The evolution of social dominance through reinforcement learning

Olof Leimar

Department of Zoology, Stockholm University, SE-106 91 Stockholm, Sweden.

E-mail: olof.leimar@zoologi.su.se.

Abstract

Groups of social animals are often organised into dominance hierarchies that are formed through pairwise interactions. There is much experimental data on hierarchies, examining such things as winner, loser, and bystander effects, as well as the linearity and replicability of hierarchies, but there is a lack evolutionary analyses of these basic observations. Here I present a game-theory model of hierarchy formation in which individuals adjust their aggressive behaviour towards other group members through reinforcement learning. Individual traits such as the tendency to generalise learning between interactions with different individuals, the rate of learning, and the initial tendency to be aggressive are genetically determined and can be tuned by evolution. I find that evolution favours individuals with high social competence, making use of individual recognition, bystander learning and, to a limited extent, generalising learned behaviour between opponents when adjusting their behaviour towards other group members. The results are in good agreement with experimental data, for instance in finding weaker winner effects compared to loser effects.

Introduction

The empirical study of dominance hierarchies is nearly a century old (Schjelderup-Ebbe, 1922) and has identified important phenomena such as winner, loser, and bystander effects. Work in the field has also pointed to potential challenges for the understanding of hierarchy formation, for instance by noting that hierarchy replicability is sometimes low even though hierarchies are nearly linear (e.g. Chase et al., 2002).

A number of mathematical models of dominance hierarchy formation have been developed, both mechanistic and evolutionary models (reviewed by Mesterton-Gibbons et al., 2016). By assuming an updating of internal variables when individuals win or lose interactions (e.g. Dugatkin, 1997), or when they are bystanders to wins and losses by others (Dugatkin, 2001), mechanistic models have had some success in describing hierarchy formation, for instance in verifying that bystander effects promote hierarchy linearity. Because of the modelling complexities introduced by aggressive interactions between individuals that may differ in their fighting ability and previous experience, it has proven difficult to achieve evolutionary analyses of such mechanistic models. Nevertheless, winner and loser effects have been found in game-theory models by looking at simplified situations for groups of three individuals (Mesterton-Gibbons, 1999; van Doorn et al., 2003a,b).

To overcome these difficulties I here model the behavioural mechanisms of social hierarchy formation as reinforcement learning (e.g. Sutton and Barto, 2018). Recent work in neuroscience provides experimental support for the general idea that social dominance relations develop through processes that are similar to reinforcement learning (Kumaran et al., 2016; Ligneul et al., 2016; Qu et al., 2017; Zhou et al., 2018). Furthermore, observational learning, which is used as the mechanism underlying bystander effects in the model, is now part of the general neuroscience

framework of reinforcement learning (Burke et al., 2010; Olsson et al., 2020). In addition to work in neuroscience and experimental psychology, there are thoroughly-studied reinforcement learning algorithms, with application in computer science and machine learning (Sutton and Barto, 2018), making the theory a good candidate for the integration of function and mechanism in game theory for biology (as discussed by McNamara and Leimar, 2020).

For evolutionary analysis of the model, the parameters of an individual's learning rule are assumed to be genetically determined traits, and the evolution of these traits are studied through individual-based evolutionary simulations. An example of such a trait is an individual's tendency to generalise its learned behaviour from one group member to other individuals, which gives rise to winner and loser effects during hierarchy formation. Other examples are learning rates, including bystander learning rates, and an individual's tendency to be aggressive at the start of hierarchy formation. By performing such simulations for different sizes of social groups, as well as by keeping certain parameters at fixed values, which would correspond to evolutionary constraints, I investigate the evolution of social dominance behaviour in different contexts. This provides an understanding of the nature of social competence (Taborsky and Oliveira, 2012) for social dominance behaviour, by identifying the traits that are adaptive for an individual to adjust its behaviour during hierarchy formation. The analysis can reveal, for instance, whether individuals adapted to larger group sizes should show stronger winner and loser effects compared to those adapted to smaller group sizes, or under which circumstances winner effects should be weaker than loser effects.

The model is meant to be general, but still makes assumptions about cognitive capacities and how fitness depends on social interactions during a life history. As a possible example, one might consider interactions in groups of fowl (Schjelderup-Ebbe, 1922), in particular groups of male fowl (Pizzari and McDonald, 2019), in which individuals both display and fight during hierarchy formation.

To make contact with experimental work, I present model simulations illustrating potential results from experiments investigating winner, loser, and bystander effects, as well as experiments measuring hierarchy replicability and linearity. These results are reported both for evolved learning traits and for other values of some of these traits, such as the degree of generalisation and the bystander learning rate, in order to illustrate the influence of these traits in experiments. I also simulate an experimental set-up for estimating the amount of aggressive behaviour in interactions between a bystander and either the winner or loser of the bystander-observed interaction. The set-up tests the hypothesis that bystander effects differ between the cases, such that bystander effects are larger when a bystander interacts with the former winner (Earley and Dugatkin, 2002). I discuss whether the main model results are in qualitative agreement with observations of social dominance and I end by recapitulating the general game-theory approach used.

Methods

Model description

In each generation, there are a number of rounds or time steps of interaction, $t = 1, \dots, T$, and in each round there is a dominance interaction between two randomly selected individuals from a group of size g_s . Each such interaction has the same general reward structure (see below), and group members can learn from the perceived rewards, which depend on their qualities q_i (fighting abilities). At the start of a generation, individual qualities are randomly drawn from a normal distribution with mean μ_q and standard deviation σ_1 , and they are assumed not to vary between rounds of the game. As a choice of zero point of the scale for q , let us assume that $\mu_q = 0$ (e.g., if fighting ability would be fully determined by size, q could be the logarithm of size, with the mean size in the population equal to one).

Concerning what is ‘known’ by group members, assume that initially they do not have any particular information, including about their own fighting ability, but that they observe differences during dominance interactions and learn which actions to use through the rewards they perceive. The situation is then similar to traditional instrumental or operant conditioning.

For learning, the model uses the actor-critic method, described in section 13.5 of Sutton and Barto (2018). In this approach, an individual maintains and updates action preferences and estimates of values. The action preference (the ‘actor’) determines the behaviour to use in a given situation, which in the model is an interaction with a particular other individual, and the estimated value (the ‘critic’) is an estimate of the perceived reward from the interaction. The model makes a distinction between perceived rewards, which direct an individual’s learning, and evolutionary costs and benefits, which determine the action of natural selection. In addition to the description here, further details and explanation are found in Appendix A.

Observations/states and actions

A real dominance interaction can consist of a sequence of several behaviours, including threat displays and fighting. The model simplifies this into two stages. In the first stage, interacting individuals, but not bystanders, make an ‘observation’ that acts like a state. Individuals observe some aspect ξ of relative fighting ability at the start of a dominance interaction, and also observe the opponent’s identity (individual recognition). The observation ξ might correspond to information from displays of some kind. The observation by each individual is statistically related to its quality (q_i) and that of the opponent (q_j). For the interaction between individuals i and j in round t , the observation is assumed to be

$$\xi_{ijt} = a_1 q_i - b_1 q_j + \epsilon_{ijt}, \quad (1)$$

where ϵ_{ijt} is an error of observation, assumed to be normal with mean zero and standard deviation σ . There is a similar expressions for j against i . By adjusting the parameters σ_1 and a_1 , b_1 , and σ from equation (1) one can make the information about relative quality more or less accurate. The observation/state (ξ_{ij}, j) is followed by a second stage, where individual i chooses an action, and similarly for individual j . The model simplifies to only two actions, A and S, corresponding to aggressive and submissive behaviour.

