**Homo Entropicus, THE EMOTIONAL AGENT AND SOCIETIES OF NEURAL NETWORKS**

A PREPRINT

Felipp Alves
falves@if.usp.br

Nestor Caticha
ncaticha@usp.br
Instituto de Física
Universidade de São Paulo
CP 66318, CEP 05315-970, São Paulo, SP, Brazil

October 20, 2021

**ABSTRACT**

A neural network with a learning algorithm optimized by information theory entropic dynamics is used to build an agent dubbed *Homo Entropicus*. The algorithm can be described at a macroscopic level in terms of aggregate variables interpretable as quantitative markers of proto-emotions. We use systems of such interacting neural networks to construct a framework for modeling societies that show complex emergent behavior. A few applications are presented to investigate the role the interactions of opinions about multidimensional issues and trust on the information source play on the state of the agent society. These include the case of a class of *N* agents learning from a fixed teacher; two dynamical agents; panels of three agents modeling the interactions that occur in decisions of the US Court of Appeals, where we quantify how politically biased are the agents, how trustful of other agents-judges of other parties, how much the agents follow a common understanding of the law. Finally we address under which conditions ideological polarization follows or precedes affective polarization in large societies and how simpler versions of the learning algorithm may change these relations.

**Keywords**: Agent Based Models, Entropic Dynamics, Neural Networks agents, distrust, affective and ideological polarization

1 Introduction

Jager [2017], inspired by Rosaria Conte, has made the case against keeping agents models too simple and “enhancing the realism of simulations” (EROS). Realism means psychological realism. Here we try to address the unneeded gap brought by extreme simplicity. A possible strategy could consider constructing models of agents which behave according to a minimal list of psychological traits. There is a need to include in the agent’s model some complexity that at least remotely resembles some crude emotions but not being able to quantify emotions *a priori*, we have to follow a different path. We don’t put emotions by hand, but they emerge as natural ingredients in the learning algorithms deduced from optimizing information processing capabilities of agents.

The agents interact by exchanging opinions represented by binary for/against options or by a continuous variable, about issues that are multidimensional. A natural choice is to let the agent be an adaptive neural network. The information exchange elicits learning and we describe optimized learning algorithms deduced from considerations of entropic dynamics in information theory. While mathematically intricate when described by microscopic variables living in possibly high dimensional spaces, there are simple aggregate or macroscopic variables that estimate what can be described in terms of words such as surprise, distrust of other agents, confidence in the opinions, performance
monitoring. These agents are capable of learning or reversal of learning. Estimating these quantitative proto-emotions markers, which emerge from impositions of better information processing, requires extra machinery (calculations) which can be thought as modules added in an evolutionary selection sequence that enhances information processing obtained by simpler algorithms. By damaging them and measuring performance in learning a fixed rule we can see how the modules influence each other, suggesting evolutionary paths towards the complexification of the information processing agent.

Homo entropicus, the agent formed by a neural network architecture and the learning algorithm, is constructed in order to optimally learn a fixed rule from information received from another agent, the emitter, in the form of examples pairs: an issue, parsed into an input array and an opinion about it. Thus its optimization occurs in the case of small number of agents, in parallel to our evolution in small societies in the savannas of Africa. This course of action is also related to our incapability to do the necessary mathematics to optimize in larger groups. Again, similarly to humans, we also will take Homo entropicus from the entropical savannas, its optimizing environment, and ask it to engage in interaction with large groups of similar agents.

The mathematical deduction of the agent’s learning algorithm, which is our main contribution, is presented very briefly, but we extensively discuss its interpretation in terms of macroscopic aggregate variables and their proposed role as proto-emotions quantifiers.

The interacting agents framework we have constructed is then applied to different situations, which we now describe. In order to warm up, we study a class of $N$ students learning form a fixed (non-dynamical) teacher. The dynamical evolution of distrust of the agents towards the static teacher, depends on initial values of the combined effect of previous knowledge about the subject matter and initial distrust towards the teacher. We don’t identify the nature, e.g. race, gender or dislike of subject matter, of the initial distrust. Here we start to see how distrust of the source leads to reversal learning and eventual negation of the source’s message.

Next we analyze the case of the US Courts of appeals in order to inquire as did Sunstein et al. [2006] whether judges act politically. This is an interesting question since judges are appointed by elected politicians that belong to either one of two parties. They have collected and analyzed data for the US federal judiciary, where panels of three judges interact and emit binary opinions about several types of cases. We simulate panels of three interacting agents and measure their opinions about synthetic cases. Different behaviors emerge as we consider the initial conditions, not only of distrust and their confidence about it, but about the initial knowledge about the law and about their political ideology. As discussed in Sunstein et al. [2006], judges turn out to act as if politically motivated, but their political bias, in addition to their own political party, depends on the other 2 judges in the panel. We find under which conditions our agents behave in a statistically analogous form. If our agents are to replicate the data, they have not only to trust judges of their same party, but need to extend this courtesy to judges of the other party. In addition, they have to be quite confident about their trust on judges on the opposite camp. For agents that trust and are certain about their trust, we can analyze the behavior when they don’t follow a common law, but were only politically motivated and also when they only follow the law without any political bias. We conclude that there is no evidence of those behaviors that can be gleaned from the data.

Polarization in societies, which is the last topic investigated, can come in different flavors as pointed out in Klar et al. [2018a, Iyengar et al. [2019], Jost et al. [2009]. In addition to empirical measurement, theoretical analysis of their cross influences is under intense debate. What are the mechanisms of interaction between affective and ideological polarization? Do humans dislike each other because they disagree on the issues or is it ideological differences that precede and induce dislike? We show that the complexity of the agenda under discussion has an influence on the order in which the polarizations occur. When many issues are being discussed ideological polarization precedes affective polarization, but for smaller agendas fast affective polarization is established before ideological polarization. The richness of the agenda also can induce long transients with glassy dynamics and persistent breakdown of balanced norms of trust [Heider [1958] Cartwright and Harari [1956] in triads of agents, as well as breakdown of balance of ideological alignment, which will be fully discussed elsewhere. Among other interesting topics that we will study elsewhere using our framework, we mention societies where some agents feel disgust or political disdain towards those that vocally express their political opinions, independently of their ideological alignment; agents with for, against or don’t know opinions; the crystallization of discrimination, the effect of different initial conditions of education on inequalities and the study of dictators in these toy societies.

### 2 Proto-emotions in neural networks

We could use a neural network with any architecture with a binary for/against, ±1 output, however while we want some elements of EROS, this should be tempered with a little KISS and thus we choose as a model for each agent the neural network of simplest architecture, the single layer perceptron with an optimized algorithm. Linearly separable models,
in some manner similar to the Rescorla-Wagner model [Rescorla and Wagner 1972] have been shown to be useful in describing human performance in several cases.

