Physica A 400 (2014) 124–138

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Moral foundations in an interacting neural networks society: A statistical mechanics analysis

R. Vicente [a,*], A. Susemihl [b], J.P. Jericó [c], N. Caticha [c]

[a] Department of Applied Mathematics, Instituto de Matemática e Estatística, Universidade de São Paulo, 05508-090, São Paulo-SP, Brazil

[b] Artificial Intelligence Group, Technical University Berlin, Franklinstraße, 28/29, D-10587 Berlin, Germany

[c] Dep. de Física Geral, Instituto de Física, Universidade de São Paulo, CP 66318, 05315-970, São Paulo-SP, Brazil

HIGHLIGHTS

- Moral foundations statistics depend on peer pressure and cognitive style.
- An order–disorder transition is found in the peer pressure-novelty seeking plane.
- A mean field theory is constructed and confirms numerical simulations.
- Dynamical properties are consistent with cognitive/political affiliation groups.

ARTICLE INFO

Article history:
Received 24 July 2013
Received in revised form 22 October 2013
Available online 20 January 2014

Keywords:
Sociophysics
Social interactions
Opinion dynamics
Neural networks
Moral foundations theory

ABSTRACT

The moral foundations theory supports that people, across cultures, tend to consider a small number of dimensions when classifying issues on a moral basis. The data also show that the statistics of weights attributed to each moral dimension is related to self-declared political affiliation, which in turn has been connected to cognitive learning styles by the recent literature in neuroscience and psychology. Inspired by these data, we propose a simple statistical mechanics model with interacting neural networks classifying vectors and learning from members of their social neighbourhood about their average opinion on a large set of issues. The purpose of learning is to reduce dissension among agents when disagreeing. We consider a family of learning algorithms parametrized by δ, that represents the importance given to corroborating (same sign) opinions. We define an order parameter that quantifies the diversity of opinions in a group with homogeneous learning style. Using Monte Carlo simulations and a mean field approximation we find the relation between the order parameter and the learning parameter δ at a temperature we associate with the importance of social influence in a given group. In concordance with data, groups that rely more strongly on corroborating evidence sustain less opinion diversity. We discuss predictions of the model and propose possible experimental tests.

1. Introduction

Sociophysics, the approach to mathematical modelling of social science, is still maturing as a scientific field [1]. Opinion dynamics, voting, social influence and contagion models have been thoroughly studied [2,3], patterns in social data have been identified (e.g. Ref. [4] or Ref. [5] and references therein) and some successful predictions have been achieved (e.g. Ref. [6]).

In this paper our aim is to present a data driven statistical mechanics model for the formation of opinions about morality. We would like to verify if we can explain features of social data by considering a stylized model for neurocognitive processes

* Corresponding author. Tel.: +55 11 22943492.
E-mail addresses: rvicente@ime.usp.br (R. Vicente), nestor@if.usp.br (N. Caticha).

0378-4371/$ – see front matter
http://dx.doi.org/10.1016/j.physa.2014.01.013

that are well described in the literature. Clearly practical limits to such a goal have to be considered. At the scale of individuals, neurocognitive data inspiring any modelling are always exposed to ecological validity issues with multiple uncontrolled causes. At the social scale, we also have to keep in mind the sheer complexity of human nature and human relationships. By stylized model we here mean a model to be used mainly to connect pieces of empirical evidence, to help the identification of important variables and as an aid to formulate new empirical questions. Furthermore, we would also like to have a testable model, namely, a model that makes some predictions after fitting a few key parameters to empirical data.

We argue here and in our previous work [7] that the evidence available on the moral classification problem can be accommodated by assuming agents that are conformist classifiers adapting to their social neighbourhood by reinforcement learning. Empirical evidence regarding different cognitive styles can then be represented in the model as distinct learning algorithms following the now established tradition of the statistical mechanics of learning [8]. Studies on social psychology [9] allow the further simplification of assuming that social influence only takes place between individuals perceived as similar. As a first approximation we thus assume that the social network can be partitioned into homogeneous social influence subnetworks, each one with a given cognitive style or learning algorithm.

