dependent risk-taking by predators in systems with  
300 defended prey,” *Oikos*, vol. 103, no. 1, pp. 93–100, 2003.

- [14] E. Hagen, R. Sullivan, R. Schmidt, G. Morris, R. Kempter, and P. Hammerstein, "Ecology and neurobiology of toxin avoidance and the paradox of drug reward," *Neuroscience*, vol. 160, no. 1, pp. 69 – 84, 2009.
- [15] G. Gamberale-Stille and B. S. Tullberg, "Fruit or aposematic insect? context-dependent colour preferences in domestic chicks," *Proceedings of the Royal Society B: Biological Sciences*, vol. 268, pp. 2525–2529, 2001.
- [16] S. Lev-Yadun and K. Gould, "What do red and yellow autumn leaves signal?," *Botanical Review*, vol. 73, no. 4, pp. 279–289, 2007. cited By (since 1996) 30.
- [17] Ø. H. Holen, "Disentangling taste and toxicity in aposematic prey," *Proceedings of the Royal Society B: Biological Sciences*, vol. 280, p. 20122588, 2013.
- [18] C. R. Darst, "Predator learning, experimental psychology and novel predictions for mimicry dynamics," *Animal Behaviour*, vol. 71, no. 4, pp. 743 – 748, 2006.
- [19] R. H. MacArthur and E. R. Pianka, "On optimal use of a patchy environment," *American Naturalist*, vol. 100, pp. 603–609, 1966.
- [20] D. W. Stephens and J. R. Krebs, *Foraging theory*. Princeton University Press, 1987.
- [21] A. Sih and B. Christensen, "Optimal diet theory: when does it work, and when and why does it fail?," *Animal Behaviour*, vol. 61, no. 2, pp. 379 – 390, 2001.
- [22] G. H. Pyke, "Optimal foraging theory: a critical review," *Annual review of ecology and systematics*, vol. 15, pp. 523–575, 1984.
- [23] G. Perry and E. R. Pianka, "Animal foraging: past, present and future," *Trends in Ecology & Evolution*, vol. 12, no. 9, pp. 360–364, 1997.
- [24] J. M. McNamara and A. I. Houston, "Optimal foraging and learning," *Journal of Theoretical Biology*, vol. 117, no. 2, pp. 231 – 249, 1985.
- [25] A. I. Houston and J. McNamara, "A sequential approach to risk-taking," *Animal Behaviour*, vol. 30, pp. 1260 – 1261, 1982.
- [26] D. W. Stephens and E. L. Charnov, "Optimal foraging: some simple stochastic models," *Behavioral Ecology and Sociobiology*, vol. 10, no. 4, pp. 251–263, 1982.
- [27] M. Mangel and C. W. Clark, "Towards a unified foraging theory," *Ecology*, vol. 67, pp. 1127 – 1138, 1986.

- 336 [28] J. M. McNamara, R. F. Green, and O. Olsson, "Bayes' theorem and its  
337 applications in animal behaviour," *Oikos*, vol. 112, no. 2, pp. 243–251,  
338 2006.
- 339 [29] C. W. Clark and M. Mangel, *Dynamic State Variable Models in Ecology:  
340 Methods and Applications: Methods and Applications*. Oxford University  
341 Press, USA, 2000.
- 342 [30] Y. Niv, "Reinforcement learning in the brain," *Journal of Mathematical  
343 Psychology*, vol. 53, no. 3, pp. 139–154, 2009.
- 344 [31] G. S. Berns, S. M. McClure, G. Pagnoni, and P. R. Montague, "Predictabil-  
345 ity modulates human brain response to reward," *The Journal of Neuro-  
346 science*, vol. 21, no. 8, pp. 2793–2798, 2001.
- 347 [32] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of predic-  
348 tion and reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, 1997.
- 349 [33] P. R. Montague, S. E. Hyman, and J. D. Cohen, "Computational roles for  
350 dopamine in behavioural control," *Nature*, vol. 431, no. 7010, pp. 760–767,  
351 2004.
- 352 [34] N. Daw, K. Doya, *et al.*, "The computational neurobiology of learning and  
353 reward," *Current opinion in neurobiology*, vol. 16, no. 2, pp. 199–204, 2006.
- 354 [35] P. Dayan and Y. Niv, "Reinforcement learning: the good, the bad and the  
355 ugly," *Current opinion in neurobiology*, vol. 18, no. 2, pp. 185–196, 2008.
- 356 [36] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*.  
357 Cambridge Univ Press, 1998.
- 358 [37] C. Watkins, *Learning from delayed rewards*. PhD thesis, King's College,  
359 Cambridge, 1989.
- 360 [38] A. G. Barto, R. S. Sutton, and C. W. Anderson, "Neuronlike adaptive  
361 elements that can solve difficult learning control problems," *IEEE Trans-  
362 actions on systems, man, and cybernetics*, vol. 13, no. 5, pp. 834–846, 1983.
- 363 [39] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3,  
364 pp. 279–292, 1992.
- 365 [40] T. Guilford, "'go-slow" signalling and the problem of automimicry," *Journal  
366 of theoretical biology*, vol. 170, no. 3, pp. 311–316, 1994.
- 367 [41] J. Ollason, "Learning to forage - optimally?," *Theoretical Population Biol-  
368 ogy*, vol. 18, no. 1, pp. 44 – 56, 1980.
- 369 [42] E. L. Charnov, "Optimal foraging, the marginal value theorem," *Theoretical  
370 population biology*, vol. 9, no. 2, pp. 129–136, 1976.