It is natural to consider online learning since opinions about issues are not considered in batches, but only one at a time. The deduction of the general algorithm can be found in Caticha [2020]. These methods can be applied, in principle, to more complex architectures, but it is sensible to start with this already rich case since, in addition to being analytically tractable, it offers the possibility of discussing it with high level aggregate variables interpretable as proto-emotions.

Questions that need to be addressed are: (a) What is it that we mean by a proto-emotion? We will answer this retrospectively, once we have an algorithm to analyze. (b) Can a neural network learn without proto-emotions? The answer is yes, very simple algorithms do indeed work, but (c) is the learning efficient? No. This is the reason proto-emotions arise in neural networks. Optimization of an algorithm depends on the context in which the NN is going to perform and thus different algorithms may result. For a given context, some variables, which will be associated to a proto-emotion, have to be estimated in order to determine the learning step size and direction of change in an optimal way. They inform modulation systems that set the correct scale of the changes.

The society evolves under a discrete time dynamics. At a given time, two agents are selected to interact and randomly assigned the emitter and receiver roles. From a set, that can be called the agenda, an issue/question is chosen. These are the prominent questions of policy which occupy the public attention. The emitter sends its opinion about the issue to the receiver, which changes its internal state using the learning algorithm. An agent’s internal state includes microscopic variables that determine the ideological opinion about the issue and others that estimate the level of certainty about the “correctness” or about how certain the agent is about this first set. These are called the variables of the ideological/opinion sector. Our agents have evolved in an environment where concealed cheating is a possibility, which means that agents have an advantage if they attribute to other agents a measurement of distrust. Defining what is trust in a human context is beyond our capabilities, see Yamagishi [2011] Shockley et al. [2016] for a discussion. We just stress that we mean something close to distrust in an individual and not mistrust of institutions. So in addition, there are variables that describe how distrustful the receiver agent is about the other emitter and also how certain it is about that distrust attribution. These are the variables of the affective/distrust sector. From a technical point of view, this means the agents are optimized to receive information through a noisy communication channel and distrust is the theoretical answer to the problem of dealing with corrupted data in a robust way. This topic has a long tradition in this area of inference in general and has been studied in several publications [Biehl et al. 1995] Copelli et al. [1996] Copelli et al. [1997] from the point of view of distrust of the information source. For neural networks agents it was studied by Alves and Caticha [2016], Caticha and Alves [2019].

There are several conceivable ways to choose the emitter/receiver pair of agents. Borrowing language from biology, Alves and Caticha [2016] use the term allopatric group formation, when an effective communication barrier is created by repeated disagreement and this pair of agents cease to interact, which permits the dynamical grouping with intra-faction interactions, but not inter-faction interactions. In allopatric group formation the probability of choosing an emitter depends on the distrust the receiver holds toward it, in addition to any other neighborhood constraints that want to be imposed. Alternatively, sympatric group formation is introduced as a process where agents keep interacting despite holding opposing views on the set of issues. The deduced dynamics of distrust permits sympatric group formation through the possibility of anti-learning, i.e learning the reverse of the emitter’s opinion.

The optimal learning algorithm in general is discussed in Caticha [2020]. For opinion and distrust is described below without any deduction details. The interested reader should consult Alves and Caticha [2021] and Alves [2020]. The output of the receiver agent’s neural network classifier at time $t$ is

$$\sigma_r = \text{sign}(\mathbf{w} \cdot \mathbf{x}_t).$$

where $\mathbf{x}_t$ is the issue under consideration, i.e. an assertion parsed into an array of numbers and $\mathbf{w}$ are part of the ideological sector variables. Both $\mathbf{x}_t$ and $\mathbf{w}$ have the same dimension $K$, which can be chosen as needed for a given application. For example, in modeling Haidt and collaborators [Haidt 2001] Moral Foundations theory with 5 moral foundations, Caticha and Vicente [2011] used $K = 5$. In this context, each component of $\mathbf{x}_t$ represents a quantitative estimate of the moral content of an assertion with respect to a given moral foundation. Each component of the agent ideological state $\mathbf{w}$ represents the importance of a foundation for that agent. From a Bayesian point of view, $\mathbf{w}$ is the location parameter of the distribution of probabilities of the weights. This distribution has a covariance matrix $C$, which will represent the uncertainty about the $\mathbf{w}$ variables, which will lead to the uncertainty about the opinion about an issue. For the distrust sector the incomplete information available to the agent is represented by a distribution and thus two distrust/affective variables are introduced, the expected value $\mu_{x|r}$, that takes value on the reals, and its standard deviation, $\sigma_{x|r}$. We make the convention that negatives values of distrust means that the emitter is trusted (liked) by the receiver. It is significant that scalar measures of uncertainty for the opinion and the affinity sectors $\gamma_c$ and $\gamma_v$.
respectively, can be defined and calculated. While technically important, semantically the details are not:

\[ \gamma_C = \sqrt{1 + x \cdot \mathbf{C}_t x}, \quad \gamma_V = \sqrt{1 + V_{e|r}}. \]  

(2)

These set the scale in a convenient manner so that the sectors can be compared during a receiver agent interaction. To do so, we need the ideological and affective fields for the receiver agent, interacting with the emitter that sends information \( \sigma_e \) about issue \( x \), defined by:

\[ h_w = \frac{\tilde{w} \cdot x \sigma_e}{\gamma_C}, \quad h_\mu = \frac{\mu_{e|r}}{\gamma_V}. \]

(3)

so that \( h_w > 0 \) the emitter and the receiver agree on the issue. If \( |h_w| \) is large, the receiver is very sure about its opinion. If \( h_\mu > 0 \) the receiver distrusts the emitter and if \( |h_\mu| \) is large, it is very confident about the distrust. Given these preliminaries, the interaction dynamics, leads to the new state of the receiver:

\[ \tilde{w}_{t+1} = \tilde{w}_t + \frac{1}{\gamma_C} F_w C_t x \sigma_e \]

(4)

\[ C_{t+1} = C_t + \frac{1}{\gamma_C} F_C C_t x x^T C_t \]

(5)

\[ \mu_{e|r}(t+1) = \mu_{e|r}(t) + \frac{1}{\gamma_V} F_m V_{e|r}(t) \]

(6)

\[ V_{e|r}(t+1) = V_{e|r}(t) + \frac{1}{\gamma_V} F_V V_{e|r}(t)^2 \]

(7)