Action preferences and estimated values

For an individual i interacting with j (with $j \neq i$) in round t , l_{ijt} denotes the preference for A. The probability that i uses A is then

$$p_{ijt} = \frac{1}{1 + \exp(-l_{ijt})}, \quad (2)$$

so that l_{ijt} is the logit of the probability of using A. The model uses a linear (intercept and slope) representation of the effect of ξ_{ijt} on the preference, and write l_{ijt} as the sum of three components

$$l_{ijt} = h_{iit} + h_{ijt} + \gamma_{0i}\xi_{ijt} = f_i\theta_{iit} + (1 - f_i)\theta_{ijt} + \gamma_{0i}\xi_{ijt}. \quad (3)$$

Here $h_{iit} = f_i\theta_{iit}$ is a contribution from generalisation of learning from all dominance interactions, $h_{ijt} = (1 - f_i)\theta_{ijt}$ is a contribution specifically from learning from interactions with a particular opponent j , and $\gamma_{0i}\xi_{ijt}$ is a contribution from the current observation of relative quality. Note that for $f_i = 0$ the learning about each opponent is a separate thing, with no generalisation between opponents, and for $f_i = 1$ the intercept component of the action preference is the same for all opponents, so that effectively there is no individual recognition (although the observations ξ_{ijt} could still differ between opponents). One can similarly write the estimated value

$\hat{v}_{ijt}$ of an interaction as a sum of three components:

$$\hat{v}_{ijt} = f_i w_{iit} + (1 - f_i) w_{ijt} + g_{0i} \xi_{ijt}. \quad (4)$$

The actor-critic method updates θ_{iit} , θ_{ijt} , w_{iit} , and w_{ijt} in these expressions based on perceived rewards, whereas f_i , γ_{0i} , and g_{0i} are genetically determined.

Exploration in learning

The trade-off between exploration and exploitation is a basic challenge for reinforcement learning algorithms (Sutton and Barto, 2018). For learning to be efficient over longer time spans there must be exploration (variation in actions), in order to discover beneficial actions. Learning algorithms, including the actor-critic method, might not provide sufficient exploration, because learning tends to respond to short-term rewards. In the model, exploration is implemented as follows: if the probability in equation (2) is less than 0.01 or greater than 0.99, the actual choice probabilities are assumed to stay within these limits, i.e. is 0.01 or 0.99, respectively. In principle the degree of exploration could be genetically determined and evolve to an equilibrium value, but for simplicity this is not implemented in the model.

Perceived rewards

An SS interaction is assumed to have zero rewards, $R_{ijt} = R_{jit} = 0$. For an AS interaction, the aggressive individual i ‘wins’, which is associated with a perceived (primary) reward $R_{ijt} = v_i$ for individual i , which is genetically determined and can evolve. The reward for the submissive individual j is zero, $R_{jit} = 0$, and vice versa for SA interactions. If both individuals use A, some form of costly dominance display or fight ensues, with perceived costs (negative rewards or penalties) that are influenced by the qualities of the two individuals. The perceived rewards of an AA

interaction are assumed to be

$$R_{ijt} = v_i - C \exp(-q_i + q_j + e_{ijt}) \quad (5)$$

$$R_{jit} = v_j - C \exp(-q_j + q_i + e_{jit}),$$

where e_{ijt} is a normally distributed random influence on the perceived penalty, with mean zero and standard deviation σ_p , and similarly for e_{jit} . Note that it is not assumed that there is a ‘winner’ of the aggressive interaction. The reward components v_i and v_j are interpreted as innate motivation to perform aggressive behaviour. For simplicity, the parameter C is the same as for the evolutionary cost in equation (6) below. In effect, this gives a scale for the perceived rewards v_i and v_j .

Evolutionary costs and benefits

The model assumes that there are evolutionary costs of AA interactions, because of injuries or exhaustion. For individual i interacting with j this is given by

$$D(q_i, q_j) = C \exp(-q_i + q_j). \quad (6)$$

For simplicity, the dependence on the qualities has the same form as the perceived reward in equation (5). There are also evolutionary benefits from resources that become available (acquiring these resources is the ultimate function of dominance interactions). For simplicity, assume that resources are available after each interaction, and that the expected benefit for an individual that acquires the resource is V . If both individuals use A, the model assumes that they are preoccupied with aggressive behaviour, and neither gets the resource (it might for instance be that some other, unspecified agent gets the resource, or, if it is a potential mate, that the resource moves on). If one of a pair uses A and the other S, the aggressive individual

is the winner and obtains the resource, and if both use S, they are equally likely to get the resource, or alternatively they share it (these give the same expected payoff).

Learning updates

Let us first deal with updates for the individuals directly involved in a dominance interaction (see below for bystanders). The updates are driven by the prediction error (TD error)

$$\delta_{ijt} = R_{ijt} - \hat{v}_{ijt}, \quad (7)$$

which is the difference between the actual perceived reward R_{ijt} and the estimated value $\hat{v}_{ijt}$. The learning updates for the θ parameters are given by

$$\begin{aligned} \theta_{ii,t+1} &= \theta_{ii,t} + \alpha_{\theta_i} f_i \zeta_{ijt} \delta_{ijt} \\ \theta_{ij,t+1} &= \theta_{ij,t} + \alpha_{\theta_i} (1 - f_i) \zeta_{ijt} \delta_{ijt}, \end{aligned} \quad (8)$$

where

$$\zeta_{ijt} = \begin{cases} 1 - p_{ijt} & \text{if A was chosen} \\ -p_{ijt} & \text{if S was chosen} \end{cases} \quad (9)$$

is referred to as a policy-gradient factor and α_{θ_i} is the preference learning rate for individual i . Note that ζ_{ijt} will be small if p_{ijt} is close to one and individual i performed action A, which slows down learning, with a corresponding slowing down if p_{ijt} is close to zero and S is chosen. There are also learning updates for the w

parameters given by

$$\begin{aligned} w_{ii,t+1} &= w_{ii,t} + \alpha_{wi} f_i \delta_{ijt} \\ w_{ij,t+1} &= w_{ij,t} + \alpha_{wi} (1 - f_i) \delta_{ijt}, \end{aligned} \tag{10}$$

where α_{wi} is the value learning rate for individual i . See SI for a derivation of these learning updates, using so-called policy gradient methods (following chapter 13 in Sutton and Barto, 2018). The values for the parameters at the start of a generation are also needed; let these be $\theta_{ii0} = \theta_{ij0} = \theta_{0i}$ and $w_{ii0} = w_{ij0} = w_{0i}$, where θ_{0i} and w_{0i} are assumed to be genetically determined. Note that θ_{0i} determines how aggressive individual i is at the start of dominance hierarchy formation, which is an important aspect of social dominance behaviour.

Bystander updates

To achieve a realistic description of social dominance, bystander effects are modelled as observational learning. In reinforcement learning theory, observation learning is conceptualised in a very similar way as other learning (Burke et al., 2010; Olsson et al., 2020), but without explicit specification of rewards. Instead, an observer can assign higher or lower significance to a particular event, implemented as the magnitude of an observational learning rate.