- 371 [43] D. R. Williams and H. Barry, “Counter conditioning in an operant conflict  
372 situation.,” *Journal of comparative and physiological psychology*, vol. 61,  
373 no. 1, p. 154, 1966.
- 374 [44] A. P. Blaisdell, J. C. Denniston, H. I. Savastano, and R. R. Miller, “Counter-  
375 conditioning of an overshadowed cue attenuates overshadowing.,” *Journal*  
376 *of Experimental Psychology: Animal Behavior Processes; Journal of Ex-*  
377 *perimental Psychology: Animal Behavior Processes*, vol. 26, no. 1, p. 74,  
378 2000.
- 379 [45] J. Mazur and T. Ratti, “Choice behavior in transition: Development of pref-  
380 erence in a free-operant procedure,” *Animal Learning & Behavior*, vol. 19,  
381 pp. 241–248, 1991.
- 382 [46] T. Matsushima, A. Kawamori, and T. Bem-Sojka, “Neuro-economics in  
383 chicks: Foraging choices based on amount, delay and cost,” *Brain Research*  
384 *Bulletin*, vol. 76, no. 3, pp. 245 – 252, 2008.

## 385 AppendixA. Q-learning algorithm

386 Q-learning is a simple algorithmic implementation of reinforcement learning.  
 387 Particularly, it is a model free method which allows to learn about Markovian  
 388 environments from experienced rewards without the necessity of building rep-  
 389 resentations of the environment. Instead, the algorithm uses moving target  
 390 values.

391 The predator learns from iterative interactions with its environment. We  
 392 term the current iteration subscript  $k$ . At each iteration  $k$  the predator finds  
 393 itself in state  $s_k$  of its environment, accordingly,  $s_k$  is the encounter with a  
 394 particular type of prey in our model. The actual learning process targets the  
 395 predator's reward prediction following action  $a_k$  (respectively, attacking conspic-  
 396 uous or alternative prey) in state  $s_k$  termed the action-value function  $Q(s_k, a_k)$ .  
 397 This action-value function is an approximation of the actual function  $Q^*(s, a)$ .  
 398 Consequently, the aim of the learning process is to find  $Q(s_k, a_k) \approx Q^*(s, a)$ .  
 399 The predator is basing its decision process on  $Q(s_k, a_k)$  following a decision  
 400 policy  $\pi(s_k, Q(s_k, a_k))$ , effectively knowing all of the current  $Q$  values gives the  
 401 probability that we choose to attack or not for the next encounter. This involves  
 402 an iterative update process which is typically formulated in an algorithmic rep-  
 403 resentation because of its origin in computing, as follows:

$$Q'(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha \underbrace{\left( \overbrace{r_{k+1} + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1})}^{\text{target}} - Q(s_k, a_k) \right)}_{\text{TD error}}. \quad (\text{A.1})$$

404 The iterative algorithm expands as follows: at iteration  $k$ , the predator in-  
 405 teracts with the environment of state  $s_k$  which is a realisation from the state  
 406 space  $S$ . Following a certain decision policy  $\pi$ , the predator takes action  $a_k$  out  
 407 of the action space  $A$ . As a result of this interaction at iteration  $k$ , the predator  
 408 experiences an immediate reward  $r_{k+1}$ . The terminology refers to the experi-  
 409 enced reward at the subsequent iteration  $k+1$  which emphasis that the reward  
 410 is in consequence of the predator's action. Next, the predator forms a target  
 411 value which is a composition of the experienced reward  $r_{k+1}$  and discounted fu-  
 412 ture rewards. Thereby, future rewards are a prevailing estimate  $Q(s_{k+1}, a_{k+1})$   
 413 which is known as *bootstrapping*. The difference between the target value and  
 414 the estimate at iteration  $k$  gives the *temporal-difference (TD) error*. Finally, the  
 415 Q-learning algorithm updates the estimate  $Q(s_k, a_k)$  to  $Q'(s_k, a_k)$  towards the  
 416 formed target value, subsequently reducing the TD error. As the Q-learning al-  
 417 gorithm uses bootstrapping, these targets are moving ones. Hence, the update  
 418 process should progress slowly with  $\alpha$ , the learning rate, being a small posi-  
 419 tive constant. Figure A.6 shows a possible implementation of the Q-learning  
 420 algorithm as pseudo-code.

```

421  $Q \leftarrow 0$ 
422  $s_k \leftarrow s_0$ 
423 WHILE learning DO
424      $a_k \leftarrow \pi(s_k, Q)$ 
425      $s_{(k+1)} \leftarrow f(s_k, a_k)$ 
426      $Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha (r_{(k+1)} +$ 
427          $\gamma \max_a Q(s_{(k+1)}, a) - Q(s_k, a_k) )$ 
428      $s_k \leftarrow s_{(k+1)}$ 
429

```

Figure A.6: Q-learning algorithm in pseudo-code