Equation 4 describes the changes of the ideological sector variables. First note the term \( x \sigma_e \), this is called the Hebbian term in the Statistical Physics community of neural networks. If we forget the prefactor, it just adds to \( \tilde{w}_t \) the issue vector \( x \) if the emitter’s opinion is positive or the reversed sign issue \( -x \), if the emitter’s opinion is negative. This is the simplest learning algorithm, which when applied to learning a fixed rule, leads to a decay of the probability of disagreeing with the emitter, i.e. the generalization error, asymptotically proportional to \( 1/\sqrt{P} \) after learning \( P \) independent issues. This is an algorithm that treats every issue in the same manner, there are no emotions. With the full algorithm the behavior saturates theoretical bounds, with an asymptotic generalization error decaying as \( 1/P \). So, let’s not forget the whole prefactor. The Hebbian term is changed to \( C_t x \sigma_e \). This is called a tensorial update. The covariance matrix rotates and stretches the issue \( x \) in a non intuitive way. The covariance is changed, as described in 5, by the addition of a rank 1 matrix, formed by the vector modifying \( \tilde{w}_t \). These elements give a geometrical recipe to the update of the ideological sector variables.

Given this geometrical structure, the magnitude of the changes is modulated by the functions \( F_w \) and \( F_C \) which we discuss in detail. Since equations 6 and 7, which give the update of the mean and standard deviation of the affinity sector, also show modulations functions \( F_m \) and \( F_V \), we discuss them together. They are functions of the the ideological \( h_w \) and affective \( h_\mu \) fields given by equation 3.

\[ F_w(h_w, h_\mu) = (1 - 2\Phi(h_\mu)) \frac{g(h_w)}{Z}, \]

(8)

\[ F_C(h_w, h_\mu) = -F_w(F_w + h_w), \]

(9)

\[ F_\mu(h_w, h_\mu) = (1 - 2\Phi(h_w)) \frac{g(h_\mu)}{Z}, \]

(10)

\[ F_V(h_w, h_\mu) = -F_\mu(F_\mu + h_\mu), \]

(11)

where \( g(u) = (2\pi)^{-1/2}e^{-\frac{1}{2}u^2} \) is a standard Gaussian. \( Z \), called the Bayesian evidence for the model of the receiver, is

\[ Z = \Phi(h_w) + \Phi(h_{\mu|r}) - 2\Phi(h_w)\Phi(h_{\mu|r}) \]

(12)

and \( \Phi(s) = \int_{-\infty}^{\infty} du g(u) \) is the cumulative distribution of the Gaussian. Equations 8 to 11 can be used to see a symmetry between the two sectors, but only when appropriately scaled by the \( \gamma \)’s

\[ F_w(h_w, h_\mu) = F_\mu(h_\mu, h_w), \]

(13)

\[ F_C(h_w, h_\mu) = F_V(h_\mu, h_w). \]

(14)

These expressions, deduced in [Alves and Caticha 2021], are quite interesting and reasonable as we show next. For fixed distrust, the opinion sector has the modulation function obtained for learning optimally in the Tree Committee and the Parity machine neural networks [Copelli and Caticha 1995] [Simonetti and Caticha 1996] obtained by different mathematical arguments. While remarkable, this is a posteriori not surprising, in view of the mathematical equivalence of noise and hidden units in some machines [Copelli et al. 1996].
3 The modulation functions: What should be blamed for a surprise?

Figure 1: The modulation functions for the affective sector (left) and the opinion sector (right) in the space of agreement field $h_w$ and distrust field $h_\mu$. In the floor the contour plot of the evidence $Z$. Dark regions correspond to corroboration and light regions to surprises. Note that intense learning only occurs in the surprise regions and essentially only in the sector whose field has the smallest absolute value and it is thus blamed for the surprise. Negative values for $F_w$ mean reversal of learning.

Figure 2: Similar to figure 1 but for the uncertainty modulation functions. The uncertainty increases near the diagonal where $h_w \approx -h_\mu$. A surprise where the other sector is blamed results in a slight decrease in uncertainty.

As shown in figures 1 and 2, the modulation functions for the two sectors are only large in the surprise regions where $h_w h_\mu > 0$, the receiver disagrees with a trusted emitter ($h_\mu < 0$, $h_w < 0$) or agrees with a distrusted emitter ($h_\mu > 0$, $h_w > 0$). But this doesn’t mean that both sectors will be strongly modulated. There is a crossover region near the diagonal $h_w = h_\mu$, with a transition from blaming the opinion sector for the surprise, when $|h_w| < |h_\mu|$, to blaming the affective sector when $|h_w| > |h_\mu|$. We stress that the crossover is clearly seen only when the fields are correctly scaled (equation 3) so that their values can be compared. At the diagonal, the crossover of blame attribution leads to a strong signal increasing the uncertainties, not knowing which sector to blame for the surprise boost uncertainty about opinions and mistrust.

The scaling of equation 3 by describing the two sectors in comparable units, permits seeing this transition, which appears as very natural, at least in a non quantitative manner. However it is quite instructive to redraw the modulation function $\gamma_C F_w$ as a function of the raw stability field $h_{\sigma_e} = x \cdot \hat{w}_{\sigma_e}$, for different values of the uncertainty scale in the opinion sector, as in figure 5, which is known from the student-teacher scenario in Statistical Mechanics. 

\[ \text{[Kinouchi]} \]
Figure 3: Similar to figure [1]. The modulation functions are positive (respect. negative) in the red (blue) regions.

Figure 4: Blame attribution can be seen for the uncertainty modulation functions shown in figure [2]. The green line diagonal is the border of blame. The uncertainty increases near the diagonal where $h_w \approx h_\mu$. A surprise, when a sector is blamed, results in a slight decrease in uncertainty, shown in the light blue regions, since there is effective learning. Near the blame cross-over no update in the uncertainty occurs, although both sectors change their weights. After the blame cross-over, there is a large increase, near the diagonal, in the red-yellow regions.