Applying this to dominance interactions, individuals other than the interacting pair i and j , for instance an individual k , could use the outcome to update the learning parameters. Assume that individual k only performs this updating if i and j use AS or SA (because there is no clear ‘winner’ in AA and SS interactions, and bystanders do not perceive the costs of AA interactions). The probabilities for individuals i and j to use A are p_{ijt} and p_{jit} , from equation (2). These are ‘true’ values and are not known by individual k . However, given that the outcome is either

AS or SA, one readily derives that the logit of the probability that it is AS is

$$\text{logit Pr(AS|AS or SA)} = l_{ijt} - l_{jit}. \quad (11)$$

From equation (3) one can see that this involves various learning parameters for i and j . For bystander learning an assumption is needed about how an individual k represents this logit. A simple assumption is that k represents the logit as

$$\text{logit Pr(AS|AS or SA)} = -\theta_{kit} + \theta_{kjt}, \quad (12)$$

which entails that k does not use any information about q_i or q_j . The assumption is reasonable in that a large θ_{kit} means that k behaves as if individual i is weak, and similarly for θ_{kjt} . Using the notation

$$P_{kijt} = \frac{1}{1 + \exp(\theta_{kit} - \theta_{kjt})}, \quad (13)$$

the bystander updates by k is assumed to be

$$\begin{aligned} \theta_{ki,t+1} &= \theta_{kit} - (1 - P_{kijt})\beta_k \\ \theta_{kj,t+1} &= \theta_{kjt} + (1 - P_{kijt})\beta_k \end{aligned} \quad (14)$$

is i wins (outcome is AS) and

$$\begin{aligned} \theta_{ki,t+1} &= \theta_{kit} + P_{kijt}\beta_k \\ \theta_{kj,t+1} &= \theta_{kjt} - P_{kijt}\beta_k \end{aligned} \quad (15)$$

if j wins (outcome is SA). The parameter β_k is a measure of how salient or significant a bystander observation is for individual k , and this parameter is assumed to be genetically determined and can evolve. These bystander updates are similar to the

direct-learning updates of the actor component of the actor-critic model.

Forgetting

Memory is an important cognitive capacity, which in practice will be more or less limited. Despite its importance, for simplicity the model does not implement a sophisticated memory mechanisms (including such things as interference between memories). Instead, we assume that there is a memory factor m such that the θ and w learning parameters decay towards their initial values with a rate m per time step.

Evolutionary simulations

A number of parameters/traits in the model are assumed to be genetically determined and can evolve. The traits for an individual i are as follows: influence of states/observations on preference and value functions, from equations (3, 4), f_i , γ_{0i} , and g_{0i} ; perceived reward from performing A, from equation (5), v_i ; learning rates, from equations (8, 10, 14, 15), $\alpha_{\theta i}$, α_{wi} , and β_i ; and initial values for the preference and value learning parameters, θ_{0i} and w_{0i} . In individual-based evolutionary simulations, each trait was implemented as determined by an unlinked diploid locus with additive alleles. For each generation, a population of $N = 8000$ individuals were randomly divided into groups of size g_s equal to 4, 8, or 16. Each individual was assigned a quality q_i , independently drawn from a normal distribution with mean zero and standard deviation σ_1 . Dominance-hierarchy formation then took place in each group, over a number of time steps such that each individual participated in an expected number of 200 dominance interactions. During a generation group members accumulated fitness payoffs as follows: a basic increment V_0 per round of interaction, plus a payoff V for each resource acquired, and minus a cost as in equation (6) for each AA round. The next generation was formed by randomly

selected parents from the entire population, with probabilities proportional to an individual's accumulated payoff.

For each case reported in Table B1, simulations were performed over 5000 generations, repeated in sequence at least 100 times, to estimate mean and standard deviation of traits at an evolutionary equilibrium. These rather extensive simulations were performed to average out statistical fluctuations in the distribution of trait values over evolutionary time. As pointed out by, e.g., Bürger and Lande (1994), for mutation-drift-selection equilibrium dynamics under stabilising selection there can be substantial autocorrelation over generations of the population dynamical stochastic process.

Standard parameter values

For several of the evolutionary simulations the following ‘standard values’ of parameters were used: fitness pay-offs, $V_0 = 0.5$, $V = 0.25$, $C = 0.2$; distribution of individual quality variables, $a_1 = 0.707$, $b_1 = 0.707$, $\sigma_1 = 0.50$; observations of relative quality, $\sigma = 0.50$; perceived penalty variation, $\sigma_p = 0.25$; memory factor, $m = 0.999$. For these parameter values, around 50% of the variation in the observation ξ_{ij} in equation (1) is due to variation in $q_i - q_j$.

Code availability

C++ source code for the individual-based simulations is available at GitHub, together with instructions for compilation on a Linux operating system:

<https://github.com/oleimar/socdom>.

Results

To illustrate the model (described in the Methods section), I present an example of the formation of a dominance hierarchy in a group of four individuals (Fig. 1). In the example, the learning traits are chosen to be at an approximate evolutionary equilibrium under the specified conditions, except that for simplicity there is no bystander learning in this example.

In each time step there is one round of dominance interaction between two randomly selected group members. In a round each individual first obtains an observation (with error) of the difference in fighting ability between itself and the opponent (not shown in Fig. 1; see equation (1)), for instance from display behaviours, and also observes the identity of the opponent. This is followed by a choice of actions, A (aggressive) or S (submissive), after which the round ends. An individual i perceives an internal reward v_i from performing A, which can be thought of as a motivation to be aggressive. If both individuals choose A there is a fighting interaction, with perceived negative rewards (penalties) that depend on the fighting abilities (qualities) q_i and q_j of individuals i and j (equation (6)). From these each group member learns whether to use A or S against another group member. Panels (a) to (d) in Fig. 1 correspond to group members 1 to 4 and show the probabilities of using A against another group member (colour coded). Individual 1 is strongest and continues to almost always use A against the other group members throughout the time in the group (Fig. 1a). Individual 4 is the next strongest and initially tends to use A against all other group members, but eventually becomes submissive towards individual 1 (Fig. 1d). The other two individuals also start out aggressive but end up in lower dominance positions (Fig. 1b,c). The outcome is a linear hierarchy with dominance positions arranged according to fighting ability.

The model uses the actor-critic learning algorithm (Sutton and Barto, 2018). In

this approach, the logit of an individual's probability of using A is referred to as its preference for A and is expressed as a sum of components (equations (2, 3)).

Figure 1: Illustration of the reinforcement-learning model for a group of size $g_s = 4$, interacting over a total of $T = 400$ time steps. The first 120 time steps are shown, during which a stable dominance hierarchy is established. Panels (a) to (d) show the probability, from equation (2), of using the aggressive action A for individual i interacting with individual j , with one panel for each i and with colour coding for j as given in the inserted legend in panel (a). The inserted legend also gives the individual quality values q_i . Panels (e) to (h) show the corresponding preference components h_{ij} , from equation (3), with the colour-coded solid lines giving the component h_{ij} for a specific opponent j , and the dotted lines showing the generalised component h_{ii} , with one panel for each individual i . The parameter values for learning are as follows: $f_i = 0.19$, $\beta_i = 0.0$, $\alpha_{\theta i} = 175.9$, $\alpha_{wi} = 0.09$, $\theta_{0i} = 5.45$, $w_{0i} = 0.002$, $\gamma_{0i} = 2.57$, $g_{0i} = 0.025$, and $v_i = 0.15$. These are at an approximate equilibrium from evolutionary simulations, with $g_s = 4$ and other parameters having the ‘standard values’ given in Methods (case 1 in Table B1 except that, for simplicity, bystander effects, from equation (15), are removed by putting $\beta_i = 0$).