But what do these conformist agents classify? We assume that any issue under debate can be parsed into a discrete set of independent attributes or dimensions. The modern theory of moral foundations [10] suggests that, as far as morality is concerned, these dimensions are not many more than five, namely: (a) harm/violence; (b) justice/fairness; (c) in-group loyalty; (d) respect for authority; and (e) purity or sanctity. For our modelling effort it is, however, sufficient that morality can be parsed into a discrete number of identifiable dimensions. As a starting point we do not consider the origin of these dimensions, its particular meanings or the practical issues that may be involved in trying to parse a given subject into these dimensions. These five dimensions have been found empirically to be sufficient to characterize political orientations along the liberal–conservative spectrum. The need for a sixth dimension, (f) liberty/oppression, has also been discussed to extend the description to include libertarians, but this is outside the scope of the data we have analyzed.

We thus consider that the moral content of an issue may be represented by a direction in a unit radius D-dimensional hypersphere $\mathbf{x} \in \mathbb{S}^D$. In the course of daily social relationships an individual j will be exposed to a variety of issues of diverse moral content parsed as $\mathbf{x}_j^\mu$ with $\mu = 1, 2, \ldots$. For each of these issues an opinion $\sigma_j^\mu \in [-1, 1]$ with a sign and an amplitude $|\sigma_j^\mu|$ is displayed. The sign can be interpreted as providing a for/against information and the amplitude as carrying information on how convict individual j is. A way to describe a classification task of this sort is by assuming that $\sigma_j^\mu = \mathbf{x}_j^\mu \cdot \mathbb{J}_j$, where $\mathbb{J}_j$ is an adaptive internal representation, inaccessible to other individuals, used by individual j to perform moral classification tasks. For simplicity we will study the case where all moral vectors are normalized to unit length $\mathbb{J}_j \in \mathbb{S}^D$. This also implies that differences in moral values are not interpreted as any type of moral superiority and that no moral shallowness is implied by the differences. Thus only the direction the moral vector points is considered as important, removing a layer of complexity in the interpretation of the model.

A conformist individual will then seek agreement with social neighbours in moral classifications by adjusting the internal representation $\mathbb{J}_j$. Employing the statistical mechanics of learning jargon, we are supposing that model agents are interacting *normalized linear perceptrons* [8] (for previous studies of interacting neural networks see Refs. [11–13]).

We suppose that in their daily lives agents interact within a social neighbourhood and exchange opinions about a large set of P issues. We assume that every issue has a *subjective* representation in the space of moral foundations $\mathbf{x}_j^\mu = \mathbf{x}^\mu + \epsilon_j^\mu$, with an idiosyncratic component ϵ_j^μ and an *objective* component $\mathbf{x}^\mu$. To further simplify the model we assume that agents do not adapt to each issue separately, instead, agents decrease their cognitive load by reducing the whole set of opinions of a neighbour to a single opinion about an average *objective* issue $\mathbb{Z}$ (that we call the *Zeitgeist* vector). Another layer of information compression can be added by assuming that agents are consistent in their moral classifications, to say, knowing the opinion about $\mathbf{x}_j^\mu$ determines the opinion about $-\mathbf{x}_j^\mu$ and agents only consider one of these alternatives at a time. With this restriction the opinion field would then be given by

$$h_j = \left(\frac{\sum_{\mu=1}^{P} \mathbf{x}_j^\mu}{\left\| \sum_{\mu=1}^{P} \mathbf{x}_j^\mu \right\|} \right) \cdot \mathbb{J}_j. \tag{1}$$

We also assume that there are no relevant biases or correlations in the individual parsing through the social network, and, by the law of large numbers, that idiosyncratic components cancel out as the number of issues discussed grows. We then write the opinion field as $h_j = \mathbb{Z} \cdot \mathbb{J}_j$, where the mean issue

$$\mathbb{Z} = \frac{\sum_{\mu=1}^{P} \mathbf{x}_j^\mu}{\left\| \sum_{\mu=1}^{P} \mathbf{x}_j^\mu \right\|}, \tag{2}$$