In this case the student-receiver trusts the teacher-emitter with $\mu = -3$. Since the answer of the receiver is $\sigma_r = \text{sign}(h)$, $h$ measures how certain the student is in its opinion about the issue in the sense of how robust is a classification under small perturbations of the issue or of its weight vector. We can now see that for large uncertainties, the student is quite indifferent to whether it agrees or not with the teacher: the modulation under strong uncertainty is oblivious to the magnitude of the internal $h$ and the students treats information independently of its answer. If the student is very uncertain about the rule, an error is not surprising. As learning of the teacher’s rule proceeds and the uncertainty decreases, concurring examples will elicit smaller changes, but disagreement ($h < 0$) generates a larger scale modulation. This surprise persists until there is the cross-over to blaming the teacher for the error and the student remains unchanged. For these agents a surprise only occurs if there is some certainty about a prediction. Technically the neural network has gone from learning by correlations in the initial stages to learning by strong error correction, modulated by a measure of the surprise, with a protection against a drunk teacher. This optimal rule learning scheme depends not only on the Hebbian correlation term ($x_\sigma_\epsilon$), but on extra machinery that can be thought of as added hardware, by hand or by evolution, if these machines were under pressure to generalize better. One new module to measure surprises and hence correct errors, another to estimate performance in order to estimate uncertainty. It is interesting that measuring only surprise and not performance leads to a better generalization than the
Homo Entropicus, the emotional agent and societies of Neural Networks

4 Societies of Neural Networks

See [Metzler et al., 2000] for early work on interacting perceptrons. Our model is already sufficiently complex and that some empirical facts can be understood doing this intermediate step justifies that we don’t look directly at richer agent models. Nevertheless, our claim is that the natural way to construct realistic agents that decide and influence others is based on information processing machines, which find their current paradigm in neural networks. In the following examples we study first a class of $N$ students learning form a teacher, which remains fixed. The students don’t talk to each other. This is not a realistic but, a necessary step to understand the behavior in simple cases. This is a warm up for a more realistic application, where three agents model a panel of judges. What makes it interesting is the availability of data, collected by [Sunstein et al., 2006]. It permits quantification of the political bias and the degree to which judges, collectively follow the law. In the future, this will be extended to the analysis of specific US appeal courts. We work with US data for two reasons, the availability of the data and the simplicity of analysis brought forward by their simple two party culture. A preliminary report appeared in [Caticha and Alves, 2019]. The last application discusses how the different sectors of our agents lead to introducing two types of polarization and the breakdown of balanced affections and ideologies, contributing to modeling the current discussion on affective versus ideological polarization, studied by the political science community [Klar et al., 2018b], [Iyengar et al., 2019] by identifying conditions where one or the other polarization is established first.

4.1 A class of $N$ students and one teacher

In this section we show an example from a student-teacher scenario. At every step of the dynamics one of $N$ students is uniformly chosen independently of anything else and it learns the opinion of a fixed teacher about a vector randomly chosen from the learning set. We see that some student agents learn and other don’t, despite having similar previous knowledge of a subject. In order to build into the model initial conditions that describe the agent we proceed as follows. First all vectors that appear on the right hand side of the expressions below have unit length. The ones that appear at the right are also normalized to one. The teacher or emitter would like to teach a curricular core vector $V_c$. Students will be graded along their dynamics based on their similarity to $V_c$: the overlap $\rho_i = w_i \cdot V_c / |w_i|$. The learning set has $P$ vectors $V_{a}$. Each $V_{a}$ is constructed by adding to $V_c$ a neutral contribution $N$ orthogonal to $V_c$ and a random vector $\eta_{V}$:

$$V_{a} \propto N + \alpha_{c} V_c + \eta_{a}.$$
Homo Entropicus, the emotional agent and societies of Neural Networks

To model diversity in the student’s previous knowledge about the subject being taught, their initial weight vectors receive a contribution $v_\theta$ of a random vector making an angle $\theta$ with $V_c$. A random vector $\eta_i$ of unit length is added to model peculiarities of each student:

$$w_i(t=0) \propto v_\theta + \alpha_P \eta_i,$$

where $\alpha_P$ measures the relative importance of such contributions. The teacher is fixed at:

$$w_T \propto \eta + \alpha_T V_c.$$  \hfill (17)

Figure 6: An example of a simulation of class of $N = 300$ student agents interacting with a teacher network. Color coded by ranking of final grade. Color goes from violet to red as overlap of student-teacher decreases from high to low values. Trajectories for every student in the $\varepsilon, \rho$ space. Left: $P = 1$, right: $P = 300$.

Figure 7: Same case as in left side of figure 6. Time evolution of distribution of grades for initial high distrust ($\varepsilon > 1/2$, left), low distrust ($\varepsilon < 1/2$, right). As time increases the inequality of the distribution of grades stems from the initial distrust of the teacher.
The parameters $\alpha_V$ and $\alpha_T$ control the importance of the subject on the issues being discussed and on the teacher respectively. The choice of the initial values of the set of distrusts $\{\mu_r\}$ define the initial distrusts

$$\varepsilon_{e|r} = \Phi(\mu_{e|r}),$$

which are chosen iid uniformly in the interval $(0, 1)$. The informational interpretation of $\varepsilon$ is that it is the probability of the "correct" answer being flipped in the noisy channel. If $\mu_{e|r} < 0$ (respectively $> 0$) then $\varepsilon < 1/2$ ($> 1/2$) and the information/source/teacher is initially trusted (distrusted) by the student.

Results from a simulation in figure 6 show typical flows in the space $(\rho, \varepsilon)$, with an unstable hyperbolic fixed point at $\rho = 0, \varepsilon = 1/2$. The effect of changing the number of issues $P$ is simply to go from quite smooth hyperbolic trajectories for one example, to a more disordered flux as $P$ is increased to 300. Regardless of $P$, we can clearly see that the class polarizes into two groups. Students with good grades and small distrust in one group and bad grades and large distrust for the teacher in the other. Interestingly both the certainty about the answers and for the distrust increase in average, students become more sure about their opinions independently of their grades. This is of course due to the crossover blame attribution that emerges from the symmetry of the modulation functions. The expected grades histograms as a function of time is shown in figure 7. The students are separated into groups according to initial distrust of the teacher and the histograms show the evolution in time of the class. Grades are reported on a 0 ($\rho = -1$) to 10 ($\rho = 1$) scale.

From the point of view of the interpretation it is interesting that a strong distrust for a teacher can lead an initially capable student to have a terrible final performance. Symmetrically, an initially very bad student can achieve good grades if the initial distrust is weak. These scenario can also be interpreted as a model for the development of negation despite receiving positive information. We will present elsewhere applications to climate and vaccine negationism, as well as differences of performance in testing by different ethnic groups. It is clear that we can model priming the students by boosting the initial distrust about the information source in either way. Variations on the theme, with different parameters and conditions, will be studied elsewhere.