One could loosely compare with how effects are combined in logistic regression, but for the actor-critic method the preference components are changed through learning

updates, driven by so-called prediction errors (equations (7, 8)). Two of the components are shown in panels (e) to (h) in Fig. 1; these correspond to panels (a) to (d). The solid lines (colour coded) show components specific to a particular opponent, whereas the dotted lines show the generalised component, which contributes to the preference for any opponent. Note that the opponent-specific components vary much more compared to the generalised component, and tend to change in an abrupt manner, producing rapid shifts from mostly aggressive to mostly submissive behaviour. A further illustration of the components over the entire time in the group appears in Fig. B1.

The reason for the large jumps in preferences seen in Fig. 1 is that the evolved preference learning rate is very high. Such high learning rates are a characteristic feature of the evolution of social dominance in the model. There are two learning rates in the model, a preference learning rate α_θ and a value learning rate α_w . The former is a parameter of the actor part of actor-critic learning and the latter belongs to the critic part; the critic learning rate does not evolve to high values. Comparing with real animals, the high preference learning rate could correspond to special cognitive processing of social dominance behaviour. Its function might be to allow rapid shifts from aggressive to submissive behaviour.

The degree of generalisation between opponents and the bystander learning rate are also important in social dominance, which is illustrated in Fig. 2. The learning traits are approximately at an evolutionary equilibrium (established by individual-based evolutionary simulations) for the cases in the figure. The comparison between blue and grey points show that suboptimally high generalisation (for the modelled situation) and a removal of bystander effects slows down learning about relative fighting ability and leads to a less sharp decrease in fights over time in the group. These effects are more pronounced in larger groups and they illustrate the nature of social competence in dominance-hierarchy formation, such as limited generalisation between opponents and substantial bystander learning rates,

Figure 2: Influence of social competence on the learning about relative fighting ability and the decrease over time of aggression during hierarchy formation. Blue dots illustrate cases where the learning parameters, including the generalisation parameter (f_i) and the bystander learning rate (β_i), are at an approximate evolutionary equilibrium. The grey dots show the same except that generalisation is fixed at a suboptimally high value ($f_i = 0.5$) and there is no bystander learning ($\beta_i = 0$), corresponding to lower social competence. In panels (a) and (b), group size is $g_s = 4$ with interactions over a total of $T = 400$ time steps, and in (c) and (d), $g_s = 8$ with interactions over a total of $T = 800$ time steps. Panels (a) and (c) show the probability p_{ijt} (equation (2)) of using the aggressive action A by an individual i plotted against the quality difference $q_i - q_j$ between itself and its opponent j , for times t corresponding to around 20 round per pair: each point gives the mean of p_{ijt} over time steps t corresponding to 20 ± 1 rounds per pair. The solid (for blue dots) and dashed (for grey dots) lines represent linear regressions on a logit scale. Panels (b) and (d) show the proportion of interactions that are AA (i.e. are fights) as a function of the number of rounds per pair. The dots overlap and represent averages for one simulated group over an interval of time steps. The evolved learning parameter values are given in Table B1, cases 1 and 7 for panels (a) and (b), and cases 2 and 8 for panels (c) and (d). Other parameters have the ‘standard values’ given in Methods.

which are higher for individuals adapted to larger group sizes (see values of β for cases 1, 2, and 3 in Table B1). Pronounced generalisation between opponents, implying strong winner and loser effects, is detrimental to the formation of dominance hierarchies, as is illustrated in Fig. B2.

For help in interpreting the evolved learning traits, Fig. 3 shows simulated experiments examining winner, loser, and bystander effects. In Fig. 3a, winner and loser effects are shown in darker blue for learning traits that evolved in groups of four, in red for different imposed values of the generalisation factor f (see equation (3)), and in light blue when the initial preference for A is kept at zero, corresponding to 50% chance of using A at the start of learning. As expected, the strength of winner and loser effects increases with the tendency to generalise between opponents ($f = 0$ means that there is no generalisation and $f = 1$ means that all opponents are treated as the same). Because of the experimental design, where the focal individual is tested against a matched opponent, there are substantial winner and loser effects for the evolved learning traits, although as seen from Fig. 1 the generalised preference component has rather little influence on hierarchy formation. The loser effects deviate more from $p_A = 0.5$ than the winner effects in Fig. 3a, except for the case of intermediate initial aggressiveness (light blue points), where the difference is small. This reveals a reason for weaker winner effects in the model. If individuals initially have a high preference for A, the actor learning updates from prediction errors while using A will be smaller than those from using S (equations (8, 9)), which can be compared with well-established results in the study of operant conditioning.

We can also compare with the extreme case of the grey points in Fig. B2c,d (case 11 in Table B1, with a negative θ_{0i}), where there is no real hierarchy formation. For this case winner effects of the kind shown in Fig. 3 would be very strong ($p_A = 0.99$), as would be loser effects ($p_A = 0.02$). Such strong winner effects are not to be expected for individuals well adapted to life in social hierarchies.

Figure 3: Winner, loser and bystander effects, as they might be observed in experiments. Panel (a) shows results from simulated winner- and loser-effect experiments. A focal individual either first experiences a win against a weaker or a loss against a stronger opponent (over at least 200 rounds and with $\Delta q = 0.5$), and then interacts with another, matched opponent, scoring a win or a loss. The darker blue points (with 95% binomial confidence intervals) represent the probability of focal winning from 6400 replicated experiments of each kind, with learning parameters as in Fig. 2a (case 1 in Table B1). The red points illustrate the effects of changing the generalisation parameter f_i from its evolutionary equilibrium, and the lighter blue points similarly illustrate setting the initial preference θ_{0i} for action A to zero. Panel (b) shows results from simulated bystander-effect experiments, where a focal individual (with $q = 0$) first either observes or does not observe (control) a dominance interaction between a pair of individuals (over at least 10 rounds and with $q = -0.25$ and $q = 0.25$, respectively). The focal then interacts with the winner of the pair (solid lines and points), or the loser (dashed horizontal lines), and the number of rounds with fighting (AA rounds) is scored. The red, blue, and green horizontal lines (solid and dashed) show the overall mean for the control treatment, with different learning parameters (corresponding to cases 1, 2, and 3 in Table B1, with group sizes 4, 8 and 16, respectively). The vertical descending solid arrows show the bystander-winner effects (marked ‘bwe’) when the observer interacts with the winner, positioned at the evolved values of the bystander learning rates β_i . The red, blue, and green points (with 95% bootstrap confidence intervals) illustrate the consequences of changing the bystander learning rate, with 6400 replicates for each point. Other parameters have the ‘standard values’ given in Methods.

The simulated experiments on bystander effects are inspired by previous experimental approaches (Johnsson and Åkerman, 1998; Earley and Dugatkin, 2002). The main result (Fig. 3b) is that bystander-winner effects, where a bystander meets the winner of an observed pair, are quite pronounced, and more so for individuals adapted to larger group sizes, while bystander-loser effects are small and not statistically detectable. Only the control results for bystander-loser effects are shown in Fig. 3b (dashed horizontal lines); the treatment results had confidence intervals overlapping the control, in spite of a very large sample size, so any bystander-loser effect in this example is very weak. Note also that the evolved bystander learning rates in Fig. 3b are higher when adapted to larger groups.

One way of investigating the influence of individual fighting abilities on dominance hierarchy formation is to let the same group form a hierarchy twice, with sufficient time in between so that individuals have no memory of the first at the second time (Chase et al., 2002). The simulated experiments of replicated hierarchy formation (Fig. 4a) show, as expected, that dominance relations are uncorrelated between replicates if there is very little variation in fighting ability q , but the correlation becomes high (but less than perfect) for the amount of variation that the learning traits are adapted to. There is little influence of bystander learning.

However, as seen in Fig. 4b, bystander learning has a major influence on hierarchy linearity. Strikingly, when there is bystander learning, the linearity of hierarchies can be high even when replicates are essentially uncorrelated. Linearity is above random also without bystander learning (orange and light blue in Fig. 4b), but with the evolved bystander learning rates, linearity is essentially perfect both in groups of 4 and 8 individuals. To assess linearity the classical measure developed by Landau (1951) was used, taking into account the modifications by de Vries (1995) to account for possible ties.

In addition to the effects in Fig. 4, there are a number of other consequences of changing the amount of variation in fighting ability q in experimental populations.

Figure 4: Replicability of dominance relations, in panel (a), and the linearity of dominance hierarchies, in panel (b), as a function the amount of variation in fighting ability in experimental populations, with and without bystander learning. The darker red and blue points (with 95% bootstrap confidence intervals) are for experiments with individuals with evolved learning parameters from cases 1 and 2 in Table B1, with group sizes $g_s = 4$ and $g_s = 8$, respectively. The orange and light blue points are for the corresponding evolved learning parameters when the bystander learning rate is fixed at $\beta_i = 0$ (cases 4 and 5 in Table B1). In a simulated experiment, a group of size of either 4 or 8 is taken from an experimental population with some given SD of fighting ability q (values of SD: 0.01, 0.1, 0.2, 0.3, 0.4, 0.5; points are shifted left and right for clarity), ranging from nearly matched individuals in a group up to the variation in q to which the evolved learning parameters are adapted (i.e. an SD of 0.5). For each of these cases, and for each value of SD, the simulated data consists of 50 groups and 2 replicates per group. Group members are assumed to forget any learning between replicates. Panel (a) shows the correlation between replicates of the pairwise differences of the probabilities to use A from equation (2), $p_{ij} - p_{ji}$, at the end of the time spent in the group (400 or 800 time steps). Panel (b) shows the Landau (1951) hierarchy index h , for which a value of 1 corresponds to a linear hierarchy. The orange and light blue horizontal lines indicate the expected values of the index for random dominance relations in groups of 4 and 8 individuals, respectively (given by eq. (29) in Landau, 1951). Other parameters have the ‘standard values’ given in Methods.

Many of these effects are quite intuitive, such as a stronger correlation between fighting ability and final dominance position when there is more variation in q , or more fighting interactions before hierarchy stability with less variation in q . For instance, for the cases in Fig. 4 of groups of size $g_s = 4$ with bystander learning, the mean number of AA interactions per individual increases by a factor of 3, from around 21 to around 63, when the SD of q decreases from 0.5 to 0.01. For groups of size $g_s = 8$ this effect is similar, with the mean number of AA interactions per individual increasing by a factor of 3, from around 29 to around 87, when the SD of q decreases from 0.5 to 0.01. It is thus more costly to settle dominance in groups with less variation in fighting ability.

Discussion

By using the behavioural mechanisms of reinforcement learning, the model recovers and potentially explains several of the established aspects of dominance-hierarchy formation, such as winner, loser, and bystander effects. Among these are new explanations for why winner effects are weaker than loser effects and why bystander-winner effects are stronger than bystander-loser effects.

The parameter f in the model represents the strength of generalisation between opponents during hierarchy formation. This provides a link both to a long tradition in experimental psychology (Ghirlanda and Enquist, 2003) and to reinforcement learning theory (e.g. ch. 9 in Sutton and Barto, 2018). The parameter describes how a learning individual associates its experiences in dominance interactions with different characteristics of an opponent, like markings or colour patterns, ranging from those unique to each opponent ($f = 0$) to those shared by all opponents ($f = 1$). This aspect of the model thus explains winner and loser effects using the animal-psychology concept of stimulus generalisation. In principle this general idea could be tested experimentally through manipulation of characteristics mediating

individual recognition in social groups (Tibbetts and Dale, 2007), examining if the availability of different characteristics influences winner and loser effects.

The model (Fig. 3a) provides a potentially general explanation of why winner effects are often found to be weaker than loser effects (Chase et al., 1994; Rutte et al., 2006; Oliveira et al., 2011). The explanation is that the magnitude of learning updates to the preference for A, given in equations (8, 9), depends on which action, A or S, is actually chosen. The update will be small if the action chosen had a high probability of being chosen. An interpretation is that the update is smaller when the choice is less ‘surprising’. This particular reason for a slowing down of learning appears not to have been tested experimentally, but it is in accordance with previous results and ideas about persistence in responding in operant conditioning, involving two-factor learning theory (Mowrer, 1951) and actor-critic learning theory (e.g. Maia, 2010). The explanation could be tested experimentally: the prediction would be that individuals showing high aggressiveness at the start of learning should show relatively weaker winner effects compared to loser effects.

The model implements observational learning (Burke et al., 2010; Olsson et al., 2020), sometimes referred to as vicarious reinforcement (Bandura, 1986), as a psychological mechanism underlying bystander effects. The bystander-learning updates, given in equations (14, 15), only change the opponent-specific components of the observer’s action preferences. These updates do not take into account information an observer might have about its own dominance status, which might be a simplification compared to what has been observed experimentally (Abril-de-Abreu et al., 2015).

The model simulations showed stronger bystander-winner than bystander-loser effects (Fig. 3b), which has been found in experiments (Earley and Dugatkin, 2002). The explanation for the model result is related to the explanation for stronger loser than winner effects. When a former observer interacts with the (stronger) winner of an observed pair, the former observer will have a somewhat reduced preference

for A, and if S is used instead of A, the former observer rapidly learns to become submissive. On the other hand, when the former observer interacts with the (weaker) loser, the same explanation as the one above for a weak winner effect applies: just as in control conditions, the former observer has a strong tendency to be aggressive and does not change its preference for A to a noticeable degree.

The finding in Fig. 4, that hierarchies can be close to linear even when dominance relations are uncorrelated between replicates, is in accordance with experimental results (Chase et al., 2002). Note that, for the model, hierarchy linearity is a by-product of selection for traits influencing reinforcement learning; there is no direct selection for linearity. In the same way, for the model there is no direct selection for winner or loser effects – these are instead by-products of the evolved learning traits, including the strength of generalisation. Thus, the modelling does not assume that dominance-hierarchy formation can be decomposed into winner and loser effects of different strengths. It is even the case that winner and loser effects, which act through the generalised component of the action preference (Fig. 1e-d), can play a rather small role during hierarchy formation.

The simulated experiments illustrated in Fig. 4 showed that more fighting per capita, and thus greater cost, is required to establish dominance in groups with less variation in fighting ability. This is similar to the traditional idea/observation that it takes longer to settle contests with smaller differences in fighting ability. It is also more costly for individuals to settle dominance in larger groups. This prediction could easily be tested and is broadly consistent with findings on size-based hierarchies (e.g. Buston and Cant, 2006; Ang and Manica, 2010), that social groups that are large or contain individuals of similar size can become unstable because of costs of settling dominance relations.