5 US court of appeals: shared knowledge of the law and political bias.

Three member panels of USA Courts of Appeals judges presents a good opportunity to apply our model to the three agent case, since besides its intrinsic interest, there is data Sunstein et al. [2006] kindly shared by A. Sawicki. They investigated the influence on decisions of political party alliances, i.e the party that appointed them. We construct statistical signatures of behaviors that can be measured in both the theoretical and real systems and simulate the model for different sets of initial conditions scenarios. The scenarios can be manipulated by changing (a) the magnitude of the influence of the Law that is common to all agents, (b) the magnitude of the party’s influence on the initial state of an agent, (c) the relative contribution of its idiosyncratic component, see figure 8, and (d) the initial attribution of distrust and its uncertainty about judges of the same and of the other party. Preliminary results can be found in Caticha and Alves [2019].

Figure 8: Low dimensional representation of the initial states of the agents, $w_0^A$ and $w_0^B$ made up by three contributions. A vector $L$ (Law) is common to all agents. The direction of the party contribution depends on the party being $A$ or $B$. A random contribution that is unique to each agent models the idiosyncratic part of a judge.
From Sunstein et al. [2006] (Figure 2-2. Voting Patterns for Case Types with Both Party and Panel Effects) we obtain the empirical mean percentages of liberal decisions by judges in three member panels, under different conditions of panel composition for each of the 15 areas of case types ranging from very ideological to not ideological. Judges, who are appointed by an executive officer, are labeled by the officer’s party. When focusing on the decisions of a republican indicated judge in a panel of two republicans and one democrat, we use the notation \( v = Rrd \), capitalizing the initial of the party of the judge under observation. There are six different types of votes: \( Rrr, Rrd, Rdd, Drr, Drd, Ddd \). The data supports their three working hypotheses, that there is (i) Ideological voting: “Republican appointees vote very differently from Democratic appointees”; (ii) Ideological dampening: a judge in the minority party of a panel will be less ideological; and (iii) Ideological amplification: a judge in a pure party panel will be more ideological. Hence they are describing the interactions of the judges in the panel.

We represent their data as a set of 15 dimensional vectors \( J_v \), one vector for each \( v \). A component of \( J_v \) is the percentage above 50% a judge in a panel \( v \) votes in favor of the liberal position for a particular type of case. The angles \( \theta(v, v') \) between these vectors measure differences between judges in different panels. For instance \( \theta(Rrr, Ddd) \) measures the difference between Republicans and Democrats in pure panels, hence permitting to assess hypothesis (i) of Ideological voting. Comparing \( \theta(Ddd, Rdd) \) and \( \theta(Ddd, Drr) \) informs about how liberal is a Republican sitting with two democrats and whether it is more so than a Democrat sitting with two republicans, hence probing hypothesis (ii). The angle \( \theta(Rdd, Rrr) \) measures the differences of judges in the minority or in a pure panel, relevant for hypothesis (iii). The angles \( \theta(Rrr, Rrd) \) or \( \theta(Drr, Drd) \) inform about the differences that occur in panels where a companion judge from one party is changed to the other party.

The introduction of the arrays \( J_v \) and the angles between them gives a quantitative meaning to the idea of alignment of views. The main reason to introduce \( J_v \) is that it can be constructed from readily observable quantities. We can define angles between the vectors that represent the state of the agents \( w \), but these are not empirically available for the judges since \( w \) states are only indirectly hinted from voting patterns.

For the model we consider a two parties (A and B) system. Three adaptive agents, each representing a judge, interact by exchanging their opinions about a particular issue to be judged. The initial state \( w_{ij}(t = 0) \), where \( I = A \) (respect. \( I = B \)) for agents appointed by party A (respect. by party B), of an agent at the beginning of a discussion reflects three main characteristics the judges ought to have: a common knowledge of the law; shouldn’t have: an ideological bias that depends on the political party \( I \) of the executive officer that made the appointment; and simply have: an idiosyncratic random characteristic. Then

\[
\begin{align*}
w_{ij}(t = 0) &= \alpha_L L + \alpha_P P + \alpha_I I,
\end{align*}
\]

as illustrated in figure 8. The first term \( L \) represents knowledge of the Law, common to all agents. If this were the only term, agents would have identical opinions on every issue. The second term \( P \) represents ideological party lines, perpendicular to \( L \). The plus and minus signs indicates an agent appointed by party A or B, respectively. The third term \( I \) is a vector independently chosen uniformly at random on the unit sphere in \( K \) dimensions for each agent. The parameters \( \alpha_L, \alpha_P \) and \( \alpha_I \) control the relative importance of the Law, the party and personality of the agents. The cases \( \alpha_L = 0 \) can be called Lawless cases and if \( \alpha_P = 0 \) party-less cases.

For the initial distrust attribution we consider four different scenarios, see table I. By courteous we mean initially the agents, independently of party, extend the courtesy of attributing a low distrust value of \( \mu_{a|a} = \mu_{b|a} = \mu_{a|b} = \mu_{b|b} = -1 \) to each other. Discourteous means that agents initially trust same party agents \( \mu_{a|a} = \mu_{b|b} = -1 \) and distrust agents from the other party \( \mu_{a|b} = \mu_{b|a} = 1 \). A low \( V_0 = 0.1 \) means the agents are quite certain of an attribution of distrust and a high value \( V_0 = 5 \) that agents are uncertain of the attribution.

| Distrust Scenarios     | Courteous | Discourteous |
|------------------------|-----------|--------------|
| Certain                | \( \mu_{a|a} = -1, V_0 = 0.1 \) | \( \mu_{a|a} = 1, V_0 = 0.1 \) |
| Uncertain              | \( \mu_{a|a} = -1, V_0 = 5.0 \) | \( \mu_{a|a} = 1, V_0 = 5.0 \) |

Table 1: Four different distrust initial conditions scenarios were considered. \( \mu_{a|b} \) refers to agents of different parties. For all pairs of agents of the same party \( \mu = -1 \) and \( V_0 = 0.1 \)

An issue, characterized by its angle \( \phi \) with the Law vector, is chosen and the judges engage in the exchange of opinions. We repeat this for \( n_{case} = 15 \) different \( \phi \) angles, taken uniformly in the interval \([-\pi, \pi] \). This is repeated for a few hundred sets of initial conditions. For each run we record the voting patterns, and the averages are used to construct

1The data is from rulings in 15 areas of the law: Affirmative action, NEPA, 11th Amendment, NLRB, Sex discrimination, ADA, Campaign Finance, Piercing corporate veil, EPA, Obscenity, Title VII, Desegregation, FCC, Contract Clause, Commercial speech.
the $n_{case}$ dimensional $J_v$ vectors. Then, repeat for all $v$ environments. For simplicity, we froze the $C$ dynamics at $C_{ab} = 2\delta_{ab}$ since it has been shown to be related to cognitive style, which we expect to be frozen when the agents enter adulthood [Caticha and Vicente 2011], [Caticha et al. 2015].