The model is likely to be an oversimplification of any real case of dominance-hierarchy formation. The assumptions about evolutionary costs and benefits of dominance interactions, described in the Methods section, entail that reproductive

success on average increases linearly with the dominance position, from lowest to highest. The effect of the dominance position on reproductive success might well vary between species and situations, and is often not empirically estimated in detail, but the model assumptions are at least broadly consistent with what is known about male fowl (McDonald et al., 2017; Pizzari and McDonald, 2019). The result for male fowl that experimentally induced wins or losses had rather little influence on subsequent hierarchy formation (Favati et al., 2017) is also broadly consistent with the model.

When comparing the model results with empirical observations, it is important to keep in mind that aggressive interactions, including those where winner and loser effects have been found, can represent adaptations to other situations than the formation of dominance hierarchies in social groups. Thus, contest over territories, or simply contests over resources that individuals encounter, often show effects of previous experiences of aggressive interactions (reviewed by Hsu et al., 2006). An example might be the strong winner effects found for a solitary parasitic wasp (Goubault and Decuignière, 2012) for which there are no social hierarchies, but where individuals can occur at high population densities and frequently contest resources. There are also model results dealing with effects of wins and losses in contests in large populations (e.g. Fawcett and Johnstone, 2010). Although there will be similarities between how the regulation of aggression becomes adapted to different situation, the current model applies specifically to the formation of hierarchies in social groups.

Evolutionary modelling

By assuming several of the parameters of actor-critic learning to be genetically determined traits, one can study their evolution through individual-based simulation. In this way I found, for instance, that evolved rates of bystander learning are higher for individuals that are adapted to larger social group size (Table B1). The reason is

that, in a larger group, each individual has proportionally fewer interactions with a particular other individual, making bystander learning more important. Compared to previous game-theory models of social dominance (Mesterton-Gibbons, 1999; van Doorn et al., 2003a,b), the current model is closer in spirit to previous mechanistic models (e.g. Bonabeau et al., 1996, 1999; Dugatkin, 1997, 2001).

Perhaps the most striking result from the evolutionary analysis is the extremely high evolved preference learning rates (Figs. 1, 2, Table B1). High rates of learning, including so-called one-trial learning, have for instance been found in taste aversion learning (Garcia et al., 1966; Shettleworth, 2010). For taste aversion learning the suggested explanation is that the perceived consequence (e.g. severe nausea) from ingesting unsuitable food is highly salient to the individual, and this is adaptive because individuals need to quickly learn to avoid poisonous or otherwise unsuitable food. There is not one-trial learning in social dominance, but the explanation for high rates of learning could be similar. For individuals to learn to form dominance relations in a reasonable amount of time, in particular to learn to be submissive for either of two closely matched individuals, large changes in the preference for A from fairly small prediction errors could be needed, which might explain the phenomenon. The high preference learning rates in the model could correspond to special cognitive (neural) processing of dominance interactions for real animals, which is in accordance with recent experimental results on social dominance (Kumaran et al., 2016; Ligneul et al., 2016; Qu et al., 2017; Zhou et al., 2018). This possible link between social dominance and reinforcement learning is one reason to think that the model could be relevant to social animals.

The general approach to modelling of social interactions used here is to introduce plausible behavioural mechanisms, and to treat parameters of these mechanisms as traits that can be tuned by evolution. As discussed by McNamara and Leimar (2020), this approach stands in contrast to much traditional game theory in biology, where an individual's information about the environment, including about strategies

of social partners, is represented as probability distributions over unknown aspects of the environment. Traditionally in game theory, individuals use Bayesian updates, such as from prior to posterior distributions, to take into account information that is obtained. This can be referred to as a ‘small-worlds’ approach, because it is realistic mainly when an animal’s world, including the decision-making machinery of social partners, in some way is small and simple. The real world is a large world, especially for animals that interact and gain information about others over time. In such situations the traditional small-worlds approach becomes too challenging for modellers to achieve. More importantly, evolution is unlikely to have implemented an enormously complex decision-making machinery that properly performs the required Bayesian operations on probability distributions in high-dimensional spaces. Instead, evolution has led to ‘large-worlds’ adaptations, in the form of behavioural mechanisms that work reasonably well in a range of circumstances. From this point of view, game-theory models of social behaviour ought to implement mechanisms that have strong empirical support, such as those from animal psychology and neuroscience (McNamara and Leimar, 2020).

Social hierarchies could be an area for which the large-worlds approach is particularly suited, by allowing stronger contact between modelling and observation. Using this approach here I have been able to outline important aspects of social competence for social dominance behaviour. These include individual recognition, good memory of previous interactions, some but rather limited generalisation between opponents, very high rates of preference learning, and efficient bystander learning. In the end it will be experiments that decide whether the mechanisms in the model are realistic, or if they must be changed to correspond to real dominance hierarchy formation.

Acknowledgements

I thank John McNamara and Redouan Bshary for helpful comments. This work was supported by a grant (2018-03772) from the Swedish Research Council to the author.

Appendix A: Model details

This section contains further information and derivations, with the aim of explaining how the model relates to the broader treatment of reinforcement learning as presented by, e.g., Sutton and Barto (2018). As mentioned in the main text, the model considers the special case of only two actions, A and S. Another important simplification is that each round of interaction is assumed to be a so-called episode, meaning that there is no discounting of rewards over several time periods. Instead, the rewards (and penalties) from an interaction are determined by evolved characteristics, i.e. they are so-called primary rewards. This means that an individual does not learn that being dominant in an interaction is rewarding: the trait v_i is an innate, evolved characteristic of the individual. What an individual learns in the model is instead whether to use A or S in interactions with a particular other individual.

Each round of interaction then starts with an observation or state, consisting of the observation ξ_{ij} from equation (1) and the identity of the other individual. The model follows established knowledge about individual recognition (Tibbetts and Dale, 2007) in assuming that recognition is mediated by certain traits, for instance a pattern of colouration on the head or body of another individual. In order to introduce the crucial phenomenon of generalisation between stimuli, the model makes use of the general idea of function approximation for preference and value functions as developed in chapter 9 of Sutton and Barto (2018), in particular the linear methods. In this approach a state (i.e. another individual in the model) is represented using a number of features. The weights on the features of a state, such as the θ and w parameters in the model, express the required function approximation. Different states can share features to a greater or lesser degree, and a common approach is to use basis functions, such as Gaussian functions to express similarity between contin-

uously varying stimulus components (referred to as radial basis functions by Sutton and Barto, 2018). This is similar to how generalisation is often described in animal psychology (Ghirlanda and Enquist, 2003). The model simplifies generalisation into two categories of traits: those that are shared by all opponents and those that are unique to a particular opponent. The generalisation factor f_i in equations (3, 4) describes how strongly individual i takes into account the former in comparison to the latter when responding to opponents. The generalisation factor is assumed to be genetically determined, so the extent of generalisation can be tuned by evolution.

Actor-critic learning updates

The principle used to formulate learning updates for the actor-critic method (see chapter 13 in Sutton and Barto, 2018) is to modify parameters θ and w of preference and value functions such that the expected reward from an episode is increased, which in the model corresponds to the expected reward from a round of dominance interaction. The actor-critic method implements this as approximate gradient ascent and is a policy-gradient method, inspired by the so-called policy-gradient theorem (see chapter 13 of Sutton and Barto, 2018). The updates to the policy parameters θ involves derivatives of the logarithm of the probability of choosing an action with respect to the parameters. Using equation (2), one can introduce the notation

$$\zeta_{ijt} = \begin{cases} \frac{d \log \Pr(A)}{d l_{ijt}} = 1 - p_{ijt} & \text{if A was chosen} \\ \frac{d \log \Pr(S)}{d l_{ijt}} = -p_{ijt} & \text{if S was chosen} \end{cases} \quad (\text{A1})$$

for the derivative of the logarithm of the probability of choosing an action, A or S, with respect to the preference for A, which corresponds to equation (9). The term policy-gradient factor can be used for ζ_{it} , which is related to the concept of eligibility of the performed action for learning, as used in the study of actor-critic methods

(Barto et al., 1983; Sutton and Barto, 2018). From equation (3) it follows that

$$\begin{aligned}\frac{\partial l_{ijt}}{\partial \theta_{iit}} &= f_i \\ \frac{\partial l_{ijt}}{\partial \theta_{ijt}} &= 1 - f_i\end{aligned}\tag{A2}$$

and from this one obtains the learning updates of the θ parameters as in equation (8). The actor-critic method also specifies updates of the w parameters of the value function. From equation (4) it follows that

$$\begin{aligned}\frac{\partial \hat{v}_{ijt}}{\partial w_{iit}} &= f_i \\ \frac{\partial \hat{v}_{ijt}}{\partial w_{ijt}} &= 1 - f_i\end{aligned}\tag{A3}$$

and from this one obtains the learning updates of the w parameters as in equation (10). These learning updates correspond to the one-step actor-critic updates described in chapter 13 of Sutton and Barto (2018), applied to the model.

Appendix B: Additional results

Figure B1 complements Fig. 1 in the main text by showing the dynamics of the preference components over the entire 400 time steps of the interaction. As seen in Fig. B1, there are occasional and rather large jumps in the components. They occur when an individual happens to select another action than the one indicated by its current action preferences.

Figure B1: The panels (a) to (d) correspond to panels (e) to (h) in Fig. 1, but cover the entire 400 time steps of the dominance interactions and show a much greater range of preference values. The panels show the preference components h_{ij} , from equation (3), with the colour-coded solid lines giving the component h_{ij} for a specific opponent j , and the dotted lines showing the generalised component h_{ii} , with one panel for each individual i . The inserted legend also gives the individual quality values q_i . The big jumps in the preference components, as seen in the panels, occur when an individual ‘explores’ by choosing a different action (e.g. A instead of S). The parameter values for learning are as follows: $f_i = 0.19$, $\beta_i = 0.0$, $\alpha_{\theta i} = 175.9$, $\alpha_{wi} = 0.09$, $\theta_{0i} = 5.45$, $w_{0i} = 0.002$, $\gamma_{0i} = 2.57$, $g_{0i} = 0.025$, and $v_i = 0.15$.

Such deviations represent exploration by individuals, and they are important for reinforcement learning. The consequence of exploration is typically that the current behavioural tendencies of an individual become stronger, in the sense that the jumps

typically go in a direction that enhances an emerging dominance ranking.

Figure B2 shows a comparison like the one in Fig. 2 in the main text, but for two different cases, without bystander effects.

Figure B2: Influence of social competence on the learning about relative fighting ability and the decrease over time of aggression during hierarchy formation. There are no bystander effects ($\beta_i = 0$). Blue dots illustrate cases where the other learning parameters, including the generalisation parameter (f_i), are at an approximate evolutionary equilibrium. The grey dots show the same except that generalisation is fixed at its maximal value ($f_i = 1$), in effect preventing individual recognition, entailing very low social competence. Other details are as in Fig. 2. The learning parameter values are given in Table B1, cases 4 and 10 for panels (a) and (b), and cases 5 and 11 for panels (c) and (d).

For the grey dots in Fig. B2c,d (case 11 in Table B1), note that individuals do not learn to adjust their behaviour and thereby form a dominance hierarchy, but instead tend to use observations, as in equation (1), to settle interactions.

Table B1: Parameter values for 9 different cases of individual-based evolutionary simulations of social dominance interactions.

case	g_s	T	f_i	β_i	$\alpha_{\theta i}$	α_{wi}
1	4	400	0.19 ± 0.01	1.12 ± 0.27	175.9 ± 20.0	0.09 ± 0.01
2	8	800	0.16 ± 0.01	2.09 ± 0.32	428.2 ± 37.4	0.18 ± 0.02
3	16	1600	0.14 ± 0.01	3.26 ± 0.20	704.0 ± 25.6	0.31 ± 0.04
4	4	400	0.10 ± 0.02	0.0	110.6 ± 11.3	0.06 ± 0.01
5	8	800	0.07 ± 0.01	0.0	440.9 ± 42.1	0.11 ± 0.01
6	16	1600	0.06 ± 0.01	0.0	956.4 ± 27.7	0.21 ± 0.04
7	4	400	0.5	0.0	65.8 ± 3.8	0.21 ± 0.04
8	8	800	0.5	0.0	73.2 ± 5.0	0.37 ± 0.05
9	16	1600	0.5	0.0	74.3 ± 9.2	0.71 ± 0.09
10	4	400	1.0	0.0	23.8 ± 2.5	0.18 ± 0.04
11	8	800	1.0	0.0	21.8 ± 3.4	0.25 ± 0.08
12	16	1600	1.0	0.0	22.3 ± 2.8	0.34 ± 0.12

table continued

case	θ_{0i}	w_{0i}	γ_{0i}	g_{0i}	v_i
1	5.45 ± 0.21	0.002 ± 0.004	2.57 ± 0.20	0.03 ± 0.004	0.15 ± 0.005
2	5.98 ± 0.17	-0.01 ± 0.005	2.78 ± 0.20	0.03 ± 0.004	0.15 ± 0.004
3	5.67 ± 0.24	-0.04 ± 0.007	2.92 ± 0.33	0.03 ± 0.005	0.15 ± 0.004
4	4.99 ± 0.21	-0.01 ± 0.004	2.86 ± 0.27	0.03 ± 0.004	0.12 ± 0.005
5	5.05 ± 0.15	-0.03 ± 0.004	2.61 ± 0.21	0.02 ± 0.003	0.11 ± 0.005
6	4.52 ± 0.27	-0.06 ± 0.005	3.33 ± 0.69	0.02 ± 0.003	0.10 ± 0.007
7	3.60 ± 0.12	-0.01 ± 0.009	2.71 ± 0.16	0.03 ± 0.004	0.15 ± 0.002
8	2.63 ± 0.18	-0.06 ± 0.016	3.16 ± 0.22	0.04 ± 0.005	0.13 ± 0.003
9	-2.12 ± 0.44	-0.04 ± 0.08	12.59 ± 1.10	0.08 ± 0.012	0.10 ± 0.007
10	2.58 ± 0.49	-0.03 ± 0.034	8.58 ± 0.66	0.07 ± 0.009	0.09 ± 0.004
11	-2.87 ± 0.56	-0.02 ± 0.075	17.42 ± 1.96	0.09 ± 0.015	0.09 ± 0.007
12	-5.14 ± 0.31	0.05 ± 0.123	23.09 ± 1.19	0.11 ± 0.020	0.11 ± 0.012

References

- Abril-de-Abreu, R., Cruz, A. S., and Oliveira, R. F. 2015. Social dominance modulates eavesdropping in zebrafish. *Royal Society Open Science*, 2:150220.
- Ang, T. Z. and Manica, A. 2010. Unavoidable limits on group size in a body size-based linear hierarchy. *Behavioral Ecology*, 21(4):819–825.
- Bandura, A. 1986. *Social foundations of thought and action: a social cognitive theory*. Prentice-Hall.
- Barto, A. G., Sutton, R. S., and Anderson, C. W. 1983. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846.
- Bonabeau, E., Theraulaz, G., and Deneubourg, J. L. 1996. Mathematical model of self-organizing hierarchies in animal societies. *Bulletin of Mathematical Biology*, 58(4):661–717.
- Bonabeau, E., Theraulaz, G., and Deneubourg, J. L. 1999. Dominance orders in animal societies: The self-organization hypothesis revisited. *Bulletin of Mathematical Biology*, 61(4):727–757.
- Burke, C. J., Tobler, P. N., Baddeley, M., and Schultz, W. 2010. Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences*, 107(32):14431–14436.
- Buston, P. M. and Cant, M. A. 2006. A new perspective on size hierarchies in nature: Patterns, causes, and consequences. *Oecologia*, 149(2):362–372.
- Bürger, R. and Lande, R. 1994. On the distribution of the mean and variance of a quantitative trait under mutation-selection-drift balance. *Genetics*, 138(3):901–912.