Figure 9: Voting alignment as represented by the matrix of angles $\theta_{vv'}/\pi$ between vectors $J_v$ and $J_{v'}$. Initial values $\alpha = (1, 1, 1)$. Top: four different initial conditions for the affective sector lead to the steady state figures, which color code for the angles between agent-judges in the possible panel compositions. Far right: the same representation of the empirical data for human-judges. Bottom: quantitative difference between the different scenarios and the empirical data. Visually and quantitatively, the courteous/certain scenario are closest to the empirical data.

Both from the empirical data and from the simulations we can construct vectors $J_v$ for each of the six types of votes. The angles between two vectors indicate how two agents (or how two judges) are aligned in their views. Similar voting patterns will result in small angles.

In figure 9, we present, the angles between the vectors $J_v$ obtained from the voting patterns of the agents in the four different trust initial conditions, see Table 1 and the same statistic for the judges obtained from the empirical data. In this color map representation of the matrix of angles $\theta_{(v,v')}$ between the vectors $J_v$ and $J_{v'}$, the light entries represent large angles and different voting patterns, while dark colors mean small angles or very aligned voting patterns. For these simulations the original agents weights are generated with $\alpha_L = \alpha_P = \alpha_N = 1$. In general we see that party A and party B appointees agents are very different. It is clear that the courteous-certain scenario (4th from left, figure 9) is qualitatively much closer to the empirical data (far right image). The first three images are obviously visually different from the empirical data. This conclusions are quantitatively supported by the mean quadratic error for over all panel pairs, shown in the bottom of figure 9 for the four scenarios. The conclusion is that the model behaves similarly to the Appellate Courts only if the agents trust each other and are quite certain about this trust at the beginning of the interactions.

The courteous-certain scenario is summarized in table 2. Even in this optimistic scenario, we see that judges appointed by party A behave differently from those appointed by party B, hence we see evidence of ideological voting, a reminiscent behavior of the party dependent initial conditions.

We also see evidence of Ideological dampening, for example $\theta(Rrr, Ddd) > \theta(Rrd, Ddd)$ meaning that the difference between a Republican in a pure republican panel and a Democrat in a pure democratic panel is larger that of the same Democrat and a Republican who is interacting with one republican and one democrat. Interestingly a Republican in the presence of two democrats acts as more liberal that a Democrat in the presence of two republicans. The same results hold if we change R (and D) for A (and B) in the courteous-certain scenario. Also Democrats in the company of one republican and another democrat are more similar to Republicans in the presence of two democrats than to Democrats accompanied by two republicans, which again holds for agents.

Furthermore, we can ask how panels of purely ideological or party-less judges would look, by making $\alpha_L \gg \alpha_P \geq 0$, and at purely non ideological judges or law-less, by taking $0 \leq \alpha_L \ll \alpha_P$. The results, shown in figure 10, show that judges are neither totally ignorant of the law nor free from ideological bias.
Figure 10: Same as figure [2]. Final states of voting alignment \( \theta_{vv'}/\pi \), for agent/judges with, at left no initial component of the law \( \alpha = (0, 1, 1) \) and, center, no initial political component, \( \alpha = (1, 0, 1) \). Right: Empirical values. Neither initial condition leads to anything similar to the empirical data. Sadly, judges are political, but they show some common understanding of the law.

With this tool, we obtain a similarity between the empirical signatures and the model signatures by making the choices that the agents are (i’) quite courteous towards those of another party and certain about this trust, (ii’) that there must be a sizable contribution of the common vector \( L \) to the initial conditions, (iii’) that there is also a sizable contribution of the party bias \( P \).

A better numerical fit could be done by changing the values of the \( \alpha \) parameters, but we restrain from trying to read into the model a more realistic replication of the system than what should at this point be considered reasonable. Quantitative evaluation of the components of the Law and the Party can be done for individual courts and is currently under study.

| Hypothesis          | Example                  | Interpretation                                      |
|---------------------|--------------------------|-----------------------------------------------------|
| (i) ideological voting | \( \theta_{Aaa,Bbb} \) is the largest angle | Largest differences between \( Rrr \) and \( Ddd \). |
| (ii) ideological dampening   | \( \theta_{Baa,Aaa} < \theta_{Abb,Aaa} \) \( \theta_{Abb,Bbb} < \theta_{Baa,Bbb} \) | \( D \) in \( Drr \) is more conservative than \( R \) in \( Rdd \). |
| (iii) ideological amplification | \( \theta_{Bbb,Bab} < \theta_{Bbb,Baa} \) \( \theta_{Aaa,Aab} < \theta_{Aaa,Abb} \) | \( D \) in \( Drd \) is more liberal than \( D \) in \( Drr \). |

Table 2: Some of the evidence for the three hypotheses in the Courteous and Certain scenario. See figure [9] for the simulation and for empirical data obtained from Sunstein et al. [2006].

6 A society of \( N \) agents

Our next topic is a society of \( N \) agents, interacting on the full graph, exchanging \( \pm 1 \) opinions about \( P \) issues \( x \) in \( K \) dimensions, with components independently and randomly chosen with standard Gaussian distributions and then normalizing to \( |x| = 1 \). As in the case of two agents a variety of behaviors may occur and by no effort was done to exhaust the possible interesting cases.
Figure 11: $B_A$ versus $B_I$ for different sizes of the agenda, as a function of time $\tau$ measured in units of the number of degrees of freedom of the society $N(N + K - 1)$. On the left, agents use the full covariance (tensorial update). For small agendas the affective polarization sets in before the ideological polarization. This is reversed for larger agendas. Right: agents use the simple covariance proportional to the identity matrix ($C = cI$, vector update). For small agendas the system lingers in a disordered ideological state without polarization, but reestablishes affective balance and therefore polarization. For large agendas the system remains in the glassy unbalanced situation in both sectors.

We are interested in the polarization of the society into factions and see that two possible types occur, grouping by affinity of ideology or by affinity of affection. These are also discussed in the literature of political science [Klar et al. (2018a), Iyengar et al. (2019)] and are related to the problem of current interest: Do people disagree because they hate each other or hate each other because they disagree? Interestingly, the answer is it depends. The dynamical process for each sector can have different characteristic times depending on the “complexity” of the agenda under discussion measured in a simple manner by $P$.