- Chase, I. D., Bartolomeo, C., and Dugatkin, L. A. 1994. Aggressive interactions and inter-contest interval: how long do winners keep winning?. *Animal Behaviour*, 48(2):393–400.
- Chase, I. D., Tovey, C., Spangler-Martin, D., and Manfredonia, M. 2002. Individual differences versus social dynamics in the formation of animal dominance hierarchies. *Proceedings of the National Academy of Sciences*, 99(8):5744–5749.
- de Vries, H. 1995. An improved test of linearity in dominance hierarchies containing unknown or tied relationships. *Animal Behaviour*, 50(5):1375–1389.
- Dugatkin, L. A. 1997. Winner and loser effects and the structure of dominance hierarchies. *Behavioral Ecology*, 8(6):583–587.
- Dugatkin, L. A. 2001. Bystander effects and the structure of dominance hierarchies. *Behavioral Ecology*, 12(3):348–352.
- Earley, R. L. and Dugatkin, L. A. 2002. Eavesdropping on visual cues in green swordtail (*Xiphophorus helleri*) fights: A case for networking. *Proceedings of the Royal Society B*, 269(1494):943–952.
- Favati, A., Løvlie, H., and Leimar, O. 2017. Individual aggression, but not winner-loser effects, predicts social rank in male domestic fowl. *Behavioral Ecology*, 28(3):874–882.
- Fawcett, T. W. and Johnstone, R. A. 2010. Learning your own strength: winner and loser effects should change with age and experience. *Proceedings of the Royal Society B*, 277(1686):1427–1434.
- Garcia, J., Ervin, F. R., and Koelling, R. A. 1966. Learning with prolonged delay of reinforcement. *Psychonomic Science*, 5(3):121–122.
- Ghirlanda, S. and Enquist, M. 2003. A century of generalization. *Animal Behaviour*, 66(1):15–36.

- Goubault, M. and Decuignière, M. 2012. Previous experience and contest outcome: winner effects persist in absence of evident loser effects in a parasitoid wasp. *The American Naturalist*, 180(3):364–371.
- Hsu, Y., Earley, R. L., and Wolf, L. L. 2006. Modulation of aggressive behaviour by fighting experience: Mechanisms and contest outcomes. *Biological Reviews of the Cambridge Philosophical Society*, 81(1):33–74.
- Johnsson, J. I. and Åkerman, A. 1998. Watch and learn: preview of the fighting ability of opponents alters contest behaviour in rainbow trout. *Animal Behaviour*, 56(3):771–776.
- Kumaran, D., Banino, A., Blundell, C., Hassabis, D., and Dayan, P. 2016. Computations underlying social hierarchy learning: distinct neural mechanisms for updating and representing self-relevant information. *Neuron*, 92(5):1135–1147.
- Landau, H. G. 1951. On dominance relations and the structure of animal societies: I. Effect of inherent characteristics. *The Bulletin of Mathematical Biophysics*, 13(1):1–19.
- Ligneul, R., Obeso, I., Ruff, C. C., and Dreher, J.-C. 2016. Dynamical representation of dominance relationships in the human rostromedial prefrontal cortex. *Current Biology*, 26(23):3107–3115.
- Maia, T. V. 2010. Two-factor theory, the actor-critic model, and conditioned avoidance. *Learning & Behavior*, 38(1):50–67.
- McDonald, G. C., Spurgin, L. G., Fairfield, E. A., Richardson, D. S., and Pizzari, T. 2017. Pre- and postcopulatory sexual selection favor aggressive, young males in polyandrous groups of red junglefowl. *Evolution*, 71(6):1653–1669.
- McNamara, J. M. and Leimar, O. 2020. *Game Theory in Biology: Concepts and Frontiers*. Oxford University Press, Oxford (in press).

- Mesterton-Gibbons, M. 1999. On the evolution of pure winner and loser effects: A game-theoretic model. *Bulletin of Mathematical Biology*, 61(6):1151–1186.
- Mesterton-Gibbons, M., Dai, Y., and Goubault, M. 2016. Modeling the evolution of winner and loser effects: A survey and prospectus. *Mathematical Biosciences*, 274:33–44.
- Mowrer, O. H. 1951. Two-factor learning theory: summary and comment. *Psychological Review*, 58(5):350–354.
- Oliveira, R. F., Silva, J. F., and Simões, J. M. 2011. Fighting zebrafish: characterization of aggressive behavior and winner–loser effects. *Zebrafish*, 8(2):73–81.
- Olsson, A., Knapska, E., and Lindström, B. 2020. The neural and computational systems of social learning. *Nature Reviews Neuroscience*, 21(4):197–212.
- Pizzari, T. and McDonald, G. C. 2019. Sexual selection in socially-structured, polyandrous populations: Some insights from the fowl. In Naguib, M., Barrett, L., Healy, S. D., Podos, J., Simmons, L. W., and Zuk, M., editors, *Advances in the Study of Behavior*, volume 51, pages 77–141. Academic Press.
- Qu, C., Ligneul, R., Van der Henst, J.-B., and Dreher, J.-C. 2017. An integrative interdisciplinary perspective on social dominance hierarchies. *Trends in Cognitive Sciences*, 21(11):893–908.
- Rutte, C., Taborsky, M., and Brinkhof, M. W. G. 2006. What sets the odds of winning and losing? *Trends in Ecology & Evolution*, 21(1):16–21.
- Schjelderup-Ebbe, T. 1922. Beiträge zur Sozialpsychologie des Haushuhns. *Zeitschrift für Psychologie*, 88:225–252.
- Shettleworth, S. J. 2010. *Cognition, evolution, and behavior, second edition*. Oxford University Press, Oxford, UK.

- Sutton, R. S. and Barto, A. G. 2018. *Reinforcement learning: An introduction second edition*. MIT Press, Cambridge, MA.
- Taborsky, B. and Oliveira, R. F. 2012. Social competence: an evolutionary approach. *Trends in Ecology & Evolution*, 27(12):679–688.
- Tibbetts, E. A. and Dale, J. 2007. Individual recognition: it is good to be different. *Trends in Ecology & Evolution*, 22(10):529–537.
- van Doorn, G. S., Hengeveld, G. H., and Weissing, F. J. 2003a. The evolution of social dominance I : Two-player models. *Behaviour*, 140(10):1305–1332.
- van Doorn, G. S., Weissing, F. J., and Hengeveld, G. M. 2003b. The evolution of social dominance II: Multi-player models. *Behaviour*, 140(10):1333–1358.
- Zhou, T., Sandi, C., and Hu, H. 2018. Advances in understanding neural mechanisms of social dominance. *Current Opinion in Neurobiology*, 49:99–107.