In addition to polarization we probe into the breakdown or establishment of balanced norms of trust [Heider (1958), Cartwright and Harari (1956)] and balance of opinions in triads of agents. With the same definitions as before for $\varepsilon_{jil} = \Phi(h_{ij})$, define $\nu_{jil} = 1 - 2\varepsilon_{jil}$, which is positive/negative for trust/distrust. For any three agents [Heider (1958)], the trust relation is balanced if $b_{ijk} := \nu_{jil}\nu_{kji}\nu_{lij} > 0$. This is also studied in physics of disorder materials and receives the name frustration if negative. Since these are directed graphs and the odd permutation $b_{ikj}$ may be of opposite sign, consider an odd permutation of indices as a different triplet. Then the affective balance, the average value over the population

$$B_A = \frac{1}{2N_T} \sum_{(ijk)} (b_{ijk} + b_{ikj}),$$

(20)

informs the state of affective frustration of a society. $N_T$ is the number of triplets. A second type of frustration can be studied, related to the opinion alignment measured by the symmetric overlap $\rho_{ij}$. Opinion balance, which characterizes
the state of the society is defined by the average over the population

\[ B_I = \frac{1}{N_T} \sum_{(i,j,k)} \rho_{ij} \rho_{jk} \rho_{ki}. \] (21)

An ordered society (approximate consensus) or a polarized society will have \( B_I \) and \( B_A \) close to one but these parameters are close to zero for a highly frustrated society. These are seen in figure 11 where we show the time progression of the balances for two different societies, according to the type of learning algorithm the agents use. For agents using the complete EDNNA algorithm, with a full covariance \( C \), the result depends on the complexity of the agenda. For a simple agenda the society polarizes rapidly into two factions in the affective sector and then the affective polarization drives the ideological polarization. These agents “hate” each other before they disagree on the issues. But with time they come to believe different things, also polarizing in the opinion sector, for no other reason than they learn from whom they like and have reversed learning from those they don’t. For more complex agendas the ideological polarization sets in before and then drives the affinity polarization. This is seen in the dark lines, which run below the diagonal in figure 11-Left. The time to achieve a balanced society increases as \( P \) increases. Times \( \tau \) measures the number of learning interactions per degrees of freedom of the system (\( \varphi = N(N + K - 1) \)) so that we can compare simulations with different parameters.

On the right side of figure 11, we show the same graphs but now the agents use the simpler algorithm obtained with the covariance \( C = cI \), so the geometric Hebbian term is not distorted, but simply scaled by an overall uncertainty scalar \( c \). This dependence on the agenda disappears when the agents use this simpler algorithm. For them affinity polarization always seem to occur first and with slow behavior annealing form the glassy state to the ideological polarized state.

7 Discussion and Conclusions

There is an understandable appeal for the use of simple models of agents and the intuitive collective properties that their societies show. However there is a tacit understanding that psychological elements should inform any attempt to increase an agent’s realism. Emotions have to enter the mathematical models. When studying the learning models here exposed, we were not trying to include any type of emotions into the final results. Statistical Physics doesn’t deal with emotions. However, the modern formulation of Statistical Physics, shows it to be a framework for analyzing systems that attributes probabilities to the different states in a manner that tries to be as ignorant as possible while respecting experimental or informational constraints, thus refraining from incorporating unwarranted evidence. It is thus cast as an example of Information theory. It is not about the physical reality itself of a system but what we can infer about it, based on the available information. The analysis of the learning algorithm, although in a simple way, finds variables that have a flavor of proto-emotions, primitive precursors of what might be called emotions. We can probe the agent and identify that if a dissonance occurs, it elicits changes in at least one of the sectors. Brains evolved under a myriad environmental pressures, some of them related to efficient information processing. Our agents are optimized by adhering to entropic dynamics and thus process information in their simple world in an optimal way. It is plausible to roughly describe the parallel as metaphoric or to ascribe it to general principles of information processing. It allows to speculate on some of the reasons of the origin of emotions from an evolutionary point of view: without emotion modulation, learning is deficient. This is totally different from explaining why we feel a surprise, which is not our aim, which needs a totally different set of theoretical tools that we don’t have. We clearly remained tied to a third person perspective of the description of emotions.

The entropic dynamics for neural networks architectures, EDNNA analysis provides one way to obtain such optimized learning algorithm. But they can be obtained with much more labor as the result of an evolutionary process, similar to that shown in [Neirotti and Caticha (2003)], which deals only with the opinion sector. Of course we have dealt with a small group scenario, that of rule to be learned and followed by members of this group, leading to the opinion sector and a mechanism to emit opinions about issues. Allowing for cheating by the emitters results in a defense mechanism that appears in the assignment of a level of distrust to other members of the group. Once the two agent interaction is defined, we construct a society.

Since Bayesian probability is simply the name of probability theory from an information point of view, we stress that this not the Bayesian algorithm. There is no such thing. Different information scenarios lead to different Bayesian algorithms. Failure to understand this may suggest that humans can’t be described by Bayesian modeling. The problem in modeling humans is not that information theory does not apply but that the informational structure is poorly determined in the model. The simpler covariance model results in figure 11-Right, where the covariance is just a multiple of the identity, leads to algorithms that are still just as Bayesian, but under a different set of constraints.

Statistical description of the behavior of the judiciary agents, such as [Sunstein et al. (2006)] holds the promise of identifying extreme and maybe intolerable political biases. We have restrained from trying to make precise estimates of
the amount of political bias or the contribution of the law, trying only to present a strategy to future quantification of the individual courts. It remains clear that without some amount of political bias and without some amount of common law, the agents behave very differently from human judges. It may come as a welcome surprise that the model indicates that only agents that trust agents from not only their party, but the opposing one, and in addition are quite certain of that, have some similarity to the statistical behavior of the human judges.

Questionnaires are a central method in the humanities to gather information and these neural network agents are capable to include such scenarios [Caticha and Vicente [2011]].

Lincoln said [Lincoln [1953]] in 1848, “The process is this: Three, four or half a dozen questions are prominent at a given time, the party selects its candidate, and he takes his position on each of these questions.” Then opposing parties can form around polarized simple ideological agendas, in his case, the prominent issue under discussion was whether the new states would be free from slavery. For the agents, affective affinities establish quickly and proceed to push on a slower time scale to shared opinions and its consequent polarization around ideologies. As the agenda becomes more complex, larger $P$, there is a reversal. The ideological polarization occurs first, driving the slower affective polarization. The reduction of times to polarize that accompanies the reduction of the agenda is in accordance with examples discussed in [Fiorina and Abrams [2008]] where rapid partisan polarization is not accompanied by rapid changes in positions on economic policy issues. But this doesn’t happen for the simpler scalar covariance agents. Affective polarization sets in first and ideological polarization comes later at a much slower scale.

We have used distrust as a portmanteau for several associated concepts and a refined analysis is beyond our capacity, see [Yamagishi [2011]] Shockley et al. [2016] for a extended treatment of these subjects.

The neural networks we use are certainly simple and limited, despite the rich learning algorithms they use. The EDNNA formalism applies to any architecture, however it becomes intractable fast with architecture complexity. We believe that with agents with more complex architectures and their appropriate EDNNA learning algorithms, and a larger set of microscopic variables, will display more characteristics that could pose as psychological traits. While this might be true, it would be difficult to identify them and even harder to interpret them without the acquired experience from the simple case architecture. To conclude we stress that proto-emotion markers appear as necessary for efficient learning.

Acknowledgment: We thank A. Caticha, O. Kinouchi, R. Vicente, M. Copelli, JP Neirotti for discussions. We thank A Martins for pointing out references about ABMs during FA's PhD defense. FA received financial support from a Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) PhD fellowship. This work received partial support from CNAIPS-NAP USP.

References

Wander Jager. Enhancing the realism of simulation (eros): On implementing and developing psychological theory in social simulation. Journal of Artificial Societies and Social Simulation, 20(3):14, 2017. ISSN 1460-7425. doi:10.18564/jasss.3522 URL http://jasss.soc.surrey.ac.uk/20/3/14.html

C R Sunstein, D Schkade, L Ellman, and A Sawicki. Are judges political?: an empirical analysis of the federal judiciary. Brookings Institution Press, Washington, D.C., 2006.

Samara Klar, Yanna Krupnikov, and John Barry Ryan. Affective Polarization or Partisan Disdain?: Untangling a Dislike for the Opposing Party from a Dislike of Partisanship. Public Opinion Quarterly, 82(2):379–390, 05 2018a.

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. The origins and consequences of affective polarization in the united states. Annual Review of Political Science, 22(1):129–146, 2019. doi:10.1146/annurev-polisci-051117-073034

JT Jost, CM Federico, and JL Napier. Political ideology: its structure, functions, and elective affinities. Annual review of psychology, 60:307–37, January 2009.

F. Heider. The Psychology of Interpersonal Relations. J. Wiley and Sons, New York, 1958.

D. Cartwright and F. Harari. Structural balance: a generalization of heider’s theory. The Psychological Review, 63, 1956.

RA Rescorla and AR Wagner. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Prokasy WF Black AH, editor, Classical Conditioning II: Current Research and Theory, pages 64–99. New York: Appleton Century Crofts, 1972.

Nestor Caticha. Entropic dynamics in neural networks, the renormalization group and the hamilton-jacobi-bellman equation. Entropy, 22, 2020.

Toshio Yamagishi. Trust: The Evolutionary Game of Mind and Society. Springer, 2011.
E Shockley, TMS Neal, LM PytlikZillig, and BH Bornstein (Eds.). *Interdisciplinary Perspectives on Trust Towards Theoretical and Methodological Integration*. Springer, 2016.

M Biehl, P Riegler, and M Stechert. Learning from noisy data: an exactly solvable model. *Physical Review E*, 52(5): R4624–R4626, 1995. ISSN 1550-2376.

Mauro Copelli, Osame Kinouchi, and Nestor Caticha. Equivalence between learning in noisy perceptrons and tree committee machines. *Phys. Rev. E*, 53:6341–6352, Jun 1996. doi:10.1103/PhysRevE.53.6341 URL https://link.aps.org/doi/10.1103/PhysRevE.53.6341

M. Copelli, R. Eichhorn, O. Kinouchi, M. Biehl, R. Simonetti, P. Riegler, and N. Caticha. Noise robustness in multilayer neural networks. *Europhysics Letters*, 37:427–432, 1997.

Felippe Alves and Nestor Caticha. Sympatric multiculturalism in opinion models. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering 35*, volume 1757, page 060005. AIP Publishing LLC, 2016.

Nestor Caticha and Felippe Alves. Trust, law and ideology in a neural network agent model of the US-Appellate Courts. *ESANN 2019 - Proceedings 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2019.

Felippe Alves and Nestor Caticha. Frustration, glassy behavior and dynamical annealing in societies of neural networks, arxiv 2108.06571 physics.soc-ph. arXiv, 2021.

Felippe Alves. *Community formation in agent based models of societies*. text, Universidade de São Paulo, July 2020. URL https://www.teses.usp.br/teses/disponiveis/43/43134/tde-20082020-140035/

J Haidt. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834, 2001. doi:10.1037/0033-295X.

Nestor Caticha and Renato Vicente. Agent-based social psychology: from neurocognitive processes to social data. *Advances in Complex Systems*, 14(05):711–731, 2011.

M Copelli and N Caticha. On-line learning in the committee machine. *Journal of Physics A*, 28:1615, 1995.

Robert Simonetti and Nestor Caticha. On-line learning in parity machines. *J. Phys. A: Math. Gen.*, 29:4859–4867, 1996.

O Kinouchi and N Caticha. Lower bounds for generalization with drifting rules. *Journal of Physics A*, 26:6161, 1993.

Nestor Caticha and Osame Kinouchi. Time ordering in the evolution of information processing and modulation systems. *Philosophical Magazine B*, 77(5):1565–1574, 1998. doi:10.1080/13642819808205049 URL https://doi.org/10.1080/13642819808205049

Juan Pablo Neirotti and Nestor Caticha. Dynamics of the evolution of learning algorithms by selection. *Phys. Rev. E*, 67:041912, Apr 2003. doi:10.1103/PhysRevE.67.041912

R Metzler, W Kinzel, and I Kanter. Interacting neural networks. *Physical Review E*, 62(2):2555, 2000.

Samara Klar, Yanna Krupnikov, and John Barry Ryan. Affective Polarization or Partisan Disdain?: Untangling a Dislike for the Opposing Party from a Dislike of Partisanship. *Public Opinion Quarterly*, 82(2):379–390, 05 2018b. ISSN 0033-362X. doi:10.1093/poq/nfy014 URL https://doi.org/10.1093/poq/nfy014

Nestor Caticha, Jonatas Cesar, and Renato Vicente. For whom will the Bayesian agents vote? *Frontiers in Physics*, 3 (25), 2015. ISSN 2296-424X. doi:10.3389/fphy.2015.00025

Abraham Lincoln. *Speech in U. S. House of Representatives on the Presidential Question 1848, Collected works Vol 1*. *The Abraham Lincoln Association, Springfield, Illinois*. R. P. Basler, ed. New Brunswick, N.J: Rutgers University Press, 1953.

Morris P. Fiorina and Samuel J. Abrams. Political polarization in the american public. *Annual Review of Political Science*, 11(1):563–588, 2008.